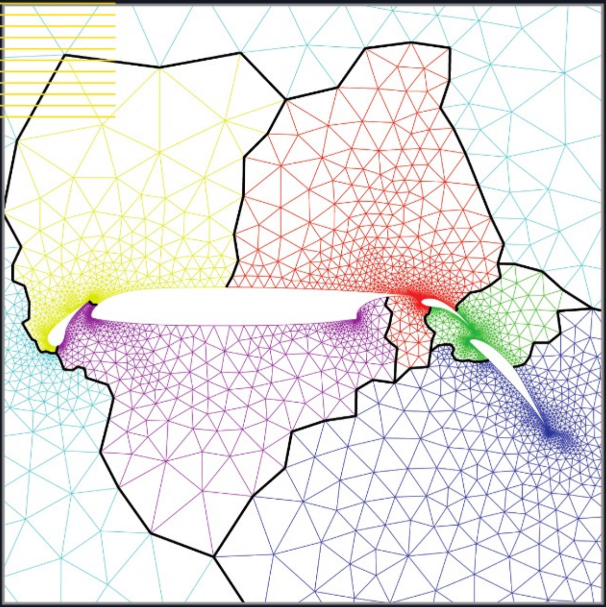
Lecture Notes in Computational Science and Engineering

61



Tarek P. A. Mathew

Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations

Springer

Lecture Notes
in Computational Science
and Engineering

61

Tarek P. A. Mathew

Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations

With 40 Figures and 1 Table

Springer

Tarek Poonithara Abraham Mathew
tmathew@poonithara.org

ISBN 978-3-540-77205-7 e-ISBN 978-3-540-77209-5

Lecture Notes in Computational Science and Engineering ISSN 1439-7358

Library of Congress Control Number: 2008921994

Mathematics Subject Classification (2000): 65F10, 65F15, 65N22, 65N30, 65N55, 65M15, 65M55, 65K10

Cover design: WMX Design GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

spinger.com

In loving dedication to my (late) dear mother,
and to my dear father and brother

Preface

These notes serve as an introduction to a subject of study in computational mathematics referred to as *domain decomposition methods*. It concerns *divide and conquer* methods for the numerical solution and approximation of partial differential equations, primarily of elliptic or parabolic type. The methods in this family include *iterative algorithms* for the solution of partial differential equations, techniques for the discretization of partial differential equations on *non-matching grids*, and techniques for the *heterogeneous approximation* of partial differential equations of heterogeneous character. The divide and conquer methodology used is based on a *decomposition* of the *domain* of the partial differential equation into smaller subdomains, and by design is suited for implementation on *parallel* computer architectures. However, even on serial computers, these methods can provide flexibility in the treatment of complex geometry and heterogeneities in a partial differential equation.

Interest in this family of computational methods for partial differential equations was spawned following the development of various high performance multiprocessor computer architectures in the early eighties. On such parallel computer architectures, the execution time of these algorithms, as well as the memory requirements per processor, *scale* reasonably well with the size of the problem and the number of processors. From a computational viewpoint, the divide and conquer methodology based on a decomposition of the domain of the partial differential equation, yields algorithms having *coarse granularity*, i.e., a significant portion of the computations can be implemented concurrently on different processors, while the remaining portion requires communication between the processors. As a consequence, these algorithms are well suited for implementation on MIMD (multiple instruction, multiple data) architectures. Currently, such parallel computer architectures can alternatively be simulated using a cluster of workstations networked with high speed connections using communication protocols such as MPI (Message Passing Interface) [GR15] or PVM (Parallel Virtual Machines) [GE2].

The mathematical roots of this subject trace back to the seminal work of H. A. Schwarz [SC5] in the nineteenth century. Schwarz proposed an iterative method, now referred to as the *Schwarz alternating method*, for constructing *harmonic* functions on regions of *irregular* shape which can be expressed as the union of subregions of regular shape (such as rectangles and spheres). His motivation was primarily theoretical, to establish the existence of harmonic functions on irregular regions, and his method was not used in computations until recently [SO, MO2, BA2, MI, MA37, DR11, LI6, LI7, BR18].

A general development of domain decomposition methodology for partial differential equations occurred only subsequent to the development of parallel computer architectures, though divide and conquer methods such as Kron's method for electrical circuits [KR] and the substructuring method [PR4] in structural engineering, pre-date domain decomposition methodology. Usage of the term "domain decomposition" seems to have originated around the mid-eighties [GL2] when interest in these methods gained momentum. The first international symposium on this subject was held in Paris in 1987, and since then there have been yearly international conferences on this subject, attracting interdisciplinary interest from communities of engineers, applied scientists and computational mathematicians from around the globe.

Early literature on domain decomposition methods focused primarily on *iterative* procedures for the solution of partial differential equations. As the methodology evolved, however, techniques were also developed for *coupling* discretizations on subregions with *non-matching* grids, and for constructing *heterogeneous* approximations of complicated systems of partial differential equations having heterogeneous character. The latter approximations are built by solving local equations of different character. From a mathematical viewpoint, these diverse categories of numerical methods for partial differential equations may be derived within several frameworks. Each decomposition of a domain typically suggests a reformulation of the original partial differential equation as an equivalent *coupled* system of partial differential equations posed on the subdomains with boundary conditions chosen to match solutions on adjacent subdomains. Such equivalent systems are referred to in these notes as *hybrid formulations*, and provide a framework for developing novel domain decomposition methods. Divide and conquer algorithms can be obtained by numerical approximation of hybrid formulations. Four hybrid formulations are considered in these notes, suited for equations primarily of elliptic type:

- *The Schwarz formulation.*
- *The Steklov-Poincaré (substructuring or Schur complement) formulation.*
- *The Lagrange multiplier formulation.*
- *The Least squares-control formulation.*

Alternative hybrid formulations are also possible, see [CA7, AC5].

The applicability and stability of each hybrid formulation depends on the underlying partial differential equation and subdomain decomposition. For instance, the Schwarz formulation requires an overlapping decomposition, while the Steklov-Poincaré and Lagrange multiplier formulations are based on a non-overlapping decomposition. The least squares-control method can be formulated given overlapping or non-overlapping decompositions. Within each framework, novel iterative methods, discretizations schemes on non-matching grids, and heterogeneous approximations of the original partial differential equation, can be developed based on the associated hybrid formulations.

In writing these notes, the author has attempted to provide an accessible introduction to the important methodologies in this subject, emphasizing a matrix formulation of algorithms. However, as the literature on domain decomposition methods is vast, various topics have either been omitted or only touched upon. The methods described here apply primarily to equations of *elliptic* or *parabolic* type, and applications to hyperbolic equations [QU2], and spectral or p-version elements have been omitted [BA4, PA16, SE2, TO10]. Applications to the equations of elasticity and to Maxwell's equations have also been omitted, see [TO10]. Parallel implementation is covered in greater depth in [GR12, GR10, FA18, FA9, GR16, GR17, HO4, SM5, BR39]. For additional domain decomposition theory, see [XU3, DR10, XU10, TO10]. A broader discussion on *heterogeneous* domain decomposition can be found in [QU6], and on FETI-DP and BDDC methods in [TO10, MA18, MA19]. For additional bibliography on domain decomposition, see *http://www.ddm.org*.

Readers are assumed to be familiar with the basic properties of elliptic and parabolic partial differential equations [JO, SM7, EV] and traditional methods for their discretization [RI, ST14, CI2, SO2, JO2, BR28, BR]. Familiarity is also assumed with basic numerical analysis [IS, ST10], computational linear algebra [GO4, SA2, AX, GR2, ME8], and elements of optimization theory [CI4, DE7, LU3, GI2]. Selected background topics are reviewed in various sections of these notes. Chap. 1 provides an overview of domain decomposition methodology in a context involving *two* subdomain decompositions. Four different hybrid formulations are illustrated for a model coercive 2nd order elliptic equation. Chapters 2, 3 and 4 describe the matrix implementation of *multisubdomain* domain decomposition iterative algorithms for traditional discretizations of self adjoint and coercive elliptic problems. These chapters should ideally be read prior to the other chapters. Readers unfamiliar with constrained minimization problems and their saddle point formulation, may find it useful to review background in Chap. 10 or in [CI4], as saddle point methodology is employed in Chaps. 1.4 and 1.5 and in Chaps. 4 and 6. With a few exceptions, the remaining chapters may be read independently.

The author expresses his deep gratitude to the anonymous referees who made numerous suggestions for revision and improvement of the manuscript. Deep gratitude is also expressed to Prof. Olof Widlund who introduced the author to this subject over twenty years ago, to Prof. Tony Chan, for his kind encouragement to embark on writing a book extending our survey paper on this subject [CH11], and to Prof. Xiao-Chuan Cai, Prof. Marcus Sarkis and Prof. Junping Wang for their research collaborations and numerous insightful discussions over the years. The author deeply thanks Prof. Timothy Barth for his kind permission to use the figure on the cover of this book, and for use of Fig. 5.1. To former colleagues at the University of Wyoming, and to professors Myron Allen, Gastao Braga, Benito Chen, Duilio Conceição, Max Dryja, Frederico Furtado, Juan Galvis, Etereldes Gonçalves, Raytcho Lazarov, Mary Elizabeth Ong, Peter Polyakov, Giovanni Russo, Christian Schaerer, Shagi-Di Shih, Daniel Szyld, Panayot Vassilevski and Henrique Versieux, the author expresses his deep gratitude. Deep appreciation is also expressed to the editors of the LNCSE series, Dr. Martin Peters, Ms. Thanh-Ha LeThi, and Mr. Frank Holzwarth for their patience and kind help during the completion of this manuscript. Finally, deep appreciation is expressed to Mr. Elumalai Balamurugan for his kind assistance with reformatting the text. The author welcomes comments and suggestions from readers, and hopes to post updates at *www.poonithara.org/publications/dd.*

January 2008 *Tarek P. A. Mathew*

Contents

1

Decomposition Frameworks

In this chapter, we introduce and illustrate several principles employed in the formulation of domain decomposition methods for an elliptic equation. In our discussion, we focus on a *two* subdomain decomposition of the domain of the elliptic equation, into overlapping or non-overlapping subdomains, and introduce the notion of a *hybrid formulation* of the elliptic equation. A hybrid formulation is a *coupled* system of elliptic equations which is *equivalent* to the original elliptic equation, with unknowns representing the true solution on each subdomain. Such formulations provide a natural framework for the construction of *divide and conquer* methods for an elliptic equation. Using a hybrid formulation, we heuristically illustrate how novel divide and conquer *iterative* methods, *non-matching grid* discretizations and *heterogeneous* approximations can be constructed for an elliptic equation.

We illustrate four alternative hybrid formulations for an elliptic equation. Each will be described for a decomposition of the domain into *two* subdomains, either overlapping or non-overlapping. We shall describe the following:

- Schwarz formulation.
- Steklov-Poincaré formulation.
- Lagrange multiplier formulation.
- Least squares-control formulation.

For each hybrid formulation, we illustrate how iterative methods, non-matching grid discretizations and heterogeneous approximations can be formulated for the elliptic equation based on its two subdomain decomposition. In Chap. 1.1, we introduce notation and heuristically describe the structure of a hybrid formulation. Chap. 1.2 describes a two subdomain Schwarz hybrid formulation, based on overlapping subdomains. Chap. 1.3 describes the Steklov-Poincaré formulation, based on two non-overlapping subdomains. The Lagrange multiplier formulation described in Chap. 1.4 applies only for a self adjoint and coercive elliptic equation, and it employs two non-overlapping subdomains. Chap. 1.5 describes the least squares-control formulation for a two subdomain overlapping or non-overlapping decomposition.

1.1 Hybrid Formulations

Given a subdomain decomposition, a hybrid formulation of an elliptic equation is an *equivalent* coupled system of elliptic equations involving unknowns on each subdomain. In this section, we introduce notation on an elliptic equation and heuristically describe the structure of its two subdomain hybrid formulation. We outline how divide and conquer iterative methods, non-matching grid discretizations, and heterogeneous approximations can be constructed for an elliptic equation, using an hybrid formulation of it. Four commonly used hybrid formulations are described in Chap. 1.2 through Chap. 1.5.

1.1.1 Elliptic Equation

We shall consider the following 2nd order elliptic equation:

$$\begin{cases} L\,u \equiv -\nabla\cdot(a(x)\nabla u) + \mathbf{b}(x)\cdot\nabla u + c(x)\,u = f, & \text{in } \Omega \\ \hfill u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

for $\Omega \subset \mathrm{I\!R}^d$. The coefficient $a(x)$ will be assumed to satisfy:

$$0 < a_0 \le a(x), \qquad \forall x \in \Omega,$$

while $\mathbf{b}(x)$ and $c(x) \ge 0$ will be assumed to be smooth, and $f(x) \in L^2(\Omega)$. Additional restrictions will be imposed on the coefficients as required.

1.1.2 Weak Formulation

A weak formulation of (1.1) is typically obtained by multiplying it by a sufficiently smooth test function $v(x)$ and integrating the diffusion term by parts on Ω. It will seek $u \in H_0^1(\Omega)$ satisfying:

$$\begin{cases} \mathcal{A}(u,v) = F(v), \qquad\qquad \forall v \in H_0^1(\Omega), \text{ where} \\ \mathcal{A}(u,v) \equiv \int_\Omega \left(a(x)\,\nabla u\cdot\nabla v + (\mathbf{b}(x)\cdot\nabla u)\,v + c(x)\,u\,v\right)dx \\ \quad F(v) \equiv \int_\Omega f\,v\,dx, \end{cases} \tag{1.2}$$

where the Sobolev space $H_0^1(\Omega)$ is formally defined as below [NE, LI4, JO2]:

$$H_0^1(\Omega) \equiv \left\{v \in H^1(\Omega) \,:\, v = 0 \text{ on } \partial\Omega\right\},$$

while the space $H^1(\Omega)$ is defined as:

$$\begin{cases} H^1(\Omega) \equiv \left\{v \in L^2(\Omega) \,:\, \|v\|_{1,\Omega}^2 < \infty\right\}, \quad \text{where} \\ \|v\|_{1,\Omega}^2 \equiv \int_\Omega \left(v^2 + |\nabla v|^2\right)\,dx, \end{cases}$$

for $\nabla v \equiv \left(\frac{\partial v}{\partial x_1}, \ldots, \frac{\partial v}{\partial x_d}\right)$. The bilinear form $\mathcal{A}(.,.)$ will be *coercive* if:

$$\mathcal{A}(u,u) \ge \alpha\,\|u\|_{1,\Omega}^2, \qquad \forall\, v \in H_0^1(\Omega),$$

for some $\alpha > 0$ independent of u. Coercivity of $\mathcal{A}(.,.)$ is guaranteed to hold by the Poincaré-Freidrichs inequality, see [NE].

1.1.3 Discretization

A finite element discretization of (1.1) is obtained by Galerkin approximation of (1.2). Let $\mathcal{T}_h(\Omega)$ denote a triangulation of Ω with elements of size h and let V_h denote the space of continuous piecewise linear finite element functions on $\mathcal{T}_h(\Omega)$, see [ST14, CI2, JO2, BR28, BR]. If $\{\phi_1, \ldots, \phi_n\}$ forms a basis for $V_h \cap H_0^1(\Omega)$, then the finite element discretization of (1.1) will yield the system:

$$A\,\mathbf{u} = \mathbf{f},$$

where $A_{ij} = \mathcal{A}(\phi_i, \phi_j)$ for $1 \leq i,\, j \leq n$ and $\mathbf{f}_i = F(\phi_i)$ for $1 \leq i \leq n$.

1.1.4 Subdomain Decompositions

We shall employ the following notation, see Fig. 1.1.

Definition 1.1. *A collection of two open subregions $\Omega_i \subset \Omega$ for $i = 1, 2$ will be referred to as a non-overlapping decomposition of Ω if the following hold:*

$$\begin{cases} \overline{\Omega}_1 \cup \overline{\Omega}_2 = \overline{\Omega}, \\ \Omega_1 \cap \Omega_2 = \emptyset. \end{cases}$$

Boundaries of the subdomains will be denoted $\partial\Omega_i$ and their interior and exterior segments by $B^{(i)} \equiv \partial\Omega_i \cap \Omega$ and $B_{[i]} \equiv \partial\Omega_i \cap \partial\Omega$, respectively. We will denote the common interface by $B \equiv \partial\Omega_1 \cap \partial\Omega_2$.

Definition 1.2. *A collection of two open subregions $\Omega_i^* \subset \Omega$ for $i = 1, 2$ will be referred to as an overlapping decomposition of Ω if the following holds:*

$$\Omega_1^* \cup \Omega_2^* = \Omega.$$

Boundaries of the subdomains will be denoted $B_i \equiv \partial\Omega_i^*$ and their interior and exterior segments by $B^{(i)} \equiv \partial\Omega_i^* \cap \Omega$ and $B_{[i]} \equiv \partial\Omega_i^* \cap \partial\Omega$, respectively.

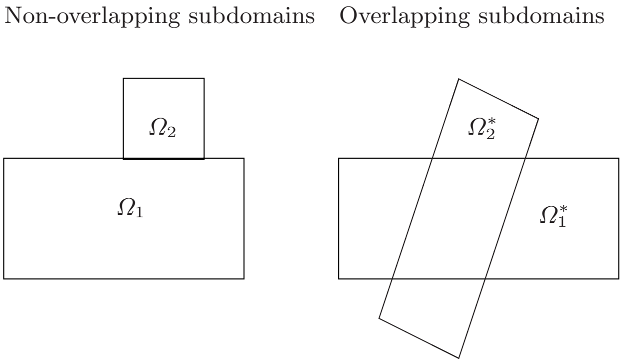

Non-overlapping subdomains Overlapping subdomains

Fig. 1.1. Two subdomain decompositions

Remark 1.3. In applications, a decomposition of Ω into subdomains can be chosen based either on the geometry of Ω or on the regularity of the solution u (if known). An overlapping subdomain Ω_i^* can, if desired, be constructed from a nonoverlapping subdomain Ω_i by extending it to include all points in Ω within a distance $\beta > 0$ of Ω_i, yielding *uniform* overlap.

1.1.5 Partition of Unity

A partition of unity subordinate to the overlapping subdomains Ω_1^* and Ω_2^* consists of smooth functions $\chi_1(x)$ and $\chi_2(x)$ satisfying:

$$\begin{cases} \chi_i(x) \geq 0, & \text{in } \overline{\Omega}_i^* \\ \chi_i(x) = 0, & \text{in } \Omega \backslash \overline{\Omega}_i^* \\ \chi_1(x) + \chi_2(x) = 1, & \text{in } \overline{\Omega}. \end{cases} \tag{1.3}$$

Each $\chi_i(.)$ may be non-zero on $B_{[i]}$. In applications, each $\chi_i(x)$ may be required to satisfy a bound of the form $|\nabla \chi_i(x)| \leq C\, h_0^{-1}$, where h_0 denotes the diameter of each subdomain Ω_i^*.

Heuristically, a continuous partition of unity subordinate to Ω_1^* and Ω_2^* can be computed as follows. Let $d_i(x)$ denote the distance function:

$$d_i(x) = \begin{cases} \text{dist}\left(x, B^{(i)}\right), & \text{if } x \in \overline{\Omega}_i^* \\ 0, & \text{if } x \notin \overline{\Omega}_i^*, \end{cases} \tag{1.4}$$

where $B^{(i)} \equiv (\partial \Omega_i^* \cap \Omega)$. Then, formally define:

$$\chi_i(x) \equiv \frac{d_i(x)}{d_1(x) + d_2(x)}, \quad \text{for} \quad 1 \leq i \leq 2. \tag{1.5}$$

By construction, each $d_i(x)$ will be *continuous*, nonnegative, with support in $\overline{\Omega}_i^*$, and satisfy the desired properties. To obtain a smooth function $\chi_i(x)$, each $d_i(x)$ may first be *mollified*, see [ST9].

Remark 1.4. Given a *non-overlapping* decomposition Ω_1 and Ω_2 of Ω, we shall sometimes employ a *discontinuous* partition of unity satisfying:

$$\begin{cases} \chi_i(x) \geq 0, & \text{in } \overline{\Omega}_i \\ \chi_i(x) = 0, & \text{in } \Omega \backslash \overline{\Omega}_i \\ \chi_1(x) + \chi_2(x) = 1, & \text{in } \Omega. \end{cases} \tag{1.6}$$

Each $\chi_i(x)$ will be discontinuous across $B = \partial \Omega_1 \cap \partial \Omega_2$. Such a partition of unity may be constructed using $d_i(x) = 1$ on $\overline{\Omega}_i$ in (1.5).

1.1.6 Hybrid Formulation

Let Ω_1 and Ω_2 (or Ω_1^* and Ω_2^*) form a decomposition of a domain Ω. Then, a hybrid formulation of (1.1), is a coupled system of partial differential equations *equivalent* to (1.1), with one unknown function $w_i(x)$, representing the local solution, on each subdomain Ω_i (or Ω_i^*). Two requirements must be satisfied. *First*, the restriction $u_i(x)$ of the true solution $u(x)$ of (1.1) to each subdomain Ω_i (or Ω_i^*) must solve the hybrid system, i.e., $(u_1(x), u_2(x))$ must solve the hybrid formulation. *Second*, the hybrid formulation must be *well posed* as a coupled system, i.e., its solution $(w_1(x), w_2(x))$ must exist and be unique, and furthermore, it must depend continuously on the data.

The first requirement ensures that the hybrid formulation is *consistent* with the original problem (1.1), yielding $w_i(x) = u_i(x)$ for $i = 1, 2$. The second requirement ensures that the hybrid formulation is *stable* and uniquely solvable. The latter is essential for the stability of a numerical approximation of the hybrid formulation. Once the hybrid system is solved, the solution $u(x)$ of (1.1) can be expressed in terms of the local solutions $w_i(x)$ as:

$$u(x) = \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x),$$

using a partition of unity $\chi_1(x)$ and $\chi_2(x)$ appropriate for the subdomains.

Typically, a hybrid formulation consists of a *local problem* posed on each individual subdomain, along with *matching conditions* that couple the local problems. In some hybrid formulations, a global functional may be employed, whose optima is sought, or new variables may be introduced to couple the local problems. Such coupling must ensure *consistency* and *well posedness.*

Local Problems. On each subdomain Ω_i (or Ω_i^*), a hybrid formulation will require $w_i(x)$ to solve the original partial differential equation (1.1):

$$\begin{cases} L w_i & = f_i, \ \text{ on } \Omega_i \ \text{ (or } \Omega_i^*) \\ T_i(w_i, \gamma) & = g_i, \ \text{ on } B^{(i)} \\ w_i & = 0, \ \text{ on } B_{[i]} \end{cases} \qquad \text{for } \ i = 1,\, 2, \tag{1.7}$$

where $T_i(w_1, \gamma)$ denotes a boundary operator which enforces either Dirichlet, Neumann or Robin boundary conditions on $B^{(i)}$:

$$T_i(w_i, \gamma) = \begin{cases} w_i, & \text{for Dirichlet boundary conditions} \\ \mathbf{n}_i \cdot (a(x) \nabla w_i) & \text{for Neumann boundary conditions} \\ \mathbf{n}_i \cdot (a(x) \nabla w_i) + \gamma\, w_i & \text{for Robin boundary conditions.} \end{cases} \tag{1.8}$$

Here $\mathbf{n}_i$ denotes the unit exterior normal to $B^{(i)}$ and $\gamma(\cdot)$ denotes a coefficient function in the Robin boundary condition. Typically, $f_i(x)$ is $f(x)$ restricted to Ω_i (or Ω_i^*). The choice of the boundary operator $T_i(w_i, \gamma)$ may differ with each hybrid formulation. The boundary data $g_i(.)$ typically corresponds to $T_i(.)$ applied to the solution on the adjacent domain, however, it may also be a control or a Lagrange multiplier function which *couples* the local problems.

Matching Conditions. Matching conditions *couple* the different local problems (1.7) by choosing $g_i(.)$ to ensure that the hybrid formulation is equivalent to (1.1). Typically, matching conditions are equations satisfied by the true solution $u(x)$ restricted to the interfaces or regions of overlap between adjacent subdomains. For an elliptic equation, these may be either *algebraic* equations, such as the requirement of continuity of the local solutions $u_i(x)$ and $u_j(x)$ across *adjacent* subdomains:

$$\begin{cases} u_i - u_j = 0, & \text{on } \partial\Omega_i \cap \partial\Omega_j, \text{ non-overlapping case} \\ u_i - u_j = 0, & \text{on } \partial\Omega_i^* \cap \Omega_j^*, \text{ overlapping case} \end{cases}$$

or they may be *differential* constraints, such as continuity of the local *fluxes*:

$$\begin{cases} \mathbf{n}_i \cdot (a(x)\nabla u_i) + \mathbf{n}_j \cdot (a(x)\nabla u_j) = 0, & \text{on } \partial\Omega_i \cap \partial\Omega_j, \text{ non-overlapping case} \\ \mathbf{n}_i \cdot (a(x)\nabla u_i) - \mathbf{n}_i \cdot (a(x)\nabla u_j) = 0, & \text{on } \partial\Omega_i^* \cap \Omega_j^*, \text{ overlapping case} \end{cases}$$

where $\mathbf{n}_i$ denotes the unit exterior normal to $\partial\Omega_i$. Such equations specify $g_i(.)$. Other differential constraints may also be employed using linear combinations of the above algebraic and differential constraints. Matching conditions may be enforced either directly, as in the preceding constraints, or indirectly through the use of intermediary variables such as Lagrange multipliers. In the latter case, the hybrid formulation may be derived as a saddle point problem (Chap. 1.4 or Chap. 10) of an associated constrained optimization problem.

We shall express general matching conditions in the form:

$$H_i(w_1, w_2, g_1, g_2) = 0, \quad \text{for } 1 \le i \le 2, \tag{1.9}$$

for suitably chosen operators $H_i(\cdot)$ on the interface $B^{(i)}$.

Reconstruction of the Global Solution. Once a hybrid formulation consisting of local equations of the form (1.7) for $1 \le i \le 2$ together with equations of the form (1.9) has been formulated and solved, the global solution $u(.)$ may be represented in the form:

$$u(x) = \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x), \tag{1.10}$$

where $\chi_i(x)$ is a (possibly discontinuous) partition of unity subordinate to the subdomains $\overline{\Omega}_1$ and $\overline{\Omega}_2$ (or Ω_1^* and Ω_2^*).

Well Posedness of the Hybrid Formulation. To ensure that the hybrid formulation is solvable and that it may be approximated numerically by stable schemes, we require that the hybrid formulation be well posed [SM7, EV], satisfying, for $C > 0$ independent of the data, the bound:

$$(\|w_1\| + \|w_2\|) \le C\, (|||f_1||| + |||f_2||| + |||g_1||| + |||g_2|||)\,,$$

where $\|\cdot\|$ and $|||\cdot|||$ are appropriately chosen norms for the solution and data, as suggested by elliptic regularity theory [GI].

Iterative Methods. Domain decomposition iterative algorithms can be formulated for solving (1.1) by directly applying traditional *relaxation*, *descent* or *saddle point* algorithms to a hybrid formulation. For instance, each unknown w_i may be updated sequentially using a *relaxation* procedure. Given current approximations of w_1, w_2, g_1, g_2 update for w_i by solving:

$$\begin{cases} L w_i &= f_i, \ \text{ on } \Omega_i \ \text{ (or } \Omega_i^*) \\ T_i(w_i, \gamma) &= g_i, \ \text{ on } B^{(i)} \\ w_i &= 0, \ \ \text{ on } B_{[i]}, \end{cases}$$

replacing $T_i(w_i, \gamma) = g_i$ by either of the equations:

$$H_j(w_1, w_2, g_1, g_2) = 0, \quad j = 1, 2,$$

using the current iterates on the other subdomains. Alternatively, a descent or saddle point algorithm can be employed.

Discretization on a Nonmatching Grid. In various applications, it may be of interest to independently triangulate different subregions Ω_i (or Ω_i^*) with grids suited to the geometry of each subdomain. The resulting grids, however, may not match on the regions of intersection between the subdomains, and are referred to as nonmatching grids, see Fig. 1.2. On such non-matching grids, a global discretization of (1.1) may be sought by directly discretizing the hybrid formulation, namely, the the local problems and the matching conditions.

Heuristically, the construction of a global discretization of equation (1.1) on a non-matching triangulation on Ω_i (or Ω_i^*), will involve the following steps.

- Let $\mathcal{T}_{h_i}(\Omega)$ (or $\mathcal{T}_{h_i}(\Omega_i^*)$) denote independent triangulations of Ω_i (or Ω_i^*) with local grid sizes h_i, see Fig. 1.2. These grids need not match on the region of intersection or overlap between the subdomains.
- Each local problem in the hybrid formulation can be discretized as:

$$\begin{cases} A_{h_i} \mathbf{w}_{h_i} &= \mathbf{f}_{h_i}, \ \text{ on } \Omega_{h_i} \ \text{ (or } \Omega_{h_i}^*) \\ T_{h_i}(\mathbf{w}_{h_i}, \gamma_{h_i}) &= \mathbf{g}_{h_i}, \ \text{ on } B^{(i)} \\ \mathbf{w}_{h_i} &= \mathbf{0}, \ \ \ \text{ on } B_{[i]}. \end{cases}$$

Each local discretization should be a stable scheme.

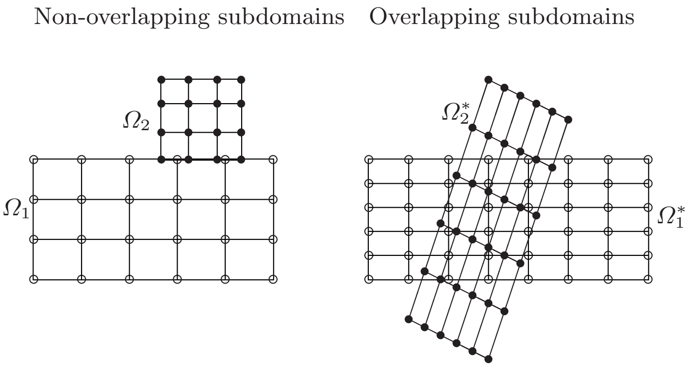

Non-overlapping subdomains Overlapping subdomains

Fig. 1.2. Nonmatching grids

- The matching conditions should also be discretized:

$$H_i^h(\mathbf{w}_{h_1}, \mathbf{w}_{h_2}, \mathbf{g}_{h_1}, \mathbf{g}_{h_2}) = 0, \quad 1 \le i \le 2.$$

  To ensure the stability and consistency of the global discretization of the hybrid formulation, care must be exercised in discretizing the matching conditions across the subdomain grids.

Such issues are described in Chap. 1.2 through 1.5, and in Chap. 11.

Heterogenous Approximation. A partial differential equation is said to be *heterogeneous* if its *type* changes from one region to another. An example is *Tricomi's* equation [JO]:

$$u_{x_1 x_1} - x_1\, u_{x_2 x_2} = f(x_1, x_2),$$

which is of hyperbolic type for $x_1 > 0$ and of elliptic type for $x_1 < 0$. In various applications, efficient computational methods may be available for the local problems involved in an heterogeneous partial differential equation. In such cases, it may be of interest to approximate a partial differential equation of *heterogeneous character* by a partial differential equation of *heterogeneous* type. We refer to such models as heterogeneous approximations.

Our discussion will be restricted to an elliptic-hyperbolic heterogeneous approximation of a singularly perturbed elliptic equation of heterogeneous character. We shall consider an advection dominated equation:

$$\begin{cases} -\epsilon\, \Delta u + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.11}$$

where $0 < \epsilon \ll 1$ is a small perturbation parameter. Depending $f(x)$, there may be a subdomain Ω_1 (or Ω_1^*) on which:

$$\epsilon\, |\Delta u| \ll |\mathbf{b}(x) \cdot \nabla u + c(x) u|, \qquad \text{for } x \in \Omega_1 \;\; (\text{or } \Omega_1^*).$$

On Ω_1 (or Ω_1^*), the restriction of elliptic equation $Lu = f$ to the subdomain, will be of hyperbolic character, approximately satisfying $L_1 u = f$, where:

$$\begin{cases} L\, u \equiv \;\;\; \epsilon L_0 u + L_1 u \\ L_0 u \equiv -\Delta u \\ L_1 u \equiv \;\;\; \mathbf{b}(x) \cdot \nabla u + c(x) u. \end{cases}$$

If Ω_2 (or Ω_2^*) denotes a complementary (*layer*) region, then equation (1.11) will be approximately of elliptic character in Ω_2 (or Ω_2^*).

Motivated by singular perturbation methodology [LA5, KE5, OM], it may be computationally advantageous to approximate elliptic equation (1.11) by an *heterogeneous approximation* involving an equation of mixed hyperbolic and elliptic character. To obtain an heterogeneous approximation of (1.11), we may approximate its hybrid formulation based on Ω_i (or Ω_i^*) for $1 \leq i \leq 2$. For instance, we may approximate (1.7) by:

$$\begin{cases} \tilde{L}_i v_i &= f_i, \quad \text{on } \Omega_i \ \text{ (or } \Omega_i^*\text{)}, \\ \tilde{T}_i(v_i, \gamma) &= \tilde{g}_i, \quad \text{on } \tilde{B}^{(i)}, \\ v_i &= 0, \quad \text{on } \tilde{B}_{[i]}, \end{cases} \qquad \text{for } i = 1, 2$$

with $v_i(x) \approx w_i(x)$, and we may approximate (1.9) by:

$$\tilde{H}_i(v_1, v_2, \tilde{g}_1, \tilde{g}_2) = 0, \qquad \text{for } i = 1, 2$$

where $\tilde{L}_i$, $\tilde{T}_i$ and $\tilde{H}_i(\cdot)$ are *heuristic* local approximations of L_i, T_i and $H_i(\cdot)$, obtained by formally omitting $\epsilon\, \Delta u$ on Ω_1 (or Ω_1^*). We refer the reader to Chap. 1.2 through Chap. 1.5 and Chap. 12 for specific examples.

Remark 1.5. Care must be exercised in the selection of approximations since each local problem must be well posed, and the global coupled system must also be well posed. For instance, if we define $\tilde{L}_1 u = L_1 u$ on Ω_1 (or Ω_1^*) then the local problem will be hyperbolic, and we must replace Dirichlet boundary conditions on $B^{(1)}$ and $B_{[1]}$ by *inflow* boundary conditions. Similarly, if we choose $\tilde{L}_2 u = Lu$ on Ω_2 (or Ω_2^*) then the local problem on Ω_2 (or Ω_2^*) will be elliptic and Dirichlet boundary or flux boundary conditions can be employed on $B^{(2)}$ and $B_{[2]}$. Often, approximate matching conditions for a heterogeneous problem can also be derived *heuristically* by a *vanishing viscosity* approach, see Chap. 1.3 and Chap. 12.

1.2 Schwarz Framework

The framework that we refer to as the *Schwarz hybrid* formulation is based on the earliest known domain decomposition method, formulated by H. A. Schwarz [SC5] in 1870. Schwarz formulated an iterative method, now referred to as the *Schwarz alternating method*, which solves Laplace's equation on an irregular domain that is the union of regular regions (such as rectangular and circular regions). Although Schwarz's motivation was to study the existence of harmonic functions on irregular regions, the *hybrid formulation* underlying Schwarz's iterative method, applies to a wider class of elliptic equations, and it enables the formulation of other divide and conquer approximations.

In this section, we describe the hybrid formulation underlying the Schwarz alternating method for a two subdomain *overlapping* decomposition of Ω. We let Ω_1^* and Ω_2^* denote the overlapping subdomains, and let $B^{(i)} = \partial\Omega_i^* \cap \Omega$ and $B_{[i]} = \partial\Omega_i^* \cap \partial\Omega$ denote the interior and exterior boundary segments of Ω_i^*, respectively, see Fig. 1.3. Using the hybrid formulation, we illustrate the formulation of *iterative* methods, *non-matching* grid discretizations, and *heterogeneous* approximations for elliptic equation (1.1).

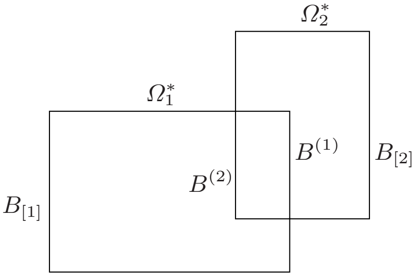


Fig. 1.3. Boundary segments for an overlapping decomposition

1.2.1 Motivation

To derive the *hybrid formulation* underlying Schwarz's method, let $u(x)$ denote the solution of (1.1). Define $w_i(x) = u(x)$ on Ω_i^* for $1 \le i \le 2$. Then, by construction $L\, w_i = f$ in Ω_i^*. Furthermore, the continuity of u will yield matching of w_1 and w_2 on $\Omega_1^* \cap \Omega_2^*$. It will therefore hold that:

$$\begin{cases} Lw_1 = f, & \text{in } \Omega_1^* \\ \;w_1 = w_2, & \text{on } B^{(1)} \\ \;w_1 = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} Lw_2 = f, & \text{in } \Omega_2^* \\ \;w_2 = w_1, & \text{on } B^{(2)} \\ \;w_2 = 0, & \text{on } B_{[2]}. \end{cases}$$

Importantly, *if* the above coupled, decomposed system for $w_1(x)$ and $w_2(x)$, is *well posed*, then by solving it, the original solution can be recovered with $u(x) = w_i(x)$ on Ω_i^* for $i = 1, 2$. We have the following uniqueness result.

Theorem 1.6. *Suppose the following assumptions hold.*

1. *Let $c(x) \ge 0$ and $\nabla \cdot \mathbf{b}(x) \le 0$.*
2. *Let $u(x)$ denote a sufficiently smooth solution of equation (1.1).*
3. *Let $w_1(x)$ and $w_2(x)$ be sufficiently smooth solutions of the following system of coupled elliptic equations:*

$$\begin{cases} Lw_1 = f, & \textit{in } \Omega_1^* \\ \;w_1 = 0, & \textit{on } B_{[1]} \\ \;w_1 = w_2, & \textit{on } B^{(1)} \end{cases} \quad \textit{and} \quad \begin{cases} Lw_2 = f, & \textit{in } \Omega_2^* \\ \;w_2 = 0, & \textit{on } B_{[2]} \\ \;w_2 = w_1, & \textit{on } B^{(2)}. \end{cases} \tag{1.12}$$

Then the following result will hold:

$$u(x) = \begin{cases} w_1(x), & \textit{on } \overline{\Omega}_1^* \\ w_2(x), & \textit{on } \overline{\Omega}_2^*. \end{cases}$$

Proof. If $u(x)$ is a solution of equation (1.1) and $w_1(x) \equiv u(x)$ in Ω_1^* and $w_2(x) \equiv u(x)$ in Ω_2^*, then $w_1(x)$ and $w_2(x)$ will satisfy (1.12) by construction.

To prove the converse, suppose that $w_1(x)$ and $w_2(x)$ satisfy (1.12). We will first show that $w_1(x) = w_2(x)$ on $\Omega_1^* \cap \Omega_2^*$. To this end, note that $w_1(x) - w_2(x)$

has zero boundary conditions on $\partial(\Omega_1^* \cap \Omega_2^*)$. Additionally, by construction $w_1(x) - w_2(x)$ will be L-harmonic. By uniqueness of L-harmonic functions for $c(x) \geq 0$ and $\nabla \cdot \mathbf{b}(x) \leq 0$, it will follow that $w_1(x) - w_2(x) = 0$ in $\Omega_1^* \cap \Omega_2^*$. This yields that $w_1(x) = w_2(x)$ on $\Omega_1^* \cap \Omega_2^*$. Now let $\chi_1(x)$ and $\chi_2(x)$ denote a sufficiently smooth partition of unity subordinate to the cover Ω_1^* and Ω_2^*. If we define $u(x) = \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x)$, then $u(x)$ will satisfy (1.1), since $w_1 = w_2$ in $\Omega_1^* \cap \Omega_2^*$ and since $Lw_i = f$ in Ω_i^*. □

Remark 1.7. The above result suggests that given a partition of unity $\chi_1(x)$ and $\chi_2(x)$ subordinate to Ω_1^* and Ω_1^*, respectively, a solution to elliptic equation (1.1) may be obtained by solving (1.12) and defining:

$$u(x) \;=\; \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x).$$

This yields an equivalence between (1.1) and (1.12).

Remark 1.8. The preceding theorem yields equivalence between sufficiently smooth solutions to (1.1) and (1.12). It is, however, not a result on the well posedness (stability) of formulation (1.12) under perturbations of its data. The latter requires that the perturbed system:

$$\begin{cases} L\tilde{w}_1 = \tilde{f}_1, & \text{in } \Omega_1^* \\ \tilde{w}_1 = 0, & \text{on } B_{[1]} \\ \tilde{w}_1 = w_2 + \tilde{r}_1, & \text{on } B^{(1)} \end{cases} \quad \text{and} \quad \begin{cases} L\tilde{w}_2 = \tilde{f}_2, & \text{in } \Omega_2^* \\ \tilde{w}_2 = 0, & \text{on } B_{[2]} \\ \tilde{w}_2 = w_1 + \tilde{r}_2, & \text{on } B^{(2)}, \end{cases} \tag{1.13}$$

be uniquely solvable and satisfy a bound of the form:

$$(|||\tilde{w}_1||| + |||\tilde{w}_2|||) \;\leq\; C\, \left(\|\tilde{f}_1\| + \|\tilde{f}_2\| + \|\tilde{r}_1\| + \|\tilde{r}_2\| \right),$$

in appropriate norms. See Chap. 15 for maximum norm well posedness.

1.2.2 Iterative Methods

The iterative method proposed by H. A. Schwarz is a very popular method for the solution of elliptic partial differential equations, see [SO, MO2, BA2] and [MI, MA37, DR11, LI6, LI7, BR18]. It is robustly convergent for a large class of elliptic equations, and can be motivated heuristically using the block structure of (1.12). If $w_i^{(k)}$ denotes the k'th iterate on subdomain Ω_i^*, it can be updated by solving the block equation of (1.12) posed on subdomain Ω_i^* with boundary conditions $w_1 = w_2$ on $B^{(1)}$ or $w_2 = w_1$ on $B^{(2)}$ approximated by the current iterate on its adjacent subdomain:

$$\begin{cases} Lw_1^{(k+1)} = f, & \text{in } \Omega_1^* \\ w_1^{(k+1)} = w_2^{(k)}, & \text{on } B^{(1)} \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} Lw_2^{(k+1)} = f, & \text{in } \Omega_2^* \\ w_2^{(k+1)} = w_1^{(k+1)}, & \text{on } B^{(2)} \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]}. \end{cases}$$

The resulting algorithm is the *Schwarz alternating method.* It is *sequential* in nature and summarized below.

Algorithm 1.2.1 *(Schwarz Alternating Method)*
Let $v^{(0)}$ denote the starting global approximate solution.

```
1. For $k = 0, 1, \cdots$ , until convergence do:
2.     Solve for $w_1^{(k+1)}$ as follows:
           $$\begin{cases} Lw_1^{(k+1)} = f_1, & \text{in } \Omega_1^* \\ w_1^{(k+1)} = v^{(k)}, & \text{on } B^{(1)} \\ w_1^{(k+1)} = g, & \text{on } B_{[1]}, \end{cases}$$
       Define $v^{(k+1/2)}$ as follows:
           $$v^{(k+1/2)} \equiv \begin{cases} w_1^{(k+1)}, & \text{on } \Omega_1^* \\ v^{(k)}, & \text{on } \Omega \backslash \Omega_1^*. \end{cases}$$
3.     Solve for $w_2^{(k+1)}$ as follows:
           $$\begin{cases} Lw_2^{(k+1)} = f_2, & \text{in } \Omega_2^* \\ w_2^{(k+1)} = g, & \text{on } B_{[2]} \\ w_2^{(k+1)} = v^{(k+1/2)}, & \text{on } B^{(2)} \end{cases}$$
       Define $v^{(k+1)}$ as follows:
           $$v^{(k+1)} \equiv \begin{cases} w_2^{(k+1)}, & \text{on } \Omega_2^* \\ v^{(k+1/2)}, & \text{on } \Omega \backslash \Omega_2^*. \end{cases}$$
4. Endfor
Output: $v^{(k)}$
```

Remark 1.9. The iterates $v^{(k+\frac{1}{2})}$ and $v^{(k+1)}$ in the preceding algorithm are continuous *extensions* of the subdomain solutions $w_1^{(k+1)}$ and $w_2^{(k+1)}$, to the entire domain Ω. Under suitable assumptions on the coefficients of the elliptic equation and overlap amongst the subdomains Ω_i^*, the iterates $v^{(k)}$ converge geometrically to the true solution u of (1.1), see Chap. 2.5 when $\mathbf{b}(x) = \mathbf{0}$.

The preceding Schwarz algorithm is sequential in nature, requiring the solution of one subdomain problem prior to another. Below, we describe an unaccelerated parallel Schwarz algorithm which requires the concurrent solution of subdomain problems. It is motivated by a popular parallel method, referred to as the additive Schwarz algorithm [DR11], which is employed typically as a preconditioner. The algorithm we describe is based on a partition of unity $\chi_1(x)$ and $\chi_2(x)$ subordinate to the overlapping subdomains Ω_1^* and Ω_2^*, respectively, see [DR11, CA19, MA33, FR8, TA5]. Let $w_i^{(k)}$ denote the k'th iterate on Ω_i^* for $1 \leq i \leq 2$. Then, new iterates are computed as follows.

Algorithm 1.2.2 *(Parallel Partition of Unity Schwarz Method)*
Let $w_1^{(0)}$, $w_2^{(0)}$ *denote starting local approximate solutions.*

```
1. For $k = 0, 1, \cdots$, until convergence do:
2.     For $i = 1, 2$ determine $w_i^{(k+1)}$ in parallel:
```

$$\begin{cases} Lw_i^{(k+1)} = f, & \text{in } \Omega_i^* \\ w_i^{(k+1)} = \chi_1(x)\, w_1^{(k)}(x) + \chi_2(x)\, w_2^{(k)}(x), & \text{on } B^{(i)} \\ w_i^{(k+1)} = 0, & \text{on } B_{[i]}, \end{cases}$$

```
3.     Endfor
4. Endfor
```

Output: $(w_1^{(k)}, w_2^{(k)})$

If $c(x) \geq c_0 > 0$ and there is sufficient overlap, the iterates $v^{(k)}$ defined by:

$$v^{(k)} \equiv \chi_1(x)\, w_1^{(k)}(x) + \chi_2(x)\, w_2^{(k)}(x),$$

will converge geometrically to the solution u of (1.1), see Chap. 15.

Remark 1.10. In practice, given a discretization of (1.1), discrete versions of the above algorithms must be applied. Matrix versions of Schwarz algorithms are described in Chap. 2. There the multisubdomain case is considered, and coarse space correction is introduced, which is essential for robust convergence. In Chap. 2 it is observed that the matrix version of the Schwarz alternating method corresponds to a generalization (due to overlap) of the traditional block Gauss-Seidel iterative method. The *additive* Schwarz method [DR11] is also introduced there, corresponding to a generalized block Jacobi method.

1.2.3 Global Discretization

An advantage of the hybrid formulation (1.12) is that novel discretizations of (1.1) may be obtained by discretizing (1.12). Each subdomain Ω_i^* may be independently triangulated, resulting in a possibly *non-matching grid*, see Fig. 1.4. Furthermore, each local problem may be discretized using traditional techniques suited to the local geometry and properties of the solution. The resulting solution, however, may be *nonconforming* along the internal boundaries $B^{(i)}$ of the subdomains, and care must be exercised in discretizing the *matching* conditions to ensure that the global discretization is stable.

Below, we outline the construction of a global finite difference discretization of (1.12) based on a two subdomain decomposition of Ω, as in Fig. 1.4, using finite difference schemes on the subdomains. For details, see Chap. 11. We triangulate each subdomain Ω_i^* for $1 \leq i \leq 2$ by a grid $\mathcal{T}_{h_i}(\Omega i^*)$ of size h_i as in Fig. 1.4. The local triangulation can be suited to the geometry and regularity of the solution on Ω_i^*. On each subdomain, we block partition the

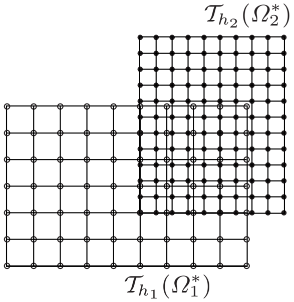

$\mathcal{T}_{h_2}(\Omega_2^*)$
$\mathcal{T}_{h_1}(\Omega_1^*)$

Fig. 1.4. Nonmatching overset grids

local discrete solution $\mathbf{w}_{h_i}$ on $\mathcal{T}_{h_i}(\Omega_i^*)$ as:

$$\mathbf{w}_{h_i} = \left(\mathbf{w}_I^{(i)}, \mathbf{w}_{B^{(i)}}^{(i)}, \mathbf{w}_{B_{[i]}}^{(i)}\right)^T, \qquad \text{for } i = 1, 2$$

corresponding to the grid points in the interior and the boundary segments $B^{(i)}$ and $B_{[i]}$, respectively. Let n_i, m_i and l_i denote the number of grid points of triangulation $\mathcal{T}_{h_i}(\Omega_i^*)$ in the interior of Ω_i^*, on $B^{(i)}$ and $B_{[i]}$, respectively. By assumption on the boundary values of $\mathbf{w}_{h_i}$ on $B_{[i]}$, it will hold that $\mathbf{w}_{B_{[i]}}^{(i)} = \mathbf{0}$. Next, for $i = 1, 2$ discretize the elliptic equation $Lw_i = f_i$ on Ω_i^* by employing a stable scheme on $\mathcal{T}_{h_i}(\Omega_i^*)$ and denote the discretization as:

$$A_{II}^{(i)}\mathbf{w}_I^{(i)} + A_{IB^{(i)}}^{(i)}\mathbf{w}_{B^{(i)}}^{(i)} = \mathbf{f}_{h_i}, \quad \text{for } 1 \le i \le 2.$$

Next, on each boundary segment $B^{(i)}$, discretize the *inter-subdomain* matching conditions $w_1 = w_2$ on $B^{(1)}$ and $w_2 = w_1$ on $B^{(2)}$ by applying appropriate *interpolation* stencils or by discretizing its weak form. If interpolation stencils are employed, then the value $w_{h_1}(x)$ at a grid point x on $B_{h_1}^{(1)}$ may be expressed as a weighted average of nodal values of $w_{h_2}(\cdot)$ on the grid points of $\Omega_{h_2}^*$. We denote the discretized matching conditions as:

$$\mathbf{w}_{B^{(1)}}^{(1)} = I_{h_2}^{h_1}\mathbf{w}_{h_2} \quad \text{and} \quad \mathbf{w}_{B^{(2)}}^{(2)} = I_{h_1}^{h_2}\mathbf{w}_{h_1}.$$

Here $I_{h_2}^{h_1}$ will denote a matrix of size $m_1 \times (n_2 + m_2 + l_2)$ and $I_{h_1}^{h_2}$ will denote a matrix of size $m_2 \times (n_1 + m_1 + l_1)$. If the local grids match on each segment $B_{[i]}$, then this discretization step would be trivial. However, for nonmatching grids care must be exercised to ensure stability of the global scheme.

The global discretization now will have the following block matrix form:

$$\begin{cases} A_{II}^{(1)}\mathbf{w}_I^{(1)} + A_{IB^{(1)}}^{(1)}\mathbf{w}_{B^{(1)}}^{(1)} = \mathbf{f}_{h_1}, \\ \qquad\qquad\qquad\quad \mathbf{w}_{B^{(1)}}^{(1)} = I_{h_2}^{h_1}\,\mathbf{w}_{h_2} \\ A_{II}^{(2)}\mathbf{w}_I^{(2)} + A_{IB^{(2)}}^{(2)}\mathbf{w}_{B^{(2)}}^{(2)} = \mathbf{f}_{h_2}, \\ \qquad\qquad\qquad\quad \mathbf{w}_{B^{(2)}}^{(2)} = I_{h_1}^{h_2}\,\mathbf{w}_{h_1}. \end{cases} \tag{1.14}$$

This algebraic system can be solved by the Schwarz alternating method.

Remark 1.11. If $c(x) \geq c_0 > 0$ and the local discretizations satisfy a discrete maximum principle, if the inter-grid interpolations $I_{h_1}^{h_2}$ and $I_{h_2}^{h_1}$ are convex weights, and if the overlap is sufficiently large so that a certain contraction property holds, see Chap. 11, then the above discretization can be shown to be stable and convergent of optimal order in the maximum norm.

1.2.4 Heterogeneous Approximation

A heterogeneous approximation of a partial differential equation is a model system of partial differential equations in which the problems posed on different subdomains are not all of the same *type*. Such approximations may be useful if there is a reduction in computational costs resulting from the use of a heterogeneous model. Here, we illustrate the construction of an elliptic-hyperbolic approximation of an advection dominated elliptic equation:

$$\begin{cases} L^{\epsilon}\, u \equiv -\epsilon\, \Delta u + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.15}$$

where $0 < \epsilon \ll 1$ is a perturbation parameter. In this case, depending on the solution u, the singularly perturbed elliptic equation may be approximately of *hyperbolic* character on some subregions and of *elliptic* character elsewhere, motivating a heterogeneous approximation.

Suppose the overlapping subdomain Ω_1^* can chosen such that:

$$|\epsilon \Delta u(x)| \ll |\mathbf{b}(x) \cdot \nabla u(x) + c(x)\, u(x)| \qquad \text{for } x \in \overline{\Omega}_1^*.$$

Then, on Ω_1^* the term $L^{\epsilon} u$ may be approximated by $L_0\, u$ defined by:

$$L_0\, u \equiv \mathbf{b}(x) \cdot \nabla u + c(x) u.$$

Motivated by singular perturbation theory [LA5, KE5], a global heterogeneous approximation of the singularly perturbed equation (1.15) may be sought by replacing the elliptic equation $L^{\epsilon}\, w_1 = f_1$ on Ω_1^* by the hyperbolic equation $L_0\, w_1 = f_1$ within the Schwarz hybrid formulation (1.12).

To ensure well posedness of the local subproblems, however, the Dirichlet boundary value problem on Ω_1^* must be replaced by suitable *inflow* boundary conditions, due to the hyperbolic nature of $L_0\, w_1 = f_1$:

$$\begin{cases} L_0\, w_1 = f_1, & \text{in } \Omega_1^* \\ \quad\; w_1 = 0, & \text{on } B_{[1],in}, \\ \quad\; w_1 = w_2, & \text{on } B_{in}^{(1)}, \end{cases}$$

where, the inflow boundary segments are defined by:

$$\begin{cases} B_{[1],in} \equiv \{x \in B_{[1]} \,:\, \mathbf{b}(x) \cdot \mathbf{n}(x) < 0\} \\ B_{in}^{(1)} \quad \equiv \{x \in B^{(1)} :\, \mathbf{b}(x) \cdot \mathbf{n}(x) < 0\}, \end{cases}$$

where $\mathbf{n}(x)$ denotes the exterior unit normal to $\partial\Omega_1^*$ at x. The resulting global heterogeneous approximation will be:

$$\begin{cases} L_0\, w_1 = f_1, & \text{in } \Omega_1^* \\ w_1 = 0, & \text{on } B_{[1],in} \\ w_1 = w_2, & \text{on } B_{in}^{(1)} \end{cases} \quad \text{and} \quad \begin{cases} L w_2 = f_2, & \text{in } \Omega_2^* \\ w_2 = 0, & \text{on } B_{[2]} \\ w_2 = w_1, & \text{on } B^{(2)}. \end{cases} \tag{1.16}$$

This heterogeneous system can be discretized, and the resulting algebraic system can be solved by the Schwarz alternating method, see Chap. 12.

Remark 1.12. Well posedness of this heterogeneous system, as well as bounds on the error resulting from such approximation are discussed in Chap. 15.

1.3 Steklov-Poincaré Framework

The hybrid formulation that we refer to as the Steklov-Poincaré framework is motivated by a principle in physics referred as a *transmission condition*, employed in the study of electric fields in conductors [PO, ST8, LE12, AG, QU5]. The underlying principle states that across any interface within a conducting medium, the electric *potential* as well as the *flux* of electric current must match, i.e., be continuous. The mathematical version of this principle suggests a hybrid formulation for a 2nd order elliptic equation given a two subdomain *non-overlapping* decomposition of its domain, separated by an interface.

1.3.1 Motivation

Consider elliptic equation (1.1) posed on Ω:

$$\begin{cases} L\, u \equiv -\nabla \cdot (a(x)\, \nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.17}$$

Let Ω_1, Ω_2 denote a non-overlapping decomposition of Ω, as in Fig. 1.5, with interface $B = \partial\Omega_1 \cap \partial\Omega_2$ separating the two subdomains and $B_{[i]} \equiv \partial\Omega_i \cap \partial\Omega$. Let $\mathbf{n}_i(x)$ denote the unit outward normal vector to $\partial\Omega_i$ at the point $x \in B$. For $i = 1, 2$, denote the solution on each subdomain Ω_i by $w_i(x) \equiv u(x)$. Then, the following *transmission conditions*, which are derived later in this section, will hold on the interface B for smooth solutions:

$$\begin{cases} w_1 = w_2, & \text{on } B \\ \mathbf{n}_1 \cdot (a \nabla w_1 - \mathbf{b}\, w_1) = \mathbf{n}_1 \cdot (a \nabla w_2 - \mathbf{b}\, w_2)\,, & \text{on } B. \end{cases} \tag{1.18}$$

The first condition requires the subdomain solutions w_1 and w_2 to match on B, while the second condition requires the local fluxes $\mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1)$ and $\mathbf{n}_1 \cdot (a\nabla w_2 - \mathbf{b}\, w_2)$ associated with w_1 and w_2 to also match on B.

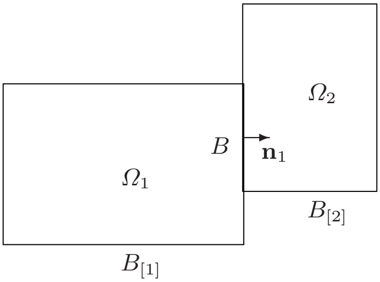


Fig. 1.5. A two subdomain non-overlapping decomposition

Combining the transmission conditions with the elliptic equation on each subdomain, yields the following *hybrid formulation* equivalent to (1.1):

$$\begin{cases} L w_1 = f, & \text{in } \Omega_1 \\ w_1 - w_2 = 0, & \text{on } B \\ w_1 = 0, & \text{on } B_{[1]} \\ L w_2 = f, & \text{in } \Omega_2 \\ \mathbf{n}_1 \cdot (a \nabla w_1 - \mathbf{b}\, w_1) + \mathbf{n}_2 \cdot (a \nabla w_2 - \mathbf{b}\, w_2) = 0, & \text{on } B \\ w_2 = 0, & \text{on } B_{[2]}. \end{cases}$$

In this section, we shall outline how this hybrid formulation can be employed to formulate novel domain decomposition iterative methods, discretization methods and heterogeneous approximations for (1.1).

Remark 1.13. If the coefficient $\mathbf{b}(x)$ in elliptic equation (1.1) is continuous, then the flux boundary condition may also be equivalently stated as:

$$\mathbf{n}_1 \cdot (a \nabla w_1) + \mathbf{n}_2 \cdot (a \nabla w_2) = 0, \qquad \text{on } B,$$

by taking linear combinations of (1.18), since $w_1(x) = w_2(x)$ on B and since $\mathbf{n}_1(x) = -\mathbf{n}_2(x)$ on B. In particular, the following equivalent flux transmission condition is *preferred* in several domain decomposition methods:

$$\mathbf{n}_1 \cdot \left(a \nabla w_1 - \frac{1}{2} \mathbf{b}\, w_1 \right) + \mathbf{n}_2 \cdot \left(a \nabla w_2 - \frac{1}{2} \mathbf{b}\, w_2 \right) = 0, \qquad \text{on } B,$$

for continuous $\mathbf{b}(x)$, see [QU6, GA14, AC7, RA3].

Equivalence of the Steklov-Poincaré hybrid formulation is shown next.

Theorem 1.14. *Suppose the following assumptions hold.*

1. *Let $L\,u$ be defined by (1.1) with smooth coefficient $\mathbf{b}(x)$ and solution u.*
2. *Let $w_1(x)$ and $w_2(x)$ be smooth solutions of the following coupled system of partial differential equations:*

$$\begin{cases} L w_1 = f, & \text{in } \Omega_1 \\ w_1 = 0, & \text{on } B_{[1]} \\ w_1 = w_2, & \text{on } B \\ L w_2 = f, & \text{in } \Omega_2 \\ w_2 = 0, & \text{on } B_{[2]} \\ \mathbf{n}_1 \cdot (a\nabla w_2 - \mathbf{b}\, w_2) = \mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1), & \text{on } B. \end{cases} \tag{1.19}$$

Then, the following result will hold.

$$\begin{cases} w_1(x) = u(x), & \text{on } \overline{\Omega}_1 \\ w_2(x) = u(x), & \text{on } \overline{\Omega}_2. \end{cases}$$

Proof. Suppose u is a smooth solution to (1.1) and $w_i \equiv u$ on $\overline{\Omega}_i$, we will verify that (w_1, w_2) solves (1.19). By construction, $Lw_i = f$ in Ω_i and $w_i = 0$ on $B_{[i]}$. By continuity of u (or an application of the trace theorem), we obtain that $w_1 = w_2$ on B. To verify that the local *fluxes* match on B, employ the following weak formulation of (1.1), and express each integral on Ω as a sum of integrals on Ω_1 and Ω_2, to obtain:

$$\sum_{i=1}^{2} \int_{\Omega_i} (a\nabla w_i \cdot \nabla v - w_i\, \nabla \cdot (\mathbf{b}\, v) + c\, w_i\, v)\; dx = \sum_{i=1}^{2} \int_{\Omega_i} f\, v\, dx,$$

for $v \in C_0^\infty(\Omega)$. If v is chosen to be of compact support in Ω and not identically zero on B, then integration by parts yields:

$$\begin{cases} \sum_{i=1}^{2} \int_{\Omega_i} -\nabla \cdot (a\nabla w_i)\; v + (\mathbf{b} \cdot \nabla w_i)\; v + c\, w_i\, v\, dx \\ \qquad - \int_B \mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1 - a\nabla w_2 + \mathbf{b}\, w_2)\; v\, ds_x = \sum_{i=1}^{2} \int_{\Omega_i} f\, v\, dx, \end{cases}$$

for $v \in C_0^\infty(\Omega)$. Substituting that $Lw_i = f$ on Ω_i, it follows that:

$$\int_B \mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1 - a\nabla w_1 + \mathbf{b}\, w_1)\; v\, ds_x = 0, \qquad \forall v \in C_0^\infty(\Omega),$$

yielding the result that $\mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1) = \mathbf{n}_1 \cdot (a\nabla w_2 + \mathbf{b}\, w_2)$ on B. The converse can be verified analogously. □

Remark 1.15. The above result only demonstrates the equivalence of solutions to both systems. It does not guarantee well posedness of hybrid formulation (1.19). This may be demonstrated using elliptic regularity theory in appropriately chosen norms (however, we shall omit this). When system (1.19) is well posed, given a solution (w_1, w_2) to (1.19), we may define:

$$u \equiv \begin{cases} w_1 & \text{in } \overline{\Omega}_1 \\ w_2 & \text{in } \overline{\Omega}_2, \end{cases}$$

thus yielding a solution u to (1.1).

We now introduce an operator, referred to as a Steklov-Poincaré operator, which represents hybrid formulation (1.19) more compactly.

Definition 1.16. *Given sufficiently smooth Dirichlet boundary data $g(\cdot)$ on the interface B, we define a Steklov-Poincaré operator $\mathcal{S}(g, f_1, f_2)$ as follows:*

$$\mathcal{S}\left(g, f_1, f_2\right) \equiv \mathbf{n}_1 \cdot \left(a \nabla w_1 - \mathbf{b}\, w_1\right) - \mathbf{n}_1 \cdot \left(a \nabla w_2 - \mathbf{b}\, w_2\right),$$

where $w_1(\cdot)$ and $w_2(\cdot)$ are solutions to the following problems:

$$\begin{cases} L w_1 = f_1, & \textit{in } \Omega_1 \\ \quad w_1 = 0, & \textit{on } B_{[1]} \\ \quad w_1 = g, & \textit{on } B, \end{cases} \textit{ and } \begin{cases} L w_2 = f_2, & \textit{in } \Omega_2 \\ \quad w_2 = 0 & \textit{on } B_{[2]} \\ \quad w_2 = g, & \textit{on } B. \end{cases} \tag{1.20}$$

If the local forcing terms $f_1(\cdot)$ and $f_2(\cdot)$ are nonzero, then the action of the Steklov-Poincaré operator $\mathcal{S}(g, f_1, f_2)$ on $g(\cdot)$ will be *affine* linear. It will map the Dirichlet data $g(\cdot)$ on B to the *jump* in the local fluxes (Neumann data) across interface B using (1.20). Importantly, if an interface function $g(\cdot)$ can be found which yields *zero* jump in the flux across B, i.e.

$$\mathcal{S}(g, f_1, f_2) = 0, \tag{1.21}$$

then, corresponding to this choice of interface data $g(\cdot)$, the local solutions $w_1(\cdot)$ and $w_2(\cdot)$ to (1.20) will satisfy:

$$\begin{cases} w_1 = w_2 \quad (= g), & \text{on } B \\ \mathbf{n}_1 \cdot \left(a \nabla w_1 - \mathbf{b}\, w_1\right) = \mathbf{n}_1 \cdot \left(a \nabla w_2 - \mathbf{b}\, w_2\right), & \text{on } B, \end{cases}$$

so that (w_1, w_2) will solve (1.19). As a result, the search for a solution (w_1, w_2) to problem (1.19) may be *reduced* to the search for interface data $g(\cdot)$ which solves the Steklov-Poincaré problem (1.21). For such interface data $g(\cdot)$, the local solutions (w_1, w_2) to (1.20) will yield the solution to (1.19) with $g(x) = u(x)$ on B. When a weak formulation is used, if X denotes the space of Dirichlet data on B, the flux or Neumann data will belong to its dual space X', where $X = H_{00}^{1/2}(B)$ for a standard subdomain decomposition and $X = H^{1/2}(B)$ for an immersed subdomain decomposition.

Remark 1.17. For computational purposes, the Steklov-Poincaré operator $\mathcal{S}$ may be expressed as the sum of two subdomain operators:

$$\mathcal{S}(g, f_1, f_2) \equiv \mathcal{S}^{(1)}(g, f_1) + \mathcal{S}^{(2)}(g, f_2),$$

where

$$\begin{cases} \mathcal{S}^{(1)}(g, f_1) \equiv \mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1) \\ \mathcal{S}^{(2)}(g, f_2) \equiv \mathbf{n}_2 \cdot (a\nabla w_2 - \mathbf{b}\, w_2)\,, \end{cases}$$

for w_1 and w_2 defined by (1.20). By definition, each operator $\mathcal{S}^{(i)}$ will require only subdomain information and will be affine linear.

Remark 1.18. Both $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ map the *Dirichlet* interface data $g(\cdot)$ prescribed on B to the corresponding *Neumann* flux data $\mathbf{n}_1 \cdot (a\nabla w_1 - \mathbf{b}\, w_1)$ and $\mathbf{n}_2 \cdot (a\nabla w_2 - \mathbf{b}\, w_2)$ on B, respectively, obtained by solution of the local problems (1.20). As a result, the maps $\mathcal{S}^{(i)}$ are commonly referred to as local *Dirichlet to Neumann* maps. These Dirichlet to Neumann maps are *not* differential operators since the solutions w_i to (1.20) have representations as integral operators acting on the data g. They are referred to as *pseudo-differential operators*, and for the correct choice of Dirichlet interface data $g(\cdot)$ on B, the jump in the Neumann data on B will be zero for the local solutions.

In the rest of this section, we outline how iterative methods, global discretizations and heterogeneous approximations can be constructed for the original problem (1.1) using the Steklov-Poincaré formulation (1.19).

1.3.2 Iterative Methods

The block structure of the Steklov-Poincaré system (1.19) suggests various iterative algorithms for its solution. For instance, if $w_1^{(k)}$ and $w_2^{(k)}$ denote the k'th iterates on subdomains Ω_1 and Ω_2, respectively, then the system of equations posed on subdomain Ω_i in (1.19) can be solved to yield updates $w_i^{(k+1)}$ for the local solutions, with boundary conditions chosen using preceding iterates. The resulting iterative algorithm sequentially enforces either the *continuity* or *flux* transmission boundary conditions on B, and is referred to as a Dirichlet-Neumann algorithm as it requires the solution of Dirichlet and Neumann boundary value problems. In the following, suppose that $\mathbf{b}(x) = \mathbf{0}$ in Ω, and let $0 < \theta < 1$ denote a *relaxation* parameter required to ensure convergence [BJ9, BR11, FU, MA29].

Algorithm 1.3.1 *(Dirichlet-Neumann Algorithm)*
Let $v_2^{(0)}$ (where $v_2^{(0)} \equiv w_2^{(0)}$ on B) denote a starting guess.

```
1. For $k = 0, 1, \cdots$, until convergence do:
2.     Solve for $w_1^{(k+1)}$ as follows:
           $\begin{cases} L w_1^{(k+1)} = f_1, & \text{in } \Omega_1 \\ w_1^{(k+1)} = v_2^{(k)}, & \text{on } B \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]}, \end{cases}$
3.     Solve for $w_2^{(k+1)}$ as follows:
           $\begin{cases} L w_2^{(k+1)} = f_2, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \mathbf{n}_2 \left( a \nabla w_2^{(k+1)} \right) = \mathbf{n}_2 \left( a \nabla w_1^{(k+1)} \right), & \text{on } B. \end{cases}$
4.     Update: $v_2^{(k+1)} = \theta\, w_2^{(k+1)} + (1 - \theta) v_2^{(k)}$ on $B$.
5. Endfor
```

Output: $(w_1^{(k)}, w_2^{(k)})$

Remark 1.19. In step 2, the local solution $w_1^{(k+1)}$ matches $v_2^{(k)}$ on B (however, the local fluxes may not match on B). This step requires the solution of an elliptic equation on Ω_1 with Dirichlet conditions on $B_{[1]}$ and B. In step 3, the flux of $w_2^{(k+1)}$ matches the flux of $w_1^{(k+1)}$ on B (though $w_2^{(k+1)}$ may not match $w_1^{(k+1)}$ on B). This step requires the solution of an elliptic equation on Ω_2 with Dirichlet conditions on $B_{[2]}$ and Neumann conditions on B. A matrix formulation of this algorithm is given in Chap. 3.

Remark 1.20. Under restrictions on the coefficients (such as $\mathbf{b}(x) \equiv 0$ and $c(x) \geq 0$), and additional restrictions on the parameter $0 < \theta < 1$, the iterates $w_i^{(k)}$ in the Dirichlet-Neumann algorithm will converge geometrically to the true local solution w_i of (1.19) as $k \to \infty$, see [FU, MA29].

The preceding Dirichlet-Neumann algorithm has sequential steps. Various algorithms have been proposed which solve subdomain problems in parallel, see [BO7, DE3, DR18, MA14, DO13, QU6, GA14, AC7, RA3]. Multidomain matrix versions of such algorithms are described in Chap. 3. Below, we describe a two fractional step algorithm, each step requiring the solution of subdomain problems in parallel [DO13, DO18, YA2]. We assume $\mathbf{b}(x) = \mathbf{0}$.

Algorithm 1.3.2 *(A Parallel Dirichlet-Neumann Algorithm)*
Let $w_1^{(0)}$ and $w_2^{(0)}$ denote a starting guess on each subdomain.
Let $0 < \theta, \delta, \beta, \alpha < 1$ denote relaxation parameters.

```
1. For $k = 0, 1, \cdots$, until convergence do:
2.     Compute $\begin{cases} \boldsymbol{\mu}^{(k+\frac{1}{2})} = \theta\, \mathbf{n}_1 \cdot \left(a \nabla w_1^{(k)}\right) + (1-\theta)\, \mathbf{n}_1 \cdot \left(a \nabla w_2^{(k)}\right), & \text{on } B \\ \mathbf{g}^{(k+\frac{1}{2})} = \delta\, w_1^{(k)} + (1-\delta)\, w_2^{(k)}, & \text{on } B. \end{cases}$
3.     In parallel solve for $w_1^{(k+\frac{1}{2})}$ and $w_2^{(k+\frac{1}{2})}$
       $\begin{cases} L w_1^{(k+\frac{1}{2})} = f, & \text{in } \Omega_1 \\ w_1^{(k+\frac{1}{2})} = 0, & \text{on } B_{[1]} \\ \mathbf{n}_1 \cdot \left(a \nabla w_1^{(k+\frac{1}{2})}\right) = \boldsymbol{\mu}^{(k+\frac{1}{2})}, & \text{on } B, \end{cases}$ and $\begin{cases} L w_2^{(k+\frac{1}{2})} = f, & \text{in } \Omega_2 \\ w_2^{(k+\frac{1}{2})} = 0, & \text{on } B_{[2]} \\ w_2^{(k+\frac{1}{2})} = \mathbf{g}^{(k+\frac{1}{2})}, & \text{on } B, \end{cases}$
4.     Compute $\begin{cases} \boldsymbol{\mu}^{(k+1)} = \beta\, \mathbf{n}_2 \cdot \left(a \nabla w_1^{(k+\frac{1}{2})}\right) + (1-\beta)\, \mathbf{n}_2 \cdot \left(a \nabla w_2^{(k+\frac{1}{2})}\right), \\ \quad \text{on } B \\ \mathbf{g}^{(k+1)} = \alpha\, w_1^{(k+\frac{1}{2})} + (1-\alpha)\, w_2^{(k+\frac{1}{2})}, \\ \quad \text{on } B. \end{cases}$
5.     In parallel solve for $w_1^{(k+1)}$ and $w_2^{(k+1)}$
       $\begin{cases} L w_1^{(k+1)} = f, & \text{in } \Omega_1 \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \\ w_1^{(k+1)} = \mathbf{g}^{(k+1)}, & \text{on } B, \end{cases}$ and $\begin{cases} L w_2^{(k+1)} = f, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \mathbf{n}_2 \cdot \left(a \nabla w_2^{(k+1)}\right) = \boldsymbol{\mu}^{(k+1)}, & \text{on } B, \end{cases}$
6. Endfor
Output: $(w_1^{(k)}, w_2^{(k)})$
```

Remark 1.21. Under appropriate restrictions on the coefficients $a(x)$ and $c(x)$, and the relaxation parameters $\theta, \delta, \beta, \alpha$, this parallel algorithm will converge geometrically [YA2]. For related parallel algorithms, see [DO13, DO18].

When the advection coefficient $\mathbf{b}(x) \neq 0$, a parallel algorithm, referred to as a Robin-Robin algorithm can also be used [QU6, GA14, AC7, RA3]. Let:

$$\Phi_i(w) \equiv \mathbf{n}_i \cdot \left(a(x) \nabla w - \frac{1}{2} \mathbf{b}(x)\, w \right) + z_i(x)\, w,$$

denote a local Robin boundary operator on B for $i = 1, 2$ for an appropriately chosen bounded interface function $z_i(x) > 0$. For convenience, $\tilde{i}$ will denote a complementary index to i (namely, $\tilde{i} = 2$ when $i = 1$ and $\tilde{i} = 1$ when $i = 2$). Then, the Robin-Robin algorithm has the following form.

Algorithm 1.3.3 *(A Robin-Robin Algorithm)*

```
Let $w_1^{(0)}$ and $w_2^{(0)}$ denote a starting guess on each subdomain
Let $0 < \theta < 1$ denote a relaxation parameter
1. For $k = 0, 1, \cdots$, until convergence do:
2.      For $i = 1, 2$ in parallel solve:
            $\left\{ \begin{array}{rll} L w_i^{(k+1)} &= f_i, & \text{in } \Omega_i \\ w_i^{(k+1)} &= 0, & \text{on } B_{[i]} \\ \Phi_i\left(w_i^{(k+1)}\right) &= \theta\, \Phi_i\left(w_i^{(k)}\right) + (1-\theta)\, \Phi_{\tilde{i}}\left(w_{\tilde{i}}^{(k)}\right), & \text{on } B \end{array} \right.$
3.      Endfor
4. Endfor
Output: $(w_1^{(k)}, w_2^{(k)})$
```

Remark 1.22. When $(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)) \geq \beta > 0$, the Robin-Robin iterates will converge geometrically, for a suitable choice of relaxation parameter $0 < \theta < 1$ and $z_i(x) > 0$, see [QU6, GA14, AC7, RA3].

1.3.3 Global Discretization

Hybrid formulation (1.19) can be used to construct a global discretization of (1.1). Such discretizations have not been studied extensively, however, see [AG, AG2, DO4] and in the context of spectral methods, see [MA4, PH]. A potential advantage of discretizing (1.19) is that each subdomain Ω_i can be independently triangulated, see Fig. 1.6, by methods suited to the local geometry and regularity of the local solution, and each subproblem may be discretized independently. However, care must be exercised in discretizing the transmission conditions so that the resulting global discretization is stable. Below, we *heuristically* outline the general stages that would be involved in discretizing (1.19) using finite element methods.

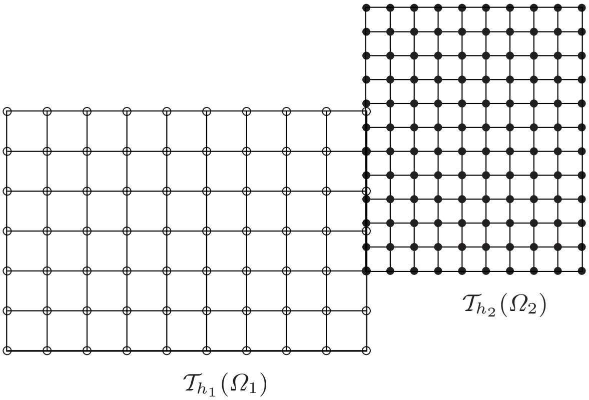


Fig. 1.6. Nonmatching local grids

On each subdomain Ω_i, generate a grid $\mathcal{T}_{h_i}(\Omega_i)$ of size h_i suited to the local geometry and solution. If the resulting local grids do not match along B, as in Fig. 1.6, they will be referred to as nonmatching grids. On each subdomain Ω_i, employ a traditional method to discretize the following Neumann problem:

$$\begin{cases} L w_i = f, & \text{in } \Omega_i \\ w_i = 0, & \text{on } B_{[i]} \\ \mathbf{n}_i \cdot (a \nabla w_i - \mathbf{b}\, w_i) = g_i, & \text{on } B, \end{cases}$$

where $\mathbf{n}_i$ denotes the exterior unit normal to $\partial\Omega_i$ and the flux data g_i is to be chosen when the transmission conditions are applied. Employing block matrix notation, denote the resulting local discretization by:

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{BI}^{(i)} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(i)} \\ \mathbf{w}_B^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{h_i} \\ \mathbf{g}_{h_i} \end{bmatrix},$$

where $\mathbf{w}_I^{(i)}$ denotes the interior unknowns on Ω_{h_i} and $\mathbf{w}_B^{(i)}$ denotes the boundary unknowns on B associated with the discrete solution on $\mathcal{T}_{h_i}(\Omega_i)$. Separately discretize the two transmission conditions on B:

$$\begin{cases} w_1 = w_2, & \text{on } B \\ \mathbf{n}_1 \cdot (a \nabla w_1 - \mathbf{b}\, w_1) = \mathbf{n}_1 \cdot (a \nabla w_2 - \mathbf{b}\, w_2), & \text{on } B. \end{cases}$$

Since the grid functions $(\mathbf{w}_I^{(i)}, \mathbf{w}_B^{(i)})$ may be nonmatching on B, care must be exercised to ensure well posedness and stability of this discretization.

Below, we indicate how each transmission condition can be discretized by a "mortar" element type method. Let n_i and m_i denote the number of unknowns in $\mathbf{w}_I^{(i)}$ and $\mathbf{w}_B^{(i)}$ respectively. Then the continuity equation $w_1 = w_2$ on B may be discretized by a Petrov-Galerkin approximation of its weak form:

$$\int_B (w_1 - w_2)\, v \, ds_x = 0, \qquad v \in X_h(B),$$

where $X_h(B)$ denotes some appropriately chosen subspace of $L^2(B)$. In a mortar element discretization, $X_h(B)$ is typically chosen as a finite element space defined on a triangulation of B inherited from either triangulation $\mathcal{T}_{h_1}(\Omega_1)$ or $\mathcal{T}_{h_2}(\Omega_2)$. Examples of such spaces are described in Chap. 11. For definiteness suppose $X_h(B) = X_{h_1}(B)$ is chosen to be of dimension m_1 based on the triangulation of B inherited from $\mathcal{T}_{h_1}(\Omega_1)$. Then, the discretized continuity transmission condition will have the following matrix form:

$$M_{11} \mathbf{w}_B^{(1)} = M_{12} \mathbf{w}_B^{(2)},$$

where M_{11} and M_{12} are $m_1 \times m_1$ and $m_1 \times m_2$ mass matrices, respectively.

The flux transmission condition on B may be similarly discretized:

$$\int_B \left(\mathbf{n}_1 \cdot (a \nabla w_1 - \mathbf{b}\, w_1) - \mathbf{n}_1 \cdot (a \nabla w_2 - \mathbf{b}\, w_2) \right) \mu \, ds_x = 0, \qquad \forall \mu \in Y_h(B),$$

where it is sufficient to choose $Y_h(B) \subset H_0^1(B)$. Again, $Y_h(B)$ may be chosen as a finite element space defined on the triangulation of B inherited from either triangulation Ω_{h_1} or Ω_{h_2}. However, to ensure that the total number of equations equals the total number of unknowns in the global system, it will be preferable that $Y_h(B)$ be chosen using the complementary triangulation. In the above example, since $X_h(B) = X_{h_1}(B)$ is of dimension m_1, we choose $Y_h(B) = Y_{h_2}(B)$ of dimension m_2 based on triangulation Ω_{h_2}. This will yield m_2 constraints, which we denote as:

$$M_{21}\left(A_{BI}^{(1)}\mathbf{w}_I^{(1)} + A_{BB}^{(1)}\mathbf{w}_B^{(1)} - \mathbf{f}_B^{(1)}\right) = -M_{22}\left(A_{BI}^{(2)}\mathbf{w}_I^{(2)} + A_{BB}^{(2)}\mathbf{w}_B^{(2)} - \mathbf{f}_B^{(2)}\right),$$

where M_{21} and M_{22} are $m_2 \times m_1$ and $m_2 \times m_2$ matrices, respectively. The interface forcing terms $\mathbf{f}_B^{(i)}$ have been added to account for the approximation resulting from integration by parts. The actual choice of subspaces $X_{h_1}(B)$ and $Y_{h_2}(B)$ will be *critical* to the stability of the resulting global discretization:

$$\begin{cases} A_{II}^{(1)}\mathbf{w}_I^{(1)} + A_{IB}^{(1)}\mathbf{w}_B^{(1)} = \mathbf{f}_{h_1} \\ M_{11}\mathbf{w}_B^{(1)} = M_{12}\mathbf{w}_B^{(2)} \\ A_{II}^{(2)}w_I^{(2)} + A_{IB}^{(2)}\mathbf{w}_B^{(2)} = \mathbf{f}_{h_2} \\ M_{21}\left(A_{BI}^{(1)}\mathbf{w}_I^{(1)} + A_{BB}^{(1)}w_B^{(1)} - \mathbf{f}_B^{(1)}\right) = -M_{22}\left(A_{BI}^{(2)}\mathbf{w}_I^{(2)} + A_{BB}^{(2)}\mathbf{w}_B^{(2)} - \mathbf{f}_B^{(2)}\right). \end{cases}$$

General theoretical results on the stability of such discretizations of (1.19) are not known to the author, and this scheme was heuristically considered only for its intrinsic interest.

Remark 1.23. If the grids $\mathcal{T}_{h_1}(\Omega_1)$ and $\mathcal{T}_{h_2}(\Omega_2)$ match on B, then $m_1 = m_2$. We would then obtain $M_{11} = M_{12}$, both square and nonsingular, yielding:

$$\mathbf{w}_B^{(1)} = \mathbf{w}_B^{(2)}.$$

Similarly, $M_{21} = M_{22}$ will be square and nonsingular yielding:

$$\left(A_{BI}^{(1)}\mathbf{w}_I^{(1)} + A_{BB}^{(1)}w_B^{(1)} - \mathbf{f}_B^{(1)}\right) = -\left(A_{BI}^{(2)}\mathbf{w}_I^{(2)} + A_{BB}^{(2)}\mathbf{w}_B^{(2)} - \mathbf{f}_B^{(2)}\right).$$

The resulting global discretization will then correspond to the standard finite element discretization of (1.1).

1.3.4 Heterogeneous Approximations

A heterogeneous approximation of a partial differential equation is a coupled system of partial differential equations which approximates the given equation, in which the approximating partial differential equations are not of the same *type* in different subregions [GA15, QU6]. In the following, motivated by classical singular perturbation approximations [KE5, LA5], we heuristically outline how an elliptic-hyperbolic heterogeneous approximation can be constructed for the following singularly perturbed elliptic equation:

$$\begin{cases} L^{\epsilon}\, u \;\equiv\; -\epsilon\, \Delta u + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad u = g, & \text{on } \partial\Omega, \end{cases} \tag{1.22}$$

where $0 < \epsilon \ll 1$ is a perturbation parameter. The Steklov-Poincaré hybrid formulation (1.19) will be employed to heuristically approximate (1.22).

Suppose Ω_1 and Ω_2 form a non-overlapping decomposition of Ω such that:

$$\epsilon\, |\Delta u| \ll |\mathbf{b} \cdot \nabla u + c\, u|\,, \qquad \text{on } \overline{\Omega}_1.$$

Then, on subdomain Ω_1, we may approximate $L^{\epsilon} u = f$ by $L_0\, u = f$, where $L_0\, u \equiv \mathbf{b}(x) \cdot \nabla u + c(x) u$. Formally, a global heterogeneous approximation of (1.22) may be obtained by substituting the preceding approximation in the hybrid formulation corresponding to (1.22), yielding:

$$\begin{cases} L_0\, w_1 = f, & \text{in } \Omega_1 \\ w_1 = 0, & \text{on } B_{[1]} \\ w_1 = w_2, & \text{on } B \\ L^{\epsilon} w_2 = f, & \text{in } \Omega_2 \\ w_2 = 0, & \text{on } B_{[2]} \\ \mathbf{n}_1 \cdot (\epsilon \nabla w_2 - \mathbf{b}\, w_2) = \mathbf{n}_1 \cdot (\epsilon \nabla w_1 - \mathbf{b}\, w_1)\,. & \text{on } B, \end{cases}$$

However, retaining the Dirichlet boundary conditions on B and $B_{[1]}$ for $w_1(.)$ will yield an *ill-posed* problem for $w_1(.)$, since $L_0 w_1$ is hyperbolic on Ω_1. Indeed, denote the *inflow* and *outflow* boundary segments on B and $B_{[1]}$ by:

$$\begin{cases} B_{in} & \equiv \{x \in B : \mathbf{n}_1 \cdot \mathbf{b}(x) < 0\} \\ B_{out} & \equiv \{x \in B : \mathbf{n}_1 \cdot \mathbf{b}(x) > 0\} \\ B_{[1],in} & \equiv \{x \in B_{[1]} : \mathbf{n}_1 \cdot \mathbf{b}(x) < 0\}. \end{cases}$$

Since $L_0 w_1 = f$ is hyperbolic, specification of Dirichlet or Neumann boundary conditions on the entire boundary $\partial\Omega_1$ will yield a locally ill posed problem. Fortunately, replacing the Dirichlet conditions by *inflow* conditions, resolves this local ill-posedness on Ω_1, see [GA15, QU6].

Thus, the boundary conditions $w_1 = 0$ on $B_{[1]}$ and $w_1 = w_2$ on B can be replaced by inflow boundary conditions $w_1 = 0$ on $B_{[1],in}$ and $w_1 = w_2$ on B_{in}, respectively. To deduce the remaining transmission boundary conditions in the heterogeneous approximation, a subdomain vanishing viscosity approach may be employed as in [GA15]. Accordingly, the elliptic equation $L^{\epsilon} u = f$ may be approximated by the discontinuous coefficient elliptic problem:

$$\begin{cases} L^{\epsilon,\eta}\, v = f, & \text{on } \Omega \\ \qquad v = 0, & \text{on } \partial\Omega, \end{cases}$$

where $L^{\epsilon,\eta}\, v \equiv -\nabla\cdot(a(x,\eta)\nabla v) + \mathbf{b}(x)\cdot\nabla v + c(x)\, v$ and $a(x,\eta)$ is defined by:

$$a(x,\eta) \equiv \begin{cases} \eta & \text{for } x \in \Omega_1 \\ \epsilon & \text{for } x \in \Omega_2. \end{cases}$$

For $\epsilon > 0$ and $\eta > 0$, the problem will be elliptic and the traditional transmission conditions should hold:

$$\begin{cases} w_1 = w_2, & \text{on } B \\ \mathbf{n}_1\cdot(\eta\nabla w_1 - \mathbf{b}\, w_1) = \mathbf{n}_1\cdot(\epsilon\nabla w_2 - \mathbf{b}\, w_2)\,, & \text{on } B. \end{cases}$$

However, letting $\eta \to 0^+$, and imposing the *inflow* condition on B_{in} yields:

$$\begin{cases} w_1 = w_2, & \text{on } B_{in} \\ -\mathbf{n}_1\cdot\mathbf{b}\, w_1 = \mathbf{n}_1\cdot(\epsilon\nabla w_2 - \mathbf{b}\, w_2)\,, & \text{on } B. \end{cases}$$

When $\mathbf{b}(x)$ is continuous, the substitution that $w_1 = w_2$ on B_{in} will yield the following additional simplifications:

$$\begin{cases} w_1 = w_2, & \text{on } B_{in} \\ 0 = \mathbf{n}_1\cdot\epsilon\nabla w_2, & \text{on } B_{in} \\ -\mathbf{n}_1\cdot\mathbf{b}\, w_1 = \mathbf{n}_1\cdot(\epsilon\nabla w_2 - \mathbf{b}\, w_2)\,, & \text{on } B_{out}. \end{cases}$$

As a result, heuristically, the global system of partial differential equations satisfied by the weak limit of the solutions $v^{\epsilon,\eta}$ as $\eta \to 0$ will be:

$$\begin{cases} L_0\, w_1 = f, & \text{in } \Omega_1 \\ w_1 = 0, & \text{on } B_{[1],in} \\ w_1 = w_2, & \text{on } B_{in} \\ L^{\epsilon} w_2 = f, & \text{in } \Omega_2 \\ \mathbf{n}_2\cdot(\epsilon\nabla w_2 - \mathbf{b}\, w_2) = -\mathbf{n}_2\cdot\mathbf{b}\, w_1, & \text{on } B_{out}, \\ \mathbf{n}_2\cdot\nabla w_2 = 0, & \text{on } B_{in}, \\ w_2 = 0, & \text{on } B_{[2]}. \end{cases}$$

Dirichlet-Neumann iterative methods can be formulated to solve the above heterogeneous approximation to (1.22), see [GA15, QU6] and Chap. 12.

Remark 1.24. For rigorous results on the well posedness of the preceding heterogeneous system, readers are referred to [GA15].

1.4 Lagrange Multiplier Framework

The framework we refer to as the Lagrange multiplier formulation [GL, GL7], underlies a variety of *non-overlapping* domain decomposition methods. It is employed in the FETI (Finite Element Tearing and Interconnection) method (a constrained optimization based parallel iterative method [FA16, FA15]), the mortar element method (a method for discretizing elliptic equations on nonmatching grids [MA4, BE22, BE18, BE6, BE4, WO4, WO5]), and in non-overlapping Schwarz iterative methods [LI8, GL8]. In this section, we illustrate its application to formulate *iterative* algorithms, *non-matching grid* discretizations and *heterogeneous* approximations.

The Lagrange multiplier framework is applicable only when there is an *optimization* principle associated with the elliptic equation. Thus, the solution u must optimize some *energy functional* $J(\cdot)$. For such a property to hold, the elliptic equation (1.1) must be *self adjoint* and *coercive*, requiring that $\mathbf{b}(x) = 0$ and $c(x) \geq 0$. Accordingly, in this section we shall consider:

$$\begin{cases} L\,u \equiv -\nabla \cdot (a(x)\,\nabla u) + c(x)\,u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\quad\; u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.23}$$

with $c(x) \geq 0$. It is well known that the solution u minimizes an energy $J(.)$, see (1.24) and (1.25) within $H_0^1(\Omega)$. Given any non-overlapping subdomain decomposition of Ω, we will show that the optimization problem (1.24) can be reformulated as a *constrained optimization* problem based on the subdomains. The Lagrange multiplier hybrid formulation will be the saddle point problem associated with this constrained minimization problem.

1.4.1 Motivation

Let Ω_1 and Ω_2 form a *non-overlapping* decomposition of the domain Ω of elliptic equation (1.23), see Fig. 1.7. Using this decomposition of Ω, we may decompose the energy functional $J(\cdot)$ associated with (1.23) as a sum of energy contributions $J_i(\cdot)$ from each subdomain Ω_i. The resulting sum of local energies will be well defined even if the local displacement functions are discontinuous across the interface $B = \partial\Omega_1 \cap \partial\Omega_2$. It is thus an *extended* energy functional.

A constrained minimization problem equivalent to the minimization of $J(.)$ can be obtained by minimizing this extended energy functional, subject to the constraint that the local displacements *match* on the interface B. The Lagrange multiplier hybrid formulation is the saddle point problem associated with this constrained minimization problem. We outline the steps below.

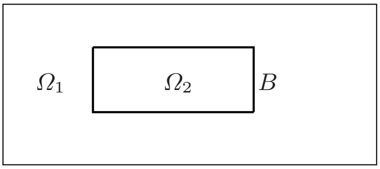


Fig. 1.7. An immersed non-overlapping decomposition

Minimization Formulation. It is well known, see [ST14, CI2, JO2, BR28], that the solution u to (1.23) minimizes the energy $J(\cdot)$ associated with (1.23):

$$J(u) = \min_{w \in X} J(w), \tag{1.24}$$

where

$$\begin{cases} J(w) & \equiv \frac{1}{2}\mathcal{A}(w,w) - F(w), \\ \mathcal{A}(v,w) & \equiv \int_\Omega (a\nabla v \cdot \nabla w + c\, vw)\, dx, \quad \text{for } v,w \in X \\ F(w) & \equiv \int_\Omega f w dx, \quad \text{for } w \in X, \\ X & \equiv H_0^1(\Omega). \end{cases} \tag{1.25}$$

Constrained Minimization Formulation. Let $\{\Omega_i\}_{i=1}^2$ be a non-overlapping decomposition of Ω. Suppose $w_i \equiv w$ on $\overline{\Omega}_i$ for $1 \le i \le 2$. We may express the energy $J(w) = J_{\mathcal{E}}(w_1, w_2) \equiv J_1(w_1) + J_2(w_2)$, where:

$$\begin{cases} J_{\mathcal{E}}(w_1,w_2) \equiv J_1(w_1) + J_2(w_2), & \text{for } w_i \in X_i \\ J_i(w_i) \equiv \frac{1}{2}\mathcal{A}_i(w_i,w_i) - F_i(w_i), & \text{for } w_i \in X_i \\ \mathcal{A}_i(v_i,w_i) \equiv \int_{\Omega_i} (\nabla v_i \cdot a\nabla w_i + c v_i w_i)\, dx, & \text{for } v_i, w_i \in X_i \\ F_i(w_i) \equiv \int_{\Omega_i} f w_i dx, & \text{for } w_i \in X_i, \\ X_i \equiv \left\{ v \in H^1(\Omega_i) : v = 0 \text{ on } B_{[i]} \right\}. \end{cases}$$

Here $J_{\mathcal{E}}(w_1, w_2)$ is defined even when $w_1 \neq w_2$ on B. To obtain a *constrained* minimization problem equivalent to (1.24), we minimize $J_{\mathcal{E}}(v_1, v_2)$ within the larger (extended) class of functions $X_1 \times X_2$ defined above, but subject to the *weak* constraint that the subdomain functions *match* on B:

$$m\left((v_1,v_2),\mu\right) \equiv \int_B (v_1 - v_2)\, \mu\, ds_x = 0, \quad \forall \mu \in Y,$$

where $Y \equiv H_{00}^{-1/2}(B)$ (the dual space of $H_{00}^{1/2}(B)$). Problem (1.24) will thus be formally equivalent to the following constrained minimization problem:

$$J_1(w_1) + J_2(w_2) = \min_{(v_1,v_2) \in \mathcal{K}} J_1(v_1) + J_2(v_2), \tag{1.26}$$

where

$$\mathcal{K} \equiv \left\{ (v_1,v_2) \in X_1 \times X_2 \,:\, m\left((v_1,v_2),\mu\right) = 0, \ \forall \mu \in Y \right\}.$$

Saddle Point Formulation. By optimization theory, see [CI4] and Chap. 10, the solution (w_1, w_2) to the constrained minimization problem (1.26) can be expressed as components in the saddle point $((w_1, w_2), \mu)$ of an associated *Lagrangian* functional $\mathcal{L}(\cdot, \cdot)$, where $\mu \in Y$ denotes an artificially introduced variable referred to as a Lagrange multiplier. We define the Lagrangian function for $((v_1, v_2), \eta) \in X_1 \times X_2 \times Y$ as:

$$\mathcal{L}\left((v_1,v_2),\eta\right) \equiv J_1(v_1) + J_2(v_2) + m\left((v_1,v_2),\eta\right). \tag{1.27}$$

At the saddle point $((w_1, w_2), \mu) \in X_1 \times X_2 \times Y$ of $\mathcal{L}(\cdot)$, we obtain:

$$\mathcal{L}((w_1, w_2), \eta) \leq \mathcal{L}((w_1, w_2), \mu) \leq \mathcal{L}((v_1, v_2), \mu) \tag{1.28}$$

for any choice of $(v_1, v_2) \in X_1 \times X_2$ and $\eta \in Y$. Requiring the first order variation at the saddle point $((w_1, w_2), \mu)$ to be zero yields:

$$\begin{cases} \sum_{i=1}^{2} \mathcal{A}_i(w_i, v_i) + m((v_1, v_2), \mu) = \sum_{i=1}^{2} F_i(v_i), & \text{for } v_i \in X_i \\ m((w_1, w_2), \eta) \qquad\qquad\qquad\qquad = 0, & \text{for } \eta \in Y. \end{cases} \tag{1.29}$$

The above system is referred to as a saddle point problem.

Hybrid Formulation. If we integrate the weak form (1.29) by parts, we can express it in terms of partial differential equations involving $w_1(.)$, $w_2(.)$ and the Lagrange multiplier variable $\mu(.)$ as follows. We seek (w_1, w_2, μ) satisfying:

$$\begin{cases} L w_1 = & f, \ \text{in } \Omega_1 \\ w_1 = & 0, \ \text{on } B_{[1]} \\ \mathbf{n}_1 \cdot (a \nabla w_1) = & -\mu, \ \text{on } B \\ L w_2 = & f, \ \text{in } \Omega_2 \\ w_2 = & 0, \ \text{on } B_{[2]} \\ \mathbf{n}_2 \cdot (a \nabla w_1) = & \mu, \ \text{on } B \\ w_1 = & w_2, \ \text{on } B \end{cases} \tag{1.30}$$

where $B_{[i]} \equiv \partial\Omega_i \cap \partial\Omega$ is the exterior boundary and $\mathbf{n}_i$ is the unit exterior normal to $\partial\Omega_i$ for $i = 1, 2$. For each choice of Neumann data $\mu(\cdot)$, each subdomain problem for $w_i(.)$ will be *uniquely* solvable provided $B_{[i]} \neq \emptyset$. We must choose the Lagrange multiplier $\mu(.)$ (representing the flux on B) so that $w_1 = w_2$ on B. The next result indicates the equivalence of (1.30) to (1.23).

Theorem 1.25. *Suppose the following assumptions hold.*

1. *Let u be a solution to (1.23).*
2. *Let (w_1, w_2, μ) be a solution to the hybrid formulation (1.30).*

Then $u(x) = w_1(x)$ in $\overline{\Omega}_1$ and $u(x) = w_2(x)$ in $\overline{\Omega}_2$.

Proof. The equivalence follows since (1.23) is equivalent to (1.19), and since (1.30) is equivalent to (1.19) for the substitution $\mu = \mathbf{n}_2 \cdot (a \nabla u)$ on B. □

Remark 1.26. The preceding result only asserts the equivalence between solutions of (1.23) and (1.30). It does not demonstrate the well posedness of (1.30). The latter can be demonstrated for (1.30) by employing general results on the well posedness of the saddle point problem (1.29) associated with it [GI3].

1.4.2 Iterative Methods

Since the Lagrange multiplier $\mu(.)$ determines $w_1(.)$ and $w_2(.)$ in (1.30), an iterative method for solving (1.23) can be obtained by applying a saddle point iterative algorithm such as Uzawa's method, see Chap. 10, to update the Lagrange multiplier function $\mu(\cdot)$, as described below.

Algorithm 1.4.1 *(Uzawa's Method)*
Let $\mu^{(0)}$ denote a starting guess with chosen step size $\tau > 0$.

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     Determine $w_1^{(k+1)}$ and $w_2^{(k+1)}$ in parallel:
           $\begin{cases} -\nabla \cdot \left(a\nabla w_1^{(k+1)}\right) + c\, w_1^{(k+1)} = f, & \text{in } \Omega_1 \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \\ \mathbf{n}_1 \cdot \left(a\nabla w_1^{(k+1)}\right) = -\mu^{(k)}, & \text{on } B, \\ -\nabla \cdot \left(a\nabla w_2^{(k+1)}\right) + c\, w_2^{(k+1)} = f, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \mathbf{n}_2 \cdot \left(a\nabla w_2^{(k+1)}\right) = \mu^{(k)}, & \text{on } B. \end{cases}$
3.     Update $\mu^{(k+1)}$ as follows:
           $\mu^{(k+1)}(x) = \mu^{(k)}(x) + \tau \left( w_1^{(k+1)}(x) - w_2^{(k+1)}(x) \right), \qquad \text{for } x \in B.$
4. Endfor
Output: $(w_1^{(k)}, w_2^{(k)})$
```

Remark 1.27. The map $\mu^{(k)} \rightarrow \left(w_1^{(k)} - w_2^{(k)}\right)$ will be compact, and thus the iterates will converge geometrically to the true solution for sufficiently small $\tau > 0$. Discrete versions of Uzawa's algorithm are described in Chap. 10.

Remark 1.28. The FETI method [FA16, FA15], see Chap. 4, is also based on updating the Lagrange multiplier μ. However, it generalizes the preceding saddle point iterative algorithm to the multisubdomain case, where the rate of convergence may deteriorate with increasing number of subdomains, and where the local problems may be singular.

An *alternative* hybrid formulation equivalent to (1.30) can be obtained by replacing the Lagrangian functional $\mathcal{L}(\cdot,\cdot)$ by an *augmented* Lagrangian $\mathcal{L}_\delta(\cdot,\cdot)$, where an additional non-negative functional is added to the original Lagrangian functional with a coefficient $\delta > 0$, see [GL7, GL8]:

$$\mathcal{L}_\delta\left((v_1, v_2), \mu\right) \equiv J_1(v_1) + J_2(v_2) + m\left((v_1, v_2), \mu\right) + \frac{\delta}{2}\|v_1 - v_2\|^2_{L^2(B)}.$$

The augmented term $\frac{\delta}{2}\|v_1 - v_2\|^2_{L^2(B)}$ will be zero when the constraint $v_1 = v_2$ is satisfied on B. As a result, both formulations will be equivalent, and the saddle point of the augmented Lagrangian will also yield the desired solution. Applying an *alternating directions implicit* (ADI) method to determine the saddle point of the augmented Lagrangian functional, will yield the following algorithm, referred to as the *non-overlapping Schwarz* method [LI8, GL8].

Algorithm 1.4.2 *(Non-Overlapping Schwarz Method)*
Let $w_1^{(0)}$, $w_2^{(0)}$ denote starting guesses.
Let $\delta > 0$ be a chosen parameter.

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     Solve in parallel:
$$\begin{cases} -\nabla \cdot \left(a\nabla w_1^{(k+1)}\right) + c w_1^{(k+1)} = f, & \text{in } \Omega_1 \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \\ \mathbf{n}_1 \cdot \left(a\nabla w_1^{(k+1)}\right) + \delta w_1^{(k+1)} = \mathbf{n}_1 \cdot \left(a\nabla w_2^{(k)}\right) + \delta w_2^{(k)}, & \text{on } B, \\ -\nabla \cdot \left(a\nabla w_2^{(k+1)}\right) + c w_2^{(k+1)} = f, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \mathbf{n}_2 \cdot \left(a\nabla w_2^{(k+1)}\right) + \delta w_2^{(k+1)} = \mathbf{n}_2 \cdot \left(a\nabla w_1^{(k)}\right) + \delta w_1^{(k)}, & \text{on } B. \end{cases}$$
3. Endfor
Output: $(w_1^{(k)}, w_2^{(k)})$
```

Remark 1.29. In practice, a careful choice of parameter $\delta > 0$ will be necessary for optimal convergence [LI8, GL8].

1.4.3 Global Discretization

In principle, a discretization of (1.23) can be obtained by discretizing (1.30). Each subdomain can be triangulated independently without requiring the local triangulations to match on B. However, to ensure that the resulting discretization yields a constrained minimization problem, it is advantageous to employ a Galerkin approximation of the saddle point problem (1.29). An extensive literature exists on such nonmatching grid discretization techniques, see [MA4, BE22, DO4, BE4, WO4, WO5]. The resulting discretization is referred to as a mortar element method, see also Chap. 11.

Triangulate each subdomain Ω_i by a grid $\mathcal{T}_{h_i}(\Omega_i)$ of size h_i suited to the local geometry and solution for $1 \leq i \leq 2$, see Fig. 1.8. Let $X_{h_i} \subset X_i$ denote a traditional finite element space defined on the triangulation $\mathcal{T}_{h_i}(\Omega_i)$. Select a triangulation of interface B inherited either from $\mathcal{T}_{h_1}(\Omega_1)$ or $\mathcal{T}_{h_2}(\Omega_2)$. For definiteness, suppose that $\mathcal{T}_{h_1}(\Omega_1)$ is chosen. Construct a finite element space $Y_{h_1}(B) \subset L^2(B) \subset Y$ consisting of piecewise polynomial functions defined on

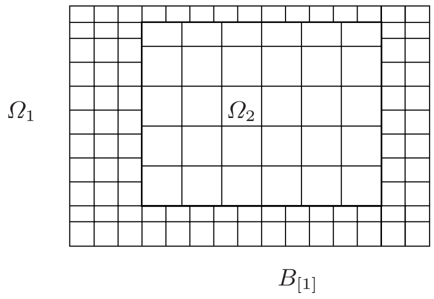


Fig. 1.8. Non-overlapping nonmatching grids

the triangulation of B inherited from $\mathcal{T}_{h_1}(\Omega_1)$. The dimension of Y_{h_1} should equal the dimension of $X_{h_1} \cap H_0^1(B)$. See Chap. 11 for multiplier spaces $Y_{h_1}(B)$.

Discretization of the saddle point formulation (1.29) using the subspaces $X_{h_1} \times X_{h_2} \times Y_{h_1}(B)$ will yield a linear system of the form:

$$\begin{bmatrix} A^{(1)} & 0 & M^{(1)^T} \\ 0 & A^{(2)} & -M^{(2)^T} \\ M^{(1)} & -M^{(2)} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_{h_1} \\ \mathbf{w}_{h_2} \\ \boldsymbol{\mu}_h \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{h_1} \\ \mathbf{f}_{h_2} \\ \mathbf{0} \end{bmatrix},$$

where:

$$\begin{cases} \mathcal{A}_i(w_{h_i}, w_{h_i}) &= \mathbf{w}_{h_i}^T A^{(i)} \mathbf{w}_{h_i}, & \text{for } 1 \le i \le 2 \\ F(w_{h_i}) &= \mathbf{w}_{h_i}^T \mathbf{f}_{h_i}, & \text{for } 1 \le i \le 2 \\ m\left((w_{h_1}, w_{h_2}), \mu_h\right) &= \boldsymbol{\mu}_h^T \left(M^{(1)} \mathbf{w}_{h_1} - M^{(2)} \mathbf{w}_{h_2}\right). \end{cases}$$

Here we have used w_{h_i} and μ_h to denote finite element functions and $\mathbf{w}_{h_i}$ and $\boldsymbol{\mu}_h$ as their vector representations with respect to some fixed basis.

If each nodal vector $\mathbf{w}_{h_i}$ is block partitioned as $\mathbf{w}_{h_i} = \left(\mathbf{w}_I^{(i)}, \mathbf{w}_B^{(i)}\right)^T$ corresponding to the unknowns in the interior of each subdomain and on the interface B, then matrices $A^{(i)}$ and $M^{(i)}$ will have the block structure:

$$A^{(i)} = \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \quad \text{and} \quad M^{(i)} = \begin{bmatrix} 0 & M_B^{(i)} \end{bmatrix}, \quad \text{for } 1 \le i \le 2$$

where $\mathbf{w}_I^{(i)}$ and $\mathbf{w}_B^{(i)}$ are of size n_i and m_i. Substituting, we obtain:

$$\begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} & 0 & 0 & 0 \\ A_{IB}^{(1)^T} & A_{BB}^{(1)} & 0 & 0 & M_B^{(1)^T} \\ 0 & 0 & A_{II}^{(2)} & A_{IB}^{(2)} & 0 \\ 0 & 0 & A_{IB}^{(2)^T} & A_{BB}^{(2)} & -M_B^{(2)^T} \\ 0 & M_B^{(1)} & 0 & -M_B^{(2)} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_B^{(1)} \\ \mathbf{w}_I^{(2)} \\ \mathbf{w}_B^{(2)} \\ \boldsymbol{\mu}_h \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B^{(2)} \\ \mathbf{0} \end{bmatrix}.$$

If the dimension of the space $Y_{h_1}(B)$ is m_1, then matrix $M_B^{(1)}$ will be square and *invertible* of size m_1. In this case, we may parameterize the solution space of the interface constraints as:

$$\mathbf{w}_B^{(1)} \equiv R_{12}\mathbf{w}_B^{(2)} \quad \text{where} \quad R_{12} \equiv M_B^{(1)^{-1}} M_B^{(2)}.$$

The local unknowns can then be represented as $\mathbf{w}_I^{(1)}$, $\mathbf{w}_B^{(1)} = R_{12}\mathbf{w}_B^{(2)}$, $\mathbf{w}_I^{(2)}$, and $\mathbf{w}_B^{(2)}$. Substituting this representation into the discrete energy $J_{h_1}(\mathbf{w}_I^{(1)}, R_{12}\mathbf{w}_B^{(2)}) + J_{h_2}(\mathbf{w}_I^{(2)}, \mathbf{w}_B^{(2)})$ and applying first order stationarity conditions for its minimum yields the following linear system:

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)} R_{12} \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ R_{12}^T A_{IB}^{(1)^T} & A_{IB}^{(2)^T} & R_{12}^T A_{BB}^{(1)} R_{12} + A_{BB}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_I^{(2)} \\ \mathbf{w}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ R_{12}^T \mathbf{f}_B^{(1)} + \mathbf{f}_B^{(2)} \end{bmatrix}.$$

If both grids match, then $R_{12} = I$ and the above discretization reduces to the traditional conforming finite element discretization of (1.23).

Mortar element spaces $Y_{h_i}(B)$ are described in Chap. 11. They include piecewise polynomial functions which are continuous across elements as well as piecewise polynomial functions which are discontinuous across elements [MA4, BE22, BE18, BE6, BE4]. In the latter case, a basis for $Y_{h_i}(B)$ can be constructed so that matrix $M_B^{(i)}$ is diagonal [WO4, WO5]. The resulting global discretization will be stable and convergent of optimal order.

1.4.4 Heterogeneous Approximations

When elliptic equation (1.23) is *singularly perturbed*, its Lagrange multiplier formulation (1.30) can be employed to *heuristically* study an heterogeneous approximation of it. Below, we illustrate two alternative approximations of the following singularly perturbed, self adjoint elliptic equation [KE5]:

$$\begin{cases} -\nabla \cdot (\epsilon \nabla u) + c(x)\, u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad u = g(x), & \text{on } \partial\Omega, \end{cases} \tag{1.31}$$

where $0 < \epsilon \ll 1$ is a small perturbation parameter and $c(x) \geq c_0 > 0$. Suppose Ω_1 and Ω_2 form a nonoverlapping decomposition of Ω, such that:

$$|\epsilon \Delta u| \ll |c(x)\, u|, \qquad \text{for } x \in \overline{\Omega}_1.$$

Then, Ω_2 must enclose the boundary layer region of the solution.

To obtain an heterogeneous approximation of (1.31), we heuristically apply the subdomain vanishing viscosity method as in [GA15]:

$$\begin{cases} -\nabla \cdot (a_{\epsilon,\eta}(x) \nabla u) + c(x)\, u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad u = g(x), & \text{on } \partial\Omega, \end{cases} \tag{1.32}$$

where

$$a_{\epsilon,\eta}(x) \equiv \begin{cases} \eta & \text{for } x \in \Omega_1 \\ \epsilon & \text{for } x \in \Omega_2. \end{cases}$$

For $\epsilon > 0$ and $\eta > 0$, the above problem is elliptic and coercive. However, as $\eta \to 0^+$, *formally* the limiting system (1.30) becomes:

$$\begin{cases} c(x)\, w_1 = f(x), & \text{in } \Omega_1 \\ w_1 = g(x), & \text{on } B_{[1]} \\ 0 = \mu, & \text{on } B \\ -\epsilon\, \Delta w_2 + c(x)\, w_2 = f(x), & \text{in } \Omega_2 \\ w_2 = g(x), & \text{on } B_{[2]} \\ \epsilon \frac{\partial w_2}{\partial n} = \mu, & \text{on } B \\ w_1 = w_2, & \text{on } B. \end{cases}$$

Two alternative approximations may be constructed. Either the *transmission* condition $w_1 = w_2$ or $\epsilon \frac{\partial w_2}{\partial n} = 0$ can be enforced, but *not* both, since $w_1(.)$ formally satisfies a *zeroth order* equation in Ω_1. Since $c(x) \geq c_0 > 0$, the limiting equation on Ω_1 for $w_1(x)$ can be solved to formally yield:

$$w_1(x) = \frac{f(x)}{c(x)}, \qquad \text{on } \Omega_1.$$

If $B_{[1]} \neq \emptyset$ and the boundary data $g(x)$ is not compatible with the formal solution $\frac{f(x)}{c(x)}$, i.e., if $g(x) \neq \frac{f(x)}{c(x)}$ on $B_{[1]}$, then the local solution may be ill posed, indicating a poor choice of subdomain Ω_1.

If a *continuous* (or $H^1(\cdot)$) solution is sought, then continuity of the local solutions must be enforced and the flux transmission condition needs to be *omitted*, yielding the following system:

$$\begin{cases} c(x)\, w_1 = f(x), & \text{in } \Omega_1 \\ w_1 = g(x), & \text{on } B_{[1]} \\ -\epsilon\, \Delta w_2 + c(x)\, w_2 = f(x), & \text{in } \Omega_2 \\ w_2 = w_1, & \text{on } B \\ w_2 = g(x), & \text{on } B_{[2]}. \end{cases}$$

If a *discontinuous* approximation is sought, then the continuity transmission condition can be *omitted*, and the flux transmission condition can be enforced, yielding the alternative system:

$$\begin{cases} c(x)\, w_1 = f(x), & \text{in } \Omega_1 \\ w_1 = g(x), & \text{on } B_{[1]} \\ -\epsilon\, \Delta w_2 + c(x)\, w_2 = f(x), & \text{in } \Omega_2 \\ \epsilon\, \frac{\partial w_2}{\partial n} = 0, & \text{on } B \\ w_2 = g(x), & \text{on } B_{[2]}. \end{cases}$$

In this case, the subproblems for w_1 and w_2 are formally *decoupled.* In both cases, the limiting solutions may not minimize the energy functional $J_{\epsilon,\eta}(\cdot)$ associated with (1.32) as $\eta \to 0^+$.

Remark 1.30. Since (1.30) is equivalent to (1.19), rigorous results on the well posedness of the above approximation may be deduced from [GA15].

Remark 1.31. Similar heuristics may be applied to construct an approximation of the singularly perturbed *anisotropic* elliptic equation using (1.30):

$$\begin{cases} -\epsilon\, u_{x_1x_1} - u_{x_2x_2} - u_{x_3x_3} + c(x)\, u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad u = g(x), & \text{on } \partial\Omega, \end{cases}$$

for which the limiting problem is a degenerate elliptic equation. In this case, both transmission conditions can be retained in the limiting problem.

1.5 Least Squares-Control Framework

The least squares-control method [LI2, GL] is a general optimization method, which has various applications to partial differential equations. It results in a *constrained least squares* problem, and is based on the minimization of a *square norm* objective functional, subject to *constraints.* In domain decomposition applications, see [AT, GL13, GU3, GU2], the square norm functional typically measures the difference between the subdomain solutions on the regions of overlap or intersection between the subdomains, while the constraints require the local solutions to solve the original partial differential equation on each subdomain, with appropriate boundary conditions. Since the boundary data on each subdomain boundary is *unknown*, it is regarded as a *control* function which parameterizes the local solution. The control boundary data must be determined to minimize the square norm function, hence the name least squares-control. Importantly, an optimization principle need not be associated with the underlying partial differential equation.

In this section, we describe the hybrid formulation associated with the least squares-control method for the following elliptic equation:

$$\begin{cases} Lu \equiv -\nabla \cdot (a(x)\,\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u = 0, & \text{in } \partial\Omega, \end{cases} \tag{1.33}$$

in which the domain Ω is decomposed into *two* subdomains. The subdomains can be *overlapping* or *non-overlapping*, but we focus on the overlapping case. We illustrate the formulation of *iterative* methods, *non-matching grid* discretizations, and *heterogeneous* approximations for (1.33).

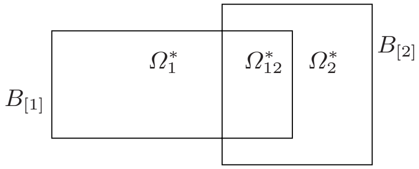


Fig. 1.9. An overlapping decomposition

1.5.1 Motivation

Let Ω_1^* and Ω_2^* form an overlapping decomposition of Ω, with $\Omega_{12}^* = \Omega_1^* \cap \Omega_2^*$, see Fig. 1.9. Let $B^{(i)} = \partial\Omega_i^* \cap \Omega$ and $B_{[i]} = \partial\Omega_i^* \cap \partial\Omega$ denote the interior and exterior segments, respectively, of the subdomain boundaries, and let $\mathbf{n}_i$ denote the unit exterior normal to $\partial\Omega_i^*$. On each subdomain Ω_i^* for $1 \le i \le 2$, we let w_i denote the approximation of the solution u to (1.33) on Ω_i^*, and let g_i denote the local Neumann data associated with w_i on $B^{(i)}$.

If $w_i(.) = u(.)$ on Ω_i^* and $g_i(.) = \mathbf{n}_i \cdot (a(x)\nabla u)$ on $B^{(i)}$, then w_i will satisfy:

$$\begin{cases} L w_i = f, & \text{in } \Omega_i^* \\ w_i = 0, & \text{on } B_{[i]} \\ \mathbf{n}_i \cdot (a\nabla w_i) = g_i, & \text{on } B^{(i)}. \end{cases}$$

Furthermore, since w_1 and w_2 will *match* on Ω_{12}^*, i.e., $w_1 = w_2$, on Ω_{12}^*, it will hold that $\|w_1 - w_2\|^2_{L^2(\Omega_{12}^*)} = 0$ and $|w_1 - w_2|^2_{H^1(\Omega_{12}^*)} = 0$. Motivated by this, define the following square norm functional $J(\cdot)$:

$$J(v_1, v_2) \equiv \frac{\gamma_1}{2}\int_{\Omega_{12}^*} (v_1 - v_2)^2\, dx + \frac{\gamma_2}{2}\int_{\Omega_{12}^*} |\nabla(v_1 - v_2)|^2\, dx. \tag{1.34}$$

Typically $(\gamma_1 = 1, \gamma_2 = 0)$, but other choices are possible. Then, it will hold:

$$J(w_1, w_2) = 0,$$

for the true subdomain solutions.

The preceding observation suggests the following *constrained minimization* problem equivalent to (1.33). Determine (w_1, w_2) which minimizes $J(\cdot)$ (with minimum value zero), within a class $\mathcal{K}$:

$$J(w_1, w_2) = \min_{(v_1,v_2)\in\mathcal{K}} J(v_1, v_2), \tag{1.35}$$

where $\mathcal{K}$ is defined by the constraints:

$$\mathcal{K} \equiv \left\{ (v_1, v_2) : \begin{array}{rl} L v_i = f, & \text{in } \Omega_i^* \\ \mathbf{n}_i \cdot (a\nabla w_i) = g_i, & \text{on } B^{(i)} \\ v_i = 0, & \text{on } B_{[i]} \end{array} \text{ for } 1 \le i \le 2 \right\}. \tag{1.36}$$

Instead of Neumann conditions on $B^{(i)}$, we may alternatively pose Robin or Dirichlet conditions. However, in the non-overlapping case, we cannot pose Dirichlet conditions on $B^{(i)}$, since the functional $J(.)$ typically measures the difference between the Dirichlet data. To avoid cumbersome notation, we often omit explicit inclusion of g_i as an argument in the definition of $J(.,.)$ and $\mathcal{K}$. In a strict sense, we must replace v_i by (v_i, g_i). Hopefully, such omission should be clear from the context.

The following equivalence will hold.

Theorem 1.32. *Suppose the following assumptions hold.*

1. *Let the solution u of (1.33) exist and be smooth.*
2. *Let (w_1, w_2) minimize (1.35) subject to the constraints (1.36).*

Then at the minimum:

$$J(w_1, w_2) = \min_{(v_1,v_2)\in\mathcal{K}} J(v_1, v_2),$$

it will hold that:

$$\begin{cases} w_1 = u, & \text{on } \Omega_1^* \\ w_2 = u, & \text{on } \Omega_2^*. \end{cases}$$

Proof. Suppose u is the solution to (1.33) and $w_i \equiv u$ on Ω_i^* for $1 \le i \le 2$. Then, (w_1, w_2) will satisfy all the required constraints (1.36). Furthermore:

$$w_1 - w_2 = u - u = 0, \qquad \text{in} \qquad \Omega_{12}^*,$$

yields that $J(w_1, w_2) = 0$ and minimizes $J(.,.) \ge 0$.

Conversely, suppose a solution to (1.35) exists, subject to constraints (1.36) and minimizes $J(v_1, v_2)$. Then this minimum value must be *zero*, since for $u_i \equiv u$ in Ω_i^* for $1 \le i \le 2$ it will hold that $(u_1, u_2) \in \mathcal{K}$ and $J(u_1, u_2) = 0$. Thus, using the definition of $J(.,.)$ and that $J(w_1, w_2) = 0$, we obtain that $w_1 = w_2$ on Ω_{12}^*. Let $\chi_1(x)$ and $\chi_2(x)$ form a partition of unity subordinate to the cover Ω_1^* and Ω_2^*. The it is easily verified that $\chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x)$ solves (1.33), since $Lw_i = f$ in Ω_i^* and since $w_1 = w_2$ in Ω_{12}^*. Thus, by the uniqueness of solutions to (1.33) it follows that:

$$u(x) \equiv \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x).$$

The desired result follows using $w_1 = w_2$ on Ω_{12}^*. □

Remark 1.33. The preceding result only demonstrates an equivalence between the solutions of (1.33) and (1.35). It does not guarantee the well posedness of (1.35) under perturbation of data. Such a result, however, will hold under appropriate assumptions (such as $\mathbf{b} = \mathbf{0}$, coercivity of (1.33)) given sufficient overlap between the subdomains.

Remark 1.34. Well posedness of the constrained minimization problem (1.35) will depend on the definition of $J(\cdot)$. For instance, when the elliptic equation (1.33) is self adjoint and coercive, $J(v_1, v_2) = \frac{1}{2}\|v_1 - v_2\|^2_{H^1(\Omega^*_{12})}$ can be shown to yield a well posed saddle point problem [GL, AT], where the term $J(v_1, v_2)$ is coercive in the constraint space $\mathcal{K}$. More generally, an augmented Lagrangian formulation [GL7] may be employed to regularize (1.35).

As mentioned earlier, the constraint set $\mathcal{K}$ in (1.36) can be *parameterized* in terms of the Dirichlet, Neumann or Robin data g_i specified on each boundary segment $B^{(i)}$, for $1 \le i \le 2$. For instance, when Neumann boundary conditions are imposed on each $B^{(i)}$, define an *affine* linear mapping $\mathcal{E}_i$ as follows:

$$\mathcal{E}_i\, g_i \equiv v_i, \quad \text{where} \qquad \begin{cases} L\, v_i = f, & \text{in } \Omega^*_i \\ \mathbf{n}_i \cdot (a \nabla v_i) = g_i, & \text{on } B^{(i)} \\ v_i = 0, & \text{on } B_{[i]}. \end{cases}$$

Then, the constraint set $\mathcal{K}$ can be represented as:

$$\mathcal{K} \equiv \{(\mathcal{E}_1 g_1, \mathcal{E}_2 g_2) : \ \text{ for } g_i \in X_i, \ \ 1 \le i \le 2\},$$

where g_1 and g_2 are regarded as *control* data. For Neumann conditions, the function space X_i for the boundary data for g_i is typically chosen for each $1 \le i \le 2$ as $X_i = (H^{1/2}_{00}(B^{(i)}))'$ or $X_i = H^{-1/2}(B^{(i)})$. This parameterization enables the reformulation of this *constrained* minimization problem (1.35) as an *unconstrained* minimization problem. Define a function $H(\cdot)$:

$$H(g_1, g_2) \equiv J(\mathcal{E}_1 g_1, \mathcal{E}_2 g_2). \tag{1.37}$$

Then, the unconstrained minimum (g_1^*, g_2^*) of $H(\cdot, \cdot)$:

$$H(g_1^*, g_2^*) \ = \ \min_{(g_1, g_2)} H(g_1, g_2), \tag{1.38}$$

will yield the *constrained* minimum of $J(.,.)$ as $(w_1, w_2) = (\mathcal{E}_1 g_1^*, \mathcal{E}_2 g_2^*)$. Thus, once g_1^* and g_2^* have been determined by minimizing $H(\cdot, \cdot)$, the desired local solutions will satisfy $w_i \equiv \mathcal{E}_i g_i^*$ for $1 \le i \le 2$. Such unconstrained minimization does not require Lagrange multipliers.

The unknown control data g_1 and g_2 can be determined by solving the system of equations which result from the application of first order stationarity conditions $\delta H = 0$ at the minimum of $H(\cdot)$. We shall omit the derivation of these equations, except to note that the calculus of variations may be applied to (1.38), or such equations may be derived by heuristic analogy with the associated discrete saddle point problem, as described in Chap. 6.

The resulting first order stationarity equations will be of the form:

$$\delta H\,(g_1, g_2) = 0 \ \Leftrightarrow \ \begin{cases} v_1(x) = 0, & \text{for } x \in B^{(1)} \\ v_2(x) = 0, & \text{for } x \in B^{(2)} \end{cases}$$

where $v_1(x)$ and $v_2(x)$ are defined in terms of $g_1(x)$ and $g_2(x)$ as follows. Solve:

$$\begin{cases} -\nabla \cdot (a\,\nabla w_i) + \mathbf{b} \cdot \nabla w_i + c\, w_i = f(x), & \text{in } \Omega_i^* \\ \qquad\qquad\qquad\qquad\qquad w_i = 0, & \text{on } B_{[i]} \qquad \text{for } i = 1, 2 \\ \qquad\qquad\qquad \mathbf{n}_i \cdot (a\nabla w_i) = g_i(x), & \text{on } B^{(i)} \end{cases}$$

for $w_1(x)$ and $w_2(x)$ using $g_1(x)$ and $g_2(x)$. Next, compute:

$$r(x) \equiv \begin{cases} w_1(x) - w_2(x), & \text{for } x \in \Omega_{12}^* \\ 0, & \text{for } x \notin \Omega_{12}^*. \end{cases}$$

Then, $v_1(x)$ and v_2 are defined as the solutions to:

$$\begin{cases} -\nabla \cdot (a\,\nabla v_i) - \nabla \cdot (\mathbf{b}\, v_i) + c\, v_i = r(x), & \text{in } \Omega_i^* \\ \qquad\qquad\qquad\qquad\qquad v_i = 0, & \text{on } B_{[i]} \qquad \text{for } 1 \le i \le 2. \\ \qquad\qquad \mathbf{n}_i \cdot (a\nabla v_i + \mathbf{b}\, v_i) = 0, & \text{on } B^{(i)} \end{cases}$$

The control data $g_1(x)$ and $g_2(x)$ must be chosen to ensure that $v_i(x) = 0$ on $B^{(i)}$ for $i = 1, 2$. Later, we shall outline a gradient method to determine g_1 and g_2 iteratively. When (1.35) is discretized, an explicit matrix representation can be derived for $H(\cdot)$ and its gradient, see Chap. 6. In this case, a preconditioned CG method can be employed to solve the resulting linear system.

Remark 1.35. If Ω is decomposed into *non-overlapping* subdomains Ω_1 and Ω_2 with common interface $B = \partial\Omega_1 \cap \partial\Omega_2$, a least squares-control formulation may be constructed as follows [GU3, GU2]. Seek (w_1, w_2) which minimizes:

$$J(w_1, w_2) = \min_{(v_1, v_2) \in \mathcal{K}} J(v_1, v_2),$$

where

$$J(v_1, v_2) \equiv \frac{1}{2}\, \|v_1 - v_2\|^2_{L^2(B)},$$

and $\mathcal{K}$ consists of all (v_1, v_2) satisfying the following constraints:

$$\begin{cases} L v_1 = & f(x), & \text{in } \Omega_1 \\ v_1 = & 0, & \text{on } B_{[1]} \\ \mathbf{n}_1 \cdot (a\nabla v_1) = & \mu(x), & \text{on } B \\ L v_2 = & f(x), & \text{in } \Omega_2 \\ v_2 = & 0, & \text{on } B_{[2]} \\ \mathbf{n}_2 \cdot (a\nabla v_2) = & -\mu(x), & \text{on } B. \end{cases}$$

Here $\mu(x)$ is a flux variable on the interface B (which can be eliminated). The above constraints will ensure that the original elliptic equation is solved on each subdomain, and that the Neumann fluxes of the two subdomain solutions match on B. In this case, the feasible set $\mathcal{K}$ can be parameterized in terms of the flux $\mu(x) = \mathbf{n}_1 \cdot (a\nabla v_1)$ on B. In applications, an alternative choice of objective functional $J(v_1, v_2) \equiv \frac{1}{2}\|v_1 - v_2\|^2_{H^{1/2}_{00}(B)}$ may also be employed, where $H^{1/2}_{00}(B)$ denotes a fractional Sobolev norm (defined in Chap. 3).

1.5.2 Iterative Methods

The solution to (1.33) can be determined iteratively, by formally applying a steepest descent method to the unconstrained minimization problem (1.38), with sufficiently small step size $\tau > 0$. Such an algorithm can be derived formally using calculus of variations, or by analogy with the discrete version of this algorithm described in Chap. 6.

Algorithm 1.5.1 *(Gradient Least Squares-Control Algorithm)*
Let $g_1^{(0)}(x)$ and $g_2^{(0)}(x)$ denote starting guesses and $\tau > 0$ a fixed step size.

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     For $i = 1, 2$ in parallel solve:
           $$\begin{cases} -\nabla \cdot (a\,\nabla v_i) + \mathbf{b} \cdot \nabla v_i + c\,v_i = f(x), & \text{in } \Omega_i^* \\ v_i = 0, & \text{on } B_{[i]} \\ \mathbf{n}_i \cdot (a\nabla v_i) = g_i^{(k)}(x), & \text{on } B^{(i)}. \end{cases}$$
3.     Endfor
4.     Compute:
           $$r(x) \equiv \begin{cases} v_1(x) - v_2(x), & \text{for } x \in \Omega_{12}^* \\ 0, & \text{for } x \notin \Omega_{12}^* \end{cases}$$
5.     For $i = 1, 2$ in parallel solve the adjoint problems:
           $$\begin{cases} -\nabla \cdot (a\,\nabla w_i) - \nabla \cdot (\mathbf{b}\,w_i) + c\,w_i = r(x), & \text{in } \Omega_i^* \\ w_i = 0, & \text{on } B_{[i]} \\ \mathbf{n}_i \cdot (a\nabla w_i + \mathbf{b}\,w_i) = 0, & \text{on } B^{(i)}. \end{cases}$$
6.     Endfor
7.     Update:
           $$\begin{cases} g_1^{(k+1)}(x) = g_1^{(k)}(x) - \tau\,w_1(x), & \text{for } x \in B^{(1)} \\ g_2^{(k+1)}(x) = g_2^{(k)}(x) + \tau\,w_2(x), & \text{for } x \in B^{(2)}. \end{cases}$$
8. Endfor
Output: $(g_1^{(k)}, g_2^{(k)})$
```

Alternative divide and conquer iterative algorithms can be formulated for (1.33) using its saddle point formulation. However, the resulting algorithm may require more computational resources. For instance, suppose that:

$$J(v_1, v_2) = \frac{1}{2} \|v_1 - v_2\|^2_{L^2(\Omega^*_{12})},$$

and that Neumann boundary conditions are imposed on $B^{(i)}$. Then, as described in Chap. 10, a constrained minimization problem such as (1.35) with (1.36), can be equivalently formulated as a saddle point problem, and saddle point iterative algorithms can be formulated to solve it.

Indeed, if λ_1 and λ_2 denote the Lagrange multipliers, then the saddle point problem associated with (1.35) would formally be of the form:

$$\begin{cases} \chi_{\Omega_{12}}\,(w_1 - w_2) + L_1^* \lambda_1 = 0, \\ -\chi_{\Omega_{12}}\,(w_1 - w_2) + L_2^* \lambda_2 = 0, \\ L_1 \tilde{w}_1 = f_1, \\ L_2 \tilde{w}_2 = f_2. \end{cases} \tag{1.39}$$

Here $L_i \tilde{w}_i = f_i$ formally denotes the operator equation associated with $L\, w_i = f$ in Ω^*_i with Neumann conditions $\mathbf{n}_i \cdot (a\, \nabla w_i) - g_i = 0$ on $B^{(i)}$ and homogeneous Dirichlet boundary conditions $w_i = 0$ on $B_{[i]}$, with $\tilde{w}_i = (w_i, g_i)$. The operator L^*_i formally denotes the adjoint of L_i. Here, $\chi_{\Omega^*_{12}}(x)$ denotes the characteristic (indicator) function of Ω^*_{12}. We omit elaborating on such a saddle point problem here, except to note that, it may be obtained by *heuristic* analogy with the discrete saddle point problems described in Chap. 10. The $\lambda_i(x)$ corresponds to Lagrange multiplier functions, see [GL, AT]. In this saddle point problem, the Lagrange multiplier variables will not be unique, and an augmented Lagrangian formulation would be preferable.

1.5.3 Global Discretization

Hybrid formulation (1.35) or (1.38) can, in principle, be employed to discretize (1.33) on a nonmatching grid such as in Fig. 1.10. Such discretizations have not been considered in the literature, however, a *heuristic* discussion of such a discretization is outlined here for its intrinsic interest, employing formulation (1.38). We employ finite element discretizations on the subdomains.

A nonmatching grid discretization of (1.38) will require discretizing $J(\cdot)$:

$$J(v_1, v_2) = \frac{1}{2} \|v_1 - v_2\|^2_{H^1(\Omega^*_{12})},$$

and this will involve two overlapping non-matching grids. In the following, we heuristically outline a mortar element discretization of $J(v_1, v_2)$ on Ω^*_{12}, and employ this to construct a global non-matching grid discretization of (1.33), with Dirichlet boundary controls on each subdomain boundary $B^{(i)}$. Each subdomain problem will involve only a conforming grid.

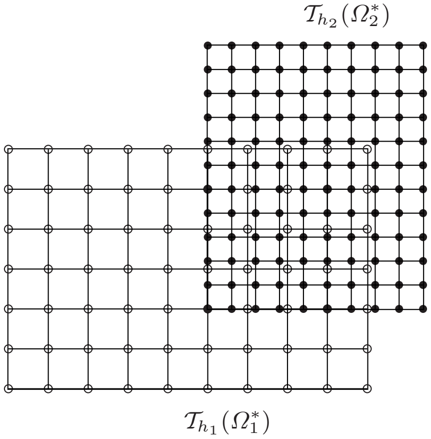

$\mathcal{T}_{h_2}(\Omega_2^*)$
$\mathcal{T}_{h_1}(\Omega_1^*)$

Fig. 1.10. Overlapping nonmatching grids

Remark 1.36. If $J(v_1, v_2)$ is replaced by $J_B(v_1, v_2) \equiv \frac{1}{2}\|v_1 - v_2\|_B^2$ where $B = \partial\Omega_1 \cap \partial\Omega_2$ and Ω_i^* is an extension of a non-overlapping decomposition Ω_i, such a discretization would be considerably simpler.

Local Triangulation. For $1 \leq i \leq 2$ triangulate each subdomain Ω_i^* by a grid $\mathcal{T}_{h_i}(\Omega_i^*)$ according to the local geometry and regularity of the solution, see Fig. 1.10. We shall assume that at least one of the local grids triangulates the region of overlap Ω_{12}^*. For definiteness assume that triangulation $\mathcal{T}_{h_1}(\Omega_1^*)$ triangulates Ω_{12}^*. Let n_i and m_i denote the number of nodes of grid $\mathcal{T}_{h_i}(\Omega_i^*)$ in the interior of Ω_i^* and on $B^{(i)}$, respectively. Additionally, let l_i denote the number of nodes of triangulation $\mathcal{T}_{h_i}(\Omega_i^*)$ in $\overline{\Omega}_{12}^*$.

Local Discretizations. For $1 \leq i \leq 2$, employ Dirichlet boundary conditions on $B^{(i)}$ in (1.36) and discretize the resulting local problems using a finite element space $X_{h_i} \subset X_i$ based on triangulation $\mathcal{T}_{h_i}(\Omega_i^*)$:

$$X_i \equiv \left\{ v_i \in H^1(\Omega_i^*) \; : \; v_i = 0 \quad \text{on} \quad B_{[i]} \right\}.$$

Block partition the unknowns $\mathbf{w}_{h_i} = (\mathbf{w}_I^{(i)}, \mathbf{w}_B^{(i)})^T$ according to the interior unknowns and the unknowns on the boundary $B^{(i)}$ respectively. Denote the block partitioned linear system for the discretized Dirichlet problem as:

$$\begin{cases} A_{II}^{(i)}\mathbf{w}_I^{(i)} + A_{IB}^{(i)}\mathbf{w}_B^{(i)} = \mathbf{f}_I^{(i)}, \\ \qquad\qquad\qquad\quad \mathbf{w}_B^{(i)} = \mathbf{g}_B^{(i)}. \end{cases}$$

Weak Matching on Ω_{12}^*. Choose a finite element space:

$$Y_h(\Omega_{12}^*) \subset L^2(\Omega_{12}^*)$$

based on the triangulation of Ω_{12}^* inherited from $\mathcal{T}_{h_1}(\Omega_1^*)$, of dimension l_1. Define the weak matching condition on Ω_{12}^* as:

$$\int_{\Omega_{12}^*} (w_{h_1} - w_{h_2})\ \mu_{h_1}\, dx = 0, \qquad \text{for } \mu_{h_1} \in Y_{h_1}(\Omega_{12}^*),$$

enforced using the subspace $Y_{h_1}(\Omega_{12}^*)$. Denote its matrix form as:

$$M_{11}\mathbf{w}_{h_1} - M_{12}\mathbf{w}_{h_2} \;=\; \mathbf{0},$$

where M_{11} is invertible of size l_1. Define an oblique projection $P_1 \equiv M_{11}^{-1}M_{12}$.

Discrete Functional $J(\cdot,\cdot)$. Let $A^{(12)}$ be the stiffness matrix associated with $J(\cdot)$ on the triangulation $\mathcal{T}_{h_1}(\Omega_{12}^*)$. The quadratic functional $J(\cdot)$ can be discretized using $A^{(12)}$ and the projection P_1 as follows:

$$\begin{cases} J\left(v_{h_1}, v_{h_2}\right) \equiv \frac{1}{2}\left\|v_{h_1} - v_{h_2}\right\|_{H^1(\Omega_{12}^*)}^2 \\ \qquad\qquad\qquad \approx \frac{1}{2}\left(\mathbf{v}_{h_1} - P_1\mathbf{v}_{h_2}\right)^T R_{12}^T A^{(12)} R_{12}\left(\mathbf{v}_{h_1} - P_1\mathbf{v}_{h_2}\right) \\ \qquad\qquad\qquad \equiv J_h\left(\mathbf{v}_{h_1}, \mathbf{v}_{h_2}\right). \end{cases}$$

Here R_{12} is a restriction map onto the nodes of $\overline{\Omega}_{12}^*$ from Ω_1^*, see Chap. 6. The reduced functional $H_h(\cdot)$ can be discretized using:

$$H_h\left(\mathbf{g}_{h_1}, \mathbf{g}_{h_2}\right) \equiv J_h\left(\mathbf{v}_{h_1}, \mathbf{v}_{h_2}\right),$$

where

$$\mathbf{v}_{h_i} = \begin{bmatrix} A_{II}^{(i)^{-1}}(\mathbf{f}_I^{(i)} - A_{IB}^{(i)}\mathbf{g}_B^{(i)}) \\ \mathbf{g}_B^{(i)} \end{bmatrix} \qquad \text{for} \qquad 1 \le i \le 2.$$

Stationarity Condition. The first order derivative conditions for the minimum of $H_h(\cdot)$ will yield the following equations for $(\mathbf{g}_B^{(1)}, \mathbf{g}_B^{(2)})$:

$$\begin{bmatrix} E_1^T R_{12}^T A^{(12)} R_{12} E_1 & -\,E_1^T R_{12}^T A^{(12)} R_{12} P_1 E_2 \\ -\,E_2^T P_1^T R_{12}^T A^{(12)} R_{12} E_1 & E_2^T P_1^T R_{12}^T A^{(12)} R_{12} P_1 E_2 \end{bmatrix} \begin{bmatrix} \mathbf{g}_B^{(1)} \\ \mathbf{g}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_B^{(1)} \\ \boldsymbol{\gamma}_B^{(2)} \end{bmatrix} \tag{1.40}$$

where

$$\begin{cases} \boldsymbol{\gamma}_B^{(1)} \equiv E_1^T R_{12}^T A^{(12)} R_{12}\left(-\boldsymbol{\mu}_I^{(1)} + P_1\boldsymbol{\mu}_I^{(2)}\right), \\ \boldsymbol{\gamma}_B^{(2)} \equiv E_2^T P_1^T R_{12}^T A^{(12)} R_{12}\left(-\boldsymbol{\mu}_I^{(1)} + P_1\boldsymbol{\mu}_I^{(2)}\right), \\ E_i \equiv \begin{bmatrix} -A_{II}^{(i)^{-1}} A_{IB}^{(i)} \\ I \end{bmatrix}, \\ \boldsymbol{\mu}_I^{(i)} \equiv \begin{bmatrix} A_{II}^{(i)^{-1}}\mathbf{f}_I^{(i)} \\ \mathbf{0} \end{bmatrix}, \\ \mathbf{w}_I^{(i)} = A_{II}^{(i)^{-1}}\left(\mathbf{f}_I^{(i)} - A_{IB}^{(i)}\mathbf{g}_B^{(i)}\right), \qquad\qquad \text{for } i = 1, 2. \end{cases}$$

Thus, a non-matching grid discretization of (1.33) based on the subdomains involves solving system (1.40) for the control boundary data $\mathbf{g}_B^{(1)}$ and $\mathbf{g}_B^{(2)}$. Subsequently, the subdomain solution $\mathbf{w}_I^{(i)}$ can be determined as:

$$\mathbf{w}_I^{(i)} = A_{II}^{(i)^{-1}} \left(\mathbf{f}_I^{(i)} - A_{IB}^{(i)} \mathbf{g}_I^{(i)} \right), \quad \text{for } 1 \le i \le 2.$$

Remark 1.37. General results on the stability and convergence properties of such discretizations are not known. However, when both local grids match on Ω_{12}^*, projection $P_1 = I$ and the global discretization will be equivalent to a traditional discretization of (1.33) on the global triangulation.

1.5.4 Heterogeneous Approximations

The least square-control formulation (1.35) provides a flexible framework for constructing heterogeneous approximations of general systems of partial differential equations of heterogeneous character [AT, GL13]. We illustrate here how an elliptic-hyperbolic approximation can be constructed for the following singularly perturbed elliptic equation:

$$\begin{cases} L^{\epsilon} u \equiv -\epsilon \, \Delta u + \mathbf{b}(x) \cdot \nabla u + c(x) \, u = f, & \text{on } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.41}$$

where $0 < \epsilon \ll 1$ is a perturbation parameter. Suppose Ω_1^* and Ω_2^* form an overlapping covering of Ω such that:

$$|\epsilon \, \Delta u| \ll |\mathbf{b}(x) \cdot \nabla u + c(x) \, u| \, , \qquad \text{in} \quad \Omega_1^*.$$

We may then heuristically approximate $L^{\epsilon} u = f$ in Ω_1^* by $L_0 u = f$ where $L_0 \, u \equiv \mathbf{b}(x) \cdot \nabla u + c(x) \, u$. To construct an elliptic-hyperbolic approximation of (1.41), replace the elliptic problem $L^{\epsilon} v_1 = f$ on Ω_1^* by the hyperbolic problem $L_0 \, v_1 = f$ within the least squares-control formulation (1.35) of (1.41). The resulting heterogeneous problem will seek (w_1, w_2) which minimizes:

$$\hat{J}(w_1, w_2) = \min_{(v_1, v_2) \in \hat{\mathcal{K}}} \hat{J}(v_1, v_2),$$

where

$$\hat{J}\,(v_1, v_2) \equiv \frac{1}{2} \|v_1 - v_2\|_{L^2(\Omega_{12}^*)}^2,$$

and $\hat{\mathcal{K}}$ consists of (v_1, v_2) which satisfy the constraints:

$$\begin{cases} L_0 \, v_1 = f, & \text{on } \Omega_1^* \\ \quad v_1 = g_1, & \text{on } B_{in}^{(1)} \\ \quad v_1 = 0, & \text{on } B_{[1],in} \end{cases} \quad \text{and} \quad \begin{cases} L^{\epsilon} \, v_2 = f, & \text{on } \Omega_2^* \\ \quad v_2 = g_2, & \text{on } B^{(2)} \\ \quad v_2 = 0, & \text{on } B_{[2]}. \end{cases} \tag{1.42}$$

Here the *inflow* boundary segments of $B^{(1)}$ and $B_{[1]}$ are defined by:

$$\begin{cases} B^{(1)}_{in} & \equiv \left\{ x \in B^{(1)} : \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0 \right\} \\ B_{[1],in} & \equiv \left\{ x \in B_{[1]} : \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0 \right\}, \end{cases}$$

where $\mathbf{n}_1(x)$ is the unit outward normal to B_1 at x.

Remark 1.38. The admissible set $\hat{\mathcal{K}}$ may be parameterized in terms of the local boundary data. An equivalent unconstrained minimization problem may then be obtained analogous to (1.37) and (1.38). See also Chap. 12.

Remark 1.39. The solution (w_1, w_2) to the above heterogeneous model may not match on Ω^*_{12} and the minimum value of $\hat{J}(\cdot)$ within the class $\hat{\mathcal{K}}$ may no longer be zero. A continuous global solution, however, may be obtained by employing a partition of unity $\chi_1(x)$ and $\chi_2(x)$ subordinate to the cover Ω^*_1 and Ω^*_2 and by defining:

$$w(x) \equiv \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x).$$

Remark 1.40. Rigorous results are not known on the well posedness of the above heterogeneous model. The above procedure has been generalized and employed to construct heterogeneous approximations to the Boltzmann, Navier-Stokes and Euler equations [AT, GL13].

2

Schwarz Iterative Algorithms

In this chapter, we describe the family of Schwarz iterative algorithms. It consists of the classical Schwarz alternating method [SC5] and several of its parallel extensions, such as the additive, hybrid and restricted Schwarz methods. Schwarz methods are based on an *overlapping* decomposition of the domain, and we describe its formulation to iteratively solve a discretization of a *self adjoint* and *coercive* elliptic equation. In contrast with iterative algorithms formulated on non-overlapping subdomains, as in Chap. 3, the computational cost per Schwarz *iteration* can exceed analogous costs per iteration on non-overlapping subdomains, by a factor proportional to the overlap between the subdomains. However, Schwarz algorithms are relatively simpler to formulate and to implement, and when there is sufficient overlap between the subdomains, these algorithms can be rapidly convergent for a few subdomains, or as the size of the subdomains decreases, provided a *coarse space* residual correction term is employed [DR11, KU6, XU3, MA15, CA19, CA17].

Our focus in this chapter will be on describing the *matrix version* of Schwarz algorithms for iteratively solving the linear system $A\mathbf{u} = \mathbf{f}$ obtained by the discretization of an elliptic equation. The matrix versions correspond to generalizations of traditional block *Gauss-Seidel* and block *Jacobi* iterative methods. Chap. 2.1 presents background and matrix notation, restriction and extension matrices. Chap. 2.2 describes the continuous version of the classical Schwarz alternating method [MO2, BA2, LI6] and derives its projection version, which involves *projection* operators onto subspaces associated with the subdomains. The projection version of the Schwarz alternating method suggests various parallel generalizations such as the additive Schwarz, hybrid Schwarz and restricted Schwarz methods. Chap. 2.3 describes the matrix version of Schwarz algorithms, which we refer to as *Schwarz subspace* algorithms [XU3]. Chap. 2.4 discusses implementational issues for applications to finite element or finite difference discretizations of elliptic equations. Specific choices of coarse spaces are also described. Chap. 2.5 describes theoretical results on the convergence of Schwarz algorithms in an energy norm.

2.1 Background

In this section, we introduce notation on the elliptic equation and its weak formulation and discretization, subdomain decompositions and block matrix partitioning of the resulting linear system, restriction and extension maps.

2.1.1 Elliptic Equation

We consider the following *self adjoint* and *coercive* elliptic equation:

$$\begin{cases} Lu \equiv -\nabla \cdot (a(x)\nabla u) + c(x)\, u = f, & \text{in } \Omega \\ u = g_D, & \text{on } \mathcal{B}_D \\ \mathbf{n} \cdot (a\nabla u) + \gamma\, u = g_N, & \text{on } \mathcal{B}_N, \end{cases} \tag{2.1}$$

on a domain $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$, with unit exterior normal $\mathbf{n}(x)$ at $x \in \partial\Omega$, Dirichlet boundary $\mathcal{B}_D \subset \partial\Omega$, and natural (Neumann or Robin) boundary $\mathcal{B}_N \subset \partial\Omega$ where $\overline{\mathcal{B}}_D \cup \overline{\mathcal{B}}_N = \partial\Omega$ and $\mathcal{B}_D \cap \mathcal{B}_N = \emptyset$. We shall assume that the diffusion coefficient $a(x)$ is piecewise smooth and for $0 < a_0 \leq a_1$ satisfies:

$$a_0|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^T\, a(x)\, \boldsymbol{\xi}, \leq a_1|\boldsymbol{\xi}|^2, \quad \forall x \in \Omega,\ \boldsymbol{\xi} \in \mathbb{R}^d.$$

To ensure the *coercivity* of (2.1), we shall assume that $c(x) \geq 0$ and $\gamma(x) \geq 0$. In most applications, we shall assume $\mathcal{B}_D = \partial\Omega$ and $\mathcal{B}_N = \emptyset$.

Remark 2.1. When $\mathcal{B}_D = \emptyset$, $\gamma(x) \equiv 0$ and $c(x) \equiv 0$, functions $f(x)$ and $g_N(x)$ will be required to satisfy compatibility conditions for solvability of (2.1):

$$\int_\Omega f(x)dx + \int_{\partial\Omega} g_N(x)ds_x = 0.$$

In this case, the general solution $u(\cdot)$ to the Neumann boundary value problem will not be unique, and will satisfy $u(x) \equiv u_*(x) + \alpha$ where $u_*(x)$ is any particular non-homogeneous solution and α is a constant.

2.1.2 Weak Formulation

The weak formulation of (2.1) is obtained by multiplying it by a test function $v(.)$ with zero boundary value on $\mathcal{B}_D$, and integrating the resulting expression by parts over Ω. The weak problem will seek $u \in H^1_D(\Omega)$ which satisfies $u(.) = g_D(.)$ on $\mathcal{B}_D$ such that:

$$\mathcal{A}(u, v) = F(v), \qquad \forall v \in H^1_D(\Omega), \tag{2.2}$$

where $\mathcal{A}(\cdot,\cdot)$, $F(\cdot)$ and $H^1_D(\Omega)$ are defined by:

$$\begin{cases} \mathcal{A}(u, v) \equiv \int_\Omega (\nabla u \cdot a\nabla v + c\, u\, v)\; dx + \int_{\mathcal{B}_N} \gamma\, u\, v\, ds_x, \\ F(v) \quad\ \equiv \int_\Omega f\, v\, dx + \int_{\mathcal{B}_N} g_N v\, ds_x, \\ H^1_D(\Omega) \equiv \left\{ v \in H^1(\Omega) : v = 0 \text{ on } \mathcal{B}_D \right\}. \end{cases} \tag{2.3}$$

Here $H^1_D(\Omega)$ denotes the space satisfying zero *Dirichlet* boundary conditions.

2.1.3 Finite Element Discretization

Let $\mathcal{T}_h(\Omega)$ denote a quasiuniform triangulation of $\Omega \subset \mathbb{R}^d$ with elements of size h. For simplicity, we assume that the elements are simplices (triangles when $d = 2$ or tetrahedra when $d = 3$) and that $V_h \subset H^1(\Omega)$ is the space of continuous *piecewise linear* finite element functions on $\mathcal{T}_h(\Omega)$. Homogeneous essential boundary conditions can be imposed in V_h by choosing $V_h \cap H^1_D(\Omega)$. The finite element discretization of (2.1), see [ST14, CI2, JO2, BR28, BR], will seek $u_h \in V_h$ with $u_h = I_h\, g_D$ on $\mathcal{B}_D$ and satisfying:

$$\mathcal{A}(u_h, v_h) = F(v_h), \qquad \forall v_h \in V_h \cap H^1_D(\Omega). \tag{2.4}$$

Here I_h denotes the nodal interpolation onto V_h, restricted to $\mathcal{B}_D$. This will yield a linear system $A_h \mathbf{u}_h = \mathbf{f}_h$. We shall often omit the subscript h.

Let n_I, $n_{\mathcal{B}_N}$ and $n_{\mathcal{B}_D}$ denote the number of nodes of triangulation $\mathcal{T}_h(\Omega)$ in the interior of Ω, the boundary segments $\mathcal{B}_N$ and $\mathcal{B}_D$, respectively. Denote by x_i for $1 \le i \le (n_I + n_{\mathcal{B}_N} + n_{\mathcal{B}_D})$ all the nodes of $\mathcal{T}_h(\Omega)$. We assume that these nodes are so ordered that:

$$\begin{cases} x_i \in \Omega, & \text{for } 1 \le i \le n_I \\ x_i \in \mathcal{B}_N, & \text{for } (n_I + 1) \le i \le (n_I + n_{\mathcal{B}_N}) \\ x_i \in \mathcal{B}_D, & \text{for } (n_I + n_{\mathcal{B}_N} + 1) \le i \le (n_I + n_{\mathcal{B}_N} + n_{\mathcal{B}_D}). \end{cases}$$

Corresponding to each node $1 \le i \le (n_I + n_{\mathcal{B}_N} + n_{\mathcal{B}_D})$, let $\phi_i(x)$ denote the continuous piecewise linear finite element nodal basis in V_h, satisfying:

$$\phi_i(x_j) = \delta_{ij}, \qquad \text{for } 1 \le i, j \le (n_I + n_{\mathcal{B}_N} + n_{\mathcal{B}_D}),$$

where δ_{ij} denotes the Kronecker delta. Given $u_h(x) \in V_h$, we expand it as:

$$\begin{cases} u_h(x) = \sum_{i=1}^{n_I} (\mathbf{u}_I)_i \phi_i(x) + \sum_{i=1}^{n_{\mathcal{B}_N}} (\mathbf{u}_{\mathcal{B}_N})_i \phi_{n_I+i}(x) \\ \qquad\qquad + \sum_{i=1}^{n_{\mathcal{B}_D}} (\mathbf{u}_{\mathcal{B}_D})_i \phi_{n_I+n_{\mathcal{B}_N}+i}(x), \end{cases}$$

where $\mathbf{u}_I$, $\mathbf{u}_{\mathcal{B}_N}$ and $\mathbf{u}_{\mathcal{B}_D}$ denote subvectors defined by:

$$\begin{cases} (\mathbf{u}_I)_i \; \equiv u_h(x_i), & 1 \le i \le n_I, \\ (\mathbf{u}_{\mathcal{B}_N})_i \equiv u_h(x_{n_I+i}), & 1 \le i \le n_{\mathcal{B}_N}, \\ (\mathbf{u}_{\mathcal{B}_D})_i \equiv u_h(x_{n_I+n_{\mathcal{B}_N}+i}), & 1 \le i \le n_{\mathcal{B}_D}. \end{cases}$$

This block partitions the vector of nodal values associated with u_h as:

$$\mathbf{u}_h = \left(\mathbf{u}_I^T, \mathbf{u}_{\mathcal{B}_N}^T, \mathbf{u}_{\mathcal{B}_D}^T\right)^T,$$

corresponding to the ordering of nodes in Ω, $\mathcal{B}_N$ and $\mathcal{B}_D$, respectively.

Employing the above block partition, the finite element discretization (2.4) of (2.1) is easily seen to have the following block structure:

$$\begin{cases} A_{II}\mathbf{u}_I \;+\; A_{I\mathcal{B}_N}\mathbf{u}_{\mathcal{B}_N} \;+\; A_{I\mathcal{B}_D}\mathbf{u}_{\mathcal{B}_D} = \mathbf{f}_I \\ A_{I\mathcal{B}_N}^T\mathbf{u}_I + A_{\mathcal{B}_N\mathcal{B}_N}\mathbf{u}_{\mathcal{B}_N} + A_{\mathcal{B}_N\mathcal{B}_D}\mathbf{u}_{\mathcal{B}_D} = \mathbf{f}_{\mathcal{B}_N} \\ \qquad\qquad\qquad\qquad\qquad\qquad \mathbf{u}_{\mathcal{B}_D} = I_h\, g_{_D}, \end{cases}$$

where the block submatrices and vectors above are defined by:

$$\begin{cases} (A_{II})_{ij} & = \mathcal{A}(\phi_i, \phi_j), & 1 \le i,\, j \le n_I \\ (A_{IB_N})_{ij} & = \mathcal{A}(\phi_i, \phi_{n_I+j}), & 1 \le i \le n_I, \quad 1 \le j \le n_{\mathcal{B}_\mathcal{N}} \\ (A_{IB_D})_{ij} & = \mathcal{A}(\phi_i, \phi_{n_I+n_{\mathcal{B}_\mathcal{N}}+j}), & 1 \le i \le n_I, \quad 1 \le j \le n_{\mathcal{B}_\mathcal{D}} \\ (A_{\mathcal{B}_\mathcal{N}\mathcal{B}_\mathcal{N}})_{ij} & = \mathcal{A}(\phi_{n_I+i}, \phi_{n_I+j}), & 1 \le i,\, j \le n_{\mathcal{B}_\mathcal{N}} \\ (A_{\mathcal{B}_\mathcal{N}\mathcal{B}_\mathcal{D}})_{ij} & = \mathcal{A}(\phi_{n_I+i}, \phi_{n_I+n_{\mathcal{B}_\mathcal{N}}+j}), & 1 \le i \le n_{\mathcal{B}_\mathcal{N}}, \quad 1 \le j \le n_{\mathcal{B}_\mathcal{D}} \\ (\mathbf{f}_I)_i & = F(\phi_i), & 1 \le i \le n_I \\ (\mathbf{f}_{\mathcal{B}_\mathcal{N}})_i & = F(\phi_{n_I+i}), & 1 \le i \le n_{\mathcal{B}_\mathcal{N}} \\ (I_h g_{_D})_i & = g_{_D}(x_{n_I+n_{\mathcal{B}_\mathcal{N}}+i}), & 1 \le i \le n_{\mathcal{B}_\mathcal{D}}. \end{cases}$$

Eliminating $\mathbf{u}_{\mathcal{B}_D}$ in the above linear system yields:

$$\begin{cases} A_{II}\mathbf{u}_I \;+\; A_{I\mathcal{B}_N}\mathbf{u}_{\mathcal{B}_N} \;= \mathbf{f}_I \;\;\; - A_{I\mathcal{B}_D} I_h g_{_D} \\ A_{I\mathcal{B}_N}^T\mathbf{u}_I + A_{\mathcal{B}_N\mathcal{B}_N}\mathbf{u}_{\mathcal{B}_N} = \mathbf{f}_{\mathcal{B}_N} - A_{\mathcal{B}_N\mathcal{B}_D} I_h g_{_D}. \end{cases} \tag{2.5}$$

In matrix notation, this yields the block partitioned linear system:

$$\begin{bmatrix} A_{II} & A_{I\mathcal{B}_N} \\ A_{I\mathcal{B}_N}^T & A_{\mathcal{B}_N\mathcal{B}_N} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_{\mathcal{B}_N} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_I \\ \tilde{\mathbf{f}}_{\mathcal{B}_N} \end{bmatrix},$$

where

$$\begin{bmatrix} \tilde{\mathbf{f}}_I \\ \tilde{\mathbf{f}}_{\mathcal{B}_N} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{f}_I - A_{I\mathcal{B}_D} I_h g_{_D} \\ \mathbf{f}_{\mathcal{B}_N} - A_{\mathcal{B}_N\mathcal{B}_D} I_h g_{_D} \end{bmatrix}.$$

Remark 2.2. If $\mathcal{B}_N = \emptyset$, then problem (2.1) will be a Dirichlet problem with $\partial\Omega = \mathcal{B}_D$. In this case, the discretization reduces to:

$$A_h\mathbf{u}_h = \mathbf{f}_h, \tag{2.6}$$

with $A_h \equiv A_{II}$ and $\mathbf{f}_h \equiv \mathbf{f}_I - A_{I\mathcal{B}} I_h g_{\mathcal{B}}$, where we have denoted $\mathcal{B} \equiv \mathcal{B}_D$.

Remark 2.3. If $\mathcal{B}_D = \emptyset$, then (2.1) will be a Robin problem if $\gamma(x) \ne 0$, or a Neumann problem if $\gamma(x) \equiv 0$. In this case $\partial\Omega = \mathcal{B}_N$ and we shall use the notation $\mathcal{B} \equiv \mathcal{B}_N$. The discretization of (2.1) will then have the form:

$$A_h\mathbf{u}_h = \mathbf{f}_h, \;\; \text{with} \;\; A_h \equiv \begin{bmatrix} A_{II} & A_{I\mathcal{B}} \\ A_{I\mathcal{B}}^T & A_{\mathcal{B}\mathcal{B}} \end{bmatrix}, \;\; \mathbf{u}_h \equiv \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_{\mathcal{B}} \end{bmatrix}, \;\; \mathbf{f}_h \equiv \begin{bmatrix} \tilde{\mathbf{f}}_I \\ \tilde{\mathbf{f}}_{\mathcal{B}} \end{bmatrix}. \tag{2.7}$$

If $\gamma(x) \equiv 0$ and $c(x) \equiv 0$, then matrix A_h will be singular, satisfying $A_h\, \mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ and $\mathbf{0}$ denote vectors of appropriate size having all entries identically 1 or 0, respectively. In this case, the forcing $\mathbf{f}_h$ in (2.7) will be required to satisfy the compatability condition $\mathbf{1}^T \mathbf{f}_h = 0$, for the linear system to be solvable. The solution space will then have the form $\mathbf{u}_h = \mathbf{u}_h^* + \alpha\, \mathbf{1}$ for $\alpha \in \mathbb{R}$, where $\mathbf{u}_h^*$ is any particular solution.

2.1.4 Multisubdomain Decompositions

We employ the following notation for multidomain decompositions, see Fig. 2.1.

Definition 2.4. *A collection of open subregions $\Omega_i \subset \Omega$ for $1 \le i \le p$ will be referred to as a nonoverlapping decomposition of Ω if the following hold:*

$$\begin{cases} \cup_{l=1}^{p} \overline{\Omega}_i = \overline{\Omega}, \\ \Omega_i \cap \Omega_j = \emptyset, \quad \text{if } \neq j. \end{cases}$$

Boundaries of the subdomains will be denoted $B_i \equiv \partial \Omega_i$ and their interior and exterior segments by $B^{(i)} \equiv \partial \Omega_i \cap \Omega$ and $B_{[i]} \equiv \partial \Omega_i \cap \partial \Omega$, respectively. We denote common interfaces by $B_{ij} \equiv B_i \cap B_j$ and $B \equiv \cup_i B^{(i)}$.

When the subdomains Ω_i are *shape regular*, we let h_0 denote its diameter. For additional notation on non-overlapping subdomains, see Chap. 3.

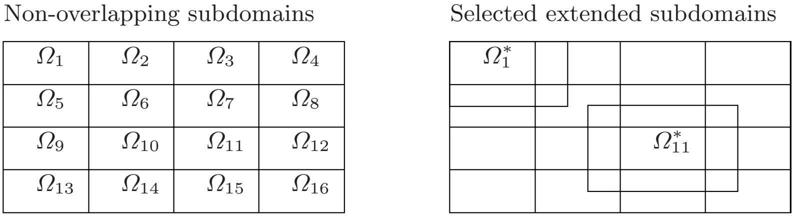
Definition 2.5. *A collection of open subregions $\Omega_i^* \subset \Omega$ for $1 \le i \le p$ will be referred to as an overlapping decomposition of Ω if the following holds:*

$$\cup_{l=1}^{p} \Omega_i^* = \Omega.$$

If $\{\Omega_l\}_{l=1}^p$ forms a non-overlapping decomposition of Ω of diameter h_0 and each $\Omega_i \subset \Omega_i^*$, then $\{\Omega_l^*\}_{l=1}^p$ will be said to form an overlapping decomposition of Ω obtained by extension of $\{\Omega_l\}_{l=1}^p$. Most commonly:

$$\Omega_i^* \equiv \Omega_i^{\beta\, h_0} \equiv \{x \in \Omega \,:\, \text{dist}(x, \Omega_i) < \beta\, h_0\} \tag{2.8}$$

where $0 < \beta < 1$ is called the overlap factor. Boundaries will be denoted $\partial \Omega_i^*$ and with abuse of notation, $B^{(i)} \equiv \partial \Omega_i^* \cap \Omega$ and $B_{[i]} \equiv \partial \Omega_i^* \cap \partial \Omega$, respectively.


Non-overlapping subdomains

Selected extended subdomains

Fig. 2.1. Multidomain overlapping and non-overlapping decompositions

2.1.5 Restriction and Extension Maps

Restriction and extension maps are *rectangular* matrices used for representing domain decomposition preconditioners. A restriction map will *restrict* a vector of nodal values to a subvector corresponding to indices in some index set $\mathcal{S}$. An extension map will *extend* a subvector of nodal values in $\mathcal{S}$ to a full vector, whose entries will be *zero* outside $\mathcal{S}$. Formally, given any subregion $\mathcal{S} \subset (\Omega \cup \mathcal{B}_N)$, order the nodes of $\mathcal{T}_h(\Omega)$ in $\mathcal{S}$ in some *local ordering*. Let $n \equiv (n_I + n_{\mathcal{B}_N})$ denote the total number of finite element unknowns, and $n_{\mathcal{S}}$ the number of nodes of $\mathcal{T}_h(\Omega)$ in $\mathcal{S}$. We shall associate an index function $\mathit{index}(\mathcal{S}, i)$ to denote the global index of the i'th local node in $\mathcal{S}$ for $1 \le i \le n_{\mathcal{S}}$. We then define an $n_{\mathcal{S}} \times n$ *restriction* matrix $R_{\mathcal{S}}$ which will map a vector in $\mathbb{R}^n$ of nodal values on the grid $\mathcal{T}_h(\Omega)$ into a subvector in $\mathbb{R}^{n_S}$ of nodal values associated with the nodes in $\mathcal{S}$ in the local ordering:

$$(R_{\mathcal{S}})_{ij} = \begin{cases} 1 & \text{if } \mathit{index}(\mathcal{S}, i) = j \\ 0 & \text{if } \mathit{index}(\mathcal{S}, i) \neq j. \end{cases} \tag{2.9}$$

The transpose $R_{\mathcal{S}}^T$ of restriction matrix $R_{\mathcal{S}}$ is referred to as an *extension* matrix. It will be an $n \times n_{\mathcal{S}}$ matrix which extends a vector in $\mathbb{R}^{n_S}$ to a vector in $\mathbb{R}^n$ with zero entries corresponding to indices not in $\mathcal{S}$.

Remark 2.6. Given a vector $\mathbf{v} \in \mathbb{R}^n$ of nodal values in $\mathcal{T}_h(\Omega)$, the vector $R_{\mathcal{S}}\,\mathbf{v} \in \mathbb{R}^{n_S}$ will denote its subvector corresponding to indices of nodes in $\mathcal{S}$ (using the local ordering of nodes in $\mathcal{S}$). Given a nodal vector $\mathbf{v}_{\mathcal{S}} \in \mathbb{R}^{n_S}$ of nodal values in $\mathcal{S}$, the vector $R_{\mathcal{S}}^T \mathbf{v}_{\mathcal{S}} \in \mathbb{R}^n$ will denote a nodal vector in $\mathcal{T}_h(\Omega)$ which extends $\mathbf{v}_{\mathcal{S}}$ to have *zero* nodal values at all nodes not in $\mathcal{S}$. To implement such maps, their action on vectors should be computed algorithmically employing suitable data structures and *scatter-gather* operations.

Remark 2.7. Given the global stiffness matrix A_h of size n, its submatrix $A_{\mathcal{S}\mathcal{S}}$ of size $n_{\mathcal{S}}$ corresponding to the nodes in $\mathcal{S}$ may be expressed formally as:

$$A_{\mathcal{S}\mathcal{S}} = R_{\mathcal{S}} A_h R_{\mathcal{S}}^T.$$

In implementations, the action of $A_{\mathcal{S}\mathcal{S}}$ on vectors should be computed algorithmically employing *scatter-gather* operations and sparse data structures.

Remark 2.8. Typical choices of $\mathcal{S}$ in Schwarz algorithms will be indices of nodes in $\Omega_i^* \cup (\mathcal{B}_N \cap \partial\Omega_i^*)$. (In Schur complement algorithms, see Chap. 3, the set $\mathcal{S}$ will correspond to indices of nodes on segments, called *globs*, of the subdomain boundaries $B^{(i)}$. The notation $\mathcal{R}_{\mathcal{S}}$ and $\mathcal{R}_{\mathcal{S}}^T$ will be used).

2.1.6 Partition of Unity

Given an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ of Ω, we shall often employ a smooth partition of unity $\chi_1(x), \ldots, \chi_p(x)$ subordinate to these subdomains. The partition of unity functions must satisfy the following requirements:

$$\begin{cases} \chi_i(x) \geq 0, \text{ in } \overline{\Omega}_i^* \\ \chi_i(x) = 0, \text{ in } \overline{\Omega} \backslash \overline{\Omega}_i^* \\ \chi_1(x) + \cdots + \chi_p(x) = 1, \text{ in } \overline{\Omega}. \end{cases} \tag{2.10}$$

As in Chap. 1.1, a *continuous* partition of unity may be constructed based on the distance functions $d_i(x) \equiv \text{dist}(x, \partial\Omega_i^* \cap \Omega) \geq 0$ as follows:

$$\chi_i(x) \equiv \frac{d_i(x)}{d_1(x) + \cdots + d_p(x)}, \quad \text{for} \quad 1 \leq i \leq p.$$

Smoother $\chi_i(x)$ may be obtained by using *mollified* $d_i(x)$, see [ST9].

2.1.7 Coarse Spaces

The convergence rate of *one-level* domain decomposition algorithms (namely, algorithms involving only subdomains problems) will typically *deteriorate* as the number p of subdomains increases. This may be understood *heuristically* as follows. Consider a rectangular domain Ω divided into p vertical strips. Each iteration, say of a Schwarz alternating method, will only transfer information between adjacent subdomains. Thus, if the forcing term is nonzero only in the first strip and the starting iterate is zero, then it will take p iterations for the local solution to be nonzero in the p'th subdomain. For elliptic equations (which have a global domain of dependence on the solution, due to the Green's function representation), the solution will typically be nonzero globally even when the forcing term is nonzero only in a small subregion. Thus, an algorithm such as the classical Schwarz alternating method (and other one-level methods) will impose limits on the speed at which information is transferred globally across the entire domain.

The preceding limitation in the rate of convergence of one-level domain decomposition iterative algorithms can be handled if a mechanism is included for the *global transfer* of information across the subdomains. Motivated by multigrid methodology [BR22, HA2, MC2] and its generalizations [DR11, XU3], such a global transfer of information can be incorporated by solving a subproblem on an appropriately chosen *subspace* of the finite element space, whose support covers the entire domain. Such subspaces are referred to as *coarse spaces*, provided they satisfy specified assumptions. A simple example would be the space of coarse grid finite element functions defined on a *coarse* triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω, as in two-level multigrid methods. In the following, we list the approximation property desired in such coarse spaces, where $0 < h_0$ represents a small parameter (typically denoting the subdomain size).

Definition 2.9. *A subspace $V_0 \subset V_h \cap H_D^1(\Omega)$ will be referred to as a coarse space having approximation of order $O(h_0)$ if the following hold:*

$$\begin{cases} \|Q_0 \, u_h\|_{H^1(\Omega)} \leq C \, \|u_h\|_{H^1(\Omega)}, & \forall u_h \in V_h \cap H_D^1(\Omega) \\ \|u_h - Q_0 \, u_h\|_{L^2(\Omega)} \leq C \, h_0 \|u_h\|_{H^1(\Omega)}, & \forall u_h \in V_h \cap H_D^1(\Omega) \end{cases}$$

where Q_0 denotes the L^2-orthogonal projection onto subspace $V_0 \cap H_D^1(\Omega)$.

Using a coarse space $V_0 \subset V_h$, information may be transferred globally across many subdomains, by solving a finite dimensional global problem, using *residual correction* as follows. Suppose w_h denotes an approximate solution of discrete problem (2.4) in $V_h \cap H_D^1(\Omega)$. An improved approximation $w_h + w_0$ of u_h may be sought by selecting $w_0 \in V_0$ so that it satisfies the following *residual* equation:

$$\mathcal{A}(w_0, v) = F(v) - \mathcal{A}(w_h, v), \qquad \forall v \in V_0. \tag{2.11}$$

It is easily verified that w_0 is the $\mathcal{A}(.,.)$-orthogonal *projection* of $u_h - w_h$ onto the subspace V_0. Once w_0 is determined, $w_h + w_0$ will provide an improved approximation of the desired solution u_h.

The preceding coarse space residual problem (2.11) can be represented in matrix terms as follows. Let n_0 denote the dimension of $V_0 \subset V_h \cap H_D^1(\Omega)$ and let $\psi_1^{(0)}(\cdot), \cdots, \psi_{n_0}^{(0)}(\cdot)$ denote a basis for V_0. If $n = (n_I + n_{\mathcal{B}_N})$ is the dimension of $V_h \cap H_D^1(\Omega)$, let $x_1, \cdots, x_n$ denote the nodes in $(\Omega \cup \mathcal{B}_N)$. Define an $n \times n_0$ matrix R_0^T whose entries are defined as follows:

$$R_0^T = \begin{bmatrix} \psi_1^{(0)}(x_1) & \cdots & \psi_{n_0}^{(0)}(x_1) \\ \vdots & & \vdots \\ \psi_1^{(0)}(x_n) & \cdots & \psi_{n_0}^{(0)}(x_n) \end{bmatrix}.$$

Let $\mathbf{w}_0 = R_0^T \boldsymbol{\alpha}$ and $\mathbf{v} = R_0^T \boldsymbol{\beta}$ denote nodal vectors representing w_0 and v above, for suitable coefficient vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{n_0}$. Then (2.11) becomes:

$$\boldsymbol{\beta}^T (R_0 A_h R_0^T) \boldsymbol{\alpha} = \boldsymbol{\beta}^T R_0 \left(\mathbf{f}_h - A_h \mathbf{w}_h\right), \quad \forall \boldsymbol{\beta} \in \mathbb{R}^{n_0}.$$

This yields the linear system $A_0 \, \boldsymbol{\alpha} = R_0 \left(\mathbf{f}_h - A_h \mathbf{w}_h\right)$, where $A_0 = (R_0 A_h R_0^T)$. The vector update to the approximate solution $\mathbf{w}_h$ will then be $\mathbf{w}_h + R_0^T \boldsymbol{\alpha}$, which may also be expressed as $\mathbf{w}_h + R_0^T A_0^{-1} R_0 \left(\mathbf{f}_h - A_h \mathbf{w}_h\right)$. Four specific coarse spaces V_0 are described in the following. Additional spaces are described in [BR15, SM2, CO8, SA7, WI6, MA17].

Coarse Triangulation Space. If domain Ω can be triangulated by a quasi-uniform triangulation $\mathcal{T}_{h_0}(\Omega)$ with elements of size $h_0 > h$, such that $\mathcal{T}_h(\Omega)$ is obtained by successive refinement of $\mathcal{T}_{h_0}(\Omega)$, then a coarse space V_0 can be defined as the space of continuous, piecewise linear finite element functions on triangulation $\mathcal{T}_{h_0}(\Omega)$. To enforce homogeneous essential boundary conditions so that $V_0 \subset V_h \cap H_D^1(\Omega)$, the Dirichlet boundary segment $\mathcal{B}_D$ must the union of boundary segments of elements of $\mathcal{T}_{h_0}(\Omega)$. Such coarse spaces are motivated by multigrid methodology.

Interpolation of a Coarse Triangulation Space. If the geometry of Ω is complex or the triangulation $\mathcal{T}_h(\Omega)$ is *unstructured*, then it may be computationally difficult, if not impossible, to construct a coarse triangulation

$\mathcal{T}_{h_0}(\Omega)$ of Ω from which to obtain $\mathcal{T}_h(\Omega)$ by successive refinement. In such cases, an alternative coarse space [CA4, CH17] can be constructed as follows, when $\mathcal{B}_N = \emptyset$. Let $\Omega^* \supset \Omega$ denote an *extension* of Ω having simpler geometry (such as a polygon). Let $\mathcal{T}_{h_0}(\Omega^*)$ denote a coarse triangulation of Ω^* having elements of size $h_0 > h$. The elements of $\mathcal{T}_{h_0}(\Omega^*)$ will in general not be the union of elements in $\mathcal{T}_h(\Omega)$. Despite this, a coarse subspace of V_h can be defined as follows. Let $V_{h_0}(\Omega^*) \subset H_0^1(\Omega^*)$ denote a finite element space on triangulation $\mathcal{T}_{h_0}(\Omega^*)$ of Ω^* with zero boundary values. Define V_0 as:

$$V_0 \equiv \{\pi_h w_{h_0}^* : w_{h_0}^* \in V_{h_0}(\Omega^*)\},$$

where π_h denotes the standard nodal interpolation onto all grid points of $\mathcal{T}_h(\Omega)$ *excluding* nodes on $\mathcal{B}_D$. By construction $V_0 \subset V_h \cap H_D^1(\Omega)$.

Interpolation of a Polynomial Space. If as in the preceding case, the geometry of Ω is complex or the triangulation $\mathcal{T}_h(\Omega)$ is unstructured, and $\mathcal{B}_D = \emptyset$, then a coarse space may be defined as follows. Let $\mathcal{P}_d(\Omega)$ denote the space of all polynomials of degree d or less on Ω. Generally $\mathcal{P}_d(\Omega) \not\subset V_h$. However, we may interpolate such polynomials onto the finite element space $V_h \cap H_D^1(\Omega)$ as follows:

$$V_0 \equiv \{\pi_h w_d(x) : w_d(x) \in \mathcal{P}_d(\Omega)\},$$

where π_h denotes the standard nodal interpolant onto the finite element space $V_h \cap H_D^1(\Omega)$. By construction $V_0 \subset V_h \cap H_D^1(\Omega)$.

Piecewise Constant Space. A more general coarse space, referred to as the *piecewise constant* coarse space [CO8, SA7, MA17, WA6], can be constructed given any nonoverlapping decomposition $\Omega_1, \ldots, \Omega_p$ of Ω as follows. Let h_0 denote the size of the subdomains and define Ω_i^* as the extension of Ω_i containing all points of Ω within a distance $\beta\, h_0$ to Ω_i. Let $\chi_1(.), \ldots, \chi_p(.)$ denote a *partition of unity* based on $\Omega_1^*, \ldots, \Omega_p^*$. This partition of unity should be constructed so that its sum is zero on $\mathcal{B}_D$ and unity on $\mathcal{B}_N$. Denote the union of subdomain interfaces as $B \equiv (\cup_{i=1}^p \partial\Omega_i) \setminus \mathcal{B}_D$.

Define a restriction map R_B which restricts any function $w(x)$ onto B:

$$R_B\, w(x) \equiv w(x), \qquad \text{for} \quad x \in B.$$

Given a function $v(x)$ defined on B, denotes its piecewise harmonic extension $Ev(x)$ into the interior of each subdomain Ω_i for $1 \le i \le p$ as:

$$\begin{cases} L\,(Ev) = 0, & \text{in } \Omega_i \\ Ev = v, & \text{on } \partial\Omega_i, \end{cases}$$

where $L\, Ev$ denotes the elliptic operator applied to Ev. The *continuous* version of the piecewise constant coarse space V_0 is now defined as:

$$V_0 \equiv \ \text{span}\,[E\, R_B\, \chi_1 \ \ldots \ E\, R_B\, \chi_p]\,.$$

A finite element version of V_0 can be constructed analogously, see Chap. 2.5, using restriction onto nodal values on B and discrete harmonic extensions into the subdomains. If the coefficient $a(.)$ in (2.1) is discontinuous of the form:

$$a(x) \equiv a_i \qquad \text{for } x \in \Omega_i, \;\; 1 \le i \le p,$$

then it will be advantageous to rescale the original partition of unity to account for large variation in $a(.)$. A new partition of unity $\hat{\chi}_1(.), \ldots, \hat{\chi}_p(.)$ will be:

$$\hat{\chi}_i(x) \equiv \frac{a_i\, \chi_i(x)}{a_1\, \chi_1(x) + \cdots + a_p\, \chi_p(x)} \qquad \text{for } \; 1 \le i \le p.$$

An alternative coarse space $\hat{V}_0$ can be constructed based on this.

2.2 Projection Formulation of Schwarz Algorithms

In this section, we describe the classical Schwarz alternating method for iteratively solving the following coercive elliptic equation:

$$\begin{cases} Lu \equiv -\nabla \cdot (a(x)\nabla) + c(x)\, u = f, & \text{in } \Omega \\ \qquad\qquad\quad \mathbf{n} \cdot (\, a\nabla u) + \gamma\, u = g_N, & \text{in } \mathcal{B}_N \\ \qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \mathcal{B}_D, \end{cases} \tag{2.12}$$

where $c(x) \ge 0$, $\gamma(x) \ge 0$, and $\mathcal{B}_D$ and $\mathcal{B}_N$ denote Dirichlet and natural boundary segments of $\partial\Omega$. The weak formulation of (2.12) seeks $u \in H^1_D(\Omega)$:

$$\mathcal{A}(u, v) = F(v), \qquad \forall v \in H^1_D(\Omega), \tag{2.13}$$

where

$$\begin{cases} \mathcal{A}(u, v) \equiv \int_\Omega \left(a(x)\nabla v \cdot \nabla v + c(x)\, u\, v\right)\, dx & \\ \qquad\qquad + \int_{\mathcal{B}_\mathcal{N}} \gamma(x)\, u\, v\, ds(x), & \text{for } u,\, v \in H^1_D(\Omega) \\ F(v) \quad\; \equiv \int_\Omega f(x)\, v(x)\, dx + \int_{\mathcal{B}_\mathcal{N}} g_N(x)\, v\, ds(x), & \text{for } v \in H^1_D(\Omega) \\ H^1_D(\Omega) \equiv \left\{ v \in H^1(\Omega) : v = 0 \text{ on } \mathcal{B}_D \right\}. & \end{cases} \tag{2.14}$$

Applying integration by parts to the *continuous* version of the multidomain Schwarz alternating method, we shall derive a formal expression for the *updates* in the iterates as involving orthogonal *projections* onto certain subspaces of $H^1_D(\Omega)$. Employing these projections, we shall derive various parallel extensions of the classical Schwarz alternating method, including the *additive* Schwarz, *hybrid* Schwarz and *restricted* Schwarz methods. Let $\Omega_1^*, \cdots, \Omega_p^*$ denote an overlapping decomposition of Ω, and let $B^{(i)} \equiv \partial\Omega_i^* \cap \Omega$ and $B_{[i]} \equiv \partial\Omega_i^* \cap \partial\Omega$ denote the interior and exterior boundary segments of Ω_i^*.

2.2.1 Classical Schwarz Alternating Method

Let $w^{(0)}$ denote a starting iterate satisfying $w^{(0)} = 0$ on $\mathcal{B}_D$. Then, the multidomain Schwarz alternating method will iteratively seek the solution to (2.12) by *sequentially* updating the iterate on each subdomain Ω_i^* in some prescribed order. Each iteration (or *sweep*) will consist of p fractional steps and we shall denote the iterate in the i'th fractional step of the k'th sweep as $w^{(k+\frac{i}{p})}$. Given $w^{(k+\frac{i-1}{p})}$ the next iterate $w^{(k+\frac{i}{p})}$ is computed as follows:

$$\begin{cases} -\nabla \cdot \left(a(x)\nabla w^{(k+\frac{i}{p})}\right) + c(x)\, w^{(k+\frac{i}{p})} = f(x), & \text{in } \Omega_i^* \\ \mathbf{n} \cdot \left(a\nabla w^{(k+\frac{i}{p})}\right) + \gamma\, w^{(k+\frac{i}{p})} = g_N, & \text{on } B_{[i]} \cap \mathcal{B}_N \\ w^{(k+\frac{i}{p})} = w^{(k+\frac{i-1}{p})}, & \text{on } B^{(i)} \\ w^{(k+\frac{i}{p})} = 0, & \text{on } B_{[i]} \cap \mathcal{B}_D. \end{cases} \tag{2.15}$$

The local solution $w^{(k+\frac{i-1}{p})}$ is then extended outside Ω_i^* as follows:

$$w^{(k+\frac{i}{p})} \equiv w^{(k+\frac{i-1}{p})}, \quad \text{on} \quad \Omega \setminus \overline{\Omega}_i^*. \tag{2.16}$$

The resulting iterates will thus be *continuous* on Ω by construction.

Algorithm 2.2.1 *(Continuous Schwarz Alternating Method)*
Input: $w^{(0)}$ starting iterate.

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     For $i = 1, \cdots, p$ solve:
           $\begin{cases} -\nabla \cdot \left(a(x)\nabla v^{(k+\frac{i}{p})}\right) + c(x)\, v^{(k+\frac{i}{p})} = f(x), & \text{in } \Omega_i^* \\ \mathbf{n} \cdot \left(a\nabla v^{(k+\frac{i}{p})}\right) + \gamma\, v^{(k+\frac{i}{p})} = g_N, & \text{on } B_{[i]} \cap \mathcal{B}_N \\ v^{(k+\frac{i}{p})} = w^{(k+\frac{i-1}{p})}, & \text{on } B^{(i)} \\ v^{(k+\frac{i}{p})} = 0, & \text{on } B_{[i]} \cap \mathcal{B}_D. \end{cases}$
       Update:
           $w^{(k+\frac{i}{p})} \equiv \begin{cases} v^{(k+\frac{i}{p})}, & \text{on } \overline{\Omega}_i^* \\ w^{(k+\frac{i-1}{p})}, & \text{on } \Omega \setminus \overline{\Omega}_i^*. \end{cases}$
3.     Endfor
4. Endfor
```

The iterates $w^{(k)}(.)$ will converge geometrically to the solution $u(.)$ with:

$$\|u - w^{(k)}\|_{H^1(\Omega)} \leq \delta^k \, \|u - w^{(0)}\|_{H^1(\Omega)}.$$

The convergence factor $0 < \delta < 1$ will generally depend on the overlap β between the subdomains, the diameters $\text{diam}(\Omega_i^*)$ of the subdomains, and the coefficients in (2.1), see Chap. 2.5 and Chap. 15.

As the number p of subdomains increases, the convergence rate typically deteriorates yielding $\delta \to 1$. This is because the true solution to (2.12) has a global domain of dependence on $f(.)$, while if $w^{(0)} = 0$ and $f(.)$ has support in only one subdomain, then since information is transferred only between adjacent subdomains during each sweep of the Schwarz iteration, it may generally take p sweeps before this information is transferred globally. Such a deterioration in the convergence, however, can often be remedied by using *coarse space* residual correction (described later).

The Schwarz alternating Alg. 2.2.1 is also known as the *multiplicative* or *sequential* Schwarz algorithm. It is sequential in nature. However, parallelizability of this algorithm can be significantly improved by grouping the subdomains into *colors* so that distinct subdomains of the same color do not intersect. Then, all subproblems on subdomains of the same color can be solved *concurrently*, since such subdomain does not intersect.

Definition 2.10. *Given subdomains $\Omega_1^*, \cdots, \Omega_p^*$, a partition $\mathcal{C}_1, \cdots, \mathcal{C}_d$ of the index set $\{1, 2, \cdots, p\}$ is said to yield a d-coloring of the subdomains if:*

$$i, j \in \mathcal{C}_k \quad \textit{with} \quad i \neq j \qquad \Longrightarrow \qquad \Omega_i^* \cap \Omega_j^* = \emptyset,$$

so that subdomains of the same color $\mathcal{C}_k$ do not intersect.

The following is the multicolor Schwarz algorithm with starting iterate $w(.)$.

Algorithm 2.2.2 *(Multicolor Schwarz Alternating Algorithm)*
Input: $w(.)$

```
1. For $k = 0, \cdots$ until convergence do:
2.     For $l = 1, \cdots, d$ do:
3.         For each $i \in \mathcal{C}_l$ solve in parallel:
```

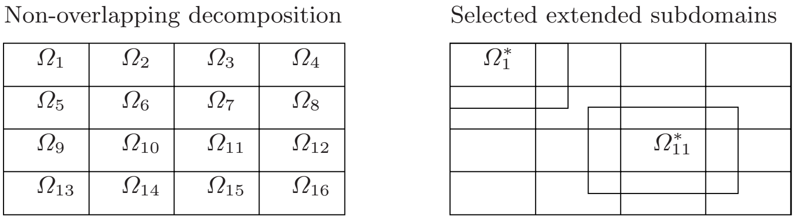
$$\begin{cases} -\nabla \cdot \left(a(x) \nabla v^{(k+\frac{i}{p})} \right) + c(x)\, v^{(k+\frac{i}{p})} = f(x), & \textit{in } \Omega_i^* \\ \mathbf{n} \cdot \left(a \nabla v^{(k+\frac{i}{p})} \right) + \gamma v^{(k+\frac{i}{p})} = g_N, & \textit{on } B_{[i]} \cap \mathcal{B}_N \\ v^{(k+\frac{i}{p})} = w, & \textit{on } B^{(i)} \\ v^{(k+\frac{i}{p})} = 0, & \textit{on } B_{[i]} \cap \mathcal{B}_D. \end{cases}$$

```
            Update:
```

$$w \leftarrow v^{(k+\frac{i}{p})}, \quad \textit{on} \quad \overline{\Omega}_i^*.$$

```
4.         Endfor
5.     Endfor
6.     $w^{(k+1)} \leftarrow w$
7. Endfor
```

Output: $w(.)$



Non-overlapping decomposition

Selected extended subdomains

Fig. 2.2. Multisubdomain overlapping decomposition

Remark 2.11. To minimize the number d of sequential steps, the number of colors d should be chosen to be as small as possible. Additionally, to ensure that the loads assigned to each processor are balanced, there should be approximately the same number of subdomains of each color, and each subdomain should be approximately of the same diameter. For instance, the subdomains $\Omega_1^*, \cdots, \Omega_{16}^*$ in Fig. 2.2 may be grouped into four colors:

$$\mathcal{C}_1 = \{1, 3, 9, 11\}, \; \mathcal{C}_2 = \{2, 4, 10, 12\}, \; \mathcal{C}_3 = \{5, 7, 13, 15\}, \; \mathcal{C}_4 = \{6, 8, 14, 16\},$$

provided the overlap β is not too large.

Remark 2.12. If q processors are available and the subdomains can be colored into d colors with approximately (p/d) subdomains of the same color, and further if (p/d) is a multiple of q, then subdomains of the same color may be partitioned into q groups and each group assigned to one of the processors. Some communication will be necessary between the different subdomains.

The updates $w^{(k+\frac{i}{p})}$ in the continuous Schwarz alternating method can be expressed in terms of certain projection operators onto subspaces of $H_D^1(\Omega)$, see [MA37, LI6]. On each Ω_i^* define a subspace V_i of $H_D^1(\Omega)$ as:

$$V_i \equiv \left\{ v \in H_D^1(\Omega) \; : \; v = 0 \text{ in } \Omega \setminus \overline{\Omega_i^*} \right\}. \tag{2.17}$$

We will employ the property that the bilinear form $\mathcal{A}(.,.)$ in (2.14) defines an inner product on $H_D^1(\Omega)$ when $\mathcal{B}_D \neq \emptyset$, see [CI2, JO2]. We define an $\mathcal{A}(.,.)$-orthogonal *projection* operator P_i onto subspace V_i of $H_D^1(\Omega)$ as follows.

Definition 2.13. *Given $w \in H_D^1(\Omega)$ define $P_i w \in V_i$ as the solution of:*

$$\mathcal{A}(P_i w, v) = \mathcal{A}(w, v), \qquad \textit{for} \qquad v \in V_i.$$

Remark 2.14. The existence and uniqueness of $P_i w$ is guaranteed by the Lax-Milgram lemma, see [CI2]. If u denotes the solution of weak formulation (2.13), then $P_i u$ can be computed without explicit knowledge of u using that $\mathcal{A}(u, v) = F(v)$, since $F(\cdot)$ is given for all $v \in V_i$.

The following result shows that the projection maps P_i can represent the updates in the *continuous* version of the Schwarz alternating method.

Lemma 2.15. *Suppose the following assumptions hold.*

1. *Let u satisfy (2.13) and let $g_N(x) \equiv \mathbf{n} \cdot (a(x)\nabla u) + \gamma(x)\, u$ on $\mathcal{B}_N$.*
2. *Given $w \in H^1_D(\Omega)$ let w_i satisfy:*

$$\begin{cases} -\nabla \cdot (a(x)\nabla w_i) + c(x)\, w_i = f(x), & \text{on } \Omega_i^* \\ \mathbf{n} \cdot (a \nabla w_i) + \gamma\, w_i = g_N, & \text{on } B_{[i]} \cap \mathcal{B}_N \\ w_i = w, & \text{on } B^{(i)} \\ w_i = 0, & \text{on } B_{[i]} \cap \mathcal{B}_D. \end{cases} \tag{2.18}$$

with $w_i \equiv w$ on $\Omega \setminus \Omega_i^*$.

Then $w_i = w + P_i\,(u - w)$.

Proof. Multiplying (2.18) by $v \in V_i \subset H^1_D(\Omega)$ (which is zero outside Ω_i^*), and integrating the resulting term by parts yields:

$$\int_{\Omega_i^*} (L w_i)\; v\, dx = \int_{\Omega} (L w_i)\; v\, dx = \mathcal{A}(w_i, v) = F(v) = \mathcal{A}(u, v), \qquad \forall v \in V_i,$$

where $w_i \notin V_i$ due to its boundary conditions. Employing the above yields:

$$\mathcal{A}(w_i - w, v) = \mathcal{A}(u - w, v), \qquad \forall v \in V_i.$$

Since $(w_i - w) = 0$ in $\Omega \setminus \Omega_i^*$ it yields $w_i - w \in V_i$ and $w_i - w = P_i(u - w)$. Thus, we obtain $w_i = w + P_i\,(u - w)$. □

The continuous version of the Schwarz alternating method may now be reformulated in terms of the projection operators P_i onto $V_i \subset H^1_D(\Omega)$. An application of Lemma 2.15 with $w_i \equiv w^{(k+\frac{i}{p})}$ and $w \equiv w^{(k+\frac{i-1}{p})}$ yields:

$$w^{(k+\frac{i}{p})} = w^{(k+\frac{i-1}{p})} + P_i\Big(u - w^{(k+\frac{i-1}{p})}\Big). \tag{2.19}$$

Substituting this representation into the Schwarz alternating method yields its projection formulation.

Algorithm 2.2.3 *(Projection Version of the Classical Schwarz Method)*
Input: $w^{(0)}$ starting iterate.

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     For $i = 1, \cdots, p$ do
           $w^{(k+\frac{i}{p})} = w^{(k+\frac{i-1}{p})} + P_i\Big(u - w^{(k+\frac{i-1}{p})}\Big).$
3.     Endfor
4. Endfor
```

Output: $w^{(k)}$

Remark 2.16. The preceding projection version of the Schwarz alternating method will also be applicable for more general subspaces $V_i \subset H^1_D(\Omega)$. To ensure convergence, however, the subspaces V_i of $H^1_D(\Omega)$ must satisfy:

$$H^1_D(\Omega) = V_1 + \cdots + V_p$$

see Chap. 2.5. For general subspaces $V_i \subset H^1_D(\Omega)$, the projections P_i may no longer involve the solution of partial differential equations on subdomains.

Subtracting the iterates in (2.19) from u and recursively applying the expression yields the following equation for the error $u - w^{(k+1)}$

$$\begin{cases} \left(u - w^{(k+1)}\right) = (I - P_p)\left(u - w^{(k+\frac{p-1}{p})}\right) \\ \qquad\qquad\quad\;\; = (I - P_p)(I - P_{p-1})\left(u - w^{(k+\frac{p-2}{p})}\right) \\ \qquad\qquad\quad\;\; \vdots \\ \qquad\qquad\quad\;\; = (I - P_p)\cdots(I - P_1)\left(u - w^{(k)}\right). \end{cases}$$

Define the *error amplification* map by $T = (I - P_p)\cdots(I - P_1)$. This map T will be a contraction (in an appropriate norm, see Chap. 2.5). Since $(I - T)$ involves only sums (or differences) of products of projections P_i, we may compute $w_* \equiv (I - T)u$ without explicit knowledge of u. For instance, when $p = 2$ we obtain that $(I - T) = P_1 + P_2 - P_2 P_1$ and $w_* = P_1 u + P_2 u - P_2 P_1 u$. Consequently, an equivalent problem for determining u is:

$$(I - T)u = w_*. \tag{2.20}$$

Equation (2.20) will be well posed since T is a contraction.

2.2.2 Additive Schwarz Method

The additive Schwarz method to solve (2.13) is a highly parallel algorithm in the Schwarz family [DR11]. It reformulates (2.13) using a *sum* of projections $P \equiv P_1 + \cdots + P_p$, where each P_i is the $\mathcal{A}(.,.)$-orthogonal projection onto V_i defined by (2.17). Formally, the solution u of (2.13) will also solve:

$$P\,u = w_*, \tag{2.21}$$

where $w_* \equiv P_1 u + \cdots + P_p u$ can be computed without explicit knowledge of u, since the terms $P_i u \in V_i$ can be computed by solving:

$$\mathcal{A}(P_i u, v) = \mathcal{A}(u, v) = F(v), \qquad \forall v \in V_i.$$

It is shown in Chap. 2.5 that the operator P is self adjoint and coercive in the Sobolev space $H^1_D(\Omega)$ equipped with the inner product $\mathcal{A}(.,.)$. Furthermore, upper and lower bounds can be calculated for the spectra of P, ensuring the well posedness of problem (2.21).

The additive Schwarz formulation of (2.13) is based on the solution of (2.21). In the discrete case, it is typically employed as a preconditioner, however, for illustrative purposes we indicate a Richardson iteration to solve (2.21). Given an iterate $w^{(k)}$, a new iterate $w^{(k+1)}$ is constructed as follows [TA5]. For $1 \le i \le p$ solve in *parallel:*

$$\begin{cases} -\nabla \cdot \left(a(x) \nabla v_i^{(k+1)} \right) + c(x)\, v_i^{(k+1)} = f(x), & \text{in } \Omega_i^* \\ \mathbf{n} \cdot \left(a \nabla v_i^{(k+1)} \right) + \gamma\, v_i^{(k+1)} = g_N, & \text{on } B_{[i]} \cap \mathcal{B}_N \\ v_i^{(k+1)} = w^{(k)}, & \text{on } B^{(i)} \\ v_i^{(k+1)} = 0, & \text{on } B_{[i]} \cap \mathcal{B}_D \end{cases}$$

and extend $v_i^{(k+1)} \equiv w^{(k)}$ on $\Omega \setminus \overline{\Omega_i^*}$. Then update:

$$w^{(k+1)} \equiv (1 - \tau\, p)\, w^{(k)} + \tau \left(v_1^{(k+1)} + \cdots + v_p^{(k)} \right),$$

where $0 < t_1 < \tau < t_2 < \frac{1}{p}$ is the step size parameter in Richardson's iteration. The resulting algorithm is summarized below in terms of projections.

Algorithm 2.2.4 *(Additive Schwarz-Richardson Iteration)*
Input: $w^{(0)}$ *(starting iterate) and* $0 < t_1 < \tau < t_2 < \frac{1}{p}$

```
1. For $k = 0, \cdots$ until convergence do:
2.     Compute in parallel:
           $w^{(k+1)} \equiv w^{(k)} + \tau\Big( P_1(u - w^{(k)}) + \cdots + P_p(u - w^{(k)}) \Big).$
3. Endfor
```

The additive Schwarz-Richardson iterates $w^{(k)}$ will converge geometrically to u for appropriately chosen τ. However, the multiplicative Schwarz iterates will generally converge more rapidly [XU3]. The matrix version of the additive Schwarz preconditioner is described in Chap.2.3. If a coarse space $V_0 \subset H_D^1(\Omega)$ is employed, then $P = (P_0 + \cdots + P_p)$ must be employed.

2.2.3 Hybrid Schwarz Method

The hybrid Schwarz method is a variant of the additive Schwarz method obtained by incorporating sequential steps from the multiplicative Schwarz method [MA15]. The resulting method yields improved convergence over the additive Schwarz method, but the algorithm is less parallelizable due to the extra sequential steps.

As in the additive Schwarz method, subspaces V_i are defined by (2.17), with associated $\mathcal{A}(.,.)$-orthogonal projections P_i for $1 \le i \le p$. Additionally, a coarse space $V_0 \subset H^1_D(\Omega)$ is employed with $\mathcal{A}(.,.)$-orthogonal projection P_0. The hybrid Schwarz formulation decomposes the solution to (2.13) as:

$$u = P_0 u + (I - P_0)u,$$

which is an $\mathcal{A}(.,.)$-orthogonal decomposition. The component $P_0 u \in V_0$ can be formally determined by solving the subproblem:

$$\mathcal{A}(P_0 u, v_0) = F(v_0), \qquad \forall v_0 \in V_0,$$

without explicit knowledge of u. The component $(I - P_0)u \in V_0^{\perp}$ can be sought, in principle, by applying an additive Schwarz method in $V_0^{\perp}$:

$$(I - P_0)\,(P_1 + \cdots + P_p)\,(I - P_0)u = g_*,$$

where $V_0^{\perp}$ denotes the orthogonal complement of V_0 in the inner product $\mathcal{A}(.,.)$. Here $g_* = (I - P_0)\,(P_1 + \cdots + P_p)\,(I - P_0)u$ can be computed without explicit knowledge of u. The preceding observations may be combined. Define:

$$\hat{P} \equiv P_0 + (I - P_0)\,(P_1 + \cdots + P_p)\,(I - P_0),$$

and formally construct the following problem equivalent to (2.13):

$$\hat{P}\,u = f_*, \tag{2.22}$$

where $f_* \equiv \hat{P}u$ can be computed explicitly. The operator $\hat{P}$ can be shown to be *self adjoint* and *coercive* in $\mathcal{A}(.,.)$ and will generally have improved spectral properties over the additive Schwarz operator $P = (P_1 + \cdots + P_p)$. Formally, the hybrid Schwarz method solves (2.22).

Remark 2.17. The forcing f_* in (2.22) can be computed explicitly as follows. Determine $u_0 \in V_0$ satisfying:

$$\mathcal{A}(u_0, v_0) = F(v_0), \qquad \forall v_0 \in V_0.$$

For $1 \le i \le p$ determine $w_i \in V_i$ satisfying:

$$\mathcal{A}(w_i, v_i) = F(v_i) - \mathcal{A}(u_0, v_i), \qquad \forall v_i \in V_i.$$

Define $w \equiv w_1 + \cdots + w_p$ and determine $\tilde{u}_0 \in V_0$ satisfying:

$$\mathcal{A}(\tilde{u}_0, v_0) = \mathcal{A}(w, v_0), \qquad \forall v_0 \in V_0.$$

Then $f_* = \hat{P}u = u_0 + (w - \tilde{u}_0)$.

In the following, we illustrate a Richardson iteration to solve (2.22).

Algorithm 2.2.5 *(Hybrid Schwarz-Richardson Iteration)*
Input: $w^{(0)}$ *starting iterate and* $0 < t_1 < \tau < t_2 < \frac{1}{p}$

```
1. For $k = 0, \cdots$ until convergence do:
2.     Compute in parallel:
           $w^{(k+1)} \equiv w^{(k)} + \tau \tilde{P}(u - w^{(k)}).$
3. Endfor
```

Remark 2.18. The balancing domain decomposition preconditioner for Schur complement matrices (in Chap. 3) is based on this principle [MA14, MA17]. In it, the exact projections P_i are replaced by approximations which require the solution of Neumann boundary value problems on non-overlapping subdomains Ω_i. For each subdomain Neumann problem to be solvable, certain compatibility conditions must be satisfied locally. In such applications, the coarse space V_0 may be constructed so that all the subdomain compatability conditions are simultaneously enforced in the orthogonal complement of V_0.

2.2.4 Restricted Schwarz Algorithm

The restricted Schwarz method is a variant of the additive Schwarz method employing a *partition of unity*, see [CA19, KU6, CA17]. Formally, it can also be motivated by a multisubdomain hybrid formulation of (2.12) based on a partition of unity $\chi_1(x), \cdots, \chi_p(x)$ subordinate to $\Omega_1^*, \cdots, \Omega_p^*$. In practice, the algorithm can be applied either as an unaccelerated iteration or as a preconditioner. In the latter case, it yields a non-symmetric preconditioner even for self adjoint problems. Given the partition of unity $\{\chi_i(.)\}_{i=1}^p$, we note that $\sum_{j \neq i} \chi_j(x) = 1$ on $B^{(i)}$ for $1 \leq i \leq p$, since $\chi_i(x) = 0$ on $B^{(i)}$. Using this, we obtain the hybrid formulation.

Theorem 2.19. *Suppose the following assumptions hold.*

1. *Let* $c(x) \geq c_0 > 0$ *in (2.12).*
2. *Let* $u(x)$ *denote a solution to (2.12).*
3. *Let* $w_1(.), \cdots, w_p(.)$ *solve the hybrid formulation for* $1 \leq i \leq p$*:*

$$\begin{cases} -\nabla \cdot (a(x)\nabla w_i) + c(x)\, w_i = f(x), & \text{in } \Omega_i^* \\ \mathbf{n} \cdot (a\nabla w_i) + \gamma\, w_i = g_N, & \text{on } B_{[i]} \cap \mathcal{B}_N \\ w_i = \sum_{j \neq i} \chi_j\, w_j, & \text{on } B^{(i)} \\ w_i = 0, & \text{on } B_{[i]} \cap \mathcal{B}_D. \end{cases} \tag{2.23}$$

Then, the following result will hold:

$$u(x) = w_i(x) \quad \text{on } \overline{\Omega}_i^*, \quad \text{for} \quad 1 \leq i \leq p.$$

Proof. See Chap. 15 for the case $\mathcal{B}_N = \emptyset$. □

The hybrid formulation (2.23) corresponds to a *fixed point* equation for the following linear mapping $\mathcal{T}$ defined by:

$$\mathcal{T}\,(v_1, \cdots, v_p) = (w_1, \cdots, w_p)$$

where for v_i satisfying $v_i = 0$ on $B_{[i]} \cap \mathcal{B}_D$ and $\mathbf{n} \cdot (a\nabla v_i) + \gamma\, v_i = g_N$ on $B_{[i]} \cap \mathcal{B}_N$ for $1 \le i \le p$, the outputs w_i satisfy:

$$\begin{cases} -\nabla \cdot (a(x)\nabla w_i) + c(x)\, w_i = f(x), & \text{on } \Omega_i^* \\ \mathbf{n} \cdot (a\nabla w_i) + \gamma\, w_i = g_N, & \text{on } B_{[i]} \cap \mathcal{B}_N \\ w_i = \sum_{j \ne i} \chi_j\, v_j, & \text{on } B^{(i)} \\ w_i = 0, & \text{on } B_{[i]} \cap \mathcal{B}_D. \end{cases} \tag{2.24}$$

Under the assumption $c(x) \ge c_0 > 0$ and $\mathcal{B}_N = \emptyset$, the mapping $\mathcal{T}$ will be a *contraction* and the Picard iterates of $\mathcal{T}$ will converge to its fixed point $(u_1, \cdots, u_p)$ where $u_i \equiv u$ on each subdomain Ω_i^*. Given local approximations $(v_1, \cdots, v_p)$ define a global approximation $v \equiv \sum_{j=1}^p \chi_j\, v_j$. Since $\chi_i(x) = 0$ for $x \in B^{(i)}$. the global approximation $v(x)$ will satisfy:

$$v(x) = \sum_{j \ne i} \chi_j(x) v_j(x), \qquad \text{on each } \; B^{(i)}.$$

Substitute this into (2.24) and apply Lemma 2.15 to w_i with $w \equiv v$ to obtain:

$$w_i = v + P_i(u - v), \qquad \text{on } \Omega_i^*, \quad \text{for } 1 \le i \le p$$

where u solves (2.13). At the fixed point of $\mathcal{T}$ where $v = w$, this yields:

$$w = \sum_i \chi_i\, w_i = \sum_i \chi_i\, (w + P_i(u - w)) = w + \sum_i \chi_i P_i(u - w). \tag{2.25}$$

The following algorithm corresponds to a Picard iteration of the map $\mathcal{T}$.

Algorithm 2.2.6 *(Restricted Schwarz Method in Projection Form)*
Input: $(w_1^{(0)}, \cdots, w_p^{(0)})$ *and* $w^{(0)}(x) \equiv \sum_{j=1}^p \chi_j(x) w_j^{(0)}(x)$

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     For $i = 1, \cdots, p$ in parallel compute:
           $w_i^{(k+1)} \equiv P_i\Big(u - w^{(k)}\Big).$
3.     Endfor
4.     Define: $w^{(k+1)}(x) \equiv w^{(k)}(x) + \sum_{i=1}^p \chi_i(x) w_i^{(k+1)}(x).$
5. Endfor
```

Under appropriate assumptions, $\mathcal{T}$ will be a contraction and the iterates $w^{(k)}$ will converge geometrically to the solution u of (2.12):

$$\|w^{(k)} - u\|_{\infty,\Omega} \leq \delta^k \, \|w^{(0)} - u\|_{\infty,\Omega},$$

where $\|\cdot\|_{\infty,\Omega}$ denotes the maximum norm, see Chap. 15 when $\mathcal{B}_N = \emptyset$.

Remark 2.20. The matrix form of preconditioner associated with the restricted Schwarz method is described in Chap. 2.3. The preceding restricted Schwarz algorithm did not employ coarse space residual correction. Consequently, as the number of subdomains is increased and their diameters decrease in size, the rate of convergence of the algorithm can deteriorate. The convergence of the preconditioner associated with the preceding algorithm can also be improved significantly if a coarse space projection term is employed additively.

2.3 Matrix Form of Schwarz Subspace Algorithms

In this section, we shall describe the matrix version of Schwarz algorithms. Our formulation will employ the finite dimensional linear space $V = \mathbb{R}^n$, endowed with a self adjoint and coercive bilinear form $\mathcal{A}(.,.)$. We shall further assume that we are given subspaces $V_i \subset V$ for $0 \leq i \leq p$ satisfying:

$$V = V_0 + V_1 + \cdots + V_p. \tag{2.26}$$

In this case, matrix expressions can be derived for the projection version of the Schwarz algorithms described in the preceding section, for problems of the form (2.13), see [MI, MA37, DR11, BR18, TA8, XU3, GR4].

Consider the finite dimensional space $V \equiv \mathbb{R}^n$ endowed with a self adjoint and coercive bilinear form $\mathcal{A}(.,.)$, which also defines an inner product on V. Given a linear functional $F(\cdot)$, we shall seek $\mathbf{u} \in V$ such that:

$$\begin{cases} \mathcal{A}(\mathbf{u},\mathbf{v}) = F(\mathbf{v}), & \text{for } \mathbf{v} \in V, \text{ where} \\ \mathcal{A}(\mathbf{v},\mathbf{w}) \equiv \mathbf{v}^T A \mathbf{w}, & \text{for } \mathbf{v},\mathbf{w} \in V \\ F(\mathbf{v}) \equiv \mathbf{v}^T \mathbf{f}, & \text{for } \mathbf{v} \in V, \end{cases} \tag{2.27}$$

where A is an $n \times n$ symmetric and positive definite matrix and $\mathbf{f} \in \mathbb{R}^n$. In matrix terms, problem (2.27) will correspond to the linear system:

$$A\mathbf{u} = \mathbf{f}. \tag{2.28}$$

We shall formulate matrix Schwarz algorithms to solve this system by analogy with the projection algorithms described in Chap. 2.2.

We shall assume that each $V_i \subset \mathbb{R}^n$ is of dimension n_i, and that it is the *column space* (Range) of an $n \times n_i$ matrix R_i^T of full rank:

$$V_i \equiv \text{Range}\left(R_i^T\right), \qquad \text{for } 0 \leq i \leq p.$$

Thus, the columns of R_i^T must form a basis for V_i. We assume that V_i satisfies (2.26). This requires that given $\mathbf{v} \in V$, there must exist $\mathbf{v}_i \in V_i$ satisfying:

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_1 + \cdots + \mathbf{v}_p.$$

An elementary rank argument will show that $(n_0 + n_1 + \cdots + n_p) \geq n$.

Remark 2.21. The matrices R_i will be referred to as *restriction* maps while their transposes R_i^T will be referred to as *extension* maps. Matrix versions of Schwarz algorithms to solve (2.28) based on the subspaces V_i can be obtained by transcribing the projection algorithms in terms of matrices. This will require a matrix representation of the projections P_i.

Definition 2.22. *Given $\mathbf{v} \in V$, we define $P_i \mathbf{v} \in V_i$:*

$$\mathcal{A}(P_i\mathbf{v}, \mathbf{w}_i) = \mathcal{A}(\mathbf{v}, \mathbf{w}_i) \qquad \forall \mathbf{w}_i \in V_i \tag{2.29}$$

as the $\mathcal{A}(.,.)$-orthogonal of $\mathbf{v} \in V$ onto V_i.

Remark 2.23. A matrix representation of P_i can be derived as follows. Since V_i is the column space of R_i^T, represent $P_i\mathbf{v} = R_i^T\mathbf{x}_i$ and $\mathbf{w}_i = R_i^T\mathbf{y}_i$ for $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{n_i}$. Substitute these representations into (2.29) to obtain:

$$\mathbf{y}_i^T \, (R_i A R_i^T) \, \mathbf{x}_i = \mathbf{y}_i^T \, R_i A \, \mathbf{v}, \qquad \forall \mathbf{y}_i \in \mathbb{R}^{n_i}.$$

Since this must hold for all $\mathbf{y}_i \in \mathbb{R}^{n_i}$, we obtain that $A_i \, \mathbf{x}_i = R_i A \mathbf{v}$, where $A_i \equiv (R_i A R_i^T)$. Solving this linear system yields $\mathbf{x}_i = A_i^{-1} R_i A \mathbf{v}$, and substituting $P_i \, \mathbf{v} = R_i^T \mathbf{x}_i$ results in the expression:

$$P_i = R_i^T A_i^{-1} R_i A \tag{2.30}$$

for the *matrix representation* of P_i. Matrix A_i^{-1} should not be assembled. Instead, an expression $\mathbf{w}_i = A_i^{-1}\mathbf{r}_i$ can be computed by solving $A_i\mathbf{w}_i = \mathbf{r}_i$.

Remark 2.24. If the rows and columns of matrix R_i are *elementary vectors*, corresponding to selected columns or rows or some identity matrix of appropriate size, then matrix $A_i = R_i A R_i^T$ will correspond to principal submatrices of A. In particular, if $(n_0 + \cdots + n_p) = n$ and R_l corresponds to the rows of an identity matrix of size n with indices in $\mathcal{I}_l$:

$$\mathcal{I}_l = \{(n_0 + \cdots + n_{l-1}) + 1, \, \ldots \, , (n_0 + \cdots + n_l)\} \, ,$$

then A_l will correspond to the diagonal block of A with indices in $\mathcal{I}_l$.

Multiplicative Schwarz Algorithm. The matrix version of Alg. 2.2.3 to solve system (2.28) instead of problem (2.2), can be obtained by replacing each update $P_i(u - w^{(k+\frac{i-1}{p})})$ by its discrete counterpart $P_i\left(\mathbf{u} - \mathbf{w}^{(k+\frac{i-1}{p})}\right)$, where $\mathbf{u}$ is the solution to (2.28). Substituting the matrix form of projection P_i

and using that $A\mathbf{u} = \mathbf{f}$ yields:

$$\begin{cases} P_i(\mathbf{u} - \mathbf{w}) = R_i^T A_i^{-1} R_i A \left(\mathbf{u} - \mathbf{w}^{(k+\frac{i-1}{p})}\right) \\ \qquad\qquad\quad = R_i^T A_i^{-1} R_i \left(\mathbf{f} - A\mathbf{w}^{(k+\frac{i-1}{p})}\right). \end{cases}$$

Thus, the matrix form of:

$$w^{(k+\frac{i}{p})} = w^{(k+\frac{i-1}{p})} + P_i \left(u - w^{(k+\frac{i-1}{p})}\right),$$

becomes:

$$\mathbf{w}^{(k+\frac{i}{p})} = \mathbf{w}^{(k+\frac{i-1}{p})} + R_i^T A_i^{-1} R_i \left(\mathbf{f} - A\mathbf{w}^{(k+\frac{i-1}{p})}\right).$$

The resulting multiplicative or sequential Schwarz algorithm is listed next.

Algorithm 2.3.1 *(Multiplicative Schwarz Method to Solve (2.28))*
Input: $\mathbf{w}^{(0)} = 0$ *(starting guess),* $\mathbf{f}$

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     For $i = 0, \cdots, p$ do:
           $\mathbf{w}^{(k+\frac{i+1}{p+1})} = \mathbf{w}^{(k+\frac{i}{p+1})} + R_i^T A_i^{-1} R_i \left(\mathbf{f} - A\mathbf{w}^{(k+\frac{i}{p+1})}\right).$
3.     Endfor
4. Endfor
```

Output: $\mathbf{w}^{(k)}$

The iterates $\mathbf{w}^{(k)}$ in this algorithm will converge to the solution of (2.28) without acceleration. If CG acceleration is employed to solve $A\mathbf{u} = \mathbf{f}$, then a symmetric positive definite preconditioner would be necessary [GO4]. The inverse of the symmetrized Schwarz preconditioner M is described below.

Algorithm 2.3.2 *(Symmetrized Schwarz Preconditioner for (2.28))*
Input: $\mathbf{w} \equiv \mathbf{0}$ *and* $\mathbf{r}$

```
1. For $i = p, \cdots, 1, 0, 1, \cdots, p$ do:
       $\mathbf{w} \leftarrow \mathbf{w} + R_i^T A_i^{-1} R_i (\mathbf{r} - A\mathbf{w}).$
2. Endfor
```

Output: $M^{-1}\mathbf{r} \equiv \mathbf{w}$

Remark 2.25. The notation A_i^{-1} was only employed for convenience in the preceding algorithms. In practice, A_i^{-1} should not be assembled. Instead, its action on a vector should be computed by solution of the associated linear system. For instance, the computation of $R_i^T A_i^{-1} R_i \mathbf{f}$ should first involve the computation of $R_i \mathbf{f}$, followed by the solution of the linear system $A_i \mathbf{v}_i = R_i \mathbf{f}$, followed by the computation $R_i^T \mathbf{v}_i$. Scatter-gather operations can be used to implement R_i^T and R_i.

Remark 2.26. In both of the preceding algorithms, the matrices $A_i = R_i A R_i^T$ can be replaced by appropriately chosen preconditioners $\tilde{A}_i = \tilde{A}_i^T > 0$. As an example, a sparse preconditioner $\tilde{A}_i$ for A_i can be obtained by ILU factorization of A_i, see [BE, AX, SA2]. If approximations are employed in the multiplicative Schwarz method, to ensure convergence without acceleration, the condition $\lambda_{\max}\left(\tilde{A}_i^{-1} A_i\right) < 2$ must be satisfied, see [XU3].

Remark 2.27. If a preconditioner is employed for A_0, an alternative symmetrization involving one additional fractional step can be used in the symmetrized Schwarz preconditioner. In step 1 of the preceding algorithm, residual corrections can be implemented for $i = p, p-1, \cdots, 1, 0, 0, 1, \cdots, p-1, p$. Both versions will be equivalent if an exact solver is employed for A_0.

Additive Schwarz Algorithm. The matrix version of the *additive* Schwarz equation $P\,u = f_*$ for solution of (2.28) has the form:

$$\left(\sum_{i=0}^{p} R_i^T A_i^{-1} R_i A\right) \mathbf{u} = \mathbf{w}_*, \tag{2.31}$$

where

$$\mathbf{w}_* \equiv \sum_{i=0}^{p} R_i^T A_i^{-1} R_i \mathbf{f}.$$

The system (2.31) for $\mathbf{u}$ corresponds to a preconditioned system of the form $M^{-1} A \mathbf{u} = M^{-1} \mathbf{f}$. This yields the additive Schwarz preconditioner as:

$$M^{-1} = \sum_{i=0}^{p} R_i^T A_i^{-1} R_i.$$

This is summarized below.

Algorithm 2.3.3 *(Additive Schwarz Preconditioner for (2.28))*
Input: $\mathbf{r}$

```
1. For $i = 0, \cdots, p$ in parallel do:
       $\mathbf{w}_i = R_i^T A_i^{-1} R_i \mathbf{r}$
2. Endfor
3. Sum:
       $\mathbf{w} \equiv \mathbf{w}_0 + \cdots + \mathbf{w}_p.$
```

Output: $M^{-1}\mathbf{r} \equiv \mathbf{w}$

Remark 2.28. When $(n_0 + n_1 + \cdots + n_p) = n$ and the columns of R_l correspond to selected columns of an identity matrix, then it is easily seen that the matrix version of the additive Schwarz preconditioner corresponds to a *block Jacobi* preconditioner, and the matrix version of the multiplicative Schwarz method corresponds to the *block Gauss-Seidel* method. When $(n_0 + \cdots + n_p) > n$ or when the columns of R_l are not columns of an identity matrix, then the multiplicative and additive Schwarz algorithms generalize the block Jacobi and block Gauss-Seidel algorithms.

Hybrid Schwarz Method. The matrix version of the hybrid Schwarz preconditioner can be derived from the hybrid Schwarz problem $\tilde{P}u = f_*$ where $\hat{P} = P_0 + (I - P_0)\,(P_1 + \cdots + P_p)\,(I - P_0)$. As this problem represents the preconditioned system $M^{-1}A\mathbf{u} = M^{-1}\mathbf{f}$, the action M^{-1} of the inverse of preconditioner M can easily be deduced to be the following.

Algorithm 2.3.4 *(Hybrid Schwarz Preconditioner for (2.28))*

```
Input: $\mathbf{r}$
1. Compute:
       $\mathbf{w}_0 = R_0^T A_0^{-1} R_0 \mathbf{r}.$
2. For $i = 1, \cdots, p$ in parallel do:
       $\mathbf{v}_i = R_i^T A_i^{-1} R_i (\mathbf{r} - A\mathbf{w}_0).$
3. Endfor
4. Sum: $\mathbf{v} = \mathbf{v}_1 + \cdots + \mathbf{v}_p.$
5. Compute:
       $\mathbf{v}_0 = R_0^T A_0^{-1} R_0 A \mathbf{v}.$
6. Compute: $\mathbf{w} = \mathbf{w}_0 + \mathbf{v} - \mathbf{v}_0.$
Output: $M^{-1}\mathbf{r} \equiv \mathbf{w}$
```

Remark 2.29. If the input residual $\mathbf{r}$ satisfies $R_0\,\mathbf{r} = \mathbf{0}$, then step 1 in the hybrid Schwarz preconditioner can be skipped, yielding $\mathbf{w}_0 = \mathbf{0}$. This suggests choosing a starting iterate $\mathbf{u}_0 \in \mathbb{R}^n$ in the conjugate gradient method so that the initial residual $\mathbf{r} = \mathbf{f} - A\,\mathbf{u}_0$ satisfies $R_0\,(\mathbf{f} - A\,\mathbf{u}_0) = \mathbf{0}$. Then, as will be shown below, all subsequent residuals in the conjugate gradient method with hybrid Schwarz preconditioner will satisfy this constraint. Note that to construct a starting iterate $\mathbf{u}_0 \in \mathbb{R}^n$, so that $R_0\,(\mathbf{f} - A\mathbf{u}_0) = \mathbf{0}$, seek it in the form $\mathbf{u}_0 = R_0^T \boldsymbol{\alpha}_0$ for some unknown coefficient vector $\boldsymbol{\alpha}_0 \in \mathbb{R}^{n_0}$. Imposing the preceding constraint will yield:

$$R_0\,(\mathbf{f} - A\,\mathbf{u}_0) = \mathbf{0} \Leftrightarrow R_0\left(\mathbf{f} - A\,R_0^T\,\boldsymbol{\alpha}_0\right) = \mathbf{0} \Leftrightarrow \boldsymbol{\alpha}_0 = A_0^{-1}\,R_0\,\mathbf{f},$$

where $A_0 = R_0 A R_0^T$. Thus, $\mathbf{u}_0 = R_0^T A_0^{-1} R_0 \mathbf{f}$. Next, to verify that $M^{-1}\mathbf{r}$ will satisfy $R_0 A M^{-1}\mathbf{r} = \mathbf{0}$ whenever $\mathbf{r} \in \mathbb{R}^n$ satisfies $R_0\mathbf{r} = \mathbf{0}$, apply $R_0 A$ to step 6 in the hybrid Schwarz preconditioner with $\mathbf{w}_0 = \mathbf{0}$ to obtain:

$$R_0 A M^{-1}\mathbf{r} = R_0 A \mathbf{v} - R_0 A R_0^T A_0^{-1} R_0 A \mathbf{v} = \mathbf{0}.$$

Thus, the computational costs in a conjugate gradient method to solve $A\mathbf{u} = \mathbf{f}$ can be reduced by splitting the solution as $\mathbf{u} = \mathbf{u}_0 + \mathbf{v}$ with $\mathbf{u}_0 = R_0^T A_0^{-1} R_0 \mathbf{f}$. To determine $\mathbf{v}$, solve the linear system $A\mathbf{v} = \mathbf{f} - A\mathbf{u}_0$ by a conjugate gradient method with a hybrid Schwarz preconditioner in which step 1 is skipped.

Remark 2.30. In Chap. 2.5, it is shown that the hybrid Schwarz preconditioned matrix $\tilde{P}$ is better conditioned than its associated additive Schwarz preconditioned matrix P.

Remark 2.31. The submatrices $A_i = R_i A R_i^T$ in the hybrid Schwarz preconditioner may be replaced by *approximations* $\tilde{A}_i$ for $1 \leq i \leq p$. In certain applications, it may even be advantageous to employ *singular* matrices $\tilde{A}_i$ whose null spaces are known. In this case, linear systems of the form $\tilde{A}_i \mathbf{v}_i = \mathbf{r}_i$ will be solvable only if a *compatibility* condition is satisfied. Indeed, if $\boldsymbol{\alpha}_i$ is an $n_i \times d_i$ matrix whose columns form a basis for the null space of $\tilde{A}_i$, then $\boldsymbol{\alpha}_i^T \mathbf{r}_i = \mathbf{0}$ must hold for solvability. Then, the solution $\mathbf{v}_i$ will not be unique, and will involve an arbitrary additive term from the null space. In such applications, a careful choice of coarse space V_0 in the hybrid Schwarz method can ensure solvability of all such local problems, and also effectively handle the arbitrariness of the local solutions. Define a coarse space $V_0 \subset \mathbb{R}^n$ as:

$$V_0 \equiv \text{Range}\left(R_0^T\right), \quad \text{where} \quad R_0^T \equiv \left[R_1^T \boldsymbol{\alpha}_1, \ldots, R_p^T \boldsymbol{\alpha}_p\right].$$

By construction of the term $\mathbf{w}_0$ in step 1 of the hybrid Schwarz preconditioner, it will hold that $R_0 \left(\mathbf{r} - A\mathbf{w}_0\right) = \mathbf{0}$. Substituting the definition of R_0 yields that $\boldsymbol{\alpha}_i^T R_i \left(\mathbf{r} - A\mathbf{w}_0\right) = \mathbf{0}$ for $1 \leq i \leq p$, so that the subproblems in step 2 of the hybrid Schwarz preconditioner are well defined when A_i is replaced by $\tilde{A}_i$. Each $\mathbf{v}_i$ in step 2 of the hybrid Schwarz preconditioner can have an arbitrary additive term of the form $R_i^T \boldsymbol{\alpha}_i \boldsymbol{\beta}_i$ with $\boldsymbol{\beta}_i \in \mathbb{R}^{d_i}$. However, the projection term $\mathbf{v} - R_0^T A_0^{-1} R_0 A\mathbf{v}$ in step 6 modifies these arbitrary terms so that $R_0 A M^{-1} \mathbf{r} = \mathbf{0}$ holds. This is the principle underlying the *balancing* domain decomposition preconditioner [MA14].

Restricted Schwarz Algorithm. Since the restricted Schwarz algorithm in Chap. 2.2 is based on a partition of unity, its general matrix version will require an algebraic partition of unity, if such can be found.

Definition 2.32. *Let* $V_i = \text{Range}(R_i^T)$ *be subspaces of* $V = \mathbb{R}^n$ *for* $1 \leq i \leq p$. *We say that matrices* $E_1, \cdots, E_p$ *form a discrete partition of unity relative to* $R_1, \cdots, R_p$ *if:*

$$E_1 R_1 + \cdots + E_p R_p = I,$$

where each E_i *is an* $n \times n_i$ *matrix for* $1 \leq i \leq p$.

The action M^{-1} of the inverse of the restricted Schwarz preconditioner to solve (2.28) is motivated by (2.25) when iterate $w = 0$. In the version given below, a coarse space correction term is included, with $E_0 \equiv R_0^T$.

Algorithm 2.3.5 *(Restricted Schwarz Preconditioner for (2.28))*
Input: $\mathbf{r}$, $0 < \alpha < 1$.

```
1. For $i = 0, 1, \cdots, p$ in parallel compute:
       $\mathbf{w}_i = E_i A_i^{-1} R_i \mathbf{r}.$
2. Endfor
```

Output: $M^{-1}\mathbf{r} \equiv \alpha\, \mathbf{w}_0 + (1-\alpha)\,(\mathbf{w}_1 + \cdots + \mathbf{w}_p)$.

Remark 2.33. Since the above preconditioner is not symmetric, it cannot be employed in a conjugate gradient method [CA19].

2.4 Implementational Issues

In this section, we remark on applying the matrix Schwarz algorithms from Chap. 2.3 to solve a discretization of (2.1). For simplicity, we only consider a finite element discretization, though the methodology (with the exception of a coarse space V_0) will typically carry over for a finite difference discretization. We shall also remark on local solvers and parallel software libraries.

2.4.1 Choice of Subdomains and Subdomain Spaces

Various factors may influence the choice of an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ of Ω. These include the geometry of the domain, regularity of the solution, availability of fast solvers for subdomain problems and heterogeneity in the coefficients. When a natural decomposition is not obvious, an automated strategy may be employed, using the *graph partitioning* algorithms discussed in Chap. 5, so that the decomposition yields approximately balanced loads, see [BE14, FO2, SI2, FA9, BA20, PO3, PO2]. Ideally, the number of subdomains p also depends on the number of processors.

Once a an overlapping decomposition $\{\Omega_l^*\}_{l=1}^p$ has been chosen, and given the finite element space $V_h \subset H_D^1(\Omega)$, we define the local spaces as:

$$V_i \equiv V_h \cap \left\{ v \in H^1(\Omega) \,:\, v = 0 \;\text{ on }\; \overline{\Omega} \backslash \overline{\Omega}_i^* \right\} \quad \text{for} \quad 1 \le i \le p.$$

Let $n_i = \dim(V_i)$ and let $index(\Omega_i^*, j)$ denote the global index of the j'th local node in $\Omega_i^* \cup \left(\mathcal{B}_N \cap B_{[i]}\right)$. Then, define R_i as an $n_i \times n$ restriction matrix:

$$(R_i)_{kj} = \begin{cases} 1, & \text{if } index(\Omega_i^*, k) = j \\ 0, & \text{if } index(\Omega_i^*, k) \ne j, \end{cases} \quad \text{for} \quad 1 \le i \le p.$$

For $1 \le i \le p$ these matrices will have zero or one entries, and at most one nonzero entry per row or column. The action of R_i and R_i^T for $1 \le i \le p$

may be implemented using scatter-gather operations and the data structure of *index*$(\Omega_i^*, \cdot)$. The subdomain submatrices A_i of size $n_i \times n_i$ defined by:

$$A_i = R_i A_h R_i^T, \quad \text{for} \quad 1 \le i \le p,$$

will be principal submatrices of A corresponding to the subdomain indices.

2.4.2 Choice of Coarse Spaces

A coarse space $V_0 \subset (V_h \cap H_D^1(\Omega))$ may be employed as described in Chap. 2.1. If $\psi_1^{(0)}(\cdot), \cdots, \psi_{n_0}^{(0)}(\cdot)$ forms a finite element basis for V_0, then an extension matrix R_0^T of size $n \times n_0$ will have the following entries:

$$\left(R_0^T\right)_{ij} = \psi_j^{(0)}(x_i), \qquad \text{for} \quad 1 \le i \le n, \quad 1 \le j \le n_0.$$

Matrix R_0 will not be a zero-one matrix, unlike R_i for $1 \le i \le p$. Furthermore, $A_0 = R_0 A_h R_0^T$ will not be a submatrix of A. In some applications, the coarse space may be omitted, without adversely affecting the rate of convergence of Schwarz algorithms. For instance, if $c(x) \ge c_0 > 0$ and coefficient $a(x)$ is *anisotropic* with a sufficiently small parameter and aligned subdomains, or for a *time stepped* problem, with sufficiently small time step and large overlap.

Remark 2.34. When the boundary segment $\mathcal{B}_D \neq \emptyset$, equation (2.12) will have a unique solution, and matrix A will be symmetric positive definite. However, when $\mathcal{B}_D = \emptyset$ and $c(x) = 0$ and $\gamma(x) = 0$ then (2.12) will be a Neumann problem. In this case, a compatability condition must be imposed for the solvability of (2.1), and its solution will be unique only up to a constant. By construction, all the subdomain matrices A_i will be nonsingular for $1 \le i \le p$ since Dirichlet boundary conditions will be imposed on $B^{(i)} \neq \emptyset$. However, matrix A_0 will be singular with $\mathbf{1}$ spanning its null space. To ensure that each coarse problem of the form $A_0 \mathbf{v}_0 = R_0 \mathbf{r}$ is solvable, it must hold that $\mathbf{1}^T R_0 \mathbf{r} = 0$. Then, the coarse solution will be nonunique, but a specific solution may be selected so that either $\mathbf{1}^T \mathbf{v}_0 = 0$, or $\mathbf{1}^T \mathbf{v} = 0$ for the global solution.

2.4.3 Discrete Partition of Unity

For the restricted Schwarz algorithm, an algebraic partition of unity consisting of matrices E_i can be constructed as follows. Let $\chi_1(\cdot), \cdots, \chi_p(\cdot)$ denote a *continuous* partition of unity subordinate to $\Omega_1^*, \cdots, \Omega_p^*$. If $x_1, \cdots, x_n$ denote the nodes of $\mathcal{T}_h(\Omega)$ in $\Omega \cup \mathcal{B}_N$, define:

$$\left(E_i\right)_{lj} = \begin{cases} \chi_i(x_l) & \text{if } \mathit{index}(\Omega_i^*, j) = l \\ 0 & \text{if } \mathit{index}(\Omega_i^*, j) \neq l \end{cases}$$

Here $1 \leq i \leq p$, $1 \leq l \leq n$ and $1 \leq j \leq n_i$. Then, by construction:

$$\sum_{i=1}^{p} E_i R_i = I.$$

Similar discrete partitions of unity are employed in [MA17]. For the coarse space, we formally define $E_0 \equiv R_0^T$.

2.4.4 Convergence Rates

For discretizations of self adjoint and coercive elliptic equations, Schwarz algorithms typically converge at a rate independent of (or mildly dependent on) the mesh size h and the subdomain size h_0, *provided* the overlap between subdomains is sufficiently large, and a coarse space V_0 is employed with an $O(h_0)$ approximation property. This is verified by both computational tests and theoretical analysis. The latter typically assumes that the overlap between subdomains is $\beta\, h_0 > 0$ and shows that the rate of convergence can depend on the coefficient $a(.)$, and mildly on the parameter β, see Chap. 2.5.

2.4.5 Local Solvers

The implementation of Schwarz algorithms requires computing terms of the form $\mathbf{w}_i = A_i^{-1} R_i \mathbf{r}$ for multiple choices of $R_i \mathbf{r}$. In practice, $\mathbf{w}_i$ is obtained by solving the associated system $A_i \mathbf{w}_i = R_i \mathbf{r}$, using a direct or iterative solver. Direct solvers are commonly employed, since they are robust and do not involve double iteration. Furthermore, efficient sparse direct solvers are available in software packages. In the following, we list several solvers.

Direct Solvers. Since $A_i = A_i^T > 0$ is sparse, a direct solver based on Cholesky factorization can be employed [GO4, GE5, DU]. Matrix A_i its Cholesky factorization $A_i = L_i L_i^T$ should be stored using a sparse format. Systems of the form $A_i \mathbf{w}_i = R_i \mathbf{r}$ can then be solved using back substitution, solving $L_i \mathbf{z}_i = R_i \mathbf{r}$ and $L_i^T \mathbf{w}_i = \mathbf{z}_i$, see [GO4]. Such algorithms are available in *LAPACK*, *SPARSPAK* and *SPARSKIT*, see [GE5, DU, GO4, SA2, AN].

Remark 2.35. The cost of employing a direct solver to solve $A_i \mathbf{w}_i = R_i \mathbf{r}$ depends on the cost of computing its Cholesky factors L_i and L_i^T, and the cost for solving $L_i \mathbf{z}_i = R_i \mathbf{r}$ and $L_i^T \mathbf{w}_i = \mathbf{z}_i$. When multiple systems of the form $A_i \mathbf{w}_i = R_i \mathbf{r}$ need to be solved, the Cholesky factors of A_i need to be determined only once and stored. The cost of computing the Cholesky factorization of A_i will depend on the sparsity of A_i, while the cost of solving $L_i \mathbf{z}_i = R_i \mathbf{r}$ and $L_i^T \mathbf{w}_i = \mathbf{z}_i$ will depend on the sparsity of L_i. These costs can be significantly reduced by *reordering* (permuting) the unknowns. For instance, if subdomain Ω_i^* is a thin strip, then a *band* solver can be efficient, provided the unknowns are reordered within the strip so that the band size is minimized. Other common orderings include the nested dissection ordering, and the Cuthill-McKee and reverse Cuthill-McKee orderings, see [GE5, DU, SA2]. Sparse software packages such as *SPARSPAK* and *SPARSKIT*, typically employ graph theoretic methods to automate the choice of a reordering so that the amount of *fill in* is approximately minimized, to reduce the cost of employing a direct solver [GE5, DU]. Such solvers typically have a complexity of $O(n_i^\alpha)$ for $1 < \alpha < 3$.

FFT Based Solvers. Fast direct solvers based on Fast Fourier Transforms (FFT's) may be available for special geometries, coefficients, triangulations and boundary conditions, see [VA4]. Such solvers will apply when the eigenvalue decomposition $A_i = F_i \Lambda_i F_i^T$ of A_i is known, where Λ_i is a diagonal matrix of eigenvalues of A_i, and F_i is a discrete Fourier (or sine or cosine) transform. Such solvers will typically have a complexity of $O(n_i \log(n_i))$.

Iterative Solvers. Each subdomain problem $A_i \mathbf{w}_i = \mathbf{r}_i$ may also be solved iteratively using a CG algorithm with a preconditioner M_i (such as ILU, Gauss-Seidel, Jacobi) in an inner loop. This will introduce *double iteration.* To ensure convergence, the fixed number of local iterations must be accurate to within the discretization error. If the number of iterations vary with each application of the local solver, then the Schwarz preconditioner may vary with each iteration, see [GO4, SA2, AX, SI3].

Remark 2.36. If an iterative local solver is employed, with fixed number of iterations and zero starting guess, this will yield a preconditioner $\tilde{A}_i$ for A_i, see [GO4, BE2, NO2, AX, MA8]. To ensure the convergence of Schwarz algorithms when approximate solvers are employed, matrices $\tilde{A}_i$ must satisfy certain assumptions. For instance, the condition number of the additive Schwarz preconditioner with inexact solver will increase at most by the factor γ:

$$\gamma \equiv \frac{\max_i \lambda_{\max}\left(\tilde{A}_i^{-1} A_i\right)}{\min_i \lambda_{\min}\left(\tilde{A}_i^{-1} A_i\right)}.$$

If inexact solvers $\tilde{A}_i$ are employed in the multiplicative Schwarz algorithm, then the spectral radius must satisfy $\rho\left(\tilde{A}_i^{-1} A_i\right) < 2$ to ensure convergence. In the hybrid Schwarz algorithm (in balancing domain decomposition [MA15]) the coarse problem must be solved *exactly.*

2.4.6 Parallelization and Software Libraries

With the exception of the sequential Schwarz algorithm without coloring, the computations on different subdomains in a Schwarz algorithm can typically be implemented concurrently. From the viewpoint of parallelization, Schwarz algorithms thus have "coarse granularity", i.e., a significant portion of the computations can be performed in parallel, with the remaining portion requiring more intensive communication between processors. As an example, consider the additive Schwarz preconditioner:

$$M^{-1}\mathbf{r} = \sum_{l=0}^{p} R_l^T A_l^{-1} R_l \mathbf{r}.$$

Suppose there are $(p+1)$ processors available, and that we assign one processor to each subproblem and distribute the data amongst the processors. Then, the action of $M^{-1}\mathbf{r}$ can be computed as follows. *First*, given $\mathbf{r}$, synchronize all the processors and communicate relevant data between the processors, so that processor l receives the data necessary to assemble $R_l\mathbf{r}$ from other processors. *Second*, let each processor solve its assigned problem $A_l\mathbf{w}_l = R_l\mathbf{r}$ in parallel. *Third*, synchronize and communicate the local solution $\mathbf{w}_l$ to other processors, as needed (processor $l = 0$ should transfer $R_l\mathbf{w}_0$ to processor l, while processor l should transfer $R_j R_l^T \mathbf{w}_l$ to processor j if $\Omega_j^* \cap \Omega_l^* \neq \emptyset$). *Fourth*, let each processor sum relevant components and store the result locally (processor l can sum $R_l\left(R_0^T\mathbf{w}_0 + R_1^T\mathbf{w}_1 + \cdots + R_p^T\mathbf{w}_p\right)$). For simplicity, processor 0 may be kept idle in this step. Other Schwarz algorithms may be parallelized similarly. The *PETSc* library contains parallelized codes in C, C++ and Fortran, for implementing most Schwarz solvers, see [BA15, BA14, BA13]. These codes employ *MPI* and *LAPACK*.

MPI. The message passing interface *(MPI)* is a library of routines for implementing parallel tasks in C, C++ and Fortran, see [PA, GR15]. It is based on the "message passing model", which assumes that different processors have separate memory addresses, and that data can be moved from one memory address to another. Using *MPI*, a parallel computer architecture can be simulated given a cluster of work stations connected by high speed communication lines. Once the *MPI* library has been installed, the same executable code of a parallel program employing the *MPI* library is stored and executed on each processor. Each processor is assigned a label (or rank). If there are p processors, then processor l is assigned rank l. Since the same executable code is to be run on each processor, parallelization is obtained by branching the programs based on the rank. The library employs protocol for synchronizing and communicating data between the different processors. Readers are referred to [PA, GR15] for details on the syntax, and for instructions on downloading and installing *MPI*. In many domain decomposition applications, however, details of *MPI* syntax may not be required if the *PETSc* parallel library is employed.

PETSc. The suite of routines called PETSc (Portable, Extensible Toolkit for Scientific Computing) is a library of routines for implementing domain decomposition iterative methods, optimization algorithms, and other algorithms used in scientific computing. The *PETSc* library is available in C, C++ and Fortran, but requires installation of the *MPI* and *LAPACK* libraries. Most Schwarz and Schur complement solvers are implemented in *PETSc*, and are coded to run on parallel computers. We refer to [BA14] for a tutorial on the syntax for this library.

2.5 Theoretical Results

In this section, we describe theoretical results on the convergence of multiplicative, additive and hybrid Schwarz algorithms in an Hilbert space norm, see [MA37, DR11, LI6, LI7, WI4, BR18, XU3]. We formulate an abstract convergence theory for Schwarz projection algorithms on a finite dimensional Hilbert space, where the convergence rate of the algorithms can be reduced to two key parameters, which depend the properties of the subspaces underlying the projections. The theoretical framework admits replacement of exact projections by approximations, in which case two additional parameters will arise in the convergence bounds. We focus first on the abstract theory before estimating the key parameters in applications to finite element discretizations of self adjoint and coercive elliptic equations. Additional analysis of Schwarz algorithms is presented in [ZH2, WA2, GR4, DR17, MA15].

Our discussion will be organized as follows. In Chap. 2.5.1 we present background and notation. Chap. 2.5.2 presents the abstract Schwarz convergence theory. Applications to finite element discretizations of elliptic equations are considered in Chap. 2.5.3. Our discussion follows [XU3, CH11] where additional results may be found. Selected results on the convergence of Schwarz algorithms in the maximum norm are presented in Chap. 15, see also [FR7, FR8].

2.5.1 Background

Let V denote a Hilbert space equipped with inner product $\mathcal{A}(.,.)$ and norm:

$$\|w\|_V \equiv \mathcal{A}(w,w)^{1/2}, \qquad \forall w \in V.$$

We consider the following problem. Find $u \in V$ satisfying:

$$\mathcal{A}(u,v) = F(v), \qquad \forall v \in V, \tag{2.32}$$

where $F(\cdot)$ is a bounded linear functional on V. The solution to (2.32) will be sought by Schwarz algorithms based on $(p+1)$ subspaces $V_0, \cdots, V_p$ of V:

$$V = V_0 + V_1 + \cdots + V_p,$$

i.e., for each $v \in V$ we can find $v_i \in V_i$ such that

$$v = v_0 + \cdots + v_p.$$

On each V_k, let $\mathcal{A}_k : V_k \times V_k \to \mathbb{R}$ be a symmetric, bilinear form defined as:

$$\mathcal{A}_k(v,w) \equiv \mathcal{A}(v,w), \qquad \forall v,w \in V_k.$$

If *inexact* projections (or solvers) are employed in the Schwarz algorithms, we let $\tilde{\mathcal{A}}_k : V_k \times V_k \to \mathbb{R}$ denote a symmetric, bilinear form corresponding to the inexact solver for the projection onto V_k.

Remark 2.37. We assume there exists parameters $0 < \omega_0 \le \omega_1$ such that:

$$\omega_0 \le \frac{\mathcal{A}_k(v,v)}{\tilde{\mathcal{A}}_k(v,v)} \le \omega_1, \qquad \forall v \in V_k \backslash \{0\} \tag{2.33}$$

for $0 \le k \le p$. If $\tilde{\mathcal{A}}_k(\cdot,\cdot) \equiv \mathcal{A}_k(\cdot,\cdot)$ for $0 \le k \le p$ we obtain $\omega_0 = \omega_1 = 1$.

Remark 2.38. If V is finite dimensional, by employing basis vectors for V and V_k, we may represent the bilinear forms $\mathcal{A}(\cdot,\cdot)$, $\mathcal{A}_k(\cdot,\cdot)$ and $\tilde{\mathcal{A}}_k(\cdot,\cdot)$ in terms of matrices A, A_k and $\tilde{A}_k$, respectively. Indeed, suppose n and n_k denote the dimensions of V and V_k, respectively, and let $\phi_1, \ldots, \phi_n$ be a basis for V and $\psi_1^{(k)}, \cdots, \psi_{n_k}^{(k)}$ a basis for V_k. Define an $n \times n$ matrix A and $n_k \times n_k$ matrices A_k and $\tilde{A}_k$ with entries $(A)_{ij} = \mathcal{A}(\phi_i, \phi_j)$ for $1 \le i, j \le n$, and $(A_k)_{ij} = \mathcal{A}_k(\psi_i^{(k)}, \psi_j^{(k)})$ and $\left(\tilde{A}_k\right)_{ij} = \tilde{\mathcal{A}}_k(\psi_i^{(k)}, \psi_j^{(k)})$ for $1 \le i, j \le n_k$. Matrix A_k may be obtained from matrix A as follows. Denote by R_k^T an $n \times n_k$ extension matrix whose i'th column consists of the coefficients obtained when expanding $\psi_i^{(k)}$ in the basis $\phi_1, \cdots, \phi_n$ for V:

$$\psi_i^{(k)} = \sum_{j=1}^{n} \left(R_k^T\right)_{ji} \phi_j, \qquad \text{for} \quad 0 \le k \le p.$$

Substituting this into the definition of A_k above, yields:

$$(A_k)_{ij} = \mathcal{A}_k(\psi_i^{(k)}, \psi_j^{(k)}) = \mathcal{A}\left(\sum_{l=1}^{n} \left(R_k^T\right)_{li} \phi_l, \sum_{q=1}^{n} \left(R_k^T\right)_{qj} \phi_q\right) = \left(R_k A R_k^T\right)_{ij}.$$

Thus $A_k = R_k A R_k^T$. Substituting $v = \sum_{j=1}^{n_k} (\mathbf{v})_j \, \psi_j^{(k)}$ into (2.33) yields:

$$\omega_0 \le \frac{\mathbf{v}^T A_k \mathbf{v}}{\mathbf{v}^T \tilde{A}_k \mathbf{v}} \le \omega_1, \qquad \forall \mathbf{v} \in \mathbb{R}^{n_k} \backslash \{\mathbf{0}\}.$$

This yields:

$$\omega_0 = \min_k \lambda_{\min}\left(\tilde{A}_k^{-1} A_k\right) \le \max_k \lambda_{\max}\left(\tilde{A}_k^{-1} A_k\right) = \omega_1,$$

corresponding to uniform lower and upper bounds for the spectra of $\tilde{A}_k^{-1} A_k$.

Remark 2.39. In applications to elliptic equation (2.12) with $\mathcal{B}_N = \emptyset$, the Hilbert space $V = H_0^1(\Omega)$ and $V_k = H_0^1(\Omega_k^*)$ for $1 \le k \le p$, the forms are:

$$\begin{cases} \mathcal{A}(u,v) & \equiv \int_{\Omega} \left(a(x) \nabla u \cdot \nabla v + c(x) u v\right) dx, \quad \text{for } u, v \in V \\ \mathcal{A}_k(u,v) & \equiv \int_{\Omega_k^*} \left(a(x) \nabla u \cdot \nabla v + c(x) u v\right) dx, \quad \text{for } u, v \in V_k. \end{cases}$$

A simple approximation $\tilde{\mathcal{A}}_k(\cdot,\cdot)$ of $\mathcal{A}_k(\cdot,\cdot)$ can be obtained by replacing the variable coefficients $a(.)$ and $c(.)$ by their values at an interior point $x_k \in \Omega_k^*$.

This can be particularly useful if Ω_k^* is a rectangular domain with a uniform grid, in which case fast solvers can be formulated for $\tilde{A}_k$:

$$\tilde{\mathcal{A}}_k(u,v) \equiv \int_{\Omega_k^*} (a(x_k)\nabla u \cdot \nabla v + c(x_k)uv)\,dx, \qquad \text{for } u,v \in V_k.$$

Provided $a(\cdot)$ and $c(\cdot)$ do not have large variation in Ω_k^* then ω_0 and ω_1 will correspond to uniform lower and upper bounds for $\frac{a(x)}{a(x_k)}$ and $\frac{c(x)}{c(x_k)}$ in Ω_k^*. In applications, $\tilde{A}_k$ can be any scaled preconditioner for A_k, such as ILU.

We now define a projection map $P_k : V \rightarrow V_k$ and its approximation $\tilde{P}_k : V \rightarrow V_k$ for $0 \le k \le p$ as follows.

Definition 2.40. *Given $u, w \in V$, we define $P_k u$ and $\tilde{P}_k w$ as the unique elements of V_k satisfying:*

$$\begin{cases} \mathcal{A}_k(P_k u, v) = \mathcal{A}(u,v), & \textit{for all } v \in V_k \\ \tilde{\mathcal{A}}_k(\tilde{P}_k w, v) = \mathcal{A}(w,v), & \textit{for all } v \in V_k. \end{cases}$$

The existence of P_k and $\tilde{P}_k$ follows by the Lax-Milgram lemma, see [CI2].

The following properties of P_k and $\tilde{P}_k$ will be employed in this section.

Lemma 2.41. *Let P_k and $\tilde{P}_k$ be as defined above. The following hold.*

1. *The matrix representations $\mathbf{P}_k$ of P_k and $\tilde{\mathbf{P}}_k$ of $\tilde{P}_k$ are given by:*

$$\mathbf{P}_k = R_k^T A_k^{-1} R_k A \;\text{ and }\; \tilde{\mathbf{P}}_k = R_k^T \tilde{A}_k^{-1} R_k A.$$

2. *The mappings P_k and $\tilde{P}_k$ are symmetric, positive semidefinite in $\mathcal{A}(\cdot,\cdot)$:*

$$\begin{cases} \mathcal{A}(P_k v, w) = \mathcal{A}(v, P_k w), & \textit{for } v, w, \in V \\ \mathcal{A}(\tilde{P}_k v, w) = \mathcal{A}(v, \tilde{P}_k w), & \textit{for } v, w, \in V \end{cases}$$

   with $\mathcal{A}(P_k v, v) \ge 0$ and $\mathcal{A}(\tilde{P}_k v, v) \ge 0$ for $v \in V$. In matrix terms, this corresponds to $A\mathbf{P}_k = \mathbf{P}_k^T A$, $A\tilde{\mathbf{P}}_k = \tilde{\mathbf{P}}_k^T A$, $\mathbf{v}^T A \mathbf{P}_k \mathbf{v} \ge 0$, $\mathbf{v}^T A \tilde{\mathbf{P}}_k \mathbf{v} \ge 0$.
3. *The projections P_k satisfy:*

$$P_k P_k = P_k, \quad P_k(I - P_k) = 0 \;\text{ and }\; \|P_k\|_V \le 1.$$

4. *The map $\tilde{P}_k$ satisfies $\|\tilde{P}_k\|_V \le \omega_1$ and also:*

$$\begin{cases} \omega_0\, \mathcal{A}(P_k u, u) \le \mathcal{A}(\tilde{P}_k u, u), & \textit{for all } u \in V \\ \mathcal{A}(\tilde{P}_k u, \tilde{P}_k u) \le \omega_1\, \mathcal{A}(\tilde{P}_k u, u), & \textit{for all } u \in V. \end{cases}$$

Proof. Properties of orthogonal projections P_k are standard, see [ST13, LA10]. The symmetry of $\tilde{P}_k$ in $\mathcal{A}(\cdot,\cdot)$ may be verified by employing the definition of $\tilde{P}_k$ and using that $\tilde{P}_k u, \tilde{P}_k v \in V_k$ for all $u, v \in V$:

$$\mathcal{A}(\tilde{P}_k u, v) = \mathcal{A}(v, \tilde{P}_k u) = \tilde{\mathcal{A}}_k(\tilde{P}_k v, \tilde{P}_k u) = \tilde{\mathcal{A}}_k(\tilde{P}_k u, \tilde{P}_k v) = \mathcal{A}(u, \tilde{P}_k v).$$

The positive semi-definiteness of $\tilde{P}_k$ in $\mathcal{A}(\cdot,\cdot)$ follows since:

$$0 \le \tilde{\mathcal{A}}_k(\tilde{P}_k v, \tilde{P}_k v) = \mathcal{A}(v, \tilde{P}_k v), \qquad \forall v \in V.$$

To obtain $\|\tilde{P}_k\|_V \le \omega_1$, apply the definition of $\tilde{P}_k$ and employ (2.33):

$$\begin{aligned} \|\tilde{P}_k u\|_V^2 = \mathcal{A}(\tilde{P}_k u, \tilde{P}_k u) &= \mathcal{A}_k(\tilde{P}_k u, \tilde{P}_k u) \\ &\le \omega_1 \tilde{\mathcal{A}}_k(\tilde{P}_k u, \tilde{P}_k u) \\ &= \omega_1 \mathcal{A}(u, \tilde{P}_k u) \\ &\le \omega_1 \|u\|_V \|\tilde{P}_k u\|_V. \end{aligned}$$

The desired bound follows. To verify the bound on $\mathcal{A}(\tilde{P}_k u, u)$, employ the matrix equivalents $\mathbf{P}_k\mathbf{u}$ and $\tilde{\mathbf{P}}_k\mathbf{u}$ of $P_k u$ and $\tilde{P}_k u$, respectively to obtain:

$$\begin{aligned} \mathcal{A}(P_k u, u) = \mathbf{u}^T A \mathbf{P}_k \mathbf{u} &= \mathbf{u}^T A R_k^T A_k^{-1} R_k A \mathbf{u} \\ &\le \tfrac{1}{\omega_0} \mathbf{u}^T A R_k^T \tilde{A}_k^{-1} R_k A \mathbf{u} \\ &= \tfrac{1}{\omega_0} \mathcal{A}(\tilde{P} u, u). \end{aligned}$$

Here, we have employed the property of symmetric positive definite matrices:

$$\omega_0 \le \frac{\mathbf{v}^T A_k \mathbf{v}}{\mathbf{v}^T \tilde{A}_k \mathbf{v}} \le \omega_1 \quad \forall \mathbf{v} \ne \mathbf{0} \;\Leftrightarrow\; \frac{1}{\omega_0} \ge \frac{\mathbf{v}^T A_k^{-1} \mathbf{v}}{\mathbf{v}^T \tilde{A}_k^{-1} \mathbf{v}} \ge \frac{1}{\omega_1} \quad \forall \mathbf{v} \ne \mathbf{0}.$$

To verify that $\mathcal{A}(\tilde{P}_k u, \tilde{P}_k u) \le \omega_1 \mathcal{A}(\tilde{P}_k u, u)$ consider:

$$\begin{aligned} \mathcal{A}(\tilde{P}_k u, u) = \mathcal{A}(u, \tilde{P}_k u) &= \tilde{\mathcal{A}}_k(\tilde{P}_k u, \tilde{P}_k u) \\ &\ge \tfrac{1}{\omega_1} \mathcal{A}_k(\tilde{P}_k u, \tilde{P}_k u) \\ &= \tfrac{1}{\omega_1} \mathcal{A}(\tilde{P}_k u, \tilde{P}_k u), \end{aligned}$$

where we have employed the definition of $\tilde{P}_k u$, and property (2.33), and the definition of $a_k(\cdot,\cdot)$. This yields the desired result. □

In the following, we shall derive properties of different Schwarz algorithms in terms of the mappings P_k or $\tilde{P}_k$, which will be used later.

Classical (Multiplicative) Schwarz Algorithm. Each sweep of the classical Schwarz algorithm to solve (2.32) based on subspaces $V_0, \cdots, V_p$ has the following representation in terms of projections (or its approximations):

```
For $i = 0, \cdots, p$ do
    $u^{(k+\frac{i+1}{p+1})} = u^{(k+\frac{i}{p+1})} + \tilde{P}_i \left( u - u^{(k+\frac{i}{p+1})} \right)$
Endfor
```

Since the solution u trivially satisfies $u = u + \tilde{P}_i\,(u - u)$ for $0 \le i \le p$, subtracting this from the above yields:

$$\begin{cases} \textit{For } i = 0, \cdots, p \;\; \textit{do} \\ \qquad u - u^{(k+\frac{i+1}{p+1})} = u - u^{(k+\frac{i}{p+1})} - \tilde{P}_i\left(u - u^{(k+\frac{i}{p+1})}\right) \\ \qquad\qquad\qquad\quad\;\; = (I - \tilde{P}_i)\left(u - u^{(k+\frac{i}{p+1})}\right) \\ \textit{Endfor} \end{cases}$$

Recursive application of the above yields the following expression:

$$(u - u^{(k+1)}) = (I - \tilde{P}_p) \cdots (I - \tilde{P}_0)(u - u^{(k)}), \tag{2.34}$$

which expresses the error $u - u^{(k+1)}$ in terms of the error $u - u^{(k)}$. This is referred to as the *error propagation map* or the *error amplification map*. The iterates $u^{(k)}$ of the multiplicative Schwarz algorithm will converge to the desired solution u in the energy norm $\|\cdot\|_V$ if $\|(I - \tilde{P}_p)\cdots(I - \tilde{P}_0)\|_V < 1$. This will be demonstrated later in this section.

Remark 2.42. If M^{-1} denotes the matrix action corresponding to one sweep of the *unsymmetrized* Schwarz Alg. 2.3.1 to solve (2.28), then we obtain:

$$I - M^{-1}A = (I - \tilde{\mathbf{P}}_p) \cdots (I - \tilde{\mathbf{P}}_0).$$

By (2.34), it would follow that the Schwarz vector iterates $\mathbf{u}^{(k)}$ will converge to $\mathbf{u}$ if $\|(I - \tilde{\mathbf{P}}_p)\cdots(I - \tilde{\mathbf{P}}_0)\|_V \le \delta$ for some $0 \le \delta < 1$.

Next, we express the preconditioned matrices $M^{-1}A$ corresponding to the additive, hybrid and symmetrized multiplicative Schwarz preconditioners with *inexact* local solvers $\tilde{A}_k$ in terms of the matrices $\tilde{\mathbf{P}}_k$.

Additive, Hybrid and Symmetrized Schwarz Preconditioners.

- The inverse M^{-1} of the additive Schwarz preconditioner satisfies:

$$M^{-1}A = \sum_{i=0}^{p} R_i^T \tilde{A}_i^{-1} R_i A,$$

where an inexact solver $\tilde{A}_k$ was assumed. This may also be expressed as:

$$\begin{cases} \tilde{\mathbf{P}} \equiv \sum_{i=0}^{p} \tilde{\mathbf{P}}_i = M^{-1}A, & \text{in matrix form} \\ \tilde{P} \equiv \sum_{i=0}^{p} \tilde{P}_i, & \text{in operator form} \end{cases}$$

where $\tilde{P}$ is self adjoint, and will be shown to be *coercive*, in the $\mathcal{A}(.,.)$ inner product. Its condition number satisfies:

$$cond(M, A) \equiv \frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)} = \frac{\lambda_{\max}(\tilde{P})}{\lambda_{\min}(\tilde{P})},$$

and it will be estimated later in this section.

- The inverse M^{-1} of the hybrid Schwarz preconditioner satisfies:

$$\begin{cases} M^{-1}A \\ \equiv R_0^T A_0^{-1} R_0 A + (I - R_0^T A_0^{-1} R_0)(\sum_{i=1}^p R_i^T \tilde{A}_i^{-1} R_i)(I - A R_0^T A_0^{-1} R_0)A \\ = \mathbf{P}_0 + (I - \mathbf{P}_0)\left(\mathbf{P}_0 + \sum_{i=1}^p \tilde{\mathbf{P}}_i\right)(I - \mathbf{P}_0) \\ = \mathbf{P}_0 + (I - \mathbf{P}_0)\tilde{\mathbf{P}}\,(I - \mathbf{P}_0), \end{cases}$$

  where $\tilde{\mathbf{P}} \equiv \mathbf{P}_0 + \tilde{\mathbf{P}}_1 + \cdots + \tilde{\mathbf{P}}_p$. Here, the local matrices A_i were replaced by approximations $\tilde{A}_i$ for $1 \le i \le p$. However, to ensure that all iterates lie in $V_0^{\perp}$, the coarse matrix A_0 should *not* be approximated. We obtain:

$$cond(M, A) \equiv \frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)} = \frac{\lambda_{\max}\left(P_0 + (I - P_0)\tilde{P}(I - P_0)\right)}{\lambda_{\min}\left(P_0 + (I - P_0)\tilde{P}(I - P_0)\right)},$$

  where $\tilde{P} = P_0 + \tilde{P}_1 + \cdots + \tilde{P}_p$ represents the additive Schwarz operator. This will be shown to be better conditioned than $\tilde{P}$.
- The symmetrized Schwarz preconditioner M satisfies:

$$M^{-1}A \equiv I - (I - \tilde{\mathbf{P}}_p)\cdots(I - \mathbf{P}_1)(I - \mathbf{P}_0)(I - \tilde{\mathbf{P}}_1)\cdots(I - \tilde{\mathbf{P}}_p).$$

  If an *approximate* coarse space projection $\tilde{\mathbf{P}}_0 \neq \mathbf{P}_0$ is employed, then the following alternative symmetrization $\tilde{A}$ may also be employed:

$$\tilde{A}^{-1}A \equiv I - (I - \tilde{\mathbf{P}}_p)\cdots(I - \mathbf{P}_1)(I - \tilde{\mathbf{P}}_0)(I - \tilde{\mathbf{P}}_0)(I - \tilde{\mathbf{P}}_1)\cdots(I - \tilde{\mathbf{P}}_p).$$

  Both symmetrizations will be equivalent if $\tilde{\mathbf{P}}_0 = \mathbf{P}_0$, though the latter involves an extra residual correction on V_0. We will analyze the latter.

Schwarz convergence analysis will be based on bounds for the preceding.

2.5.2 Convergence of Abstract Schwarz Algorithms

Our study of the convergence of Schwarz algorithms will involve the study of the operator $\tilde{P}$ associated with the additive Schwarz method, and E_p, the error propagation map of the multiplicative Schwarz method:

$$\begin{cases} \tilde{P} \equiv \tilde{P}_0 + \cdots + \tilde{P}_p, \\ E_p \equiv (I - \tilde{P}_p)\cdots(I - \tilde{P}_0), \end{cases}$$

Here, each $\tilde{P}_i$ as defined earlier, denotes an approximation of the projection P_i onto the subspace V_i. The spectra $\lambda_{\min}\left(\tilde{P}\right)$ and $\lambda_{\max}\left(\tilde{P}\right)$ of the $\mathcal{A}(\cdot,\cdot)$-self adjoint operator $\tilde{P}$ and the norm $\|E_p\|_V$ of the error propagation map E_p will be estimated. These quantities will generally depend on two parameters

K_0 and K_1 associated with the subspaces $V_0, \cdots, V_p$, and the approximate solvers $\tilde{A}_i$ for $0 \le i \le p$. Estimates of K_0 and K_1 will be described later in this section for a finite element discretization of a self adjoint and coercive elliptic equation and will also depend on the parameters ω_0 and ω_1.

Definition 2.43. *We associate a parameter $K_0 > 0$ with the spaces $V_0, \ldots, V_p$ and the forms $\tilde{\mathcal{A}}_0(.,.), \ldots, \tilde{\mathcal{A}}_p(.,.)$ if for each $w \in V$ there exists $w_i \in V_i$:*

$$w = w_0 + \cdots + w_p$$

and satisfying the bound:

$$\sum_{i=0}^{p} \tilde{\mathcal{A}}_i(w_i, w_i) \le K_0 \, \mathcal{A}(w, w).$$

Remark 2.44. In matrix form, the above may be stated that given $\mathbf{w} \in \mathbb{R}^n$, there exists $\mathbf{w}_i \in \mathbb{R}^{n_i}$ for $0 \le i \le p$ such that:

$$\mathbf{w} = R_0^T \mathbf{w}_0 + \cdots + R_p^T \mathbf{w}_p,$$

and

$$\sum_{i=0}^{p} \mathbf{w}_i^T \tilde{A}_i \mathbf{w}_i \le K_0 \, \mathbf{w}^T A \mathbf{w}.$$

The following result reduces the estimation of K_0 to a parameter C_0 in [LI6].

Lemma 2.45. *Suppose the following assumptions hold.*

1. *Let $C_0 > 0$ be a parameter such that for each $w \in V$ there exists $w_i \in V_i$ for $0 \le i \le p$ satisfying $w = w_0 + \cdots + w_p$ and:*

$$\sum_{i=0}^{p} \mathcal{A}_i(w_i, w_i) \le C_0 \mathcal{A}(w, w).$$

2. *Let $\omega_0 > 0$ be defined by (2.33).*

Then, the following estimate will hold:

$$K_0 \le \frac{C_0}{\omega_0}.$$

Proof. By assumption:

$$\sum_{i=0}^{p} \mathcal{A}_i(w_i, w_i) \le C_0 \, \mathcal{A}(w, w).$$

Substituting

$$\omega_0 \, \tilde{\mathcal{A}}_i(w_i, w_i) \le \mathcal{A}_i(w_i, w_i), \qquad \text{for } 0 \le i \le p,$$

in the above, yields the desired result. □

Definition 2.46. *Let $K_1 > 0$ be a parameter such that for all choices of $v_0, \cdots, v_p,\, w_0, \cdots, w_p \in V$ and for any collection $\mathcal{I}$ of subindices:*

$$\mathcal{I} \subset \{(i,j) : 0 \le i \le p,\ 0 \le j \le p\},$$

the following holds:

$$\sum_{(i,j)\in\mathcal{I}} \mathcal{A}\left(\tilde{P}_i v_i, \tilde{P}_j w_j\right) \le K_1 \left(\sum_{i=0}^{p} \mathcal{A}\left(\tilde{P}_i v_i, v_i\right)\right)^{1/2} \left(\sum_{j=0}^{p} \mathcal{A}\left(\tilde{P}_j w_j, w_j\right)\right)^{1/2}.$$

Remark 2.47. In matrix terms, the preceding requires that for all choices of $\mathbf{v}_0, \cdots, \mathbf{v}_p, \mathbf{w}_0, \cdots, \mathbf{w}_p$ and indices $\mathcal{I}$ the following holds:

$$\begin{aligned} &\sum\nolimits_{(i,j)\in\mathcal{I}} \mathbf{v}_i^T A R_i^T \tilde{A}_i^{-1} R_i A R_j^T \tilde{A}_j^{-1} R_j A \mathbf{w}_j \\ &\quad \le K_1 \left(\sum\nolimits_{i=0}^{p} \|R_i A \mathbf{v}_i\|^2_{\tilde{A}_i^{-1}}\right)^{1/2} \left(\sum\nolimits_{j=0}^{p} \|R_j A \mathbf{w}_j\|^2_{\tilde{A}_j^{-1}}\right)^{1/2}. \end{aligned}$$

Here we denote the norm $\|\mathbf{x}_i\|^2_{\tilde{A}_i^{-1}} = \mathbf{x}_i^T \tilde{A}_i^{-1} \mathbf{x}_i$ for $\tilde{A}_i = \tilde{A}_i^T > 0$.

The parameter K_1 can be estimated in terms of ω_1 and the spectral radius $\rho(\mathcal{E})$ of a matrix $\mathcal{E} = (\epsilon_{ij})$, whose entries ϵ_{ij} are *strengthened* Cauchy-Schwartz inequality parameters associated with each pair of subspaces V_i and V_j.

Definition 2.48. *For each index pair $i, j \in \{0, \cdots, p\}$ define the parameters $0 \le \epsilon_{ij} \le 1$ as the smallest possible coefficient satisfying:*

$$\mathcal{A}(w_i, w_j) \le \epsilon_{ij}\, \mathcal{A}(w_i, w_i)^{1/2} \mathcal{A}(w_j, w_j)^{1/2}, \qquad \forall w_i \in V_i,\ \ w_j \in V_j.$$

Matrix $\mathcal{E} \equiv (\epsilon_{ij})$ for $0 \le i,\, j \le p$.

Remark 2.49. Parameter ϵ_{ij} represents the maximum modulus of the cosine of the *angle* between all pairs of vectors in subspace V_i and V_j. If $\epsilon_{ij} < 1$ the above is called a *strengthened* Cauchy-Schwartz inequality. In particular, if the subspaces are orthogonal, i.e., each vector in V_i is orthogonal to each vector in V_j, then $\epsilon_{ij} = 0$, while if V_i and V_j share at least one nontrivial vector in common, then $\epsilon_{ij} = 1$.

Lemma 2.50. *Suppose the following assumptions hold.*

1. *Let parameter ω_1 be as defined earlier.*
2. *Let matrix $\mathcal{E}$ be as defined earlier.*

Then the following estimate will hold:

$$K_1 \le \omega_1\, \rho\,(\mathcal{E})\,.$$

Proof. Applying the strengthened Schwartz inequalities pairwise yields:

$$\begin{aligned}
&\textstyle\sum_{(i,j)\in\mathcal{I}} \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_j w_j) \\
&\quad \le \textstyle\sum_{(i,j)\in\mathcal{I}} \epsilon_{ij}\, \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_i v_i)^{1/2} \mathcal{A}(\tilde{P}_j w_j, \tilde{P}_j w_j)^{1/2} \\
&\quad \le \textstyle\sum_{i,j} \epsilon_{ij}\, \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_i v_i)^{1/2} \mathcal{A}(\tilde{P}_j w_j, \tilde{P}_j w_j)^{1/2} \\
&\quad \le \textstyle\sum_{i,j} \epsilon_{ij}\, \omega_1\, \mathcal{A}(\tilde{P}_i v_i, v_i)^{1/2} \mathcal{A}(\tilde{P}_j w_j, w_j)^{1/2} \\
&\quad \le \omega_1\, \rho(\mathcal{E})\, (\textstyle\sum_{i=0}^{p} \mathcal{A}(\tilde{P}_i v_i, v_i))^{1/2} (\textstyle\sum_{j=0}^{p} \mathcal{A}(\tilde{P}_j w_j, w_j))^{1/2}.
\end{aligned}$$

For additional details, see [XU3]. □

The following result describes alternative bounds for K_1.

Lemma 2.51. *Suppose the following assumptions hold.*

1. *Let $V_0, \cdots, V_p$ denote subspaces of V.*
2. *Let $\mathcal{E} = (\epsilon_{ij})$ denote the strengthened Cauchy-Schwartz parameters which are associated with the subspaces V_i and V_j for $0 \le i,\, j \le p$.*
3. *Denote by l_0*

$$l_0 \equiv \max_{1\le i\le p} \left(\sum_{j=1}^{p} \epsilon_{ij} \right). \tag{2.35}$$

Then the following estimate will hold:

$$K_1 \le \begin{cases} \omega_1\, l_0, & \text{if } V_0 \text{ is not employed} \\ \omega_1\, (l_0 + 1), & \text{if } V_0 \text{ is employed.} \end{cases}$$

Proof. See [XU3, TO10]. If a coarse space V_0 is not employed, let $\tilde{\mathcal{E}}$ be defined by $\tilde{\mathcal{E}}_{ij} \equiv \epsilon_{ij}$ for $1 \le i, j \le p$. We apply lemma 2.50 to estimate K_1 as:

$$K_1 \le \omega_1\, l_0, \qquad \text{if } V_0 \text{ is not employed,}$$

since $\rho(\tilde{\mathcal{E}}) \le \|\tilde{\mathcal{E}}\|_\infty = l_0$.

If a coarse space V_0 is employed, we estimate K_1 as follows. Given an index set $\mathcal{I} \subset \{(i,j) : 0 \le i, j \le p\}$ define $\mathcal{I}_{00}, \mathcal{I}_{01}, \mathcal{I}_{10}, \mathcal{I}_{11}$ as follows:

$$\begin{cases} \mathcal{I}_{00} \equiv \{(i,j) \in \mathcal{I} : i = 0, j = 0\} \\ \mathcal{I}_{01} \equiv \{(i,j) \in \mathcal{I} : i = 0,\ 1 \le j \le p\} \\ \mathcal{I}_{10} \equiv \{(i,j) \in \mathcal{I} : 1 \le i \le p,\ j = 0\} \\ \mathcal{I}_{11} \equiv \{(i,j) \in \mathcal{I} : 1 \le i \le p,\ 1 \le j \le p\}. \end{cases}$$

Let $v_i, w_i \in V_i$ for $0 \le i \le p$. Applying Lemma 2.50 yields:

$$\begin{array}{l}
(\sum_{(i,j)\in\mathcal{I}_{00}} \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_j w_j))^2 \le \omega_1^2\, \mathcal{A}(\tilde{P}_0 v_0, v_0)\, \mathcal{A}(\tilde{P}_0 w_0, w_0) \\
(\sum_{(i,j)\in\mathcal{I}_{11}} a(\tilde{P}_i v_i, \tilde{P}_j w_j))^2 \le \omega_1^2\, l_0^2\, (\sum_{i=1}^p \mathcal{A}(\tilde{P}_i v_i, v_i))(\sum_{j=1}^p \mathcal{A}(\tilde{P}_j w_j, w_j)).
\end{array}$$

Next, consider the sum over index set $\mathcal{I}_{01}$:

$$\begin{array}{l}
(\sum_{(i,j)\in\mathcal{I}_{01}} \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_j w_j))^2 = (\sum_{j:(0,j)\in\mathcal{I}_{01}} \mathcal{A}(\tilde{P}_0 v_0, \tilde{P}_j w_j))^2 \\
= \ (\mathcal{A}(\tilde{P}_0 v_0, \sum_{j:(0,j)\in\mathcal{I}_{01}} \tilde{P}_j w_j))^2 \\
\le \mathcal{A}(\tilde{P}_0 v_0, \tilde{P}_0 v_0)\, \mathcal{A}(\sum_{j:(0,j)\in\mathcal{I}_{01}} \tilde{P}_j w_j, \sum_{j:(0,j)\in\mathcal{I}_{01}} \tilde{P}_j w_j) \\
\le \omega_1\, \mathcal{A}(\tilde{P}_0 v_0, v_0)\, \mathcal{A}(\sum_{j:(0,j)\in\mathcal{I}_{01}} \tilde{P}_j w_j, \sum_{j:(0,j)\in\mathcal{I}_{01}} \tilde{P}_j w_j) \\
\le \omega_1^2\, l_0\, \mathcal{A}(\tilde{P}_0 v_0, v_0)\, (\sum_{j:(0,j)\in\mathcal{I}_{01}} \mathcal{A}(\tilde{P}_j w_j, w_j)) \\
\le \omega_1^2\, l_0\, \mathcal{A}(\tilde{P}_0 v_0, v_0)\, (\sum_{j=0}^p \mathcal{A}(\tilde{P}_j w_j, w_j)).
\end{array}$$

Similarly, we obtain for the sum over index set $\mathcal{I}_{10}$:

$$\begin{array}{l}
(\sum_{(i,j)\in\mathcal{I}_{10}} \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_j w_j))^2 = (\sum_{j:(i,0)\in\mathcal{I}_{10}} \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_0 w_0))^2 \\
= \ (\mathcal{A}(\sum_{j:(i,0)\in\mathcal{I}_{10}} \tilde{P}_i v_i, \tilde{P}_0 w_0))^2 \\
\le \mathcal{A}(\sum_{i:(i,0)\in\mathcal{I}_{10}} \tilde{P}_i v_i, \sum_{i:(i,0)\in\mathcal{I}_{10}} \tilde{P}_i v_i)\, \mathcal{A}(\tilde{P}_0 w_0, \tilde{P}_0 w_0) \\
\le \omega_1\, \mathcal{A}(\sum_{i:(i,0)\in\mathcal{I}_{10}} \tilde{P}_i v_i, \sum_{i:(i,0)\in\mathcal{I}_{10}} \tilde{P}_i v_i)\, \mathcal{A}(\tilde{P}_0 w_0, w_0) \\
\le \omega_1^2\, l_0\, (\sum_{i:(i,0)\in\mathcal{I}_{10}} \mathcal{A}(\tilde{P}_i v_i, v_i))\, \mathcal{A}(\tilde{P}_0 w_0, w_0) \\
\le \omega_1^2\, l_0\, (\sum_{i=0}^p \mathcal{A}(\tilde{P}_i v_i, v_i))\, \mathcal{A}(\tilde{P}_0 w_0, w_0).
\end{array}$$

Combining the preceding results using that $\mathcal{I} = \mathcal{I}_{00} \cup \mathcal{I}_{01} \cup \mathcal{I}_{10} \cup \mathcal{I}_{11}$ yields:

$$\begin{array}{l}
(\sum_{(i,j)\in\mathcal{I}} \mathcal{A}(\tilde{P}_i v_i, \tilde{P}_j w_j))^2 \\
\le \omega_1^2\, (1 + 2 l_0 + l_0^2)\, (\sum_{i=0}^p \mathcal{A}(\tilde{P}_i v_i, v_i))\, (\sum_{j=0}^p \mathcal{A}(\tilde{P}_i w_i, w_i)) \\
= \ \omega_1^2\ (1 + l_0)^2\ (\sum_{i=0}^p \mathcal{A}(\tilde{P}_i v_i, v_i))\, (\sum_{j=0}^p \mathcal{A}(\tilde{P}_i w_i, w_i)).
\end{array}$$

This yields the desired bound for $K_1 \le \omega_1\,(l_0 + 1)$. □

We now estimate the condition number of the additive Schwarz operator $M^{-1}A = \tilde{P} = \sum_{i=0}^p \tilde{P}_i$. Since each $\tilde{P}_i$ is symmetric in the $\mathcal{A}(.,.)$ inner product, its eigenvalues will be real, as also the eigenvalues of $\tilde{P}$. The condition number of $\tilde{P}$ will be a quotient of the maximal and minimal eigenvalues of $\tilde{P}$, and will satisfy the following Rayleigh quotient bounds:

$$K_0^{-1} \le \frac{\mathcal{A}(\tilde{P}u, u)}{\mathcal{A}(u, u)} \le K_1, \qquad u \ne 0.$$

Theorem 2.52. *The following bounds will hold for the spectra of $\tilde{P}$:*

$$K_0^{-1} \le \lambda_{\min}\left(\tilde{P}\right) \le \lambda_{\max}\left(\tilde{P}\right) \le K_1.$$

Proof. For an *upper bound*, expand $\|\tilde{P}v\|_V^2$, and apply the definition of K_1:

$$\begin{aligned}
\|\tilde{P}v\|_V^2 = \mathcal{A}\left(\tilde{P}v, \tilde{P}v\right) &= \textstyle\sum_{i=0}^{p}\sum_{j=0}^{p} \mathcal{A}\left(\tilde{P}_i v, \tilde{P}_j v\right) \\
&\le K_1 \left(\textstyle\sum_{i=0}^{p} \mathcal{A}(\tilde{P}_i v, v)\right)^{1/2} \left(\textstyle\sum_{j=0}^{p} \mathcal{A}(\tilde{P}_j v, v)\right)^{1/2} \\
&= K_1\, \mathcal{A}\left(\tilde{P}v, v\right) \\
&\le K_1\, \|\tilde{P}\, v\|_V\, \|v\|_V.
\end{aligned}$$

The upper bound $\|\tilde{P}\, v\|_V \le K_1\, \|v\|_V$ thus follows immediately.

For a *lower bound*, choose $v \in V$ and expand $v = v_0 + \cdots + v_p$ employing the decomposition guaranteed by definition of K_0. Substitute this into $\mathcal{A}\,(v, v)$ and simplify using the definition of $\tilde{P}_i$ and the Cauchy-Schwartz inequality:

$$\begin{aligned}
\mathcal{A}\,(v, v) = \textstyle\sum_{i=0}^{p} \mathcal{A}\,(v, v_i) &= \textstyle\sum_{i=0}^{p} \tilde{\mathcal{A}}_i \left(\tilde{P}_i v, v_i\right) \\
&\le \textstyle\sum_{i=0}^{p} \tilde{\mathcal{A}}_i \left(\tilde{P}_i v, \tilde{P}_i v\right)^{1/2} \tilde{\mathcal{A}}_i\,(v_i, v_i) \\
&= \textstyle\sum_{i=0}^{p} \mathcal{A}\left(v, \tilde{P}_i v\right)^{1/2} \tilde{\mathcal{A}}_i\,(v_i, v_i) \\
&\le (\textstyle\sum_{i=0}^{p} \mathcal{A}(v, \tilde{P}_i v))^{1/2} \left(\textstyle\sum_{i=0}^{p} \tilde{\mathcal{A}}_i(v_i, v_i)\right)^{1/2} \\
&= \mathcal{A}(\tilde{P}v, v)^{1/2} \left(\textstyle\sum_{i=0}^{p} \tilde{\mathcal{A}}_i(v_i, v_i)\right)^{1/2} \\
&\le \mathcal{A}(\tilde{P}v, v)^{1/2}\, K_0^{1/2}\, \|v\|_V.
\end{aligned}$$

We thus obtain $\|v\|_V \le K_0^{1/2}\, \mathcal{A}(\tilde{P}v, v)^{1/2}$. Squaring both sides yields:

$$\|v\|_V^2 = \mathcal{A}(v, v) \le K_0\, \mathcal{A}(\tilde{P}v, v),$$

which is a lower bound for the spectrum of $\tilde{P}$. See [XU3, TO10]. □

Remark 2.53. Combining the upper and lower bounds together yields:

$$\mathrm{cond}(M, A) = \frac{\lambda_{\max}(\tilde{P})}{\lambda_{\min}(\tilde{P})} \le K_0\, K_1,$$

which is a bound for the condition number of $M^{-1}A = \tilde{P}$.

Remark 2.54. If subspaces $V_0, \cdots, V_p$ form an orthogonal decomposition of V and exact solvers are employed (i.e., $\tilde{A}_k = A_k$ for all k), then it is easily verified that $K_0 = K_1 = 1$. In this case the additive Schwarz preconditioned system will have condition number of 1 and the conjugate gradient method will converge in a single iteration.

The following result concerns the optimal choice of parameter K_0.

Lemma 2.55. *If $\hat{K}_0$ is the smallest admissible choice of parameter K_0, then:*

$$\hat{K}_0^{-1} = \lambda_{\min}\left(\tilde{P}\right).$$

Proof. For any choice of admissible parameter K_0, Thm. 2.52 yields:

$$0 < K_0^{-1} \le \lambda_{\min}\left(\tilde{P}\right).$$

Thus, $\tilde{P}$ is invertible and given $v \in V$ we may construct an optimal partition. For $0 \le i \le p$ define:

$$v_i \equiv \tilde{P}_i \tilde{P}^{-1} v.$$

By construction

$$\sum_{i=0}^{p} v_i = \sum_{i=0}^{p} \tilde{P}_i \tilde{P}^{-1} v = \tilde{P} \tilde{P}^{-1} v = v.$$

For this decomposition, the definition of $\tilde{P}_i$ and that $\tilde{P}_i \tilde{P}^{-1} v \in V_i$ yields:

$$\begin{aligned}
\textstyle\sum_{i=0}^{p} \tilde{\mathcal{A}}_i\left(v_i, v_i\right) &= \textstyle\sum_{i=0}^{p} \tilde{\mathcal{A}}_i\left(\tilde{P}_i \tilde{P}^{-1} v, \tilde{P}_i \tilde{P}^{-1} v\right) \\
&= \textstyle\sum_{i=0}^{p} \mathcal{A}\left(\tilde{P}^{-1} v, \tilde{P}_i \tilde{P}^{-1} v\right) \\
&= \mathcal{A}\left(\tilde{P}^{-1} v, \textstyle\sum_{i=0}^{p} \tilde{P}_i \tilde{P}^{-1} v\right) \\
&= \mathcal{A}\left(\tilde{P}^{-1} v, \tilde{P} \tilde{P}^{-1} v\right) \\
&= \mathcal{A}\left(\tilde{P}^{-1} v, v\right) \\
&\le \tfrac{1}{\lambda_{\min}(\tilde{P})} \mathcal{A}\left(v, v\right).
\end{aligned}$$

Thus, $K_0 = \frac{1}{\lambda_{\min}(\tilde{P})}$ is an admissible parameter. □

The following result shows that the hybrid Schwarz preconditioner $\tilde{P}_*$ is better conditioned than the associated additive Schwarz preconditioner.

Lemma 2.56. *Let K_0 and K_1 be as defined above. Define:*

$$\begin{cases} \tilde{P} \equiv P_0 + \tilde{P}_1 + \cdots + \tilde{P}_p \\ \tilde{P}_* \equiv P_0 + (I - P_0)\tilde{P}(I - P_0). \end{cases}$$

Then, the spectra of $\tilde{P}_*$ will satisfy:

$$K_0^{-1} \le \lambda_{\min}\left(\tilde{P}\right) \le \lambda_{\min}\left(\tilde{P}_*\right) \le \lambda_{\max}\left(\tilde{P}_*\right) \le \lambda_{\max}\left(\tilde{P}\right) \le K_1.$$

In particular, $\kappa_2(\tilde{P}_*) \le \kappa_2(\tilde{P})$.

Proof. Expand the terms in the Rayleigh quotient associated with $\tilde{P}_*$ as:

$$\frac{\mathcal{A}(\tilde{P}_* u, u)}{\mathcal{A}(u,u)} = \frac{\mathcal{A}(P_0 u, P_0 u) + \mathcal{A}(\tilde{P}(I-P_0)u, (I-P_0)u)}{\mathcal{A}(P_0 u, P_0 u) + \mathcal{A}((I-P_0)u, (I-P_0)u)},$$

employing the $\mathcal{A}(.,.)$-orthogonality of the decomposition $u = P_0 u + (I - P_0)u$. Since the range of $(I - P_0)$ is $V_0^{\perp}$, a subspace of V, the Rayleigh quotient associated with the self adjoint operator $(I - P_0)\tilde{P}(I - P_0)$ will satisfy:

$$\lambda_{\min}\left(\tilde{P}\right) \le \min_{u \in V_0^{\perp} \backslash \{0\}} \frac{\mathcal{A}(\tilde{P}(I-P_0)u, (I-P_0)u)}{(\mathcal{A}((I-P_0)u, (I-P_0)u)},$$

and

$$\max_{u \in V_0^{\perp} \backslash \{0\}} \frac{\mathcal{A}(\tilde{P}(I-P_0)u, (I-P_0)u)}{(\mathcal{A}((I-P_0)u, (I-P_0)u)} \le \lambda_{\max}(\tilde{P}),$$

since the extrema are considered on a subspace of V. Substituting these observations in the Rayleigh quotient yields the desired result. □

We next consider norm bounds for $E_p = (I - \tilde{P}_0)\cdots(I - \tilde{P}_p)$, the error map associated with the multiplicative Schwarz method. Bounds for $\|E_p\|_V$ directly yield convergence rates for the multiplicative Schwarz method and condition number estimates for the symmetrized Schwarz preconditioner.

Remark 2.57. If inexact solvers are employed, there are two alternative possibilities for symmetrizing Schwarz sweeps. Both define $\mathbf{w} = \mathbf{0}$ initially and define $M^{-1}\mathbf{f} \equiv \mathbf{w}$ at the end of the sweeps. The first symmetrization is:

$$\begin{cases} \textit{For} \quad k = p, p-1, \cdots, 1, 0, 1, \cdots, p-1, p \;\; \textit{do} \\ \qquad\quad \mathbf{w} \leftarrow \mathbf{w} + R_k^T \tilde{A}_k^{-1} R_k \left(\mathbf{f} - A\mathbf{w}\right) \\ \textit{Endfor} \end{cases}$$

An alternative symmetrization has an additional fractional step for $k = 0$.

$$\begin{cases} \textit{For} \quad k = p, p-1, \cdots, 1, 0, 0, 1, \cdots, p-1, p \;\; \textit{do} \\ \qquad\quad \mathbf{w} \leftarrow \mathbf{w} + R_k^T \tilde{A}_k^{-1} R_k \left(\mathbf{f} - A\mathbf{w}\right) \\ \textit{Endfor} \end{cases}$$

If an *exact* solver is used for $k = 0$, then both sweeps will be mathematically equivalent. In our analysis, we consider the latter sweep.

Lemma 2.58. *Suppose the following assumptions hold.*

1. *For some* $0 \le \delta < 1$ *let* $E_p = (I - \tilde{P}_0)\cdots(I - \tilde{P}_p)$ *satisfy:*

$$\|E_p\|_V \le \delta.$$

2. *Let* M *be the symmetrized multiplicative Schwarz preconditioner with:*

$$I - M^{-1}A = \mathbf{E}_p^T \mathbf{E}_p,$$

where $\mathbf{E}_p = (I - \tilde{\mathbf{P}}_0)\cdots(I - \tilde{\mathbf{P}}_p)$ *is the matrix equivalent of* E_p.

Then the following results will hold.

1. *The maximum eigenvalue of $M^{-1}A$ will satisfy:*
$$\lambda_{\max}\left(M^{-1}A\right) \leq 1.$$
2. *The minimum eigenvalue of $M^{-1}A$ will satisfy:*
$$1-\delta^2 \leq \lambda_{\min}(M^{-1}A).$$
3. *The condition number of the preconditioned matrix will satisfy:*
$$cond(M,A) \equiv \frac{\lambda_{\max}\left(M^{-1}A\right)}{\lambda_{\min}\left(M^{-1}A\right)} \leq \frac{1}{1-\delta^2}.$$

Proof. See [XU3, TO10]. The assumption that $\|E_p\|_V \leq \delta$ is equivalent to:

$$\mathcal{A}(E_p v, E_p v) \leq \delta^2\, \mathcal{A}(u,u), \qquad \forall u \in V.$$

Since $M^{-1}A = I - \mathbf{E}_p^T \mathbf{E}_p$, we may substitute the above into the following Rayleigh quotient, with $\mathbf{v}$ denoting the vector representation of v, to obtain:

$$\frac{\mathbf{v}^T A M^{-1} A \mathbf{v}}{\mathbf{v}^T A \mathbf{v}} = \frac{\mathcal{A}\left(v,v\right) - \mathcal{A}\left(E_p v, E_p v\right)}{\mathcal{A}\left(v,v\right)}.$$

Since $0 \leq \mathcal{A}\left(E_p v, E_p v\right) \leq \delta^2 \mathcal{A}(v,v)$, the desired results follow. □

We next derive an estimate for $\|E_p\|_V$. We employ the notation:

$$\begin{cases} E_{-1} \equiv I \\ E_0 \quad \equiv (I-\tilde{P}_0) \\ E_1 \quad \equiv (I-\tilde{P}_1)(I-\tilde{P}_0) \\ \vdots \qquad \quad \vdots \\ E_p \quad \equiv (I-\tilde{P}_p)\cdots(I-\tilde{P}_0). \end{cases} \tag{2.36}$$

We derive two preliminary results.

Lemma 2.59. *The following algebraic relations will hold for E_i defined by (2.36):*

$$\begin{cases} E_{k-1} - E_k = \tilde{P}_k E_{k-1}, & \text{for } 0 \leq k \leq p \\ \quad I - E_i = \sum_{k=0}^{i} \tilde{P}_k E_{k-1}, & \text{for } 0 \leq i \leq p. \end{cases}$$

Proof. Employing the definition of E_k and substituting $E_k = (I-\tilde{P}_k)E_{k-1}$ for $0 \leq k \leq p$ yields the first identity. The second identity is obtained from by summing up the first identity and collapsing the sum. □

Lemma 2.60. *Let the parameters ω_1, K_0 and K_1 be as defined earlier. Then, for $v \in V$, the following bound will hold:*

$$\|v\|_V^2 - \|E_p v\|_V^2 \geq (2-\omega_1) \sum_{j=0}^{p} \mathcal{A}\Big(\tilde{P}_j E_{j-1} v, E_{j-1} v\Big).$$

Proof. Consider identity $E_{k-1}v - E_k v = \tilde{P}_k E_{k-1} v$ from Lemma 2.59, take $\mathcal{A}(.,.)$ inner products of both sides with $E_{k-1}v + E_k v$, and simplify:

$$\begin{aligned}\|E_{k-1}v\|_V^2 - \|E_k v\|_V^2 &= \mathcal{A}\Big(\tilde{P}_k E_{k-1}v, E_{k-1}v\Big) + \mathcal{A}\Big(\tilde{P}_k E_{k-1}v, E_k v\Big) \\ &= \mathcal{A}\Big(\tilde{P}_k E_{k-1}v, E_{k-1}v\Big) + \mathcal{A}\Big(\tilde{P}_k E_{k-1}v, (I-\tilde{P}_k)E_{k-1}v\Big) \\ &= 2\mathcal{A}\Big(\tilde{P}_k E_{k-1}v, E_{k-1}v\Big) - \mathcal{A}\Big(\tilde{P}_k E_{k-1}v, \tilde{P}_k E_{k-1}v\Big).\end{aligned}$$

By Lemma 2.41, the map $\tilde{P}_k$ is symmetric and positive semidefinite in the $\mathcal{A}(.,.)$ inner product and satisfies:

$$\mathcal{A}\Big(\tilde{P}_k E_{k-1}v, \tilde{P}_k E_{k-1}v\Big) \leq \omega_1 \mathcal{A}\Big(\tilde{P}_k E_{k-1}v, E_{k-1}v\Big).$$

Substituting this yields:

$$\begin{aligned}\|E_{k-1}v\|_V^2 - \|E_k v\|_V^2 &= 2\mathcal{A}\left(\tilde{P}_k E_{k-1}v, E_{k-1}v\right) - \mathcal{A}\left(\tilde{P}_k E_{k-1}v, \tilde{P}_k E_{k-1}v\right) \\ &\geq (2-\omega_1)\mathcal{A}\Big(\tilde{P}_k E_{k-1}v, E_{k-1}v\Big).\end{aligned}$$

Summing for $k = 0, \cdots, p$ and collapsing the sum yields the desired result:

$$\|v\|_V^2 - \|E_p v\|_V^2 \geq (2-\omega_1) \sum_{k=0}^{p} \mathcal{A}\Big(\tilde{P}_k E_{k-1} v, E_{k-1} v\Big).$$

See [XU3, TO10] for additional details. □

We are now able to derive the main result on norm bounds for E_p.

Theorem 2.61. *Let parameters ω_1, K_0 and K_1 be as defined earlier. Then for $v \in V$, the following bound will hold:*

$$\|E_p v\|_V^2 \leq \left(1 - \frac{2-\omega_1}{K_0(1+K_1)^2}\right) \|v\|_V^2 \tag{2.37}$$

for the error propagation map E_p of the multiplicative Schwarz method.

Proof. Expand $\tilde{P}v$ and substitute $v = E_{i-1}v + (I - E_{i-1})v$ to obtain:

$$\begin{aligned}\mathcal{A}\left(\tilde{P}v, v\right) &= \textstyle\sum_{i=0}^{p} \mathcal{A}\left(\tilde{P}_i v, v\right) \\ &= \textstyle\sum_{i=0}^{p} \mathcal{A}\left(\tilde{P}_i v, E_{i-1}v\right) + \sum_{i=0}^{p} \mathcal{A}\left(\tilde{P}_i v, (I-E_{i-1})v\right) \\ &= \textstyle\sum_{i=0}^{p} \mathcal{A}\Big(\tilde{P}_i v, E_{i-1}v\Big) + \sum_{i=0}^{p}\sum_{k=1}^{i} \mathcal{A}\Big(\tilde{P}_i v, \tilde{P}_k E_{k-1}v\Big).\end{aligned}$$

The last line was obtained by an application of Lemma 2.59. By Lemma 2.41, the mappings $\tilde{P}_i$ are symmetric and positive semidefinite in $\mathcal{A}(\cdot,\cdot)$. Consequently, the Cauchy-Schwartz inequality may be generalized to yield:

$$\mathcal{A}\left(\tilde{P}_i v, E_{i-1} v\right) \le \mathcal{A}\left(\tilde{P}_i v, v\right)^{1/2} \mathcal{A}\left(\tilde{P}_i E_{i-1} v, E_{i-1} v\right)^{1/2}.$$

Summing the above for $i = 0, \cdots, p$ yields:

$$\begin{aligned} \textstyle\sum_{i=0}^p \mathcal{A}\left(\tilde{P}_i v, E_{i-1} v\right) &\le \textstyle\sum_{i=0}^p \mathcal{A}\left(\tilde{P}_i v, v\right)^{1/2} \mathcal{A}\left(\tilde{P}_i E_{i-1} v, E_{i-1} v\right)^{1/2} \\ &\le \left(\textstyle\sum_{i=0}^p \mathcal{A}(\tilde{P}_i v, v)\right)^{1/2} \left(\textstyle\sum_{i=0}^p \mathcal{A}(\tilde{P}_i E_{i-1} v, E_{i-1} v)\right)^{1/2} \\ &= \mathcal{A}\left(\tilde{P} v, v\right)^{1/2} \left(\textstyle\sum_{i=0}^p \mathcal{A}(\tilde{P}_i E_{i-1} v, E_{i-1} v)\right)^{1/2}. \end{aligned}$$

Applying the definition of K_1 yields:

$$\begin{aligned} &\textstyle\sum_{i=0}^p \sum_{k=1}^i \mathcal{A}\left(\tilde{P}_i v, \tilde{P}_k E_{k-1}) v\right) \\ &\le K_1 \left(\textstyle\sum_{i=0}^p \mathcal{A}(\tilde{P}_i v, v)\right)^{1/2} \left(\textstyle\sum_{k=0}^p \mathcal{A}(\tilde{P}_k E_{k-1}) v, E_{k-1} v)\right)^{1/2} \\ &= K_1 \mathcal{A}\left(\tilde{P} v, v\right)^{1/2} \left(\textstyle\sum_{k=0}^p \mathcal{A}(\tilde{P}_k E_{k-1}) v, E_{k-1} v)\right)^{1/2}. \end{aligned}$$

Combining both these results yields:

$$\begin{aligned} \mathcal{A}\left(\tilde{P} v, v\right) &\le \mathcal{A}\left(\tilde{P} v, v\right)^{1/2} \left(\textstyle\sum_{i=0}^p \mathcal{A}(\tilde{P}_i E_{i-1} v, E_{i-1} v)\right)^{1/2} \\ &\quad + K_1 \mathcal{A}\left(\tilde{P} v, v\right)^{1/2} \left(\textstyle\sum_{k=0}^p \mathcal{A}(\tilde{P}_k E_{k-1}) v, E_{k-1} v)\right)^{1/2} \\ &= (1 + K_1) \mathcal{A}\left(\tilde{P} v, v\right)^{1/2} \left(\textstyle\sum_{k=0}^p \mathcal{A}(\tilde{P}_k E_{k-1}) v, E_{k-1} v)\right)^{1/2}. \end{aligned}$$

Canceling common terms yields:

$$\begin{aligned} \mathcal{A}\left(\tilde{P} v, v\right)^{1/2} &\le (1 + K_1) \left(\textstyle\sum_{k=0}^p \mathcal{A}(\tilde{P}_k E_{k-1}) v, E_{k-1} v)\right)^{1/2} \\ \mathcal{A}\left(\tilde{P} v, v\right) &\le (1 + K_1)^2 \textstyle\sum_{k=0}^p \mathcal{A}(\tilde{P}_k E_{k-1}) v, E_{k-1} v). \end{aligned}$$

Applying Lemma 2.60 yields:

$$\mathcal{A}\left(\tilde{P} v, v\right) \le \frac{(1 + K_1)^2}{2 - \omega_1} \left(\|v\|_V^2 - \|E_p v\|_V^2\right).$$

Finally, applying the lower bound for the eigenvalue of $\tilde{P}$ yields:

$$K_0^{-1} \|v\|_V^2 \le \mathcal{A}\left(\tilde{P} v, v\right) \le \frac{(1 + K_1)^2}{2 - \omega_1} \left(\|v\|_V^2 - \|E_p v\|_V^2\right).$$

This immediately yields the desired inequality:

$$\|E_p v\|_V^2 \le \left(1 - \frac{2 - \omega_1}{K_0 (1 + K_1)^2}\right) \|v\|_V^2.$$

See [XU3, TO10] for additional details. □

Remark 2.62. The bound (2.37) for $\|E_p\|_V$ imposes restrictions on the choice of *inexact* solvers. To ensure convergence of multiplicative Schwarz iterates, the parameter ω_1 must satisfy $\omega_1 < 2$. We will henceforth assume that inexact solvers $\tilde{A}_k$ are suitably scaled so that $\lambda_{\max}\left(\tilde{A}_k^{-1} A_k\right) = \omega_1 < 2$.

Remark 2.63. The bound (2.37) for $\|E_P\|_V$ is not *optimal.* Indeed, suppose $V_0, \cdots, V_p$ are *mutually orthogonal* subspaces which form an orthogonal decomposition of V, equipped with the $\mathcal{A}(.,.)$-inner product. Then, the multiplicative Schwarz algorithm based on exact solvers will converge in *one* iteration, yielding $\|E_p\|_V = 0$. However, theoretical estimates yield $K_0 = K_1 = 1$ and $\omega_0 = \omega_1 = 1$ so that:

$$\|E_p\|_V \le \sqrt{\frac{3}{4}},$$

which is not optimal.

2.5.3 Applications to Finite Element Discretizations

We shall now apply the preceding abstract Schwarz convergence theory to analyze the convergence of overlapping Schwarz algorithms for solving the finite element discretization (2.28) of elliptic equation (2.12) with $\mathcal{B}_D = \partial\Omega$. We shall make several simplifying assumptions and estimate the dependence of the convergence rate on the underlying mesh size h, subdomain size h_0, overlap factor $\beta\, h_0$ and the variation in the coefficient $a(.)$. Since the rate of convergence of the *multiplicative*, *additive* and *hybrid* Schwarz algorithms depend only on the parameters K_0 and K_1, we shall estimate how these parameters depend on h, h_0, $a(.)$ and β for the finite element local spaces V_i and forms $\mathcal{A}_i(.,.)$. We shall assume that $c(x) \equiv 0$ and that *exact* solvers are employed in all projections, so that $\tilde{A}_k = A_k$ for $0 \le k \le p$ and $\omega_0 = \omega_1 = 1$. We will show that K_1 is independent of h, h_0 and $a(.)$. So our efforts will focus primarily on estimating how K_0 depends on h, h_0 and $a(.)$. Readers are referred to [XU3, TO10] for additional details.

Assumption 1. We assume that the coefficient $a(.)$ is piecewise constant on subregions $S_1, \cdots, S_q$ of Ω which form a nonoverlapping decomposition:

$$a(x) = a_k > 0, \qquad \text{for } x \in S_k, \qquad \text{for } 1 \le k \le q.$$

The notation $|||a|||$ will denote the variation in $a(x)$:

$$|||a||| \equiv \frac{\max_k a_k}{\min_l a_l}.$$

For the preceding choice of coefficients, the terms $\mathcal{A}(.,.)$ and $F(.)$ in weak formulation (2.13) of (2.12) will have the form:

$$\begin{cases} \mathcal{A}(u,v) \equiv \sum_{i=1}^{q} a_i \int_{S_i} \nabla u \cdot \nabla v \, dx, & \text{for } u, v \in H_0^1(\Omega) \\ \quad F(v) \equiv \int_{\Omega} f\, v \, dx, & \text{for } v \in H_0^1(\Omega). \end{cases}$$

We next state our assumptions on the overlapping subdomains $\{\Omega_i^*\}_{i=1}^{p}$.

Assumption 2. We assume that the overlapping subdomains $\{\Omega_i^*\}_{i=1}^p$ are constructed from a non-overlapping decomposition $\{\Omega_i\}_{i=1}^p$, where each subdomain Ω_i^* is an *extension* of Ω_i of diameter h_0, with overlap $\beta\, h_0$:

$$\Omega_i^* \equiv \Omega_i^{\beta\, h_0} \equiv \{x \in \Omega : \operatorname{dist}(x, \Omega_i) < \beta\, h_0\}, \qquad 1 \le i \le p,$$

where $0 \le \beta$ denotes an overlap parameter.

We associate a $p \times p$ *adjacency* matrix G with the subdomains $\{\Omega_i^*\}_{i=1}^p$.

Definition 2.64. *Given $\Omega_1^*, \cdots, \Omega_p^*$, we define its adjacency matrix G by:*

$$G_{ij} = \begin{cases} 1, & \text{if } \Omega_i^* \cap \Omega_j^* \neq \emptyset \\ 0, & \text{if } \Omega_i^* \cap \Omega_j^* = \emptyset \end{cases} \quad \text{and} \quad g_0 \equiv \max_i \left(\sum_{j \neq i} G_{ij} \right), \tag{2.38}$$

where g_0 denotes the maximum number of neighbors intersecting a subdomain.

We assume the following about the triangulation of Ω and the subdomains.

Assumption 3. We assume a quasiuniform triangulation $\mathcal{T}_h(\Omega)$ of Ω, whose elements align with the subdomains $\{S_i\}_{i=1}^q$, $\{\Omega_i\}_{i=1}^p$ and $\{\Omega_i^*\}_{i=1}^p$. We let V_h denote the space of continuous, piecewise linear finite element functions defined on $\mathcal{T}_h(\Omega)$. The Hilbert space $V \equiv V_h \cap H_0^1(\Omega)$, while subspaces V_i for $1 \le i \le p$ are defined as $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$. If a coarse space V_0 is employed, it will assumed to satisfy $V_0 \subset V_h \cap H_0^1(\Omega)$.

We will employ a partition of unity satisfying the following assumptions.

Assumption 4. We assume there exists a smooth partition of unity $\{\chi_i(x)\}_{i=1}^p$ subordinate to the cover $\{\Omega_i^*\}_{i=1}^p$ satisfying the following conditions:

$$\begin{cases} 0 \le \chi_i(x) \le 1, & \text{for } 1 \le i \le p \\ \chi_i(x) = 0, & \text{for } x \in \Omega \backslash \Omega_i^*,\ 1 \le i \le p \\ \chi_i(x) + \cdots + \chi_p(x) = 1, & \text{for } x \in \Omega \\ \|\nabla \chi_i\|_{L^\infty(\Omega)} \le \beta\, h_0^{-1}, & \text{for } 1 \le i \le p. \end{cases} \tag{2.39}$$

If a coarse space V_0 is employed, we consider several operators which map onto this subspace. We let Q_0 denote the $L^2(\Omega)$-*orthogonal* projection onto V_0, and when applicable, we let π_0 denote a traditional *interpolation* map onto V_0, and $\mathcal{I}_0$ a *weighted* interpolation map onto V_0. The following properties will be assumed about these operators.

Assumption 5. Let the $L^2(\Omega)$-orthogonal projection Q_0 onto V_0 satisfy:

$$\begin{cases} |Q_0 v|_{H^1(\Omega)}^2 & \le c_1(Q_0, h, h_0)\, |v|_{H^1(\Omega)}^2, & \text{for } v \in C(\overline{\Omega}) \cap H^1(\Omega) \\ \|v - Q_0 v\|_{L^2(\Omega)}^2 & \le c_2(Q_0, h, h_0)\, h_0^2\, |v|_{H^1(\Omega)}^2, & \text{for } v \in C(\overline{\Omega}) \cap H^1(\Omega), \end{cases} \tag{2.40}$$

where $c_1(Q_0, h, h_0) > 0$ and $c_2(Q_0, h, h_0) > 0$ denote parameters which may depend on h, h_0 and operator Q_0, but not on the coefficients $\{a_l\}$.

When applicable, we assume that $\pi_0 : C(\overline{\Omega}) \cap H^1(\Omega) \rightarrow V_0$ (the traditional *interpolation* map) satisfies the following local bounds on each Ω_i:

$$\begin{cases} |\pi_0 v|^2_{H^1(\Omega_i)} & \le c_1(\pi_0, h, h_0)\, |v|^2_{H^1(\Omega_i)}, \quad \text{for } v \in C(\overline{\Omega}_i) \cap H^1(\Omega_i) \\ \|v - \pi_0 v\|^2_{L^2(\Omega_i)} & \le c_2(\pi_0, h, h_0)\, h_0^2\, |v|_{H^1(\Omega_i)}, \quad \text{for } v \in C(\overline{\Omega}_i) \cap H^1(\Omega_i), \end{cases} \tag{2.41}$$

where $c_1(\pi_0, h, h_0)$ and $c_2(\pi_0, h, h_0)$ denote parameters which can depend on h, h_0 and π_0, but not on the coefficients $\{a_l\}$.

If a *weighted interpolation* map can be defined, we assume $\mathcal{S}_l = \Omega_l$ for $1 \le l \le p$ with $p = q$. We assume that $\mathcal{I}_0 : C(\overline{\Omega}) \cap H^1(\Omega) \rightarrow V_0$ satisfies the following bound on each subdomain Ω_i for $v \in H^1(\Omega)$:

$$\begin{cases} |\mathcal{I}_0 v|^2_{H^1(\Omega_i)} & \le c_1(\mathcal{I}_0, h, h_0)\, \sum_{j:G_{ij}\neq 0} d_{ij}^2\, |v|^2_{H^1(\Omega_j)}, \\ \|v - \mathcal{I}_0 v\|^2_{L^2(\Omega_i)} & \le c_2(\mathcal{I}_0, h, h_0)\, h_0^2 \sum_{j:G_{ij}\neq 0} d_{ij}^2\, |v|_{H^1(\Omega_j)}, \end{cases} \tag{2.42}$$

where $c_1(\mathcal{I}_0, h, h_0)$ and $c_2(\mathcal{I}_0, h, h_0)$ denote parameters which may depend on h, h_0 and $\mathcal{I}_0$ but not on the coefficients $\{a_l\}$. The weights $d_{ij} \ge 0$ depend on the coefficients $\{a_l\}$ and satisfy $d_{ij} \le \frac{a_j}{a_i + a_j}$, so that $\left(a_i\, d_{ij}^2 / a_j\right) \le 1$.

Remark 2.65. The $L^2(\Omega)$-orthogonal projection Q_0 will typically be *global*, in the sense that $(Q_0 w)(x)$ for $x \in \Omega_j$ may depend on $w(\cdot)$ in $\Omega \backslash \Omega_j$. In contrast, interpolation map π_0 is required to be *local* on the subregions Ω_j, since $(\pi_0 w)(x)$ for $x \in \overline{\Omega_j}$ depends only on the values of $w(\cdot)$ in $\overline{\Omega_j}$.

Remark 2.66. If as in multigrid methods, the triangulation $\mathcal{T}_h(\Omega)$ is obtained by the refinement of some coarse quasiuniform triangulation $\mathcal{T}_{h_0}(\Omega)$ whose elements $\{\Omega_i\}_{i=1}^p$ have diameter h_0, then a coarse subspace $V_0 \subset V_h$ can be defined as the continuous, piecewise linear finite element functions on $\mathcal{T}_{h_0}(\Omega)$. For such a coarse space, explicit bounds are known for $c_i(Q_0, h, h_0)$, $c_i(\pi_0, h, h_0)$ and $c_i(\mathcal{I}_0, h, h_0)$ in assumption 5, as noted in the following.

The $L^2(\Omega)$-orthogonal projection Q_0 onto V_0 will satisfy:

$$\begin{cases} |Q_0 v|^2_{H^1(\Omega)} & \le c\, |v|^2_{H^1(\Omega)}, \quad \text{for } v \in H^1(\Omega) \\ \|v - Q_0 v\|^2_{L^2(\Omega)} & \le c\, h_0^2\, |v|^2_{H^1(\Omega)}, \quad \text{for } v \in H^1(\Omega) \end{cases} \tag{2.43}$$

where c is independent of h, h_0, $a(.)$, see [BR22, BR21, XU3, DR11].

The standard nodal interpolation map π_0 onto V_0 will satisfy the following bounds on each element Ω_i of Ω for $v \in C(\overline{\Omega}) \cap H^1(\Omega)$:

$$\begin{cases} |\pi_0 v|^2_{H^1(\Omega_i)} & \le c\, (1 + \log(h_0/h))\, |v|^2_{H^1(\Omega_i)}, \quad \text{for } \Omega \subset \mathbb{R}^2 \\ |\pi_0 v|^2_{H^1(\Omega_i)} & \le c\, (1 + (h_0/h))\, |v|^2_{H^1(\Omega_i)}, \quad \text{for } \Omega \subset \mathbb{R}^3 \\ \|v - \pi_0 v\|^2_{L^2(\Omega_i)} & \le c\, h_0^2\, |v|^2_{H^1(\Omega_i)}, \quad \text{for } \Omega \subset \mathbb{R}^d, \ d = 2, 3, \end{cases} \tag{2.44}$$

where c is independent of h, h_0, $a(.)$, see [CI2, JO2, DR11, BR21].

A *piecewise constant* weighted interpolation map $\mathcal{I}_0$ onto V_0 can be defined satisfying the following bounds on each element Ω_i of Ω for $v \in C(\overline{\Omega}) \cap H^1(\Omega)$:

$$\begin{cases} |\mathcal{I}_0 v|^2_{H^1(\Omega_i)} & \leq c\, \left(1 + \log^2(h_0/h)\right) \sum_{j:G_{ij}\neq 0} d_{ij}\, |v|^2_{H^1(\Omega_j)}, \\ \|v - \mathcal{I}_0 v\|^2_{L^2(\Omega_i)} & \leq c\, h_0^2 \sum_{j:G_{ij}\neq 0} d_{ij}\, |v|_{H^1(\Omega_j)}, \end{cases} \tag{2.45}$$

where c is independent of h and h_0 (and $a(x)$), and $d_{ij} \leq \frac{a_j}{a_i + a_j}$. We refer the reader to [CO8, SA7, MA17, WA6], see also Chap. 3.9.

Remark 2.67. In applications, alternative coarse spaces may be employed, see [WI6, DR10, MA17, CA18]. In particular, the *piecewise constant* coarse space [CO8, SA7, MA17] applies to general grids and yields robust convergence.

Assumption 6. We assume that the following *inverse inequality* holds with a parameter c (independent of h) such that on each element $\kappa \in \Omega_h$

$$|v|_{H^1(\kappa)} \leq C\, h^{-1}\, \|v\|_{L^2(\kappa)}, \quad \forall v \in V_h. \tag{2.46}$$

See [ST14, CI2, GI3, JO2].

Estimation of K_1

Lemma 2.68. *Let g_0 denote the maximum number of neighboring subdomains which intersects a subdomain, as in (2.38). Then, the following will hold for the subspaces V_i defined as $V_i \equiv V_h \cap H^1_0(\Omega^*_i)$ for $1 \leq i \leq p$:*

1. *The parameter l_0 defined by (2.35) will satisfy:*
$$l_0 \leq g_0.$$
2. *The parameter K_1 will satisfy:*
$$K_1 \leq \begin{cases} \omega_1\, (g_0 + 1), & \text{if } V_0 \text{ is employed} \\ \omega_1\, g_0, & \text{if } V_0 \text{ is not employed,} \end{cases}$$
where $\omega_1 = \max_i\, \lambda_{max}\left(\tilde{A}_i^{-1} A_i\right)$.

Proof. Consider the matrix $\mathcal{E} = (\epsilon_{ij})_{i,j=0}^p$ of strengthened Cauchy-Schwartz parameters associated with subspaces $V_0, V_1, \ldots, V_p$. The following observation relates the entries of $\mathcal{E}$ to the entries of the following matrix G:

$$G_{ij} = 0 \Longrightarrow \Omega^*_i \cap \Omega^*_j = \emptyset \Longrightarrow H^1_0(\Omega^*_i) \perp H^1_0(\Omega^*_j),$$

representing subdomain adjacencies. Thus, $G_{ij} = 0$ will yield $\epsilon_{ij} = 0$ for $1 \leq i, j \leq p$. Similarly, when $G_{ij} = 1$, parameter $\epsilon_{ij} = 1$ for $1 \leq i, j \leq p$. An application of Lemma 2.51 now yields the desired result. □

Remark 2.69. For a typical overlapping decomposition $\{\Omega_i^*\}_{i=1}^p$ of Ω and for sufficiently small β, the the number g_0 of adjacent subdomains is independent of h, h_0, $\|\|a\|\|$, and β. Thus K_1 is typically independent of these parameters, and the rate of convergence of a traditional *two-level* overlapping Schwarz algorithm depends primarily only on the parameter K_0 (or equivalently C_0).

In the following, we shall estimate the parameter K_0, or equivalently the partition parameter C_0 (since we assume $\omega_0 = \omega_1 = 1$) for different Schwarz algorithms, with or without a coarse space. For convenience, with some abuse of notation, C will denote a generic constant independent of h, h_0 and $a(.)$, whose value may differ from one line to the next. The next preliminary result will be employed later in this section in estimating the parameter C_0.

Estimation of K_0

Lemma 2.70. *Suppose the following conditions hold.*

1. *Let the assumptions 1 through 6 hold.*
2. *Let $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$ for $1 \le i \le p$ be local finite element spaces.*
3. *Given $w \in V_h \cap H_0^1(\Omega)$ define $w_i \equiv \pi_h \chi_i w \in V_i$ for $1 \le i \le p$.*

Then, the following results will hold.

1. *We obtain $w = w_1 + \cdots + w_p$.*
2. *For each $1 \le i \le p$ and $1 \le j \le q$ the following bound will hold:*

$$a_j \int_{S_j} |\nabla w_i|^2 \, dx \le 2 a_j \left(\int_{S_j} |\nabla w|^2 \, dx \, + \, C \beta^{-2} \, h_0^{-2} \, \|w\|_{L^2(S_j)}^2 \right),$$

where $C > 0$ is independent of h, h_0, β and $\|\|a\|\|$.

Proof. By construction $w_1 + \cdots + w_p = \pi_h (\chi_1 + \cdots + \chi_p) w = \pi_h w = w$. Consider an element $\kappa \in S_j$ and let $\overline{x}_\kappa$ be its geometric centroid. We express:

$$\begin{cases} w_i(x) = \pi_h \chi_i(x) w(x), & x \in \kappa \\ \qquad\;\; = I_h \chi_i(\overline{x}_\kappa) w(x) + \pi_h \left(\chi_i(x) - \chi_i(\overline{x}_\kappa)\right) w(x), & x \in \kappa. \end{cases}$$

Application of the triangle and arithmetic-geometric mean inequality yields:

$$|w_i|_{H^1(\kappa)}^2 \le 2 \left|\pi_h \chi_i(\overline{x}_\kappa) w\right|_{H^1(\kappa)}^2 + 2 \left|\pi_h \left(\chi_i(\cdot) - \chi_i(\overline{x}_\kappa)\right) w\right|_{H^1(\kappa)}^2 .$$

Substituting $\pi_h \chi_i(\overline{x}_\kappa) w = \chi_i(\overline{x}_\kappa) w$ on κ and the inverse inequality yields:

$$\begin{cases} |w_i|_{H^1(\kappa)}^2 \le 2 \chi_i(\overline{x}_\kappa)^2 \, |w|_{H^1(\kappa)}^2 + C h^{-2} \left|\pi_h(\chi_i(\cdot) - \chi_i(\overline{x}_\kappa)) w\right|_{L^2(\kappa)}^2 \\ \qquad\qquad \le 2 \, |w|_{H^1(\kappa)}^2 + 2 C h^{-2} \left|\pi_h \left(\chi_i(\cdot) - \chi_i(\overline{x}_\kappa)\right) w\right|_{L^2(\kappa)}^2 . \end{cases}$$

Here, we employed that $0 \leq \chi_i(\overline{x}_\kappa) \leq 1$. By Taylor expansion, we obtain:

$$\begin{cases} |\chi_i(x) - \chi_i(\overline{x}_\kappa)| = |\nabla \chi_i(\tilde{x}) \cdot (x - \overline{x}_\kappa)| \\ \qquad\qquad\qquad\qquad\quad \leq C\beta^{-1} h_0^{-1} h, \end{cases}$$

for some point $\tilde{x}$ on the line segment $(x, \overline{x}_\kappa)$. Substituting the above in the expression preceding it yields:

$$\begin{cases} |w_i|^2_{H^1(\kappa)} \leq 2\,|w|^2_{H^1(\kappa)} + Ch^{-2} \|\pi_h\left(\chi_i(\cdot) - \chi_i(\overline{x}_\kappa)\right) w\|^2_{L^2(\kappa)} \\ \qquad\qquad \leq 2\,|w|^2_{H^1(\kappa)} + 2Ch^{-2}\beta^{-2}h_0^{-2}h^2 \|w\|^2_{L^2(\kappa)} \\ \qquad\qquad = 2\,|w|^2_{H^1(\kappa)} + 2C\beta^{-2}h_0^{-2} \|w\|^2_{L^2(\kappa)}. \end{cases}$$

Here C is a generic constant independent of h, h_0, $|||a|||$ and β. Summing over all the elements $\kappa \in S_j$ and multiplying both sides by a_j yields the result. □

Remark 2.71. Without loss of generality, we may assume that the subregions $\{S_i\}_{i=1}^q$ are obtained by *refinement* of $\{\Omega_j\}_{j=1}^p$ (if needed by intersecting the S_i with Ω_j). If m_0 denotes the maximum number of subdomains Ω_j^* intersecting a subregion S_i, then it immediately follows that $m_0 \leq g_0$ where g_0 denotes the maximum number of overlapping subdomains intersecting any Ω_i^*.

In the following result, we estimate C_0 when a coarse space V_0 is not employed. Our estimate will be based on Lemma 2.70.

Lemma 2.72. *Suppose the following conditions hold.*

1. *Let the assumptions 1 through 6 hold.*
2. *Let $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$ for $1 \leq i \leq p$.*
3. *Given $w \in V_h \cap H_0^1(\Omega)$ define $w_i \equiv \pi_h \chi_i w \in V_i$ for $1 \leq i \leq p$.*

Then, for C independent of h, β, $|||a|||$ and h_0, the decomposition will satisfy:

$$\begin{cases} \sum_{j=1}^p \mathcal{A}(w_i, w_i) \leq 2g_0\, \mathcal{A}(w,w) + 2g_0\, C\beta^{-2}h_0^{-2} \sum_{j=1}^q a_j \|w\|^2_{L^2(S_j)} \\ \qquad\qquad\qquad\quad \leq 2g_0\, \left(1 + C\beta^{-2}h_0^{-2} |||a|||\right) \mathcal{A}(w,w), \end{cases} \tag{2.47}$$

yielding that parameter $C_0 \leq 2g_0\, \left(1 + C\beta^{-2}h_0^{-2}|||a|||\right)$.

Proof. By construction $w_1 + \cdots + w_p = w$. Apply Lemma 2.70 to obtain:

$$a_j \int_{S_j} |\nabla w_i|^2 dx \leq 2a_j \int_{S_j} |\nabla w|^2 dx + 2Ca_j\beta^{-2}h_0^{-2}\|w\|^2_{L^2(\Omega_j)}.$$

Since the terms on the left hand side above are zero when $S_j \cap \Omega_i^* = 0$, we only need sum the above for i such that $G_{ij} \neq 0$ to obtain:

$$\sum_{i=1}^p a_j \int_{S_j} |\nabla w_i|^2 dx \leq 2g_0 \left(a_j \int_{S_j} |\nabla w|^2 dx + C\beta^{-2}h_0^{-2} a_j \|w\|^2_{L^2(S_j)} \right).$$

Summing the above for $j = 1, \cdots, q$ yields:

$$\begin{aligned}
\sum_{i=1}^{p} \mathcal{A}(w_i, w_i) &= \sum_{i=1}^{p} \sum_{j=1}^{q} a_j \int_{S_j} |\nabla w_i|^2 dx \\
&\leq 2 g_0 \sum_{j=1}^{q} a_j \left(\int_{S_j} |\nabla w|^2 dx + C \beta^{-2} h_0^{-2} \|w\|^2_{L^2(S_j)} \right) \\
&\leq 2 g_0 \, \mathcal{A}(w, w) + 2 g_0 C \beta^{-2} h_0^{-2} \|a\|_\infty \|w\|^2_{L^2(\Omega)} \\
&\leq 2 g_0 \, \mathcal{A}(w, w) + 2 g_0 \, C \beta^{-2} h_0^{-2} \|a\|_\infty |w|^2_{H^1(\Omega)} \\
&\leq 2 g_0 \, \mathcal{A}(w, w) + 2 g_0 \, C \beta^{-2} h_0^{-2} \|a\|_\infty \|a^{-1}\|_\infty \mathcal{A}(w, w) \\
&= 2 g_0 \, (1 + C \beta^{-2} h_0^{-2} |||a|||) \mathcal{A}(w, w).
\end{aligned}$$

Here, we employed Poincaré-Freidrich's inequality to bound $\|w\|^2_{L^2(\Omega)}$ in terms of $|w|^2_{H^1(\Omega)}$. With abuse of notation, C denotes a generic constant, whose value may differ from one line to the next. □

The preceding bound for C_0 *deteriorates* as $h_0 \to 0$. This deterioration is observed in Schwarz algorithms in which information is only exchanged between adjacent subdomains each iteration. Inclusion of a *coarse space* can remedy such deterioration, as it enables transfer of some information globally each iteration. The following result estimates C_0 when a coarse subspace V_0 is employed. These bounds, derived using the projection Q_0, are independent of h_0, but not optimal with respect to coefficient variation $|||a|||$.

Theorem 2.73. *Suppose the following conditions hold.*

1. *Let assumptions 1 to 6 hold with $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$ for $1 \leq i \leq p$.*
2. *Let $V_0 \subset V_h \cap H_0^1(\Omega)$ be a coarse space for which Q_0 satisfies (2.40).*
3. *Given $v \in V_h \cap H_0^1(\Omega)$ define $v_0 = Q_0 v$ and $v_i \equiv \pi_h \chi_i \, (v - v_0)$.*

Then, the following will hold for $v_0, v_1, \ldots, v_p$:

$$\sum_{i=0}^{p} \mathcal{A}(v_i, v_i) \leq C_0 \, \mathcal{A}(v, v),$$

with $C_0 \leq C \, (g_0 + 1) \left(1 + c_1(Q_0, h, h_0) |||a||| + c_2(Q_0, h, h_0) \beta^{-2} |||a|||\right)$, where C is independent of h, h_0, $|||a|||$ and β and $c_i(Q_0, h, h_0)$ has known dependence on h and h_0 for $i = 1, 2$, see equation (2.40) in assumption 5.

Proof. By construction, it is easily verified that $v_0 + v_1 + \cdots + v_p = v$. Since the projection Q_0 satisfies (2.40), we obtain:

$$\begin{aligned}
\mathcal{A}(Q_0 v, Q_0 v) &= \sum_{j=1}^{q} a_j \int_{S_j} |\nabla Q_0 v|^2 \, dx \\
&\leq \|a\|_\infty |Q_0 v|^2_{H^1(\Omega)} \\
&\leq \|a\|_\infty \, c_1(Q_0, h_0, h) |v|^2_{H^1(\Omega)} \\
&\leq c_1(Q_0, h_0, h) \|a\|_\infty \|a^{-1}\|_\infty \sum_{j=1}^{q} a_j \int_{S_j} |\nabla v|^2 \, dx \\
&= c_1(Q_0, h, h_0) \, |||a||| \, \mathcal{A}(v, v).
\end{aligned}$$

Here, we used equation (2.40) from assumption 5. Now apply equation (2.47) from Lemma 2.72 using $w = v - v_0$ and also using $w_i \equiv v_i = \pi_h \chi_i w$ to obtain:

$$\begin{aligned}
\textstyle\sum_{j=1}^{p} \mathcal{A}(v_i, v_i) &\le 2 g_0 \left(\mathcal{A}(w,w) + C \beta^{-2} h_0^{-2} \textstyle\sum_{j=1}^{q} a_j \|w\|^2_{L^2(S_j)} \right) \\
&\le 2 g_0 \left(\mathcal{A}(w,w) + C \beta^{-2} h_0^{-2} \|a\|_\infty \|v - Q_0 v\|^2_{L^2(\Omega)} \right) \\
&\le 2 g_0 \left(\mathcal{A}(w,w) + C\, c_2(Q_0, h, h_0)\, \beta^{-2} h_0^{-2} \|a\|_\infty h_0^2 |v|^2_{H^1(\Omega)} \right) \\
&\le 2 g_0 \left(\mathcal{A}(w,w) + C c_2(Q_0, h, h_0)\, \beta^{-2} \|a\|_\infty \|a^{-1}\|_\infty \mathcal{A}(v,v) \right) \\
&= 2 g_0 \left(\mathcal{A}(w,w) + C c_2(Q_0, h, h_0)\, \beta^{-2} |||a||| \mathcal{A}(v,v) \right),
\end{aligned}$$

where C is independent of h, h_0, $|||a|||$ and β, while $c_2(Q_0, h, h_0)$ was used from equation (2.40) in assumption 5.

Since $w = v - v_0$, applying the triangle inequality yields:

$$\mathcal{A}(w,w) \le 2 \left(1 + c_1(Q_0, h, h_0)\, |||a|||\right) \mathcal{A}(v,v).$$

Substituting the above and combining the sums for $i = 0, \cdots, p$ yields:

$$\sum_{i=0}^{p} \mathcal{A}(v_i, v_i) \le (g_0 + 1) C (1 + c_1(Q_0, h, h_0) |||a||| + c_2(Q_0, h, h_0) \beta^{-2} |||a|||) \mathcal{A}(v,v),$$

where C is a generic constant independent of h, h_0, $|||a|||$ and β. □

Remark 2.74. When V_0 is the traditional coarse space of continuous, piecewise linear finite element functions defined on a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω from which $\mathcal{T}_h(\Omega)$ is obtained by successive refinement, then $c_1(Q_0, ., .)$ and $c_2(Q_0, ., .)$ are independent of h, h_0, β and $|||a|||$, see equation (2.43), yielding:

$$C_0 \le C\, (g_0 + 1) |||a||| \left(1 + \beta^{-2}\right),$$

where C is a generic constant independent of h, h_0, $|||a|||$ and β. This result shows that a Schwarz algorithm employing traditional coarse space residual correction is robust when the variation $|||a|||$ in the coefficients is *not large.*

The next result considers alternative bounds for C_0 when $|||a|||$ is large.

Theorem 2.75. *Suppose the following assumptions hold.*

1. *Let assumptions 1 to 6 hold with $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$ for $1 \le i \le p$.*
2. *Let $p = q$ and $S_j = \Omega_j$ for $1 \le j \le p$.*
3. *Let $V_0 \subset V_h \cap H_0^1(\Omega)$ be a coarse space.*
4. *Let π_0 satisfy equation (2.41).*
5. *For $v \in V_h \cap H_0^1(\Omega)$, define $v_0 = \pi_0 v$ and $v_i \equiv \pi_h \chi_i (v - v_0)$.*

Then, the following estimate will hold:

$$\sum_{i=0}^{p} \mathcal{A}(v_i, v_i) \leq C\,(g_0+1)\left(c_1(\pi_0, h, h_0) + c_2(\pi_0, h, h_0)\,\beta^{-2}\right)\,\mathcal{A}(v, v),$$

where C is independent of h, h_0, $\|\|a\|\|$ and β and $c_i(\pi_0, h, h_0)$ are defined in equation (2.41) of assumption 5.

Proof. By construction $v_0 + \cdots + v_p = v$. Apply equation (2.41) to obtain:

$$\begin{cases} \mathcal{A}(\pi_0 v, \pi_0 v) = \sum_{j=1}^{q} a_j \int_{S_j} |\nabla \pi_0 v|^2\, dx \\ \qquad\qquad\qquad \leq c_1(\pi_0, h, h_0)\, \sum_{j=1}^{q} a_j\, |v|_{H^1(S_j)}^2 \\ \qquad\qquad\qquad = c_1(\pi_0, h, h_0)\, \mathcal{A}(v, v). \end{cases}$$

Apply Lemma 2.72 with $w = v - v_0$ and $w_i \equiv v_i = \pi_h \chi_i w$, yielding (2.47):

$$\begin{aligned} \textstyle\sum_{j=1}^{p} \mathcal{A}(v_i, v_i) &\leq 2g_0 \left(\mathcal{A}(w, w) + C\beta^{-2} h_0^{-2} \textstyle\sum_{j=1}^{q} a_j \|w\|_{L^2(S_j)}^2\right) \\ &= 2g_0 \left(\mathcal{A}(w, w) + C\beta^{-2} h_0^{-2} \textstyle\sum_{j=1}^{q} a_j \|v - \pi_0 v\|_{L^2(S_j)}^2\right) \\ &\leq 2g_0 \left(\mathcal{A}(w, w) + C\beta^{-2} h_0^{-2} c_2(\pi_0, h, h_0)\, h_0^2 \textstyle\sum_{j=1}^{q} a_j |v\|_{H^1(S_j)}^2\right) \\ &= 2g_0 \left(\mathcal{A}(w, w) + C\, c_2(\pi_0, h, h_0) \beta^{-2} \mathcal{A}(v, v)\right) \end{aligned}$$

where $c_2(\pi_0, h, h_0)$ is defined in equation (2.41) and C is independent of h, h_0, $\|\|a\|\|$ and β. Since $w = v - v_0$, the triangle inequality yields:

$$\mathcal{A}(w, w) \leq 2\,(1 + c_1(\pi_0, h, h_0))\,\mathcal{A}(v, v).$$

Substituting this and combining the terms yields:

$$\sum_{i=0}^{p} \mathcal{A}(v_i, v_i) \leq C(g_0+1)\left(c_1(\pi_0, h, h_0) + c_2(\pi_0, h, h_0)\,\beta^{-2}\right)\mathcal{A}(v, v),$$

where C is independent of h, h_0, $\|\|a\|\|$ and β. □

Remark 2.76. When V_0 is a traditional finite element coarse space defined on a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω, whose successive refinement yields $\mathcal{T}_h(\Omega)$, then bounds for $c_1(\pi_0, ., .)$ and $c_2(\pi_0, ., .)$ in equation (2.44) yields:

$$C_0 \leq \begin{cases} C\,(g_0+1)\left(\log(h_0/h) + \beta^{-2}\right), & \text{if } \Omega \subset \mathbb{R}^2 \\ C\,(g_0+1)\left((h_0/h) + \beta^{-2}\right), & \text{if } \Omega \subset \mathbb{R}^3. \end{cases}$$

This result indicates that Schwarz algorithms employing traditional coarse spaces have reasonably robust theoretical bounds independent of $\|\|a\|\|$. While these bounds deteriorate in three dimensions, computational tests indicate almost optimal convergence in both two and three dimensions.

Improved bounds result if $\mathcal{I}_0$-interpolation (2.42) is used onto V_0.

Theorem 2.77. *Suppose the following assumptions hold.*

1. *Let assumptions 1 to 6 hold with $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$ for $1 \le i \le p$.*
2. *Let $p = q$ and $S_j = \Omega_j$ for $1 \le j \le p$.*
3. *Let $V_0 \subset V_h \cap H_0^1(\Omega)$ be a coarse space.*
4. *Let $\mathcal{I}_0$ satisfy equation (2.42).*
5. *For $v \in V_h \cap H_0^1(\Omega)$, define $v_0 = \mathcal{I}_0 v$ and $v_i \equiv \pi_h \chi_i (v - v_0)$.*

Then, the following estimate will hold:

$$\sum_{i=0}^{p} \mathcal{A}(v_i, v_i) \le C\,(g_0 + 1)\left(c_1(\mathcal{I}_0, h, h_0) + c_2(\mathcal{I}_0, h, h_0)\,\beta^{-2}\right)\,\mathcal{A}(v, v),$$

where C is independent of h, h_0, $|||a|||$ and β, while $c_i(\mathcal{I}_0, h, h_0)$ is defined in equation (2.42) of assumption 5.

Proof. By construction $v_0 + \cdots + v_p = v$. Apply equation (2.42) to obtain:

$$\begin{aligned}
\mathcal{A}(\mathcal{I}_0 v, \mathcal{I}_0 v) &= \textstyle\sum_{i=1}^{p} a_i \int_{\Omega_i} |\nabla \mathcal{I}_0 v|^2 \, dx \\
&\le c_1(\mathcal{I}_0, h, h_0) \textstyle\sum_{i=1}^{p} a_i \sum_{j: G_{ij} \ne 0} \frac{d_{ij}^2}{a_j}\, a_j\, |v|_{H^1(\Omega_j)}^2 \\
&\le c_1(\mathcal{I}_0, h, h_0) \textstyle\sum_{i=1}^{p} \sum_{j: G_{ij} \ne 0} \frac{a_i\, d_{ij}^2}{a_j}\, a_j\, |v|_{H^1(\Omega_j)}^2 \\
&\le\; g_0\, c_1(\mathcal{I}_0, h, h_0)\, \mathcal{A}(v, v).
\end{aligned}$$

Apply (2.47) from Lemma 2.72 with $w = v - v_0$ and $w_l \equiv v_l = \pi_h \chi_l w$:

$$\begin{aligned}
&\textstyle\sum_{l=1}^{p} \mathcal{A}(v_l, v_l) \le 2g_0\, \left(\mathcal{A}(w, w) + C\,\beta^{-2} h_0^{-2} \sum_{i=1}^{p} a_i\, \|w\|_{L^2(\Omega_i)}^2\right) \\
&=\; 2g_0\, \left(\mathcal{A}(w, w) + C\,\beta^{-2} h_0^{-2} \textstyle\sum_{i=1}^{p} a_i\, \|v - \mathcal{I}_0 v\|_{L^2(\Omega_i)}^2\right) \\
&\le 2g_0 \left(\mathcal{A}(w, w) + C\beta^{-2} h_0^{-2} c_2(\mathcal{I}_0, h, h_0) h_0^2 \textstyle\sum_{i=1}^{p} \sum_{j: G_{ij} \ne 0} \frac{a_i d_{ij}^2}{a_j} a_j |v\|_{H^1(\Omega_j)}^2\right) \\
&=\; 2g_0 \big(\mathcal{A}(w, w) + C g_0\; c_2(\mathcal{I}_0, h, h_0)\, \beta^{-2}\, \mathcal{A}(v, v)\big),
\end{aligned}$$

where C is independent of h, h_0, $|||a|||$ and β. Since $w = v - v_0$, applying the triangle inequality yields:

$$\mathcal{A}(w, w) \le 2\,(1 + c_1(\mathcal{I}_0, h, h_0))\, \mathcal{A}(v, v).$$

Substituting this and combining the terms yields:

$$\sum_{i=0}^{p} \mathcal{A}(v_i, v_i) \le C\,(g_0 + 1)\; \left(c_1(\mathcal{I}_0, h, h_0) + c_2(\mathcal{I}_0, h, h_0)\, g_0\, \beta^{-2}\right)\, \mathcal{A}(v, v),$$

where C is independent of h, h_0, $|||a|||$ and β. □

Remark 2.78. When V_0 is the *piecewise constant* coarse space defined on the subdomain decomposition $\Omega_1, \ldots, \Omega_p$, then bounds for $c_1(\mathcal{I}_0, ., .)$ and $c_2(\mathcal{I}_0, ., .)$ in equation (2.42) will satisfy:

$$C_0 \leq C\,(g_0 + 1)\big(\log^2(h_0/h) + \beta^{-2}\big), \text{ if } \Omega \subset \mathbb{R}^d,$$

for $d = 2, 3$, see [CO8, SA7, MA17, WA6]. Thus, Schwarz algorithms employing the piecewise constant coarse space will have almost optimal convergence bounds in both *two* and *three* dimensions. Sharper estimates with respect to overlap β are obtained in [DR17].

Anisotropic Problems

We next outline estimates for Schwarz algorithms applied to solve *anisotropic* elliptic equations. We consider the following model anisotropic problem:

$$\begin{cases} -\epsilon u_{x_1 x_1} - u_{x_2 x_2} + u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.48}$$

where $\Omega \subset \mathbb{R}^2$ and $0 < \epsilon \ll 1$ is a small perturbation parameter. Due to presence of the small parameter ϵ, the preceding elliptic equation will be strongly coupled along the x_2 axis, and weakly coupled along the x_1 axis. In the limiting case of $\epsilon = 0$, the elliptic equation will not be coupled along the x_1 axis. For $0 < \epsilon \ll 1$, the solution may exhibit *boundary layer* behavior near $\partial\Omega$, i.e., there may be subregions of Ω on which the solution has large gradients. If such layers need to be resolved computationally, then refinement of the grid may be necessary in such subregions.

The weak coupling along the x_1 axis suggests several heuristic choices in the formulation of the Schwarz iterative algorithm.

- Non-overlapping subdomains $\{\Omega_i\}_{i=1}^p$ can be chosen as *strips* of the form:

  $$\Omega_i \equiv \{(x_1, x_2) : \, b_i < x_1 < b_{i+1}\} \cap \Omega, \tag{2.49}$$

  for some choice of b_i. To obtain strips of width h_0, ensure that:

  $$|b_{i+1} - b_i| = O(h_0), \qquad \text{for } 1 \leq i \leq p.$$

- Extended subdomains $\{\Omega_i^*\}_{i=1}^p$ can be constructed from the strips $\{\Omega_i\}_{i=1}^p$ using an overlap factor of $\beta\, h_0$ for some $0 < \beta < 1/2$.
- If h_0 is sufficiently small, efficient direct solvers (such as band solvers) may be available for solution of the strip problems, provided the discrete unknowns within each strip are ordered horizontally, row by row, yielding a matrix with small bandsize.
- If the overlap factor is chosen so that $\beta\, h_0 \geq c\,\sqrt{\epsilon}$, then a coarse space V_0 may not be required to ensure robust convergence.

These ideas may be extended to more general anisotropic problems in two or three dimensions, provided that in the general case the subdomains be chosen as cylinders or strips whose sections are perpendicular to the axis of weak coupling of the elliptic equation.

We now estimate the convergence rate of Schwarz iterative algorithms applied to anisotropic problem (2.48).

Lemma 2.79. *Consider a finite element discretization of elliptic equation (2.48) based on a finite element space $V_h \cap H_0^1(\Omega)$.*

1. *Choose subdomains Ω_i for $1 \leq i \leq p$ of the form (2.49) with width h_0. Extend each Ω_i to Ω_i^* to have overlap $\beta\, h_0$ where $\beta < 1/2$.*
2. *Let g_0 denote the maximum number of adjacent overlapping subdomains.*
3. *Employ a Schwarz algorithm based on subspaces $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$ for $1 \leq i \leq p$, without a coarse space V_0, and use exact local solvers.*

Then the following will hold.

1. *Parameter K_1 will satisfy $K_1 \leq g_0$,for sufficiently small β.*
2. *Parameter K_0 (equivalently C_0, since $\omega_0 = \omega_1 = 1$) will satisfy:*
$$K_0 \leq C\, g_0 \big(1 + \epsilon \beta^{-2} h_0^{-2}\big),$$
for C independent of h, h_0, ϵ and β.

Proof. We outline the proof only in the continuous case. The proof involving a finite element discretization can be obtained by appropriate modification of the proof given below. Applying Lemma 2.68 yields $K_1 \leq g_0$. To estimate K_0, given the strip subdomains, we shall employ a partition of unity $\chi_1(x), \cdots, \chi_p(x)$ subordinate to the strip subdomains $\Omega_1^*, \cdots, \Omega_p^*$, such that $\chi_i(x) = \chi_i(x_1)$, i.e., each partition of unity function is solely a function of the variable x_1. We further require the smoothness assumption:

$$\left| \frac{\partial \chi_i}{\partial x_1} \right| \leq C \beta^{-1} h_0^{-1}, \qquad 1 \leq i \leq p.$$

Such a partition of unity will not satisfy $\chi_i(x) = 0$ for $x \in \partial\Omega$. However, this will not alter the construction of w_i described below, since the partition of unity functions will multiply functions which are in $H_0^1(\Omega)$.

Given such a partition of unity and $w \in H_0^1(\Omega)$ define $w_i \equiv \chi_i\, w$. Then, by construction $(w_1 + \cdots + w_p) = (\chi_1 + \cdots + \chi_p)\; w = w$. Furthermore:

$$\frac{\partial w_i}{\partial x_1} = \left(\frac{\partial \chi_i}{\partial x_1}\, w + \chi_i\, \frac{\partial w}{\partial x_1} \right) \quad \text{and} \quad \frac{\partial w_i}{\partial x_2} = \left(\chi_i \frac{\partial w}{\partial x_2} \right).$$

Employing arguments analogous to the isotropic case, we obtain:

$$\begin{cases} \mathcal{A}(w_i, w_i) = \epsilon \| \frac{\partial w_i}{\partial x_1} \|^2_{L^2(\Omega_i^*)} + \| \frac{\partial w_i}{\partial x_2} \|^2_{L^2(\Omega_i^*)} + \|w_i\|^2_{L^2(\Omega_i^*)} \\ \leq C\, \left(\epsilon \beta^{-2} h_0^{-2} \|w\|^2_{L^2(\Omega_i^*)} + \epsilon \| \frac{\partial w}{\partial x_1} \|^2_{L^2(\Omega_i^*)} + \| \frac{\partial w}{\partial x_2} \|^2_{L^2(\Omega_i^*)} + \|w\|^2_{L^2(\Omega_i^*)} \right) \\ = C\, \left(1 + \epsilon \beta^{-2} h_0^{-2}\right) \left(\epsilon \| \frac{\partial w}{\partial x_1} \|^2_{L^2(\Omega_i^*)} + \| \frac{\partial w}{\partial x_2} \|^2_{L^2(\Omega_i^*)} + \|w\|^2_{L^2(\Omega_i^*)} \right). \end{cases}$$

Summing over $1 \leq i \leq p$ yields the following bound:

$$\sum_{i=1}^{p} \mathcal{A}(w_i, w_i) \leq C g_0 \, \left(1 + \epsilon \beta^{-2} h_0^{-2}\right) \mathcal{A}(w, w).$$

Thus $C_0 \leq C g_0 \, \left(1 + \epsilon \beta^{-2} h_0^{-2}\right)$, where C will be independent of h_0 and ϵ (and h in the discrete case). □

Remark 2.80. If the overlap satisfies $\beta h_0 \geq c \sqrt{\epsilon}$, then the term $\left(1 + \epsilon \beta^{-2} h_0^{-2}\right)$ will be bounded and convergence of Schwarz algorithms will be robust without the inclusion of coarse space correction.

Time Stepping Problems

We conclude this section by considering the Schwarz algorithm for the iterative solution of the linear system arising from the implicit time stepping of a finite element or finite difference discretization of a parabolic equation:

$$\begin{cases} u_t + Lu = f, & \text{in } \Omega \times [0, T] \\ u = 0, & \text{on } \partial\Omega \times [0, T] \\ u(x, 0) = u_0(x), & \text{in } \Omega, \end{cases} \tag{2.50}$$

where $Lu \equiv -\nabla \cdot (a \nabla u)$. If $\tau > 0$ denotes the time step, then the elliptic equation resulting from an implicit time stepping of (2.50) will have the form:

$$\begin{cases} (I + \tau L) = \tilde{f}, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{2.51}$$

This elliptic equation is singularly perturbed for $\tau \to 0^+$ and may exhibit boundary layer behavior on subregions. Grid refinement may be necessary to resolve such layer regions. We will assume that the parabolic equation has been suitably discretized.

The presence of the small parameter $0 < \tau \ll 1$ enables simplification of Schwarz algorithms to solve (2.51) or its discretizations [KU3, KU6, CA, CA3].

- Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of Ω of size h_0. Let each extended subdomain Ω_i^* be constructed by extending Ω_i to include overlap of size $\beta h_0 \geq c\sqrt{\tau}$.
- The Schwarz algorithm based on the subspaces $V_i \equiv V_h \cap H_0^1\left(\Omega_i^*\right)$ will have optimal order convergence without the use of a coarse space.

The absence of coarse space residual correction can be particularly advantageous from the viewpoint of parallelization, since coarse spaces requires interprocessor communication. Estimates yield that $K_1 \leq g_0$ and:

$$K_0 \leq C g_0 \left(1 + \tau \, \beta^{-2} h_0^{-2}\right),$$

for C independent of h, h_0 and τ, see [KU3, KU6, CA, CA3] and Chap. 9.

3

Schur Complement and Iterative Substructuring Algorithms

In this chapter, we describe multi-subdomain Schur complement and iterative substructuring methods. These methods iteratively solve the linear systems arising from the discretization of a *self adjoint* and *coercive* elliptic equation, based on a decomposition of its domain into *non-overlapping subdomains*. In the *continuous case*, the solution to an elliptic equation can be parameterized in terms of its unknown Dirichlet values on the subdomain boundaries. This parameterization enables reducing the original elliptic equation to a *Steklov-Poincaré* problem for determining the solution on such boundaries. Once the reduced problem is solved, the global solution can be obtained by solving a local boundary value problem on each subdomain, in parallel.

In the *discrete case*, parameterizing the global solution in terms of its Dirichlet values on the subdomain boundaries, to obtain a reduced problem, corresponds to a block Gaussian elimination of the unknowns in the interiors of the subdomains. This reduced system, referred to as the *Schur complement system*, is *iteratively* solved by a PCG method. The Schur complement matrix is by construction a discrete approximation of the Steklov-Poincaré operator, and this property enables the formulation of various effective *preconditioners*. By contrast, the traditional *substructuring* method in structural engineering, which pre-dates domain decomposition methodology, assembles and solves the Schur complement system using a *direct* method [PR4, PR5].

Our discussion in this chapter is organized as follows. In Chap. 3.1 we introduce notations. The Schur complement system and its algebraic properties are described in Chap. 3.2, with the substructuring method. Chap. 3.3 describes FFT based fast *direct* solvers for Schur complement systems on rectangular domains with stripwise constant coefficients. Chap. 3.4 describes several preconditioners for two subdomain Schur complement matrices, while Chap. 3.5 and Chap. 3.6 describe multi-subdomain preconditioners for Schur complements in two dimensions and three dimensions. Chap. 3.7 describes the Neumann-Neumann and balancing preconditioners, while Chap. 3.8 discusses implementational issues. Chap. 3.9 describes theoretical estimates for the condition number of various Schur complement preconditioners.

3.1 Background

We consider the following *self adjoint* and *coercive* elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\,\nabla u) + c(x)\,u = f(x), & \text{in } \Omega \\ u = 0 & \text{on } \mathcal{B}_D, \\ \mathbf{n} \cdot (a\nabla u) = g_N(x), & \text{on } \mathcal{B}_N, \end{cases} \tag{3.1}$$

where $a(x) \geq a_0 > 0$ and $c(x) \geq 0$. Here $\mathcal{B}_D$ and $\mathcal{B}_D$ denote the Dirichlet and Neumann boundary segments, with $\mathcal{B}_D \cup \mathcal{B}_N = \partial\Omega$ and $\mathcal{B}_D \cap \mathcal{B}_N = \emptyset$. Given a quasiuniform triangulation $\mathcal{T}_h(\Omega)$ of Ω, we shall let V_h denote the finite element space of continuous, piecewise linear functions defined on $\mathcal{T}_h(\Omega)$. A finite element discretization of (3.1) seeks $u_h \in V_h \cap H^1_D(\Omega)$ satisfying:

$$\begin{cases} \mathcal{A}(u_h, v_h) = F(v_h), & \forall v_h \in V_h \cap H^1_D(\Omega), \text{ where} \\ \mathcal{A}(u, v) \quad \equiv \int_\Omega (a\,\nabla u \cdot \nabla v + c\,uv)\,dx, & \forall u, v \in H^1_D(\Omega) \\ F(v) \quad \equiv \int_\Omega f\,v\,dx + \int_{\mathcal{B}_N} g_N\,v\,ds_x, & \forall v \in H^1_D(\Omega) \\ H^1_D(\Omega \quad \equiv \left\{v \in H^1(\Omega) \,:\, v = 0 \text{ on } \;\mathcal{B}_D\right\}. \end{cases} \tag{3.2}$$

Let n denote the number of nodes of $\mathcal{T}_h(\Omega)$ in $(\Omega \cup \mathcal{B}_N)$. We enumerate them as $x_1, \ldots, x_n$. Then, the standard piecewise linear nodal basis functions $\{\phi_i(x)\}_{i=1}^n$ dual to these nodes will satisfy:

$$\phi_j(x_i) = \delta_{ij}, \quad 1 \leq i, j \leq n. \tag{3.3}$$

A matrix representation of the discretization (3.2) can be obtained by expanding u_h relative to this nodal basis $u_h(y) \equiv \sum_{i=1}^n u_h(x_i)\,\phi_i(y)$, and substituting this into (3.2) with $v_h = \phi_j$ for $1 \leq j \leq n$. This results in a linear system:

$$A_h \mathbf{u} = \mathbf{f}, \tag{3.4}$$

where:

$$\begin{cases} (A_h)_{ij} = \mathcal{A}(\phi_i, \phi_j), & \text{for } 1 \leq i, j \leq n \\ (\mathbf{u})_i = u_h(x_i), & \text{for } 1 \leq i \leq n \\ (\mathbf{f})_i = F(\phi_i), & \text{for } 1 \leq i \leq n. \end{cases}$$

This system will be partitioned into subblocks based on an ordering of the nodes given a decomposition of the domain into *non-overlapping* subdomains.

Definition 3.1. *We shall say that $\Omega_1, \ldots, \Omega_p$ forms a non-overlapping decomposition of Ω (see Fig. 3.1) if:*

$$\overline{\Omega} = \cup_{l=1}^p \overline{\Omega}_l \quad \text{ and } \quad \Omega_i \cap \Omega_j = \emptyset \quad \text{when } i \neq j.$$

The following notation will be employed for subdomain boundaries.

$$B \equiv \cup_{i=1}^p B^{(i)} \text{ and } B^{(i)} \equiv \partial\Omega_i \backslash \mathcal{B}_D \text{ and } B_{[i]} \equiv \partial\Omega_i \cap \mathcal{B}_D \text{ for } 1 \leq i \leq p.$$

Here $B^{(i)}$ denotes the interior and Neumann segment of $\partial\Omega_i$, $B_{[i]}$ the exterior non-Dirichlet segment, and B the interface separating the subdomains. We also let $B_{ij} \equiv B^{(i)} \cap B^{(j)}$ denote the interface between Ω_i and Ω_j.

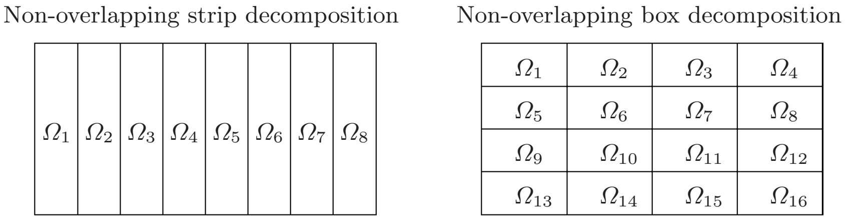


Fig. 3.1. Multidomain non-overlapping decompositions

In most applications, *box* like subdomain decompositions will be employed, though *strip* decompositions have advantages. We shall assume that the subdomains are chosen to align with the triangulation $\mathcal{T}_h(\Omega)$, and that the nodes $x_1, \ldots, x_n$ in $\mathcal{T}_h(\Omega)$ are ordered based on the subdomains $\Omega_1, \ldots, \Omega_p$ and interface B, as in Fig. 3.1. The nodes *within* each subdomain Ω_i and on the interface B may be ordered arbitrarily. Let $n_I^{(i)}$ denote the number of nodes in subdomain Ω_i and $n_B^{(i)}$ the number of nodes on $B^{(i)}$. Let n_B denote the number of nodes on B. Then, by construction it will hold $n = (n_I^{(1)} + \cdots + n_I^{(p)} + n_B)$. We shall assume that the chosen ordering of nodes satisfies:

$$\begin{cases} x_j \in \Omega_i, & \text{for } (n_I^{(1)} + \ldots n_I^{(i-1)}) + 1 \le j \le (n_I^{(1)} + \ldots n_I^{(i)}), \text{ for } 1 \le i \le p \\ x_j \in B, & \text{for } (n_I + 1) \le j \le (n_I + n_B), \end{cases}$$

where $n_I \equiv (n_I^{(1)} + \cdots + n_I^{(p)})$ denotes the total number of nodes in *subdomain* interiors. Using this ordering, system (3.4) can be block partitioned as:

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ \mathbf{f}_B \end{bmatrix}, \tag{3.5}$$

corresponding to the partition $\mathbf{u} = \left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$ and $\mathbf{f} = \left(\mathbf{f}_I^T, \mathbf{f}_B^T\right)^T$ where:

$$\begin{cases} (A_{II})_{lj} &= (A_h)_{lj}, & \text{for } 1 \le l, j \le n_I \\ (A_{IB})_{lj} &= (A_h)_{l,n_I+j}, & \text{for } 1 \le l \le n_I \text{ and } 1 \le j \le n_B \\ (A_{BB})_{lj} &= (A_h)_{n_I+l,n_I+j}, & \text{for } 1 \le l, j \le n_B \\ (\mathbf{u}_I)_j &= (\mathbf{u})_j, & \text{for } 1 \le j \le n_I \\ (\mathbf{u}_B)_j &= (\mathbf{u})_{n_I+j}, & \text{for } 1 \le j \le n_B \\ (\mathbf{f}_I)_j &= (\mathbf{f})_j, & \text{for } 1 \le j \le n_I \\ (\mathbf{f}_B)_j &= (\mathbf{f})_{n_I+j}, & \text{for } 1 \le j \le n_B. \end{cases}$$

The block submatrices A_{II} and A_{IB} in (3.5) will be further partitioned using submatrices arising from the subregions, and this will be described later.

3.2 Schur Complement System

The solution to system (3.5) can be sought formally by block Gaussian elimination. Eliminating $\mathbf{u}_I$ using the first block equation in (3.5) below:

$$\begin{cases} A_{II}\mathbf{u}_I + A_{IB}\mathbf{u}_B &= \mathbf{f}_I \\ A_{IB}^T\mathbf{u}_I + A_{BB}\mathbf{u}_B &= \mathbf{f}_B \end{cases}$$

yields $\mathbf{u}_I = A_{II}^{-1}\left(\mathbf{f}_I - A_{IB}\mathbf{u}_B\right)$ provided A_{II} is invertible. Substituting this parametric representation of $\mathbf{u}_I$ into the 2nd block equation above yields the following reduced linear system for $\mathbf{u}_B$:

$$\begin{cases} S\mathbf{u}_B = \tilde{\mathbf{f}}_B, \quad \text{where} \\ \quad S \equiv (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}) \\ \quad \tilde{\mathbf{f}}_B \equiv (\mathbf{f}_B - A_{IB}^T A_{II}^{-1}\mathbf{f}_I). \end{cases} \tag{3.6}$$

The system $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ is referred to as the Schur complement system. It corresponds to a discrete approximation of a Steklov-Poincaré problem associated with elliptic equation (3.1), but posed on the interface B. Matrix S is referred to as the Schur complement (strictly speaking, S is the Schur complement of submatrix A_{II} in A_h). The Schur complement system can be employed to determine the solution $\left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$ to (3.5) as follows. First, determine $\mathbf{u}_B$ by (iteratively or directly) solving the Schur complement system (3.6):

$$\mathbf{u}_B = S^{-1}\left(\mathbf{f}_B - A_{IB}^T A_{II}^{-1}\mathbf{f}_I\right).$$

This will be possible when matrix S is invertible. Once $\mathbf{u}_B$ has been determined, $\mathbf{u}_I$ can be obtained by solving $A_{II}\mathbf{u}_I = (\mathbf{f}_I - A_{IB}\mathbf{u}_B)$, yielding:

$$\mathbf{u}_I = A_{II}^{-1}\left(\mathbf{f}_I - A_{IB}\mathbf{u}_B\right).$$

We summarize the resulting algorithm below.

Algorithm 3.2.1 *(Schur Complement Algorithm)*

1. *Solve for* $\mathbf{w}_I$*:*
$$A_{II}\,\mathbf{w}_I \;=\; \mathbf{f}_I.$$
2. *Compute:*
$$\tilde{\mathbf{f}}_B \;=\; \mathbf{f}_B - A_{IB}^T\mathbf{w}_I.$$
3. *Solve for* $\mathbf{u}_B$*:*
$$S\,\mathbf{u}_B \;=\; \tilde{\mathbf{f}}_B.$$
4. *Solve for* $\mathbf{u}_I$*:*
$$A_{II}\,\mathbf{u}_I \;=\; (\mathbf{f}_I - A_{IB}\mathbf{u}_B).$$

Output: $\left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$.

Schur complement and iterative substructuring algorithms are motivated by the preceding algorithm. If a *direct* solver is employed to solve (3.6), then matrix S must first be *assembled*. This is the approach employed in traditional substructuring [PR4, PR5]. However, in domain decomposition applications, the Schur complement system (3.6) is typically solved using a *preconditioned* conjugate gradient *iterative* method. This does not require explicit assembly of matrix S, and instead only requires computing the action of S on different vectors. Such matrix-vector products, for instance $S\,\mathbf{w}_B$, given $\mathbf{w}_B$, may be computed by first solving $A_{II}\mathbf{w}_I = -A_{IB}\mathbf{w}_B$ in *parallel* (as discussed below) for $\mathbf{w}_I$, and by subsequently defining $S\,\mathbf{w}_B \,\equiv\, A_{BB}\mathbf{w}_B + A_{IB}^T\mathbf{w}_I$.

The preceding version of the Schur complement algorithm can be implemented in *parallel* by using the block structure of matrix A_{II}. Indeed, given a decomposition of Ω into the subdomains $\Omega_1, \ldots, \Omega_p$, and an ordering of the nodes based of this, matrix A_{II} in system (3.5) will be *block diagonal*. To see this, note that when nodes x_i and x_j belong to the interiors of different subdomains, then the nodal basis functions $\phi_i(x)$ and $\phi_j(x)$ will have support in different subdomains, yielding that $A_{ij} = \mathcal{A}(\phi_i, \phi_j) = 0$. More formally, define the index set:

$$I^{(j)} \equiv \left\{ i \,:\, (n_I^{(1)} + \cdots + n_I^{(j-1)} + 1) \,\leq\, i \,\leq\, (n_I^{(1)} + \cdots + n_I^{(j)}) \right\}.$$

By construction $x_i \in \Omega_j \,\Leftrightarrow\, i \in I^{(j)}$ and $I = I^{(1)} \cup \cdots \cup I^{(p)}$. It then follows that the diagonal blocks of $A_{II} = \mathrm{blockdiag}\left(A_{II}^{(1)}, \ldots, A_{II}^{(p)}\right)$ satisfy:

$$A_{II} = \begin{bmatrix} A_{II}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{II}^{(p)} \end{bmatrix} \quad \text{where} \quad \begin{cases} \left(A_{II}^{(j)}\right)_{lk} = (A_h)_{\tilde{l}\tilde{k}} \text{ for } 1 \leq l,\, k \leq n_I^{(j)} \\ \tilde{l} \,=\, (n_I^{(1)} + \cdots + n_I^{(j-1)}) + l \\ \tilde{k} \,=\, (n_I^{(1)} + \cdots + n_I^{(j-1)}) + k. \end{cases} \tag{3.7}$$

This block diagonal structure of A_{II} will enhance the parallelizability of Schur complement algorithms, since the action of $A_{II}^{-1} = \mathrm{blockdiag}({A_{II}^{(1)}}^{-1}$, …, ${A_{II}^{(p)}}^{-1}$) involves p separate blocks, each of which can be computed in parallel.

We next describe the *substructuring* algorithm for solving (3.5). It employs a direct method to solve (3.6), but incorporates the assembly of matrices A_h and S by a finite element *subassembly* procedure. Given the non-overlapping subdomains $\Omega_1, \ldots, \Omega_p$, let $\mathcal{A}_{\Omega_i}(.,.)$ and $F_{\Omega_i}(.)$ denote subdomain forms:

$$\begin{cases} \mathcal{A}_{\Omega_i}(u,v) \equiv \int_{\Omega_i} (a(x)\nabla u \cdot \nabla v + c(x)\, uv)\, dx, & \text{for } u,\, v \in H_D^1(\Omega) \\ \quad F_{\Omega_i}(v) \equiv \int_{\Omega_i} f\, v\, dx, & \text{for } v \in H_D^1(\Omega). \end{cases}$$

By definition, the following subassembly relation will hold:

$$\begin{cases} \mathcal{A}(u,v) = \sum_{i=1}^p \mathcal{A}_{\Omega_i}(u,v), & \text{for } u, v \in H_D^1(\Omega) \\ \quad F(v) = \sum_{i=1}^p F_{\Omega_i}(v), & \text{for } v \in H_D^!(\Omega). \end{cases} \tag{3.8}$$

If $u, v \in V_h \cap H^1_D(\Omega)$, then these local forms can be represented using matrix-vector notation. Accordingly, on each subdomain Ω_j, let $I^{(j)}$ and $B^{(j)}$ denote the index sets of nodes in Ω_j and $\partial\Omega_j \backslash \mathcal{B}_D$, respectively, each with a specified local ordering of the nodes (for instance in ascending order of indices). Let $n_I^{(j)}$ and $n_B^{(j)}$ denote the number of nodes in Ω_j and $B^{(j)}$, respectively. Given finite element functions u_h, $v_h \in V_h \cap H^1_D(\Omega)$, let $\mathbf{u}$, $\mathbf{v} \in \mathbb{R}^n$ denote its vector of nodal values, with $\mathbf{u}_I^{(j)}$, $\mathbf{v}_I^{(j)} \in \mathbb{R}^{n_j}$ and $\mathbf{u}_B^{(j)}$, $\mathbf{v}_B^{(j)} \in \mathbb{R}^{n_B^{(j)}}$ denoting subvectors corresponding to indices in $I^{(j)}$ and $B^{(j)}$ (in the local ordering of nodes). We may then represent:

$$\mathcal{A}_{\Omega_j}(u_h, v_h) = \begin{bmatrix} \mathbf{u}_I^{(j)} \\ \mathbf{u}_B^{(j)} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(j)} & A_{IB}^{(j)} \\ A_{IB}^{(j)^T} & A_{BB}^{(j)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(j)} \\ \mathbf{v}_B^{(j)} \end{bmatrix}, \; F_{\Omega_j}(v_h) = \begin{bmatrix} \mathbf{f}_I^{(j)} \\ \mathbf{f}_B^{(j)} \end{bmatrix}^T \begin{bmatrix} \mathbf{v}_I^{(j)} \\ \mathbf{v}_B^{(j)} \end{bmatrix},$$

where the submatrices and subvectors are defined by:

$$\begin{cases} \left(A_{II}^{(j)}\right)_{lk} & \equiv \mathcal{A}_{\Omega_j}\left(\phi_{\tilde{l}}, \phi_{\tilde{k}}\right), \text{ for } 1 \le l,\, k \le n_I^{(j)} \\ \left(A_{IB}^{(j)}\right)_{lk} & \equiv \mathcal{A}_{\Omega_j}\left(\phi_{\tilde{l}}, \phi_{\tilde{k}}\right), \text{ for } 1 \le l \le n_I^{(j)},\, 1 \le k \le n_B^{(j)} \\ \left(A_{BB}^{(j)}\right)_{lk} & \equiv \mathcal{A}_{\Omega_j}\left(\phi_{\tilde{l}}, \phi_{\tilde{k}}\right), \text{ for } 1 \le l,\, k \le n_B^{(j)} \\ \left(\mathbf{f}_I^{(j)}\right)_l & = F_{\Omega_i}\left(\phi_{\tilde{l}}\right), \qquad \text{for } 1 \le l \le n_I^{(j)} \\ \left(\mathbf{f}_B^{(j)}\right)_l & = F_{\Omega_i}\left(\phi_{\tilde{l}}\right), \qquad \text{for } 1 \le l \le n_B^{(j)}, \end{cases}$$

with $\tilde{l}$ and $\tilde{k}$ denoting global indices corresponding to the local indices l and k on Ω_j and $B^{(j)}$. The discrete version of *subassembly* identity (3.8) becomes:

$$\begin{cases} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I \\ \mathbf{v}_B \end{bmatrix} = \sum_{j=1}^p \begin{bmatrix} \mathbf{u}_I^{(j)} \\ \mathbf{u}_B^{(j)} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(j)} & A_{IB}^{(j)} \\ A_{IB}^{(j)^T} & A_{BB}^{(j)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(j)} \\ \mathbf{v}_B^{(j)} \end{bmatrix} \\ \begin{bmatrix} \mathbf{v}_I \\ \mathbf{v}_B \end{bmatrix}^T \begin{bmatrix} \mathbf{f}_I \\ \mathbf{f}_B \end{bmatrix} = \sum_{j=1}^p \begin{bmatrix} \mathbf{v}_I^{(j)} \\ \mathbf{v}_B^{(j)} \end{bmatrix}^T \begin{bmatrix} \mathbf{f}_I^{(j)} \\ \mathbf{f}_B^{(j)} \end{bmatrix}. \end{cases} \tag{3.9}$$

These subassembly relations may equivalently be expressed based on restriction and extension matrices, as defined below.

Definition 3.2. *For any set of indices W (such as $I^{(j)}$, $B^{(j)}$, B) let index(W, l) denote the global index associated with the l'th node in the local ordering of indices in W. If n_W denotes the number of nodes in W, we define restriction map R_W as an $n_W \times n$ matrix with entries:*

$$(R_W)_{lj} = \begin{cases} 1, & \textit{if index}(W, l) = j \\ 0, & \textit{if index}(W, l) \ne j. \end{cases}$$

Given a nodal vector $\mathbf{v} \in \mathbb{R}^n$, its restriction $R_W \mathbf{v}$ will denote a subvector of nodal values corresponding to the indices in W in the chosen local ordering of nodes. While, given $\mathbf{v}_W \in \mathbb{R}^{nw}$, its extension $R_W^T \mathbf{v}_W$ will denote a vector of size n whose entries at indices in W correspond to those of $\mathbf{v}_W$ in the local ordering, with zero values for all other entries. The subassembly relations (3.9) may now be alternatively expressed as:

$$\begin{cases} \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} = \sum_{j=1}^p \begin{bmatrix} R_I^{(j)} \\ R_B^{(j)} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(j)} & A_{IB}^{(j)} \\ A_{IB}^{(j)^T} & A_{BB}^{(j)} \end{bmatrix} \begin{bmatrix} R_I^{(j)} \\ R_B^{(j)} \end{bmatrix} \\ \begin{bmatrix} \mathbf{f}_I \\ \mathbf{f}_B \end{bmatrix} = \sum_{j=1}^p \begin{bmatrix} R_I^{(j)} \\ R_B^{(j)} \end{bmatrix}^T \begin{bmatrix} \mathbf{f}_I^{(j)} \\ \mathbf{f}_B^{(j)} \end{bmatrix}. \end{cases} \tag{3.10}$$

This *subassembly identity* relates the global stiffness matrix and load vectors to the subdomain stiffness matrices and subdomain load vectors. The following result establishes a related expression between the global Schur complement matrix S and subdomain Schur complements $S^{(i)} \equiv A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}$.

Lemma 3.3. *Suppose the following assumptions hold.*

1. *Let* $\mathbf{u} = \left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T \in \mathbb{R}^n$ *be discrete* A_h*-harmonic, i.e., satisfy:*

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{f}_B \end{bmatrix}, \tag{3.11}$$

for some vector $\mathbf{f}_B$.

2. *Let* $\mathbf{u}_I^{(i)} = R_I^{(i)} \mathbf{u}$ *and* $\mathbf{u}_B^{(i)} = R_B^{(i)} \mathbf{u}$.

Then the following results will hold.

1. *The term* $\mathbf{f}_B = S\mathbf{u}_B$ *and the Schur complement energy will satisfy:*

$$\mathbf{u}_B^T S \mathbf{u}_B = \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix}. \tag{3.12}$$

2. *The subvectors* $\mathbf{u}_I^{(i)}$ *and* $\mathbf{u}_B^{(i)}$ *will satisfy:*

$$A_{II}^{(i)} \mathbf{u}_I^{(i)} + A_{IB}^{(i)} \mathbf{u}_B^{(i)} = \mathbf{0}. \tag{3.13}$$

3. *It will hold that:*

$$\mathbf{u}_B^T S \mathbf{u}_B = \sum_{i=1}^p \mathbf{u}_B^{(i)^T} S^{(i)} \mathbf{u}_B^{(i)}, \tag{3.14}$$

where $S^{(i)} \equiv (A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)})$.

Proof. To prove that $\mathbf{f}_B = S\mathbf{u}_B$ eliminate $\mathbf{u}_I$ using (3.11) and substitute the resulting expression $\mathbf{u}_I = -A_{II}^{-1} A_{IB}\mathbf{u}_B$ into the 2nd block equation to obtain the desired result. Next, take inner product of (3.11) with $\left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$ and substitute $\mathbf{f}_B = S\mathbf{u}_B$ to obtain (3.12).

To prove (3.13), we restrict the block equation:

$$A_{II}\mathbf{u}_I + A_{IB}\mathbf{u}_B = \mathbf{0}$$

to indices in $I^{(i)}$. Apply $R_I^{(i)}$ to (3.11), using that $A_{II} = \text{blockdiag}(A_{II}^{(1)}, \ldots, A_{II}^{(p)})$ and $\mathbf{u} = (\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T}, \mathbf{u}_B^T)^T$ to obtain:

$$A_{II}^{(i)}\mathbf{u}_I^{(i)} + R_I^{(i)} A_{IB}\mathbf{u}_B = \mathbf{0}. \tag{3.15}$$

Now, for standard finite element discretizations, the nodes in Ω_i will be coupled only to nodes in Ω_i and $B^{(i)}$. This yields:

$$R_I^{(i)} A_{IB}\mathbf{u}_B = A_{IB}^{(i)}\mathbf{u}_B^{(i)}.$$

Substituting this expression into (3.15) yields the desired result.

To prove (3.14), we apply (3.13) to the local nodal vector $(\mathbf{u}_I^{(i)^T}, \mathbf{u}_B^{(i)^T})^T$:

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(i)} \\ \mathbf{u}_B^{(i)} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{f}_B^{(i)} \end{bmatrix}, \tag{3.16}$$

for some vector $\mathbf{f}_B^{(i)}$. Formally eliminating $\mathbf{u}_I^{(i)} = -A_{II}^{(i)^{-1}} A_{IB}^{(i)}\mathbf{u}_B^{(i)}$ and substituting into the 2nd block equation above yields $\mathbf{f}_B^{(i)} = S^{(i)}\mathbf{u}_B^{(i)}$ where:

$$S^{(i)} \equiv (A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}). \tag{3.17}$$

We refer to $S^{(i)}$ as a local (subdomain) Schur complement. Taking the inner product of $(\mathbf{u}_I^{(i)^T}, \mathbf{u}_B^{(i)^T})^T$ with (3.16) and employing that $\mathbf{f}_B^{(i)} = S^{(i)}\mathbf{u}_B^{(i)}$ yields:

$$\begin{bmatrix} \mathbf{u}_I^{(i)} \\ \mathbf{u}_B^{(i)} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(i)} \\ \mathbf{u}_B^{(i)} \end{bmatrix} = \mathbf{u}_B^{(i)^T} S^{(i)}\mathbf{u}_B^{(i)}. \tag{3.18}$$

Substituting expressions (3.12) and (3.18) into identity (3.9) yields (3.14). □

The subassembly identity (3.14) may be expressed equivalently using restriction and extension maps between nodal vectors on B and $B^{(j)}$, as follows.

Definition 3.4. *Given region $G \subset B$ containing n_G indices, let index(B, G, i) denote the index of the i'th local of G in the ordering of indices on B. We define an $n_G \times n_B$ matrix $\mathcal{R}_G$ as:*

$$(\mathcal{R}_G)_{il} = \begin{cases} 1, & \textit{if index}(B, G, i) = l \\ 0, & \textit{if index}(B, G, i) \neq l. \end{cases} \tag{3.19}$$

It can easily be verified that $\mathcal{R}_G = R_G R_B^T$.

Using the restriction and extension maps $\mathcal{R}_B^{(i)}$ and $\mathcal{R}_B^{(i)^T}$, respectively, the Schur complement subassembly identity (3.14) can be stated as:

$$S = \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} S^{(i)} \mathcal{R}_B^{(i)} = \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} \left(A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)} \right) \mathcal{R}_B^{(i)}, \qquad (3.20)$$

where $S^{(i)} = (A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)})$ is a subdomain Schur complement. The traditional *substructuring* algorithm solves the Schur complement system by using Cholesky factorization, and explicitly assembles the subdomain finite element stiffness matrices, load vectors and Schur complement matrices using (3.9), (3.10) and (3.20). The resulting algorithm is summarized below.

Algorithm 3.2.2 *(Substructuring Algorithm)*

```
1. For $i = 1, \cdots, p$ in parallel do:
       $\begin{cases} \text{Assemble: } A_{II}^{(i)}, A_{IB}^{(i)}, A_{BB}^{(i)}, \mathbf{f}_I^{(i)}, \mathbf{f}_B^{(i)} \\ \text{Determine the Cholesky factors: } A_{II}^{(i)} = L_I^{(i)} L_I^{(i)^T} \\ \text{Assemble: } S^{(i)} \equiv A_{BB}^{(i)} - A_{IB}^{(i)^T} L_I^{(i)^{-T}} L_I^{(i)^{-1}} A_{IB}^{(i)} \\ \text{Assemble: } \tilde{\mathbf{f}}_B^{(i)} \equiv \mathbf{f}_B^{(i)} - A_{IB}^{(i)^T} L_I^{(i)^{-T}} L_I^{(i)^{-1}} \mathbf{f}_I^{(i)}. \end{cases}$
2. Endfor
3. Assemble:
       $\begin{cases} S \equiv \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} S^{(i)} \mathcal{R}_B^{(i)} \\ \tilde{\mathbf{f}}_B = \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} \tilde{\mathbf{f}}_B^{(i)} \end{cases}$
4. Determine the Cholesky factors: $S = L_S L_S^T$ and solve:
       $\begin{cases} L_S \, \mathbf{w}_B = \tilde{\mathbf{f}}_B \\ L_S^T \, \mathbf{u}_B = \mathbf{w}_B. \end{cases}$
5. For $i = 1, \cdots, p$ in parallel solve for $\mathbf{u}_I^{(i)}$:
       $A_{II}^{(i)} \mathbf{u}_I^{(i)} = (\mathbf{f}_I^{(i)} - A_{IB}^{(i)} \mathcal{R}_B^{(i)} \mathbf{u}_B).$
6. Endfor
Output: $\left( \mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T}, \mathbf{u}_B^T \right)^T$.
```

Steps 1 and 2 in the substructuring algorithm involve the assembly of $A_{II}^{(i)}$, $A_{IB}^{(i)}$, $A_{BB}^{(i)}$, $\mathbf{f}_I^{(i)}$ and $\mathbf{f}_B^{(i)}$ on each subdomain Ω_i, followed by the computation of the subdomain Cholesky factors, modified loads and and Schur complement matrices $S^{(i)}$. The computations on different subdomains can be performed in parallel. However, the substructuring algorithm is not purely algebraic, since it employs the subdomain stiffness matrices (as they may not be available if the linear system $A\mathbf{u} = \mathbf{f}$ has already been assembled). Assembly of the global Schur complement matrix S using identity (3.20), and of the forcing term $\tilde{f}_B$ in (3.6) must be parallelized using traditional methods. Similarly, the Cholesky factorization of S and the solution of the Schur complement system yielding $\mathbf{u}_B$, must be parallelized traditionally. Once $\mathbf{u}_B$ is determined, the components $\mathbf{u}_I^{(i)}$ of $\mathbf{u}_I$ can be determined in parallel (on each subdomain). From a computational viewpoint, assembly of matrix S and its Cholesky factorization can be significant costs, since n_B can be large.

Remark 3.5. When coefficient $c(x) = 0$ and $B^{(i)} = \partial\Omega_i$, the subdomain stiffness matrices will typically be *singular*, and satisfy:

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{1} = (1, \ldots, 1)^T$ is of appropriate size. In this case, matrix $S^{(i)}$ will also be *singular* with a null vector of the form $(1, \ldots, 1)^T$. However, the submatrices $A_{II}^{(i)}$ will be invertible.

Remark 3.6. For brevity of expression, we have employed matrix inverses in the expressions for $S^{(i)}$ and $\tilde{\mathbf{f}}_B^{(i)}$ in the substructuring algorithm. However, such inverses should not be assembled explicitly [GO4], instead the action of the inverse should be computed by the solution of the associated linear system. Each subdomain Schur complement matrix $S^{(i)}$ will be of size $n_B^{(i)}$ corresponding to the number of nodes on $B^{(i)}$. Explicit assembly of $S^{(i)}$ requires the solution of $n_B^{(i)}$ linear systems involving sparse coefficient matrix $A_{II}^{(i)}$. The subdomain Schur complement matrices $S^{(i)}$ will typically *not* be sparse, however, their entries may decay in magnitude with increasing distance between the nodes.

Remark 3.7. The global Schur complement matrix S will have a block matrix structure depending on the ordering of nodes in B. If nodes x_i and x_j lie on some common subdomain boundary $B^{(k)}$, then entry S_{ij} will typically be nonzero, otherwise, the entry S_{ij} will be zero. The magnitude of a nonzero entry S_{ij} typically decreases with increasing distance between the nodes x_i and x_j. Such properties are further explored when block matrix preconditioners are constructed for S.

From a computational viewpoint, the cost of the substructuring algorithm is dominated by the cost of assembling matrix S, and the subsequent cost of solving $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ using a *direct* solver. If instead, a preconditioned *iterative* method [GO4, AX, GR2, SA2] is employed to solve $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ without assembling S, then it may be possible to reduce these computational costs provided an effective preconditioner can be found. Such a reduction in the computational costs motivates the iterative substructuring method. Preconditioners for S are considered in Chap. 3.4 through Chap. 3.7. The *iterative substructuring* method has similar steps as Alg. 3.2.2. However, matrix S is not assembled in step 3 (instead, vector $\tilde{\mathbf{f}}_B$ is assembled) and step 4 is replaced by a preconditioned CG method to solve $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ with a preconditioner M. Steps 5 and 6 remain as in Alg. 3.2.2. We summarize the resulting algorithm.

Algorithm 3.2.3 *(Iterative Substructuring Algorithm)*

```
1. For $i = 1, \cdots, p$ in parallel do:
       Assemble: $A_{II}^{(i)}, A_{IB}^{(i)}, A_{BB}^{(i)}, \mathbf{f}_I^{(i)}, \mathbf{f}_B^{(i)}$
       Determine the Cholesky factors: $A_{II}^{(i)} = L_I^{(i)} L_I^{(i)^T}$
       Assemble: $S^{(i)} \equiv A_{BB}^{(i)} - A_{IB}^{(i)^T} L_I^{(i)^{-T}} L_I^{(i)^{-1}} A_{IB}^{(i)}$
       Assemble: $\tilde{\mathbf{f}}_B^{(i)} \equiv \mathbf{f}_B^{(i)} - A_{IB}^{(i)^T} L_I^{(i)^{-T}} L_I^{(i)^{-1}} \mathbf{f}_I^{(i)}$.
2. Endfor
3. Assemble:
       $\tilde{\mathbf{f}}_B = \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} \tilde{\mathbf{f}}_B^{(i)}$
4. Solve $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ using a preconditioned CG method.
5. For $i = 1, \cdots, p$ in parallel solve for $\mathbf{u}_I^{(i)}$:
       $A_{II}^{(i)} \mathbf{u}_I^{(i)} = \mathbf{f}_I^{(i)} - A_{IB}^{(i)} \mathcal{R}_B^{(i)} \mathbf{u}_B$.
6. Endfor
Output: $\left(\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T}, \mathbf{u}_B^T\right)^T$.
```

Remark 3.8. The cost of implementing a preconditioned iterative method to solve $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ using a preconditioner M in step 4 will be proportional to the number of preconditioned iterations and to the cost per iteration. When the number of preconditioned iterations is less than $\min_i \left(n_B^{(i)}\right)$, the cumulative cost for computing matrix-vector products with S will not exceed the cost of assembling the subdomain Schur complement matrices $S^{(i)}$. Furthermore, if the cost of solving $M\,\mathbf{w}_B = \mathbf{r}_B$ is modest, then the total cost of solving $S\,\mathbf{u}_B = \tilde{\mathbf{f}}_B$ iteratively without assembling S, may be less than the cost of assembling S and solving $S\,\mathbf{u} = \mathbf{f}$ using a direct method.

Remark 3.9. The iterative substructuring method is not purely algebraic, as it employs the subdomain stiffness matrices $A_{XY}^{(i)}$ for X, $Y = I$, B. When these submatrices are available, a product with S can be computed as:

$$S\,\mathbf{w}_B = \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} \left(A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}\right) \mathcal{R}_B^{(i)}\,\mathbf{w}_B. \qquad (3.21)$$

However, when these matrices are not available (for instance, when matrix A is already assembled), then such a product can be computed using:

$$S\mathbf{w}_B = A_{BB}\mathbf{w}_B - A_{IB}^T A_{II}^{-1} A_{IB}\mathbf{w}_B. \tag{3.22}$$

This requires computing $A_{IB}\mathbf{w}_B$ first, followed by solving $A_{II}\mathbf{w}_I = -A_{IB}\mathbf{w}_B$ (in parallel, since A_{II} is block diagonal, with p diagonal blocks), followed by defining $S\mathbf{w}_B \equiv A_{BB}\mathbf{w}_B + A_{IB}^T\mathbf{w}_I$.

Remark 3.10. The iterative substructuring and Schur complement algorithms have the disadvantage that they require the solution of subdomain problems of the form $A_{II}^{(i)}\mathbf{w}_B^{(i)} = \mathbf{r}_B^{(i)}$, close to machine precision (when computing the matrix-vector product with S). An alternative approach which avoids this is to solve the original linear system $A\mathbf{u} = \mathbf{f}$ by a preconditioned CG method with a block matrix preconditioner $\tilde{A}$ for A. Indeed, suppose $\tilde{A}_{II}$ and M are preconditioners for matrices A_{II} and S, respectively, and let $\tilde{A}_{IB}$ denote an approximation of A_{IB}, then motivated by the block form (3.25) (derived later in this section), a preconditioner $\tilde{A}$ for stiffness matrix A may be constructed:

$$\tilde{A}^{-1} = \begin{bmatrix} I & -\tilde{A}_{II}^{-1}\tilde{A}_{IB} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & M^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\tilde{A}_{IB}^T & I \end{bmatrix} \begin{bmatrix} \tilde{A}_{II}^{-1} & 0 \\ 0 & I \end{bmatrix}. \tag{3.23}$$

Matrix $\tilde{A}$ will be symmetric and positive definite, and when applying a CG method, each iteration will require the solution of a linear system of the form $\tilde{A}\mathbf{z} = \mathbf{r}$, which can be obtained by formally applying the expression $\mathbf{z} = \tilde{A}^{-1}\mathbf{r}$ given above. Such an approach will have the advantage that the subdomain problems need not be exact. However, care must be exercised in the choice of matrices $\tilde{A}_{II}$ and $\tilde{A}_{IB}$ approximating A_{II} and A_{IB}, respectively, and these approximations must be scaled appropriately. Indeed, it has been shown that if $\tilde{A}_{II} \equiv \alpha\, A_{II}$ for some $\alpha \neq 1$, then the convergence rate of of the conjugate gradient method deteriorates significantly [BO4]. This approach, however, requires two subdomain solves per iteration involving coefficient matrix $\tilde{A}_{II}$.

In the remainder of this chapter, after describing properties of matrix S and FFT based direct solvers for S, we shall focus on preconditioners M for S for use in the iterative substructuring or Schur complement algorithm. These preconditioners will be grouped as two subdomain or multisubdomain preconditioners. In the latter case, we shall separately consider two dimensional and three dimensional domains, as most preconditioners depend on the geometry of the interface B. A separate section is devoted to the robust class of Neumann-Neumann and balancing domain decomposition preconditioners.

3.2.1 Properties of the Schur Complement System

From a matrix viewpoint, the block elimination of $\mathbf{u}_I$, which results in the Schur complement system, can be understood to arise from the following block matrix factorization of A, as expressed next [CO6, MA11, GO4]:

$$\begin{aligned} A \equiv \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} &= \begin{bmatrix} I & 0 \\ A_{IB}^T A_{II}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{II} & A_{IB} \\ 0 & S \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ A_{IB}^T A_{II}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{II} & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A_{II}^{-1} A_{IB} \\ 0 & I \end{bmatrix}, \end{aligned} \tag{3.24}$$

where $S \equiv \left(A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}\right)$ denotes the Schur complement matrix. In this case, matrix A^{-1} will formally have the following block factorizations:

$$\begin{aligned} A^{-1} &= \begin{bmatrix} I & -A_{II}^{-1} A_{IB} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{II}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\,A_{IB}^T A_{II}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} I & -A_{II}^{-1} A_{IB} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\,A_{IB}^T & I \end{bmatrix} \begin{bmatrix} A_{II}^{-1} & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \tag{3.25}$$

To formally determine the solution of $A\mathbf{u} = \mathbf{f}$ using this block factorization of A^{-1} requires computing the action of A_{II}^{-1} twice, and S^{-1} once. However, if iterative methods are employed, then the Schur complement matrix S need not be assembled explicitly, but its action must be computed.

The following result provides bounds for the extreme eigenvalues of S when A is a symmetric and positive definite matrix. We employ the notation $\lambda_m(C)$ and $\lambda_M(C)$ to denote the minimum and maximum eigenvalues, respectively, of a real symmetric matrix C, and let $\kappa_2(C) \equiv \lambda_M(C)/\lambda_m(C)$ denote the spectral condition number of C. Given an arbitrary matrix D, we let $\sigma_1(D)$ denote its smallest singular value.

Lemma 3.11. *Suppose the following assumptions hold.*

1. *Let A be a symmetric positive definite matrix having the block structure:*
$$A = \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix}.$$
2. *Let $S = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$ denote the Schur complement matrix.*
3. *Define $E\mathbf{w}_B \equiv -A_{II}^{-1} A_{IB} \mathbf{w}_B$ for a vector $\mathbf{w}_B$.*

Then the following results will hold:

1. *S will be symmetric and positive definite, with*
$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} 0 \\ S\mathbf{u}_B \end{bmatrix}, \tag{3.26}$$
for arbitrary $\mathbf{u}_B$.
2. *The energy associated with matrix S will satisfy:*
$$\mathbf{u}_B^T S \mathbf{u}_B = \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}, \tag{3.27}$$
for arbitrary $\mathbf{u}_B$.

3. The minimum eigenvalue of S will satisfy:

$$\lambda_m(A)\left(\sigma_1(E)^2+1\right) \le \lambda_m(S).$$

4. The maximum eigenvalue of S will satisfy:

$$\lambda_M(S) \le \left(\lambda_M(A_{BB}) - \frac{\sigma_1(A_{IB})^2}{\lambda_M(A_{II})}\right).$$

5. The Schur complement matrix S will be better conditioned than matrix A in the spectral norm:

$$\kappa_2(S) \le \kappa_2(A).$$

Proof. When A is symmetric positive definite, its diagonal block A_{II} will also be symmetric and positive definite, so that A_{II}^{-1} is well defined. Consequently, matrix $S = A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}$ will be defined and symmetric by construction. Substituting the definition of $E\mathbf{u}_B$ and computing directly yields:

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} 0 \\ S\mathbf{u}_B \end{bmatrix}.$$

To show that S is positive definite, take inner product of the above equation with $\left((E\mathbf{u}_B)^T, \mathbf{u}_B^T\right)^T$ to obtain:

$$\begin{aligned} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} &= \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} 0 \\ S\mathbf{u}_B \end{bmatrix} \\ &= \mathbf{u}_B^T S \mathbf{u}_B. \end{aligned}$$

Since A is symmetric positive definite, we obtain that:

$$\begin{aligned} \mathbf{u}_B^T S \mathbf{u}_B &= \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} \\ &\ge \lambda_m(A) \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} \\ &\ge \lambda_m(A)\left((E\mathbf{u}_B)^T E\mathbf{u}_B + \mathbf{u}_B^T\mathbf{u}_B\right) \\ &\ge \lambda_m(A)\left(\sigma_1(E)^2+1\right)\mathbf{u}_B^T\mathbf{u}_B. \end{aligned}$$

In particular, since $\sigma_1(E) \ge 0$, we immediately obtain that:

$$\mathbf{u}_B^T S \mathbf{u}_B \ge \lambda_m(A)\mathbf{u}_B^T\mathbf{u}_B,$$

and so S will be positive definite, with its lowest eigenvalue at least as large as the lowest eigenvalue of A:

$$\lambda_m(A) \le \lambda_m(S).$$

Next, employing the definition of S, we obtain that

$$\begin{aligned} \mathbf{u}_B^T S \mathbf{u}_B &= \mathbf{u}_B^T \left(A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB} \right) \mathbf{u}_B \\ &\leq \mathbf{u}_B^T A_{BB} \mathbf{u}_B - \mathbf{u}_B^T A_{IB}^T A_{II}^{-1} A_{IB} \mathbf{u}_B \\ &\leq \left(\lambda_M(A_{BB}) - \sigma_1(A_{IB})^2 \lambda_m(A_{II}^{-1}) \right) \mathbf{u}_B^T \mathbf{u}_B \\ &= \left(\lambda_M(A_{BB}) - \tfrac{\sigma_1(A_{IB})^2}{\lambda_M(A_{II})} \right) \mathbf{u}_B^T \mathbf{u}_B . \end{aligned}$$

In particular, since the eigenvalues of the principal submatrix A_{BB} of A must lie between the maximum and minimum eigenvalues of A, and since $-\frac{\sigma_1(A_{IB})^2}{\lambda_M(A_{II})} \leq 0$, we obtain:

$$\lambda_M(S) \leq \lambda_M(A).$$

Combining the upper and lower bounds for the eigenvalues of S yields:

$$\kappa_2(S) = \frac{\lambda_M(S)}{\lambda_m(S)} \leq \frac{\lambda_M(A)}{\lambda_m(A)} = \kappa_2(A),$$

which is the desired result. □

Refinements of the preceding bounds may be found in [MA11]. The next result shows that if matrix A is an M-matrix, then the Schur complement S will also be an M-matrix. This will hold even if matrix A is *non-symmetric.*

Definition 3.12. *A nonsingular matrix K is said to be an M-matrix if:*

$$\begin{cases} (K)_{ii} & > 0, \quad \forall i \\ (K)\, ij & \leq 0, \quad i \neq j \\ \left(K^{-1}\right)_{ij} & \geq 0, \quad \forall i, j, \end{cases}$$

see [VA9, SA2]. Equivalently, K is an M-matrix if it can be expressed in the form $K = r\, I - N$ where $(N)_{ij} \geq 0$ for all i, j and either $(K^{-1})_{ij} \geq 0$ entrywise or if all minors of K are positive, see [BE17].

Lemma 3.13. *Suppose the following assumptions hold.*

1. *Let matrix A be non-symmetric and block partitioned as follows:*

$$A = \begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{bmatrix}.$$

2. *Let A be an M-matrix.*

Then, $S = (A_{BB} - A_{BI} A_{II}^{-1} A_{IB})$ will also be an M-matrix.

Proof. See [CR, NA]. First note that since A is an M-matrix, it will be of the form $A = r\,I - N$ where $N \geq 0$ entrywise. Thus, submatrix A_{II} will also be an M-matrix, since $A_{II} = r\,I - N_{II}$ for $N_{II} \geq 0$ entrywise and since the minors of A_{II} will be positive. As a result, $A_{II}^{-1} \geq 0$ entrywise. Furthermore, because $A_{BI} \leq 0$, $A_{IB} \leq 0$ and $A_{II}^{-1} \geq 0$ entrywise, it will hold that $(A_{BI}A_{II}^{-1}A_{IB}) \geq 0$ entrywise. Since $A_{BB} = r\,I - N_{BB}$ where $N_{BB} \geq 0$ entrywise, it will hold that $S = (A_{BB} - A_{BI}A_{II}^{-1}A_{IB})$ has the form $r\,I - G_{BB}$ for $G_{BB} = (N_{BB} + A_{BI}A_{II}^{-1}A_{IB}) \geq 0$ entrywise. Since:

$$S^{-1} = \begin{bmatrix} 0 \; I \end{bmatrix} A^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix},$$

we obtain that $S^{-1} \geq 0$ entrywise. Thus, S will be an M-matrix [BE17]. □

We shall now consider analytic properties of the Schur complement matrix S, inherited from the underlying elliptic partial differential equation (3.1) and its discretization. These properties will be employed to construct approximations of S which serve as preconditioners. We begin by identifying a Steklov-Poincaré operator $\mathcal{S}$ whose discrete analog yields matrix S.

Remark 3.14. Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of Ω and let u_B denote a sufficiently regular function defined on interface B with zero values on $\mathcal{B}_D$. Let $Lu \equiv -\nabla \cdot (a\nabla u) + c\,u$ denote the elliptic operator underlying (3.1). Using the continuous analog of (3.26), we *heuristically* define the action of a Steklov-Poincaré operator $\mathcal{S}$ on a function u_B defined on interface B as follows:

$$\mathcal{S}u_B(x) \equiv Lw_B(x), \qquad \text{for } x \in B,$$

where $w_B = \mathcal{E}\,u_B$ denotes the piecewise L-harmonic extension of u_B on B:

$$\begin{cases} Lw_B = 0, & \text{in } \Omega_i \\ \;\; w_B = u_B, & \text{on } \partial\Omega_i, \end{cases} \quad \text{for } 1 \leq i \leq p.$$

Heuristically by analogy with (3.27), the energy associated with the Steklov-Poincaré operator $\mathcal{S}$ will satisfy:

$$(\mathcal{S}u_B, u_B)_{L^2(B)} = \mathcal{A}\,(\mathcal{E}u_B, \mathcal{E}u_B), \tag{3.28}$$

where $\mathcal{A}(.,.)$ is defined by (3.2).

We next describe bounds for the eigenvalues of S in terms of the mesh size h and coefficients in the elliptic equation. Such estimates employ properties of elliptic equation (3.1), trace theorems, fractional Sobolev norms, discrete extension theorems and also inverse inequalities for finite element spaces, see [DR2, BR12, BR15, DR14, DR10, MA17]. We shall employ the notation:

$$\begin{cases} |u|_{1,\Omega_i}^2 & \equiv \int_{\Omega_i} |\nabla u|^2 \, dx \\ \|u\|_{1,\Omega_i}^2 & \equiv \int_{\Omega_i} |\nabla u|^2 \, dx + \int_{\Omega_i} |u|^2 \, dx \\ |u|_{1/2,\partial\Omega_i}^2 & \equiv \int_{\partial\Omega_i} \int_{\partial\Omega_i} \frac{|u(x)-u(y)|^2}{|x-y|^d} \, dx \, dy, \qquad\qquad \Omega_i \subset \mathbb{R}^d \\ \|u\|_{1/2,\partial\Omega_i}^2 & \equiv \int_{\partial\Omega_i} \int_{\partial\Omega_i} \frac{|u(x)-u(y)|^2}{|x-y|^d} \, dx \, dy + \int_{\partial\Omega_i} |u|^2 \, dx. \end{cases}$$

The following result will not be optimal with respect to coefficient variation or the diameter h_0 of the subdomains.

Lemma 3.15. *Suppose the following assumptions hold with $\mathcal{B}_D = \partial\Omega$.*

1. *Let the coefficients $a(x)$ and $c(x)$ satisfy:*

$$\begin{cases} 0 < a_m \le a(x) \le a_M \\ 0 < c_m \le c(x) \le c_M. \end{cases}$$

Define $\sigma_m = \min\{c_m, a_m\}$ and $\sigma_M = \max\{c_M, a_M\}$.

2. *Let u_h denote a finite element function corresponding to a nodal vector $\mathbf{u} = \left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$ where $\mathbf{u}_I$ satisfies $\mathbf{u}_I \equiv E\mathbf{u}_B = -A_{II}^{-1} A_{IB} \mathbf{u}_B$.*
3. *Let the following inverse inequality hold for all $v_h \in V_h$*

$$\|v_h\|_{1/2,\partial\Omega_i} \le C h^{-1/2} \|v_h\|_{0,\partial\Omega_i}, \tag{3.29}$$

for $1 \le i \le p$ where C does not depend on h.

Then the following results will hold.

1. *The finite element function u_h will be piecewise discrete L-harmonic:*

$$\mathcal{A}(u_h, v) = 0, \quad \forall v \in V_h \cap H_0^1(\Omega_i), \;\; 1 \le i \le p, \tag{3.30}$$

with its energy equivalent to the Schur complement energy, as in (3.28):

$$\mathbf{u}_B^T S \mathbf{u}_B = \mathcal{A}(u_h, u_h). \tag{3.31}$$

2. *There exists $c > 0$ and $C > 0$ independent of h, σ_m and σ_M, but possibly dependent on the subdomain diameter h_0, such that:*

$$c\,\sigma_m \left(\sum_{i=1}^p \|u_h\|_{1/2,\partial\Omega_i}^2 \right) \le \mathcal{A}(u_h, u_h) \le C\,\sigma_M \left(\sum_{i=1}^p \|u_h\|_{1/2,\partial\Omega_i}^2 \right). \tag{3.32}$$

3. *There exists $c > 0$ and $C > 0$ independent of h, σ_m and σ_M, but possibly dependent on the subdomain diameter h_0, such that:*

$$c\,\sigma_m \left(\sum_{i=1}^p \|u_h\|_{0,\partial\Omega_i}^2 \right) \le \mathcal{A}(u_h, u_h) \le C\,\sigma_M \left(\sum_{i=1}^p \|u_h\|_{0,\partial\Omega_i}^2 \right) h^{-1}. \tag{3.33}$$

Proof. Applying the inner product of $\left(\mathbf{v}_B^T, (E\mathbf{v}_B)^T\right)^T$ with (3.26) yields:

$$\begin{bmatrix} \mathbf{v}_I \\ \mathbf{v}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{v}_I \\ \mathbf{v}_B \end{bmatrix}^T \begin{bmatrix} 0 \\ S\mathbf{u}_B \end{bmatrix}, \quad \forall \mathbf{v}_I, \mathbf{v}_B.$$

If $\mathbf{v}_B = \mathbf{0}$, then the right hand side will be zero. If v_h denotes the finite element function corresponding to the nodal vector $\mathbf{v} = (\mathbf{v}_I^T, \mathbf{0})^T$, we may equivalently express the preceding as:

$$\mathcal{A}(u_h, v_h) = 0, \quad \forall v_h \in V_h \cap H_0^1(\Omega_i), \quad \text{for } 1 \leq i \leq p,$$

since v_h will be zero on B. This verifies that u_h is discrete L-harmonic on each Ω_i. By choosing $\mathbf{v}_B = \mathbf{u}_B$ and $\mathbf{v}_I = E\mathbf{u}_B$, we obtain $\mathcal{A}(u_h, u_h) = \mathbf{u}_B^T S \mathbf{u}_B$.

To derive bounds for the energy $\mathcal{A}(u_h, u_h)$ associated with u_h, we employ the equivalence between the energy norm and the Sobolev norm:

$$\sigma_m \|u_h\|_{1,\Omega}^2 \leq \mathcal{A}\left(u_h, u_h\right) \leq \sigma_M \|u_h\|_{1,\Omega}^2.$$

We then decompose the Sobolev norm based on the subdomains to obtain:

$$\sigma_m \sum_{i=1}^{p} \|u_h\|_{1,\Omega_i}^2 \leq \mathcal{A}(u_h, u_h) \leq \sigma_M \sum_{i=1}^{p} \|u_h\|_{1,\Omega_i}^2. \tag{3.34}$$

Application of the trace theorem on each Ω_i yields the lower bound:

$$c \, \|u_h\|_{1/2,\partial\Omega_i}^2 \leq \|u_h\|_{1,\Omega_i}^2, \quad 1 \leq i \leq p,$$

where $c > 0$ is independent of h and the coefficients, but may depend on h_0.

To obtain an upper bound, we employ a discrete extension theorem (see Chap. 3.9) and *a prior* estimates for discrete harmonic functions to obtain:

$$\|u_h\|_{1,\Omega_i}^2 \leq C \, \|u_h\|_{1/2,\partial\Omega_i}^2, \quad \text{for } 1 \leq i \leq p,$$

for $C > 0$ independent of h and the coefficients, but possibly dependent on h_0. Substituting the above upper and lower bounds into (3.34) yields (3.32).

Combining the trivial bound $c \, \|u_h\|_{0,\partial\Omega_i}^2 \leq \|u_h\|_{1/2,\partial\Omega_i}^2$ with inverse inequality (3.29) yields:

$$c \, \|u_h\|_{0,\partial\Omega_i}^2 \leq \|u_h\|_{1/2,\partial\Omega_i}^2 \leq C \, h^{-1} \, \|u_h\|_{0,\partial\Omega_i}^2.$$

Combining the preceding bound with (3.32) yields (3.33). □

Remark 3.16. If u_h is a finite element function corresponding to the nodal vector $\mathbf{u} = \left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$, then known properties of the mass matrix [ST14, CI2] imply that $\|u_h\|_{0,\partial\Omega_i}^2$ will be equivalent, up to a scaling factor, to the Euclidean norm of the nodal vector $\mathbf{u}$ restricted to $\partial\Omega_i$. Substituting this in (3.33) yields:

$$c\,\sigma_m \le \frac{\mathbf{v}_B^T S \mathbf{v}_B}{\mathbf{v}_B^T \mathbf{v}_B} \le C\,\sigma_M h^{-1}.$$

Thus, the condition number $\kappa_2(S)$ will grow as $C\ (\sigma_M/\sigma_m)\,h^{-1}$ with decreasing mesh size h, for fixed subdomains. A refinement of this estimate yields:

$$\kappa_2(S) \le C\ (\sigma_M/\sigma_m)\,h_0^{-1}h^{-1},$$

where h_0 denotes the subdomain diameter [BR24]. These bounds compare favorably with the condition number bound of $C\ (\sigma_M/\sigma_m)\,h^{-2}$ for $\kappa_2(A)$.

Remark 3.17. If $v = 0$ on $\mathcal{B}_D$ and $\partial\Omega_i \cap \mathcal{B}_D \neq \emptyset$, then the following norm equivalence can be employed [LI4]:

$$c\,\|v\|^2_{H^{1/2}_{00}(B^{(i)})} \le \|v\|^2_{H^{1/2}(\partial\Omega_i)} \le C\,\|v\|^2_{H^{1/2}_{00}(B^{(i)})}.$$

Discrete approximations of the fractional Sobolev norm $\|v_h\|^2_{H^{1/2}_{00}(B^{(i)})}$ will be considered later in this chapter for finite element functions.

Remark 3.18. When $c(x) = 0$ in (3.1) and $a(x)$ is piecewise constant:

$$a(x) = \rho_i, \quad x \in \Omega_i, \text{for } 1 \le i \le p,$$

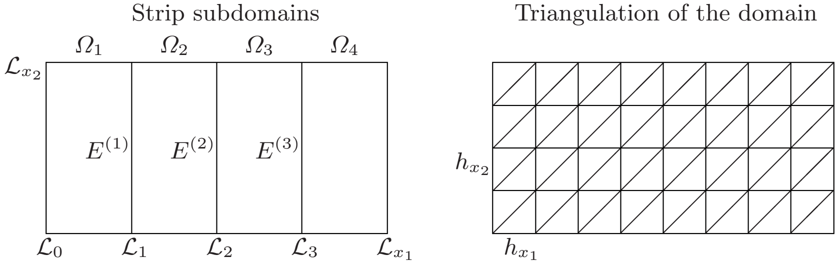
then the following equivalence will hold, see Chap. 3.9, for any finite element function $u_h \in V_h$ satisfying (3.30):

$$c\left(\sum_{i=1}^{p} \rho_i\,|u_h|^2_{1/2,\partial\Omega_i}\right) \le \mathcal{A}(u_h,u_h) \le C\left(\sum_{i=1}^{p} \rho_i\,|u_h|^2_{1/2,\partial\Omega_i}\right), \tag{3.35}$$

with $0 < c < C$ independent of h and $a(x)$. Here, seminorms replace the norms since some of the local Dirichlet energies $\mathcal{A}_{\Omega_i}(u_h,u_h)$ can become zero even when $u_h(x) \neq 0$ (for instance, when u_h is constant locally).

3.3 FFT Based Direct Solvers

For a discretization of a *separable* elliptic equation, it may be possible to construct fast Fourier transform (FFT) based direct solvers for the stiffness matrix A and the Schur complement matrix S. For such solvers to be applicable, the stiffness matrix A must have a *block* matrix structure in which each block is *simultaneously diagonalized* by a discrete Fourier transform matrix Q, see [BJ9, CH13, CH14, RE, VA4]. When this property holds, the stiffness matrix A can be transformed, using an orthogonal similarity transformation, into a block matrix with diagonal submatrices. After appropriately reordering the unknowns, this transformed system will be *block diagonal*, with band matrices along its diagonal, and it can be solved in parallel using band solvers.


Strip subdomains

Triangulation of the domain

Fig. 3.2. Strip decomposition with four subdomains

In this section, we outline the construction of such fast direct solvers, for matrix A and its Schur complement S. In the special case of a two subdomain rectangular decomposition, with a uniform grid and constant coefficients, this will yield an explicit eigendecomposition of the Schur complement S. The FFT based algorithm to solve $A\mathbf{u} = \mathbf{f}$ is summarized in Alg. 3.3.1, and algorithm to solve $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ is summarized in Alg. 3.3.2.

We shall consider the following *separable* elliptic equation posed on a two dimensional rectangular domain $\Omega = (0, \mathcal{L}_{x_1}) \times (0, \mathcal{L}_{x_2})$ for $x = (x_1, x_2)$:

$$\begin{cases} -\frac{\partial}{\partial x_1}\left(a_1(x)\frac{\partial u}{\partial x_1}\right) - \frac{\partial}{\partial x_2}\left(a_2(x)\frac{\partial u}{\partial x_2}\right) = f(x), & \text{for } x \in \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{for } x \in \partial\Omega. \end{cases} \tag{3.36}$$

Triangulate Ω using a uniform grid with $(l-1) \times (k-1)$ interior grid points having mesh spacings $h_{x_1} \equiv (\mathcal{L}_{x_1}/l)$ and $h_{x_2} \equiv (\mathcal{L}_{x_2}/k)$ as in Fig. 3.2. The grid points (ih_{x_1}, jh_{x_2}) for indices $1 \le i \le (l-1)$ and $1 \le j \le (k-1)$ will lie in the interior, and the nodal values of a finite element function u_h at these grid points will be denoted $u_{i,j} = u_h(ih_{x_1}, jh_{x_2})$. We consider a nonoverlapping decomposition $\Omega_1, \ldots, \Omega_p$ of Ω consisting of the *strip* subdomains:

$$\Omega_i \equiv (\mathcal{L}_{i-1}, \mathcal{L}_i) \times (0, \mathcal{L}_{x_2}), \qquad \text{for } 1 \le i \le p,$$

where $\mathcal{L}_0 \equiv 0 < \mathcal{L}_1 < \cdots < \mathcal{L}_p \equiv \mathcal{L}_{x_1}$.

The subdomain boundary segments $E^{(r)} \equiv \partial\Omega_r \cap \partial\Omega_{r+1}$ for $1 \le j \le (p-1)$ will be assumed to align with the triangulation, so that there are integers L_r such that $\mathcal{L}_r = L_r h_{x_1}$, for $0 \le r \le p$. The coefficients $a_1(x)$ and $a_2(x)$ in the elliptic equation will be assumed to be constant within each subdomain Ω_i:

$$\begin{cases} a_1(x) = a_1^{(i)}, & \text{for } x \in \Omega_i \\ a_2(x) = a_2^{(i)}, & \text{for } x \in \Omega_i \end{cases} \qquad \text{for } 1 \le i \le p.$$

For this choice of coefficients and triangulation, the stiffness matrix A resulting from the finite element discretization of (3.36) will have the following stencil at a gridpoint (ih_{x_1}, jh_{x_2}). We formally denote it as:

$$
(A\mathbf{u})_{i,j} = \begin{cases} \quad a_1^{(r)} \frac{h_{x_2}}{h_{x_1}} \left(2u_{i,j} - u_{i-1,j} - u_{i+1,j}\right) & \text{if } i \neq L_r, \\ + \, a_2^{(r)} \frac{h_{x_1}}{h_{x_2}} \left(2u_{i,j} - u_{i,j-1} - u_{i,j+1}\right) & \\ \quad a_1^{(r)} \frac{h_{x_2}}{h_{x_1}} \left(u_{i,j} - u_{i-1,j}\right) & \text{if } i = L_r \\ + \, a_1^{(r+1)} \frac{h_{x_2}}{h_{x_1}} \left(u_{i,j} - u_{i+1,j}\right) & \\ + \, \frac{(a_2^{(r)} + a_2^{(r+1)}) h_{x_1}}{2 h_{x_2}} \left(2u_{i,j} - u_{i,j-1} - u_{i,j+1}\right). & \end{cases} \tag{3.37}
$$

To represent (3.37) as a linear system, define subvectors $\mathbf{u}_i \equiv (u_{i,1}, \cdots, u_{i,k-1})^T$ for $1 \leq i \leq l-1$ and employ them to define a nodal vector $\mathbf{u} \equiv (\mathbf{u}_1, \cdots, \mathbf{u}_{l-1})^T$. For this ordering of nodes, the linear system $A\mathbf{u} = \mathbf{f}$ representing (3.37) is:

$$
\begin{bmatrix} T^{(1)} & -\beta^{(1)} & & & \\ -\beta^{(1)} & T^{(2)} & -\beta^{(2)} & & \\ & \ddots & \ddots & \ddots & \\ & & -\beta^{(l-3)} & T^{(l-2)} & -\beta^{(l-2)} \\ & & & -\beta^{(l-2)} & T^{(l-1)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_{l-2} \\ \mathbf{u}_{l-1} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_{l-2} \\ \mathbf{f}_{l-1} \end{bmatrix}, \tag{3.38}
$$

where $T^{(r)}$ and $\beta^{(r)}$ are submatrices of size $(k-1)$ defined for $L_{i-1} < r < L_i$:

$$
T^{(r)} \equiv \frac{a_2^{(i)} h_{x_1}}{h_{x_2}} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} + \frac{2 a_1^{(i)} h_{x_2}}{h_{x_1}} \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix},
$$

while

$$
T^{(r)} \equiv \frac{1}{2} \left(T^{(L_i - 1)} + T^{(L_i + 1)} \right), \qquad \text{for } r = L_i.
$$

Each matrix $T^{(r)}$ is symmetric, tridiagonal and Toeplitz (it is constant along each diagonal) of size $(k-1)$.

The submatrices $\beta^{(r)}$ are multiples of the identity:

$$
\beta^{(r)} \equiv \begin{cases} a_1^{(i)} \frac{h_{x_2}}{h_{x_1}} I, & \text{if } L_{i-1} < r < L_i \\ \frac{1}{2} \left(a_1^{(i+1)} + a_1^{(i)} \right) \frac{h_{x_2}}{h_{x_1}} I, & \text{if } r = L_i, \end{cases} \tag{3.39}
$$

where I denotes an identity matrix of size $(k-1)$. An important property of matrix A in (3.38) is that its submatrices $T^{(r)}$ and $\beta^{(r)}$ are diagonalized by a discrete sine transform matrix Q, as defined next.

Definition 3.19. *Given an integer $k > 2$, we define the entries of a discrete sine transform matrix Q of size $(k-1)$ as follows:*

$$Q_{ij} \equiv \sqrt{\frac{2}{k}} \sin\left(\frac{ij\pi}{k}\right), \quad \text{for } 1 \le i,j \le (k-1). \tag{3.40}$$

For $1 \le j \le (k-1)$, we let $\mathbf{q}_j$ denote the j'th column of matrix Q

$$\mathbf{q}_j = \sqrt{\frac{2}{k}} \left(\sin\left(\frac{j\pi}{k}\right), \sin\left(\frac{2j\pi}{k}\right), \cdots, \sin\left(\frac{(k-1)j\pi}{k}\right)\right)^T .$$

By construction, matrix Q is symmetric. Using trigonometric identities, it can be verified that $Q^T Q = I$ so that Q is an orthogonal matrix. Routines for fast multiplication of a vector by Q are available in most FFT packages with complexity proportional to $O(k \log(k))$, see [VA4]. To verify that each block of A is diagonalized by Q, we apply matrix $T^{(r)}$ to the j'th column vector $\mathbf{q}_j$ of Q. By direct substitution and the use of trigonometric identities it is easily verified that $\mathbf{q}_j$ is an eigenvector of matrix $T^{(r)}$:

$$T^{(r)} \mathbf{q}_j = \lambda_j^{(r)} \mathbf{q}_j,$$

corresponding to the eigenvalue $\lambda_j^{(r)}$ given by:

$$\lambda_j^{(r)} \equiv \begin{cases} \frac{2a_2^{(i)} h_{x_1}}{h_{x_2}} \left(1 - \cos\left(\frac{j\pi}{k}\right)\right) + \frac{2a_1^{(i)} h_{x_2}}{h_{x_1}}, & \text{if } L_{i-1} < r < L_i, \\ \frac{\left(a_2^{(i)} + a_2^{(i+1)}\right) h_{x_1}}{h_{x_2}} \left(1 - \cos\left(\frac{j\pi}{k}\right)\right) + \frac{\left(a_1^{(i)} + a_1^{(i+1)}\right) h_{x_2}}{h_{x_1}}, & \text{if } r = L_i, \end{cases} \tag{3.41}$$

Thus, $T^{(r)}$ has the eigendecomposition:

$$T^{(r)} = Q \Lambda^{(r)} Q^T,$$

where $\Lambda^{(r)} = \text{diag}(\lambda_1^{(r)}, \cdots, \lambda_{k-1}^{(r)})$. Since the matrices $\beta^{(r)}$ are scalar multiples of the identity, they are also trivially diagonalized by Q.

The following algebraic result shows how any block partitioned system $C\mathbf{w} = \mathbf{g}$ can be reduced to a block diagonal linear system provided all blocks of matrix C can be simultaneously diagonalized by an orthogonal matrix Q.

Lemma 3.20. *Suppose the following assumptions hold.*

1. *Let C be an invertible matrix of size $m\,n$ having an $n \times n$ block structure in which the individual blocks C_{ij} are submatrices of size m for $1 \le i,j \le n$.*
2. *Let Q be an orthogonal matrix of size m which simultaneously diagonalizes all the block submatrices of C:*

$$Q^T C_{ij} Q = D_{ij}, \qquad \text{for } 1 \le i,j \le n,$$

where each D_{ij} is a diagonal matrix of size m.

3. Let $\mathbf{w} = \left(\mathbf{w}_1^T, \ldots, \mathbf{w}_n^T\right)^T$ *with* $\mathbf{w}_i \in \mathbb{R}^m$ *denote the solution to the block partitioned linear system:*

$$\begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix}, \tag{3.42}$$

where $\mathbf{g} = \left(\mathbf{g}_1^T, \ldots, \mathbf{g}_n^T\right)^T$ *with* $\mathbf{g}_i \in \mathbb{R}^m$.

Then, the solution to system (3.42) can be obtained by solving the following block diagonal linear system:

$$\begin{bmatrix} G_{11} & & 0 \\ & \ddots & \\ 0 & & G_{mm} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \tag{3.43}$$

where G_{ii}*,* $\boldsymbol{\alpha}_i$ *and* $\boldsymbol{\mu}_i$ *are defined by:*

1. *For* $1 \le i \le m$ *matrix* G_{ii} *is of size* n *with entries defined by:*

$$(G_{ii})_{lk} \equiv (D_{lk})_{ii}\,, \quad \text{for } 1 \le i \le m,\ 1 \le l, k \le n.$$

2. *For* $1 \le i \le m$ *subvector* $\boldsymbol{\alpha}_i$ *of size* n *is defined by:*

$$(\boldsymbol{\alpha}_i)_k = \left(Q^T \mathbf{w}_k\right)_i, \quad \text{for } 1 \le i \le m,\ 1 \le k \le n.$$

3. *For* $1 \le i \le m$ *subvector* $\boldsymbol{\mu}_i$ *of size* n *is defined by:*

$$(\boldsymbol{\mu}_i)_k = \left(Q^T \mathbf{g}_k\right)_i, \quad \text{for } 1 \le i \le m,\ 1 \le k \le n.$$

Proof. Define a block diagonal matrix $\mathcal{Q} \equiv \text{blockdiag}\ (Q, \ldots, Q)$ having n diagonal blocks, using the given orthogonal matrix Q of size m. By construction $\mathcal{Q}$ will also be an orthogonal matrix. Apply $\mathcal{Q}$ to transform the linear system $C\mathbf{w} = \mathbf{g}$ into $\left(\mathcal{Q}^T C \mathcal{Q}\right)\left(\mathcal{Q}^T \mathbf{w}\right) = \left(\mathcal{Q}^T \mathbf{g}\right)$:

$$\begin{bmatrix} Q^T C_{11} Q & \cdots & Q^T C_{1n} Q \\ \vdots & & \vdots \\ Q^T C_{n1} Q & \cdots & Q^T C_{nn} Q \end{bmatrix} \begin{bmatrix} Q^T \mathbf{w}_1 \\ \vdots \\ Q^T \mathbf{w}_n \end{bmatrix} = \begin{bmatrix} Q^T \mathbf{g}_1 \\ \vdots \\ Q^T \mathbf{g}_n \end{bmatrix}. \tag{3.44}$$

Define $D \equiv \mathcal{Q}^T C \mathcal{Q}$ and let $\tilde{\mathbf{w}} \equiv \mathcal{Q}^T \mathbf{w}$ and $\tilde{\mathbf{g}} \equiv \mathcal{Q}^T \mathbf{g}$ denote the transformed vectors. Then, the transformed linear system becomes $D\tilde{\mathbf{w}} = \tilde{\mathbf{g}}$. By construction, each block submatrix $D_{ij} = Q^T C_{ij} Q$ of D will be a diagonal matrix of size m. As a consequence, components of $\tilde{\mathbf{w}}$ will be *coupled* within the transformed linear system $D\tilde{\mathbf{w}} = \tilde{\mathbf{g}}$ only when its indices differ by an integer multiple of m. Thus, a suitable reordering of the indices within the transformed system should yield a block diagonal linear system.

Accordingly, we *partition* the index set $\{1, 2, \ldots, nm\}$ into subsets such that two indices belong to the same subset only if they differ by an integer multiple of m. The resulting partition will be:

$$\{1, \ldots, n\, m\} = \{1, 1+m, \ldots, (n-1)m+1\} \cup \cdots \cup \{m, 2m, \ldots, nm\}.$$

There will be m subsets in this partition, each containing n entries ordered in ascending order. Let $\mathcal{P}^T$ denote a permutation matrix whose action on a vector reorders its entries according to the above ordering. We reorder the components of $\tilde{\mathbf{w}}$ and define $\boldsymbol{\alpha} \equiv \mathcal{P}^T \mathcal{Q}^T \mathbf{w}$. Similarly, we define $\boldsymbol{\mu} \equiv \mathcal{P}^T \mathcal{Q}^T \mathbf{g}$ as a reordering of $\tilde{\mathbf{g}}$. By construction, reordering the rows and columns of matrix D should yield $G = \text{blockdiag}(G_{11}, \ldots, G_{mm}) = \mathcal{P}^T D \mathcal{P}$ to be a block diagonal matrix. The reordered transformed system $\left(\mathcal{P}^T D \mathcal{P}\right)\left(\mathcal{P}^T \tilde{\mathbf{w}}\right) = \left(\mathcal{P}^T \tilde{\mathbf{g}}\right)$ will then correspond to the system (3.43).

Once the subproblems $G_{ii}\boldsymbol{\alpha}_i = \boldsymbol{\mu}_i$ in (3.43) have been solved in parallel, define $\mathbf{y}_k$ for $1 \leq k \leq n$ as follows:

$$(\mathbf{y}_k)_i = (\boldsymbol{\alpha}_i)_k\,, \qquad \text{for } 1 \leq i \leq m,\ 1 \leq k \leq n,$$

The original unknowns $\mathbf{w}_k$ will satisfy $\mathbf{w}_k = Q\mathbf{y}_k$ for $1 \leq k \leq n$. □

Remark 3.21. It can be easily verified that the block submatrices G_{ii} in the preceding will inherit the "block sparsity pattern" of C. For example, if C is block tridiagonal, then each submatrix G_{ii} will be a tridiagonal matrix.

FFT Based Solution of $A\mathbf{u} = \mathbf{f}$. A fast direct solver can be constructed for solving (3.38) using Lemma 3.20 and choosing $C = A$, $n = (l-1)$ with $m = (k-1)$, $\mathbf{w} = \mathbf{u}$ and $\mathbf{g} = \mathbf{f}$. Let Q denote the discrete sine transform matrix defined by (3.40). In this case, each nonzero block in $\mathcal{Q}^T A \mathcal{Q}$ will satisfy:

$$\begin{cases} Q^T T^{(r)} Q = \Lambda^{(r)}, \\ Q^T \beta^{(r)} Q = \beta^{(r)}. \end{cases}$$

We define $\mathbf{c}_j = Q^T \mathbf{u}_j$ and $\tilde{\mathbf{f}}_j = Q^T \mathbf{f}_j$ for $j = 1, \cdots, l-1$, where:

$$\mathbf{c}_j = \left(c_{1,j}, \cdots, c_{k-1,j}\right)^T, \quad \text{for } 1 \leq j \leq (l-1).$$

Since A is block tridiagonal, system $\left(\mathcal{Q}^T A \mathcal{Q}\right)\left(\mathcal{Q}^T \mathbf{u}\right) = \left(\mathcal{Q}^T \mathbf{f}\right)$ will also be block tridiagonal. Furthermore, each G_{ii} in (3.43) will be a tridiagonal matrix.

Once all the unknowns c_{ij} have been determined by *parallel* solution of the tridiagonal linear systems, the nodal values $\{u_{ij}\}$ at the grid points can be reconstructed by applying Q columnwise

$$\left(u_{1,j}, \cdots, u_{k-1,j}\right)^T = Q\left(c_{1,j}, \cdots, c_{k-1,j}\right)^T, \qquad \text{for } j = 1, \cdots, l-1.$$

Since a tridiagonal system can be solved in optimal order complexity, and since multiplication by $\mathcal{Q}$ has $O\left(l\, k\, \log(k)\right)$ complexity, the complexity of the FFT based solution algorithm will be $O\left(l\, k\, \log(k)\right)$.

Algorithm 3.3.1 *(FFT Based Solution of $A\mathbf{u} = \mathbf{f}$)*
Let $\lambda_j^{(r)}$ and $\beta^{(r)}$ be defined by (3.41) and (3.39)

```
1. For $j = 1, \cdots, l-1$ in parallel do:
2.     Compute the fast sine transform:
           $$\tilde{\mathbf{f}}_j \equiv Q\mathbf{f}_j.$$
3. Endfor
4. For $i = 1, \cdots, k-1$ in parallel do:
5.     Solve the tridiagonal system using Cholesky factorization:
           $$\begin{bmatrix} \lambda_i^{(1)} & -\beta_i^{(1)} & & & & \\ -\beta_i^{(1)} & \lambda_i^{(2)} & -\beta_i^{(2)} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\beta_i^{(l-3)} & \lambda_i^{(l-2)} & -\beta_i^{(l-2)} \\ & & & & -\beta_i^{(l-2)} & \lambda_i^{(l-1)} \end{bmatrix} \begin{bmatrix} c_{i,1} \\ c_{i,2} \\ \vdots \\ \vdots \\ c_{i,l-2} \\ c_{i,l-1} \end{bmatrix} = \begin{bmatrix} (\tilde{\mathbf{f}}_1)_i \\ (\tilde{\mathbf{f}}_2)_i \\ \vdots \\ \vdots \\ (\tilde{\mathbf{f}}_{l-2})_i \\ (\tilde{\mathbf{f}}_{l-1})_i \end{bmatrix}.$$
6. Endfor
7. For $j = 1, \cdots, l-1$ in parallel do:
8.     Compute using the fast sine transform
           $$\mathbf{u}_j \equiv Q \begin{bmatrix} c_{1,j} \\ \vdots \\ c_{k-1,j} \end{bmatrix}.$$
9. Endfor
```

Output: $\left(\mathbf{u}_1^T, \ldots, \mathbf{u}_{l-1}^T\right)^T$.

FFT based solution of $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$. Lemma 3.20 can also be applied to construct a direct solver for the Schur complement system, provided the block submatrices of S are simultaneously diagonalized by an orthogonal matrix.

Accordingly, in the following we study the block structure of the Schur complement matrix S. Given a finite element function u_h with nodal values $u_{ij} = u_h(ih_{x_1}, jh_{x_2})$ for $1 \le i \le (l-1)$ and $1 \le j \le (k-1)$, we will employ the following notation for index sets and nodal vectors associated with them.

$$\begin{aligned} I^{(r)} &\equiv \left\{(ih_{x_1}, jh_{x_2}) \, : \, L_{r-1} < i < L_r, \ 1 \le j \le (k-1)\right\}, \text{ for } 1 \le r \le p \\ I &\equiv I^{(1)} \cup \cdots \cup I^{(p)} \\ E^{(r)} &\equiv \{(L_r \, h_{x_1}, jh_{x_2}) \, : \, 1 \le j \le (k-1)\}, \qquad \text{for } 1 \le r \le (p-1) \\ B &\equiv E^{(1)} \cup \cdots \cup E^{(p-1)}. \end{aligned}$$

For convenience, we have used $E^{(r)}$ to denote interface $E^{(r)} = \partial\Omega_r \cap \partial\Omega_{r+1}$ as well as the set of indices of nodes on it. We will employ nodal subvectors $\mathbf{u}_i \equiv (u_{i,1}, \cdots, u_{i,k-1})^T$ for $1 \le i \le (l-1)$. The following additional nodal subvectors will be associated with each of the preceding index sets:

$$\begin{cases} \mathbf{u}_I^{(r)} \equiv \left(\mathbf{u}_{L_{r-1}+1}^T, \cdots, \mathbf{u}_{L_r-1}^T\right)^T, \text{ for } 1 \le r \le p \\ \mathbf{u}_I \;\;\equiv \left(u_I^{(1)^T}, \cdots, \mathbf{u}_I^{(p)^T}\right)^T \\ \mathbf{u}_E^{(r)} \equiv \mathbf{u}_{L_r}, \qquad\qquad\qquad \text{for } 1 \le r \le (p-1) \\ \mathbf{u}_B \;\equiv \left(\mathbf{u}_E^{(1)^T}, \cdots, \mathbf{u}_E^{(p-1)^T}\right)^T. \end{cases}$$

The stiffness matrix A will be block partitioned into the submatrices A_{II}, A_{IB} and A_{BB} based on the preceding index sets. Matrix A_{II} takes the form:

$$A_{II} = \begin{bmatrix} A_{II}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{II}^{(p)} \end{bmatrix}, \quad A_{II}^{(r)} = \begin{bmatrix} M^{(r)} & -\gamma^{(r)} & & & \\ -\gamma^{(r)} & M^{(r)} & -\gamma^{(r)} & & \\ & \ddots & \ddots & \ddots & \\ & & -\gamma^{(r)} & M^{(r)} & -\gamma^{(r)} \\ & & & -\gamma^{(r)} & M^{(r)} \end{bmatrix}. \quad (3.45)$$

Here $A_{II}^{(r)}$ is a block tridiagonal and block Toeplitz matrix with $d_r \times d_r$ blocks of size $(k-1)$, where $d_r \equiv (L_r - L_{r-1} - 1)$. The submatrix $\gamma^{(r)} \equiv a_1^{(r)} \frac{h_{x_2}}{h_{x_1}} I$ is of size $(k-1)$, while $M^{(r)}$ of size $(k-1)$ satisfies:

$$M^{(r)} \equiv \frac{a_2^{(r)} h_{x_1}}{h_{x_2}} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} + \frac{2a_1^{(r)} h_{x_2}}{h_{x_1}} \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}. \quad (3.46)$$

Matrix A_{IB} will be block bidiagonal with $p \times (p-1)$ blocks $X_{ij} = A_{I^{(i)}E^{(j)}}$:

$$A_{IB} = \begin{bmatrix} X_{11} & & & & & 0 \\ X_{21} & X_{22} & & & & \\ & X_{32} & X_{33} & & & \\ & & \ddots & \ddots & & \\ & & & & X_{(p-1)(p-2)} & X_{(p-1)(p-1)} \\ 0 & & & & & X_{p(p-1)} \end{bmatrix} \quad (3.47)$$

where for $2 \leq r \leq p$ and $1 \leq s \leq (p-1)$ its block submatrices are defined by:

$$X_{rr} = A_{I^{(r)}E^{(r)}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\gamma^{(r)} \end{bmatrix} \quad \text{and} \quad E_{s(s-1)} = A_{I^{(s)}E^{(s-1)}} = \begin{bmatrix} -\gamma^{(s)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.48}$$

with blocks of size $(k-1)$. Matrix A_{BB} will be a $(p-1) \times (p-1)$ block diagonal matrix whose individual blocks are each of size $(k-1)$

$$A_{BB} = \begin{bmatrix} A_{EE}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{EE}^{(p-1)} \end{bmatrix} \quad \text{where} \quad A_{EE}^{(r)} \equiv \frac{1}{2}\left(M^{(r)} + M^{(r+1)}\right). \tag{3.49}$$

Each submatrix $M^{(r)}$ is diagonalized by the sine transform matrix Q defined earlier, with $M^{(r)} = Q\Lambda^{(r)}Q^T$ and $\Lambda^{(r)} = \operatorname{diag}\left(\lambda_1^{(r)}, \ldots, \lambda_{k-1}^{(r)}\right)$, where:

$$\lambda_j^{(r)} = 2\frac{a_2^{(r)} h_{x_1}}{h_{x_2}}\left(1 - \cos(\frac{j\pi}{k})\right) + \frac{2a_1^{(r)} h_{x_2}}{h_{x_1}}, \quad \text{for } 1 \leq j \leq (k-1). \tag{3.50}$$

Since matrix $\gamma^{(r)}$ is a scalar multiple of the identity, it is trivially diagonalized by Q with eigenvalues $\left(\gamma^{(r)}\right)_j = a_1^{(r)} \frac{h_{x_2}}{h_{x_1}}$ for $1 \leq j \leq (k-1)$.

We next consider the block structure of matrix S given the ordering of nodes on B. If we substitute the block partitioned matrices (3.49), (3.47) and (3.45) for A_{BB}, A_{IB} and A_{II}, respectively, in $S = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$, then since matrices A_{BB}, A_{IB} and A_{II} are block diagonal, rectangular block bidiagonal and block diagonal, respectively, it will follow that matrix S must be block tridiagonal. Explicit expressions for the block submatrices $S_{E^{(r)}E^{(r)}}$ and $S_{E^{(r+1)}E^{(r)}}$ can be obtained by directly computing the block entries of $S = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$ using (3.49), (3.47) and (3.45). The resulting structure is summarized in the following.

Lemma 3.22. *Given the ordering of nodes on B based on the index sets $E^{(1)}, \ldots, E^{(p-1)}$ the following will hold.*

1. *The Schur complement matrix S will be block tridiagonal of the form:*

$$S = \begin{bmatrix} S_{E^{(1)}E^{(1)}} & S_{E^{(1)}E^{(2)}} & & & 0 \\ S_{E^{(1)}E^{(2)}}^T & S_{E^{(2)}E^{(2)}} & S_{E^{(2)}E^{(3)}} & & \\ & \ddots & \ddots & \ddots & \\ & & S_{E^{(p-3)}E^{(p-2)}}^T & S_{E^{(p-2)}E^{(p-2)}} & S_{E^{(p-2)}E^{(p-1)}} \\ 0 & & & S_{E^{(p-2)}E^{(p-1)}}^T & S_{E^{(p-1)}E^{(p-1)}} \end{bmatrix} \tag{3.51}$$

with block submatrices $S_{E^{(i)}E^{(j)}}$ of size $(k-1)$.

2. *For $1 \leq r \leq (p-1)$ the block submatrices $S_{E_r E_r}$ will satisfy:*

$$\begin{cases} S_{E^{(r)}E^{(r)}} = A_{E^{(r)}E^{(r)}} - A^T_{I^{(r)}E^{(r)}} A^{-1}_{I^{(r)}I^{(r)}} A_{I^{(r)}E^{(r)}} \\ \qquad\qquad\qquad - A^T_{I^{(r+1)}E^{(r)}} A^{-1}_{I^{(r+1)}I^{(r+1)}} A_{I^{(r+1)}E^{(r)}}. \end{cases}$$

3. *For $1 \leq r \leq (p-2)$ the block submatrices $S_{E^{(r+1)}E^{(r)}}$ will satisfy:*

$$S_{E^{(r+1)}E^{(r)}} = -A^T_{I^{(r+1)}E^{(r+1)}} A^{-1}_{I^{(r+1)}I^{(r+1)}} A_{I^{(r+1)}E^{(r)}}.$$

Proof. As outlined earlier. □

Since A_{II}, A_{IB} and A_{BB} can be partitioned into blocks of size $(k-1)$, each of which are diagonalizable by the discrete sine transform matrix Q, the block submatrices of $S = (A_{BB} - A^T_{IB} A^{-1}_{II} A_{IB})$ will also be diagonalizable by matrix Q. The following two results will be employed to show this, and to obtain analytical expressions for the eigenvalues of its blocks.

Lemma 3.23. *Suppose the following assumptions hold.*

1. *Let C denote a positive definite symmetric matrix of size $m\,n$ partitioned into $n \times n$ blocks C_{ij} of size m for $1 \leq i, j \leq n$.*
2. *Let Q be an orthogonal matrix of size m which simultaneously diagonalizes all the block submatrices of C:*

$$Q^T C_{ij} Q = D_{ij}, \qquad \text{for } 1 \leq i, j \leq n,$$

   where each D_{ij} is a diagonal matrix.
3. *Let $\left(\mathbf{w}_1^T, \ldots, \mathbf{w}_n^T\right)^T$ denote the solution to the block partitioned system:*

$$\begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix},$$

   where $\mathbf{w}_i \in \mathbb{R}^m$ and $\mathbf{g}_i \in \mathbb{R}^m$ for $1 \leq i \leq n$.
4. *Let $\mathbf{g}_i = \delta_i\,\mathbf{q}_t$, for scalars $\delta_i \in \mathbb{R}$ where $\mathbf{q}_t \equiv (q_{1t}, \ldots, q_{mt})^T$ denotes the t'th column of Q.*

Then, each $\mathbf{w}_i = \alpha_i\,\mathbf{q}_t$ will be a scalar multiple of $\mathbf{q}_t$ for some $\alpha_i \in \mathbb{R}$. Furthermore, the scalars $\alpha_1, \ldots, \alpha_n$ will solve the following linear system:

$$\begin{bmatrix} (D_{11})_{tt} & \cdots & (D_{1n})_{tt} \\ \vdots & & \vdots \\ (D_{n1})_{tt} & \cdots & (D_{nn})_{tt} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix}.$$

Proof. This result can be obtained by an application of Lemma 3.20. Alternatively, substitute the *ansatz* $\mathbf{w}_i = \alpha_i\,\mathbf{q}_t$ to obtain the linear system:

$$\begin{cases} C_{11}\mathbf{q}_t\,\alpha_1 + \cdots + C_{1n}\mathbf{q}_t\,\alpha_n &= \mathbf{q}_t\delta_1 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots & \vdots \\ C_{n1}\mathbf{q}_t\,\alpha_1 + \cdots + C_{nn}\mathbf{q}_t\,\alpha_n &= \mathbf{q}_t\delta_n. \end{cases}$$

Since $\mathbf{q}_t$ is an eigenvector of each matrix C_{ij} corresponding to eigenvalue $(D_{ij})_{tt}$, elimination of the common factors $\mathbf{q}_t$ yields the linear system:

$$\begin{cases} (D_{11})_{tt}\,\alpha_1 + \cdots + (D_{1n})_{tt}\,\alpha_n &= \delta_1 \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots & \vdots \\ (D_{n1})_{tt}\,\alpha_1 + \cdots + (D_{nn})_{tt}\,\alpha_n &= \delta_n. \end{cases} \tag{3.52}$$

By construction, since $(\mathbf{q}_t^T\mathbf{q}_t) = 1$, it will hold that:

$$\begin{bmatrix} \alpha_1\,\mathbf{q}_t \\ \vdots \\ \alpha_n\,\mathbf{q}_t \end{bmatrix}^T \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix} \begin{bmatrix} \alpha_1\,\mathbf{q}_t \\ \vdots \\ \alpha_n\,\mathbf{q}_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}^T \begin{bmatrix} (D_{11})_{tt} & \cdots & (D_{1n})_{tt} \\ \vdots & & \vdots \\ (D_{n1})_{tt} & \cdots & (D_{nn})_{tt} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}.$$

When C is symmetric and positive definite, both terms in the above expression will be positive for $(\alpha_1, \ldots, \alpha_n)^T \neq \mathbf{0}$, verifying that (3.52) is nonsingular. □

The next result describes the solution of a Toeplitz tridiagonal system.

Lemma 3.24. *Consider the following Toeplitz tridiagonal linear system:*

$$\begin{bmatrix} \tilde{b} & \tilde{a} & & & 0 \\ \tilde{c} & \tilde{b} & \tilde{a} & & \\ & \ddots & \ddots & \ddots & \\ & & \tilde{c} & \tilde{b} & \tilde{a} \\ & & & \tilde{c} & \tilde{b} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \vdots \\ \vdots \\ \alpha_d \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \vdots \\ \vdots \\ \mu_d \end{bmatrix},$$

where $\tilde{a}$, $\tilde{b}$, $\tilde{c} \in \mathbb{R}$ satisfies $(\tilde{b}^2 - 4\tilde{a}\,\tilde{c}) > 0$. Define ρ_1, $\rho_2 \in \mathbb{R}$ as follows:

$$\rho_1 \equiv \frac{-\tilde{b} + \sqrt{\tilde{b}^2 - 4\,\tilde{a}\,\tilde{c}}}{2\,\tilde{a}} \quad \textit{and} \quad \rho_2 \equiv \frac{-\tilde{b} - \sqrt{\tilde{b}^2 - 4\,\tilde{a}\,\tilde{c}}}{2\,\tilde{a}}. \tag{3.53}$$

Then, the following will hold:

1. *If $(\mu_1, \ldots, \mu_d)^T = (-\tilde{c}, 0, \ldots, 0)^T$, then:*

$$\alpha_i = \left(\frac{\rho_2^{d+1}\,\rho_1^i - \rho_1^{d+1}\,\rho_2^i}{\rho_2^{d+1} - \rho_1^{d+1}} \right), \qquad \textit{for} \quad 1 \le i \le d. \tag{3.54}$$

2. *If* $(\mu_1, \ldots, \mu_d)^T = (0, 0, \ldots, -\tilde{a})^T$, *then:*

$$\alpha_i = \left(\frac{\rho_1^i - \rho_2^i}{\rho_1^{d+1} - \rho_2^{d+1}} \right), \quad \text{for} \quad 1 \le i \le d. \tag{3.55}$$

Proof. Substitute the *ansatz* that $\alpha_i = \rho^i$ for $0 \le i \le (d+1)$ into the finite difference equations. This yields the following equations:

$$\left(\tilde{a}\, \rho^2 + \tilde{b}\, \rho + \tilde{c} \right) \rho^{i-1} = 0, \quad \text{for } 1 \le i \le d.$$

It can be solved simultaneously, provided ρ solves the characteristic equation:

$$\tilde{a}\, \rho^2 + \tilde{b}\, \rho + \tilde{c} = 0.$$

The roots of the characteristic polynomial are given by (3.53) and they will be real and distinct provided $(\tilde{b}^2 - 4\, \tilde{a}\, \tilde{c}) > 0$. The general discrete solution to the finite difference equations will be of the form:

$$\alpha_i = \gamma_1\, \rho_1^i + \gamma_2\, \rho_2^i, \quad \text{for each } i,$$

for arbitrary γ_1 and γ_2. To solve the first linear system, we impose the boundary condition $\alpha_0 = 1$ and $\alpha_{d+1} = 0$. Solving for γ_1 and γ_2 yields (3.54). To solve the second linear system, we impose the boundary condition $\alpha_0 = 0$ and $\alpha_{d+1} = 1$. Solving for γ_1 and γ_2 yields (3.55). □

The next result shows that each submatrix $S_{E^{(r)}E^{(s)}}$ of the Schur complement matrix S is diagonalized by the discrete sine transform Q of size $(k-1)$. Furthermore, by employing Lemma 3.23 and 3.24, we can obtain analytical expressions for the eigenvalues of $S_{E^{(r)}E^{(s)}}$.

Lemma 3.25. *Let* $\lambda_t^{(r)}$, $\gamma^{(r)}$, $\omega(r,t)$, $\rho_1(r,t)$ *and* $\rho_2(r,t)$ *be as defined below:*

$$\begin{cases} \gamma^{(r)} & \equiv a_1^{(r)} (h_{x_2}/h_{x_1}) \\ \lambda_t^{(r)} & \equiv 2a_2^{(r)} (h_{x_1}/h_{x_2}) \left(1 - \cos(\frac{t\pi}{k})\right) + 2\gamma^{(r)} \\ \omega(r,t) & \equiv \frac{a_2^{(r)} h_{x_1}^2}{a_1^{(r)} h_{x_2}^2} \left(1 - \cos(\frac{t\pi}{k})\right) + 1 \\ \rho_1(r,t) & \equiv \omega(r,t) + \sqrt{\omega(r,t)^2 - 1} \\ \rho_2(r,t) & \equiv \omega(r,t) - \sqrt{\omega(r,t)^2 - 1}. \end{cases} \tag{3.56}$$

In addition, define $d_r = (L_r - L_{r-1} - 1)$.
Then, the following results will hold.

1. *For* $1 \le r \le (p-1)$ *the vector* $\mathbf{q}_t$ *will be an eigenvector of matrix* $S_{E^{(r)}E^{(r)}}$ *corresponding to eigenvalue* $(D_{rr})_{tt}$*:*

$$S_{E^{(r)}E^{(r)}}\, \mathbf{q}_t = (D_{rr})_{tt}\, \mathbf{q}_t,$$

where $(D_{rr})_{tt}$ *is given by:*

$$\begin{cases} (D_{rr})_{tt} = -\gamma^{(r)} \left(\frac{\rho_1(r,t)^{d_r} - \rho_2(r,t)^{d_r}}{\rho_1(r,t)^{d_r+1} - \rho_2(r,t)^{d_r+1}} \right) + \frac{1}{2} \left(\lambda_t^{(r)} + \lambda_t^{(r+1)} \right) \\ \qquad\qquad -\gamma^{(r+1)} \left(\frac{\rho_1(r+1,t)^{d_{r+1}} - \rho_2(r+1,t)^{d_{r+1}}}{\rho_1(r+1,t)^{d_{r+1}+1} - \rho_2(r+1,t)^{d_{r+1}+1}} \right). \end{cases} \tag{3.57}$$

Matrix $S_{E^{(r)}E^{(r)}}$ *will be diagonalized by the discrete sine transform* Q*:*

$$Q^T S_{E^{(r)}E^{(r)}} Q = D_{rr}, \tag{3.58}$$

where D_{rr} *is a diagonal matrix of size* $(k-1)$*.*

2. *For* $1 \le r \le (p-2)$ *the vector* $\mathbf{q}_t$ *will be an eigenvector of the matrix* $S_{E^{(r)}E^{(r+1)}}$ *corresponding to the eigenvalue* $(D_{r,r+1})_{tt}$*:*

$$S_{E^{(r)}E^{(r+1)}} \mathbf{q}_t = (D_{r,r+1})_{tt}\, \mathbf{q}_t,$$

where $(D_{r,r+1})_{tt}$ *is given by:*

$$(D_{r,r+1})_{tt} = -\gamma^{(r+1)} \left(\frac{\rho_1(r+1,t) - \rho_2(r+1,t)}{\rho_1(r+1,t)^{d_{(r+1)}+1} - \rho_2(r+1,t)^{d_{(r+1)}+1}} \right). \tag{3.59}$$

Matrix $S_{E^{(r)}E^{(r+1)}}$ *will be diagonalized by the discrete sine transform* Q*:*

$$Q^T S_{E^{(r)}E^{(r+1)}} Q = D_{r,r+1},$$

where $D_{r,r+1}$ *is a diagonal matrix of size* $(k-1)$*.*

Proof. To verify that $\mathbf{q}_t$ is an eigenvector of $S_{E^{(r)}E^{(r)}}$, we shall employ the following expression for $S_{E^{(r)}E^{(r)}}\, \mathbf{q}_t$:

$$\begin{cases} S_{E^{(r)}E^{(r)}} \mathbf{q}_t = -A^T_{I^{(r)}E^{(r)}} A^{-1}_{I^{(r)}I^{(r)}} A_{I^{(r)}E^{(r)}} \mathbf{q}_t + A_{E^{(r)}E^{(r)}} \mathbf{q}_t \\ \qquad\qquad\qquad -A^T_{I^{(r+1)}E^{(r)}} A^{-1}_{I^{(r+1)}I^{(r+1)}} A_{I^{(r+1)}E^{(r)}} \mathbf{q}_t. \end{cases}$$

Each of the submatrices in the above can be block partitioned into blocks that are diagonalized by Q. By Lemma 3.23 it will follow that $\mathbf{q}_t$ is an eigenvector of each of the three matrix terms above. We will determine the eigenvalue associated with each term separately. Let θ_1 denote the eigenvalue of $-\left(A^T_{I^{(r)}E^{(r)}} A^{-1}_{I^{(r)}I^{(r)}} A_{I^{(r)}E^{(r)}} \right)$ associated with eigenvector $\mathbf{q}_t$:

$$-A^T_{I^{(r)}E^{(r)}} A^{-1}_{I^{(r)}I^{(r)}} A_{I^{(r)}E^{(r)}} \mathbf{q}_t = \theta_1\, \mathbf{q}_t.$$

An application of Lemma 3.23 will yield the following expression for θ_1:

$$\theta_1 = -\gamma^{(r)} \begin{bmatrix} \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \\ 1 \end{bmatrix}^T \begin{bmatrix} \lambda_t^{(r)} & -\gamma^{(r)} & & & \\ -\gamma^{(r)} & \lambda_t^{(r)} & -\gamma^{(r)} & & \\ & \ddots & \ddots & \ddots & \\ & & -\gamma^{(r)} & \lambda_t^{(r)} & -\gamma^{(r)} \\ & & & -\gamma^{(r)} & \lambda_t^{(r)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \\ \gamma^{(r)} \end{bmatrix}.$$

The right hand side above can be evaluated as $-\gamma^{(r)}\,\alpha_{d_r}$ in Lemma 3.24 for the choice $\tilde{a} = \tilde{c} = -\gamma^{(r)}$, $\tilde{b} = \lambda_t^{(r)}$ and $d = d_r = (L_r - L_{r-1} - 1)$. This yields:

$$\theta_1 = -\gamma^{(r)} \left(\frac{\rho_1(r,t)^{d_r} - \rho_2(r,t)^{d_r}}{\rho_1(r,t)^{d_r+1} - \rho_2(r,t)^{d_r+1}} \right).$$

The eigenvalue θ_2 of $A_{E^{(r)}E^{(r)}}$ corresponding to eigenvector $\mathbf{q}_t$ was derived earlier in this subsection as:

$$\theta_2 = \frac{1}{2}\left(\lambda_t^{(r)} + \lambda_t^{(r+1)}\right).$$

The eigenvalue θ_3 of $-\left(A^T_{I^{(r+1)}E^{(r)}} A^{-1}_{I^{(r+1)}I^{(r+1)}} A_{I^{(r+1)}E^{(r)}}\right)$ corresponding to eigenvector $\mathbf{q}_t$ can be determined as for θ_1 using Lemma 3.23 and 3.24. It results in the expression:

$$\theta_3 = -\gamma^{(r+1)} \left(\frac{\rho_1(r+1,t)^{d_{r+1}} - \rho_2(r+1,t)^{d_{r+1}}}{\rho_1(r+1,t)^{d_{r+1}+1} - \rho_2(r+1,t)^{d_{r+1}+1}} \right).$$

Combining the three terms yields an expression for the eigenvalue $(D_{rr})_{tt}$ of $S_{E^{(r)}E^{(r)}}$ corresponding to eigenvector $\mathbf{q}_t$:

$$(D_{rr})_{tt} = \theta_1 + \theta_2 + \theta_3,$$

which verifies (3.57). By construction, Q diagonalizes $S_{E^{(r)}E^{(r)}} = QD_{rr}Q^T$.

To obtain an expression for the eigenvalue $(D_{r+1,r})_{tt}$ of $S_{E^{(r+1)}E^{(r)}}$ we evaluate $-\left(A^T_{I^{(r+1)}E^{(r+1)}} A^{-1}_{I^{(r+1)}I^{(r+1)}} A_{I^{(r+1)}E^{(r)}}\right)\mathbf{q}_t$ at the eigenvector $\mathbf{q}_t$, using Lemma 3.23 and 3.24. This yields:

$$(D_{r+1,r})_{tt} = -\gamma^{(r+1)} \left(\frac{\rho_1(r+1,t) - \rho_2(r+1,t)}{\rho_1(r+1,t)^{d_{(r+1)}+1} - \rho_2(r+1,t)^{d_{(r+1)}+1}} \right).$$

By construction, matrix Q will diagonalize $S_{E^{(r+1)}E^{(r)}} = QD_{r+1,r}Q^T$. □

The preceding result shows that the block submatrices of matrix S are simultaneously diagonalized by the discrete sine transform Q. Thus, we may employ Lemma 3.20 to construct a fast direct solver for S.

We summarize the algorithm next, using matrices G_{ii} of size $(p-1)$:

$$(G_{ii})_{r,s} = (D_{r,s})_{ii} \quad \text{for } 1 \le r,s \le (p-1), \;\; 1 \le i \le (k-1),$$

where $(D_{r,s})_{ii}$ is defined by (3.57) or (3.59). Matrix G_{ii} will be tridiagonal.

Algorithm 3.3.2 *(FFT Based Solution of $S\mathbf{u}_B = \mathbf{f}_B$)*

```
Let $\mathbf{u}_B = \left(\mathbf{u}_{E^{(1)}}^T, \ldots, \mathbf{u}_{E^{(p-1)}}^T\right)^T$ and $\mathbf{f}_B = \left(\mathbf{f}_{E^{(1)}}^T, \ldots, \mathbf{f}_{E^{(p-1)}}^T\right)^T$
 1. For $i = 1, \ldots, p-1$ in parallel do:
          Compute   $\tilde{\mathbf{f}}_{E^{(i)}} \equiv Q^T \mathbf{f}_{E^{(i)}}$
 2. Endfor
 3. For $i = 1, \ldots, p-1$ do
 4.     For $j = 1, \ldots, k-1$ do
          Define   $(\mathbf{g}_j)_i \equiv \left(\tilde{\mathbf{f}}_{E^{(i)}}\right)_j$
 5.     Endfor
 6. Endfor
 7. For $j = 1, \ldots, k-1$ in parallel solve (using a tridiagonal solver):
          $G_{jj}\mathbf{c}_j = \mathbf{g}_j$
 8. Endfor
 9. For $i = 1, \ldots, p-1$ do:
10.     For $j = 1, \ldots, k-1$ do:
          Define   $\left(\tilde{\mathbf{c}}_{E^{(i)}}\right)_j = (\mathbf{c}_j)_i$
11.     Endfor
12. Endfor
13. For $i = 1, \ldots, p-1$ do:
          Compute   $\mathbf{u}_{E^{(i)}} = Q\tilde{\mathbf{c}}_{E^{(i)}}$.
14. Endfor
Output: $\mathbf{u}_B = \left(\mathbf{u}_{E^{(1)}}^T, \ldots, \mathbf{u}_{E^{(p-1)}}^T\right)^T$.
```

Remark 3.26. The loop between lines 1 and 2 requires the application of a total of $(p-1)$ fast sine transforms. The loop between lines 7 and 8 requires the solution of a total of $(k-1)$ tridiagonal linear systems, each involving $(p-1)$ unknowns. The loop between lines 13 and 14 requires the application of a total of $(p-1)$ fast sine transforms. As a result, the preceding algorithm will have a complexity of $O\left(p\, k\, \log(k)\right)$.

Remark 3.27. In the case of a *two* strip decomposition, the Schur complement matrix $S = S_{E^{(1)}E^{(1)}}$ will be diagonalized by the discrete sine transform Q:

$$S = QD_{11}Q^T.$$

Such eigendecompositions can be employed to precondition a two subdomain Schur complement matrix arising in two dimensional elliptic problems and will be considered in the next section [BJ9, CH13, CH14, RE].

Remark 3.28. In this section, we have focused solely on FFT based Schur complement solvers for discretizations of elliptic equations on *two* dimensional domains. However, the block matrix techniques that were described can also be applied to discretizations of separable elliptic equations on *three* dimensional rectangular domains with strip subdomains. In such cases, the stiffness matrix A and the Schur complement matrix S will have block tridiagonal structure, provided, the nodal vectors $\mathbf{u}_j$ correspond to nodal unknowns on planar cross sections of Ω. Matrix Q will then be a two dimensional FFT or FST matrix, and the algebraic expressions derived in this section for eigenvalues of the Schur complement blocks will remain valid provided $\lambda_j^{(r)}$ and $\gamma^{(r)}$ correspond to eigenvalues of block matrices $M^{(r)}$ and $\gamma^{(r)}$, respectively, in the three dimensional case. We omit additional details.

3.4 Two Subdomain Preconditioners

Our study of preconditioners for the Schur complement S begins with the *two* subdomain case, where the geometry of the interface B is relatively simple. In this case, S will be *dense*, however, its entries will decay in magnitude with increasing distance between the nodes. We shall describe preconditioners based either on local Schur complement matrices or on approximations of S which use properties of the Steklov-Poincaré map associated with S.

We consider a finite element discretization of elliptic equation (3.1) on a domain Ω, with Dirichlet boundary conditions on $\mathcal{B}_D = \partial\Omega$. We assume that Ω is partitioned into *two* nonoverlapping subdomains Ω_1 and Ω_2 with interface $B \equiv \partial\Omega_1 \cap \partial\Omega_2$, see Fig. 3.3, and order the nodes in Ω based on Ω_1, Ω_2 and B. Given this ordering, a nodal vector $\mathbf{u}$ can be partitioned as $\mathbf{u} = \left(\mathbf{u}_I^{(1)^T}, \mathbf{u}_I^{(2)^T}, \mathbf{u}_B^T\right)^T$, and the discretization of (3.1) will be (see Chap. 3.1):

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)} \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ A_{IB}^{(1)^T} & A_{IB}^{(2)^T} & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B \end{bmatrix}.$$

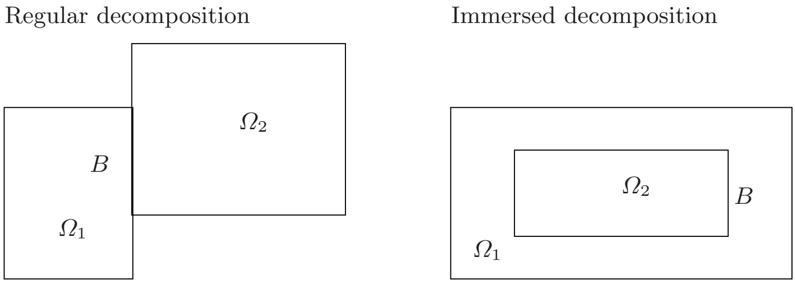


Fig. 3.3. Two subdomain decompositions

The Schur complement matrix S associated with the above system can be derived by solving $\mathbf{u}_I^{(i)} = A_{II}^{(i)^{-1}}(\mathbf{f}_I^{(i)} - A_{IB}^{(i)}\mathbf{u}_B)$ for $i = 1, 2$ and substituting this into the third block row above. This will yield the reduced system:

$$S\,\mathbf{u}_B = \left(\mathbf{f}_B - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)}\right),$$

where $S \equiv (A_{BB} - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} A_{IB}^{(1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} A_{IB}^{(2)})$ is the two subdomain Schur complement. S will be *dense*, however, its action can be computed without its assembly. We shall seek preconditioners M for S such that:

$$\mathrm{cond}(M, S) \equiv \frac{\lambda_{\max}\left(M^{-1}S\right)}{\lambda_{\min}\left(M^{-1}S\right)},$$

is significantly smaller than the condition number of S, without deterioration as $h \rightarrow 0^+$, or as the coefficient $a(x)$ and the subdomain size h_0 varies.

In this section, we shall describe three categories of Schur complement preconditioners for two subdomain decompositions:

- *Preconditioners based on subdomain Schur complements.*
- *Preconditioners based on FFT's and fractional Sobolev norms.*
- *Preconditioner based on algebraic approximations of S.*

Of these, the preconditioners based on subdomain Schur complements are more easily generalized to the many subdomain case and higher dimensions.

3.4.1 Preconditioners Based on Subdomain Schur Complements

The use of the local Schur complement $S^{(i)}$ to precondition S can be motivated by the matrix *splitting* of S by the subassembly identity (3.20):

$$S = S^{(1)} + S^{(2)}, \quad \text{where} \quad S^{(i)} = A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)},$$

since $B = B^{(1)} = B^{(2)}$ and $\mathcal{R}_B^{(i)} = I$ for $i = 1, 2$. This splitting may also be derived by substituting the identity $A_{BB} = A_{BB}^{(1)} + A_{BB}^{(2)}$, into the algebraic expression $S = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$ for the Schur complement matrix:

$$\begin{cases} S = \left(A_{BB}^{(1)} + A_{BB}^{(2)}\right) - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} A_{IB}^{(1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} A_{IB}^{(2)} \\ \quad = \left(A_{BB}^{(1)} - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} A_{IB}^{(1)}\right) + \left(A_{BB}^{(2)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} A_{IB}^{(2)}\right) = S^{(1)} + S^{(2)}. \end{cases}$$

Typically, each $S^{(i)}$ will be symmetric and positive definite, except when $c(x) = 0$ and Ω_i is *immersed* in Ω, in which case $S^{(i)}$ will be *singular*. For simplicity, however, we shall assume $S^{(i)}$ is nonsingular (see Chap. 3.7).

Matrix $S^{(i)}$ need not be assembled (and it will be *dense*, even if it were assembled). It will be important to solve the system $S^{(i)}\mathbf{v}_B^{(i)} = \mathbf{r}_B^{(i)}$ efficiently.

Fortunately, such a system can be solved without assembling $S^{(i)}$, by using the following algebraic property satisfied by $S^{(i)}$:

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(i)} \\ \mathbf{v}_B^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ S^{(i)}\mathbf{v}_B^{(i)} \end{bmatrix}. \tag{3.60}$$

This identity can be verified by block elimination of $\mathbf{v}_I^{(i)}$. It suggests that the solution to $S^{(i)}\mathbf{v}_B^{(i)} = \mathbf{r}_B^{(i)}$ can be obtained by solving (3.60) using $\mathbf{r}_B^{(i)}$ to replace $S^{(i)}\mathbf{v}_B^{(i)}$ in the right hand side, and by selecting $\mathbf{v}_B^{(i)}$. Formally:

$$S^{(i)^{-1}}\mathbf{r}_B^{(i)} = \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix} \mathbf{r}_B^{(i)}.$$

The subdomain stiffness matrix here corresponds to the discretization of an elliptic equation on Ω_i with *Neumann* boundary data on $B^{(i)}$.

$S^{(i)}$ is traditionally referred to as the *Dirichlet-Neumann* preconditioner for S, see [BJ9, BR11, FU, MA29]. Its name arises, since a Neumann problem must be solved on Ω_i and a subsequent Dirichlet problem on its complementary domain, in the Dirichlet-Neumann algorithm (Alg. 1.3.1 from Chap. 1.3). When the number of unknowns on each subdomain is approximately half the total number of unknowns, the cost of preconditioning with $S^{(i)}$ is typically less than half the cost of solving $A\mathbf{v} = \mathbf{r}$. However, the number of iterations required depends on the effectiveness of this preconditioner. It is shown in Chap. 3.9 that $\text{cond}(S^{(i)}, S) \leq c$, for $c > 0$ independent of h. In the special case where the elliptic equation and the grid is symmetric about B, it will hold that $S^{(1)} = S^{(2)}$ and $\text{cond}(S^{(i)}, S) = 1$.

Below, we list a discretization of the Steklov-Poincaré formulation (1.19) from Chap.1.3, and derive the discrete version of the Dirichlet-Neumann algorithm. Let $(\mathbf{w}_I^{(1)}, \mathbf{w}_B^{(1)})^T$ and $(\mathbf{w}_I^{(2)}, \mathbf{w}_B^{(2)})^T$ denote nodal vectors associated with finite element functions on $\overline{\Omega}_1$ and $\overline{\Omega}_2$, respectively. Then, a discretization of the Steklov-Poincaré formulation (1.19) will yield:

$$\begin{cases} A_{II}^{(1)}\mathbf{w}_I^{(1)} + A_{IB}^{(1)}\mathbf{w}_B^{(1)} = \mathbf{f}_I^{(1)} \\ \mathbf{w}_B^{(1)} = \mathbf{w}_B^{(2)} \\ A_{II}^{(2)}\mathbf{w}_I^{(2)} + A_{IB}^{(2)}\mathbf{w}_B^{(2)} = \mathbf{f}_I^{(2)} \\ A_{IB}^{(2)^T}\mathbf{w}_I^{(2)} + A_{BB}^{(2)}\mathbf{w}_B^{(2)} = -A_{IB}^{(1)^T}\mathbf{w}_I^{(1)} - A_{BB}^{(1)}\mathbf{w}_B^{(1)} + \mathbf{f}_B. \end{cases}$$

To obtain a discrete version of the Dirichlet-Neumann algorithm, let $\mathbf{v}_I^{(k)}$ and $\mathbf{v}_B^{(k)}$ denote the k'th iterate on Ω_1 and $B^{(1)}$, respectively, and $\mathbf{u}_I^{(k)}$ and $\mathbf{u}_B^{(k)}$ the k'th iterate on Ω_2 and $B^{(2)}$, respectively. Iteratively replace the transmission boundary conditions on B using a relaxation parameter $0 < \theta < 1$.

Algorithm 3.4.1 *(Dirichlet-Neumann Algorithm)*
Let $(\mathbf{v}_I^{(0)}, \mathbf{v}_B^{(0)})^T$ *and* $(\mathbf{u}_I^{(0)}, \mathbf{u}_B^{(0)})^T$ *denote starting iterates.*

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     Solve the Dirichlet problem:
           $\begin{cases} A_{II}^{(1)} \mathbf{v}_I^{(k+1)} + A_{IB}^{(1)} \mathbf{v}_B^{(k+1)} = \mathbf{f}_I^{(1)} \\ \mathbf{v}_B^{(k+1)} = \theta\, \mathbf{u}_B^{(k)} + (1-\theta)\, \mathbf{v}_B^{(k)} \end{cases}$
3.     Solve the mixed problem:
           $\begin{cases} A_{II}^{(2)} \mathbf{u}_I^{(k+1)} + A_{IB}^{(2)} \mathbf{u}_B^{(k+1)} = \mathbf{f}_I^{(2)} \\ A_{IB}^{(2)^T} \mathbf{u}_I^{(k+1)} + A_{BB}^{(2)} \mathbf{u}_2^{(k+1)} = \mathbf{f}_B - A_{IB}^{(1)^T} \mathbf{v}_I^{(k+1)} - A_{BB}^{(1)} \mathbf{v}_B^{(k+1)}. \end{cases}$
4. Endfor
Output: $\left(\mathbf{v}_I^{(k)^T}, \mathbf{v}_B^{(k)^T}\right)^T$, $\left(\mathbf{u}_I^{(k)^T}, \mathbf{u}_B^{(k)^T}\right)^T$
```

If the interior variables $\mathbf{v}_I^{(k+1)}$ and $\mathbf{u}_I^{(k+1)}$ are eliminated in the preceding algorithm, we may obtain an expression relating $\mathbf{v}_B^{(k+2)}$ to $\mathbf{v}_B^{(k+1)}$. A matrix form for this can be derived by solving for $\mathbf{v}_I^{(k+1)}$ in step 2:

$$\begin{cases} \mathbf{v}_I^{(k+1)} = A_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - A_{IB}^{(1)} \mathbf{v}_B^{(k+1)} \right) \\ \mathbf{v}_B^{(k+1)} = \theta \mathbf{u}_B^{(k)} + (1-\theta)\, \mathbf{v}_B^{(k)}, \end{cases}$$

and substituting this into the equations in step 3. Solving the resulting block system using block elimination (representing $\mathbf{u}_I^{(k+1)}$ in terms of $\mathbf{u}_B^{(k+1)}$) yields:

$$\begin{cases} S^{(2)} \mathbf{u}_B^{(k+1)} = \mathbf{f}_B - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - A_{IB}^{(1)} \mathbf{v}_B^{(k+1)} \right) \\ \qquad\qquad -A_{BB}^{(1)} \mathbf{v}_B^{(k+1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \\ \qquad\quad = \mathbf{f}_B - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} - S^{(1)} \mathbf{v}_B^{(k+1)}. \end{cases}$$

Defining $\tilde{\mathbf{f}}_B \equiv \mathbf{f}_B - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)}$, this reduces to:

$$\mathbf{u}_B^{(k+1)} = S^{(2)^{-1}} \left(\tilde{\mathbf{f}}_B - S^{(1)} \mathbf{v}_B^{(k+1)} \right).$$

Since $\mathbf{v}_B^{(k+2)}$ is defined as $\mathbf{v}_B^{(k+2)} = \theta\, \mathbf{u}_B^{(k+1)} + (1-\theta)\, \mathbf{v}_B^{(k+1)}$, this shows that the preceding Dirichlet-Neumann algorithm corresponds to an *unaccelerated* Richardson iteration to solve the Schur complement system $S\, \mathbf{u}_B = \tilde{\mathbf{f}}_B$ with $M = S^{(2)}$ as a preconditioner and θ as a *relaxation* parameter. We may also employ $M = S^{(1)}$ as a preconditioner for S. Below, we summarize the action of the Dirichlet-Neumann preconditioner $M = S^{(i)}$ on a vector $\mathbf{r}_B^{(i)}$.

Algorithm 3.4.2 *(Dirichlet-Neumann Preconditioner)*
Input: $\mathbf{r}_B$
Solve:

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(i)} \\ \mathbf{v}_B^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_B \end{bmatrix}.$$

Output: $M^{-1}\mathbf{r}_B \equiv \mathbf{v}_B^{(i)}$.

Remark 3.29. As mentioned earlier, when $c(x) = 0$ and $B^{(i)} = \partial\Omega_i$, the local stiffness matrix $A^{(i)}$ and its Schur complement $S^{(i)}$ will be *singular*, with the null space of $S^{(i)}$ spanned by $\mathbf{1} = (1, \ldots, 1)^T$. In this case, the Dirichlet-Neumann preconditioner must be modified, since $S^{(i)}\mathbf{v}_B^{(i)} = \mathbf{r}_B$ will be solvable only if the compatability condition $\mathbf{1}^T\mathbf{r}_B = 0$ is satisfied. Furthermore, the solution will be unique only up to a multiple of $\mathbf{1}$. Both issues are addressed by the *balancing* procedure in Chap. 3.7, see [MA14, MA17].

When applying the *Dirichlet-Neumann* preconditioner, a specific Schur complement matrix $S^{(i)}$ must be chosen, for $i = 1, 2$. When the geometry, coefficients, and grid on the two subdomains differ significantly, matrices $S^{(1)}$ and $S^{(2)}$ can also differ significantly. In this case, it may be more equitable to combine information from both the subdomains in the preconditioner. This motivates the *Neumann-Neumann* preconditioner. The action of the inverse of the two subdomain Neumann-Neumann preconditioner M is defined as:

$$M^{-1} \equiv \alpha\, S^{(1)^{-1}} + (1-\alpha)\, S^{(2)^{-1}},$$

where $0 < \alpha < 1$ is a scalar parameter for assigning different weights to each subdomain (though, typically $\alpha = \frac{1}{2}$). Computing the action of M^{-1} requires the solution of a discretized elliptic equation on each subdomain, and in parallel, with *Neumann* boundary conditions on B, hence, the name Neumann-Neumann preconditioner [BO7]. The action of the inverse of this preconditioner is summarized below.

Algorithm 3.4.3 *(Neumann-Neumann Preconditioner)*
Input: $\mathbf{r}_B$ *and* $0 < \alpha < 1$

1. *For* $i = 1, 2$ *in parallel solve for* $\left(\mathbf{w}_I^{(i)}, \mathbf{w}_B^{(i)}\right)^T$*:*

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(i)} \\ \mathbf{w}_B^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_B \end{bmatrix}.$$

2. *Endfor*

Output: $M^{-1}\mathbf{r}_B \equiv \alpha\,\mathbf{w}_B^{(1)} + (1-\alpha)\,\mathbf{w}_B^{(2)}$.

An advantage of the Neumann-Neumann preconditioner is that each local problem is typically easy to set up, and its algebraic form extends easily to *multisubdomain* decompositions. It also applies to subdomains with arbitrary geometry in *two or three dimensions.* Theoretical analysis in Chap. 3.9 indicates that the Dirichlet-Neumann and Neumann-Neumann preconditioners typically yield convergence rates which are independent of the mesh size h. By construction, the preceding Dirichlet-Neumann preconditioner requires only *one* subdomain solve, while the Neumann-Neumann preconditioner requires *two* subdomain solves.

Theorem 3.30. *If M denotes the Dirichlet-Neumann or Neumann-Neumann preconditioner for a two subdomain decomposition, then the condition number* $\mathrm{cond}(M, S)$ *will be bounded independent of the mesh size h.*

Proof. See [BJ9, BR11, FU, MA29] and Chap. 3.9. □

3.4.2 Preconditioners Based on FFT's and Fractional Norms

Preconditioners for S, based on FFT's and fractional Sobolev norms, can be motivated in alternate ways. In a model problem based approach, the Schur complement S on an interface B is *approximated* by the Schur complement $\tilde{S}$ of a *model* problem on another domain, whose interface $\tilde{B}$ has the same number of unknowns as on B. If the Schur complement $\tilde{S}$ in the model problem has FFT solvers, then it will provide a *heuristic* FFT based preconditioner for S. In the fractional Sobolev norm approach, an equivalence between the energy of the Schur complement S and a fractional Sobolev norm energy of its boundary data on B is employed. If M is a matrix that generates the latter fractional Sobolev norm energy, it can be employed to precondition the Schur complement S. The advantage of FFT based preconditioners is that when they are applicable, they yield almost optimal order complexity, and convergence rates independent of h. However, such methodology is primarily applicable in two dimensions, i.e., when $\Omega \subset \mathbb{R}^2$. In three dimensions, the grid on B must either be uniform, or have a multilevel structure, for applicability.

Model Problem Based Preconditioners. The model problem approach is *heuristic* in nature. Given a domain Ω with subdomains Ω_1 and Ω_2, with interface B, let $\hat{\Omega}$ be a region approximating Ω, with subdomains $\hat{\Omega}_1$, $\hat{\Omega}_2$ and interface $\hat{B}$ that approximate Ω_1, Ω_2 and B, respectively. Then, the elliptic equation (3.1) posed on Ω may be heuristically approximated by an elliptic equation posed on $\hat{\Omega}$ with (possibly modified) coefficients $\hat{a}(x)$ and $\hat{c}(x)$ approximating $a(x)$ and $c(x)$, respectively:

$$\begin{cases} \nabla \cdot (\hat{a}(x)\nabla \hat{u}) + \hat{c}(x) = \hat{f}(\hat{x}), & \text{for } \hat{x} \in \hat{\Omega} \\ \qquad\qquad\qquad\qquad \hat{u} = 0, & \text{for } \hat{x} \in \partial\hat{\Omega}. \end{cases} \tag{3.61}$$

A preconditioner $\hat{S}$ can be constructed for S, as follows. Let $\mathcal{T}_h(\hat{\Omega})$ denote a triangulation of $\hat{\Omega}$ having the same number of interior nodes in $\hat{B}$ as in B, and consider a discretization of (3.61) on this grid. Given the subdomains $\hat{\Omega}_1$, $\hat{\Omega}_2$ and interface $\hat{B}$, let $\hat{S}$ denote the Schur complement associated with the discretized model problem on $\hat{\Omega}$. Block partitioning the unknowns in $\hat{\Omega}$ based on the subregions yields the system:

$$\begin{bmatrix} \hat{A}_{II} & \hat{A}_{IB} \\ \hat{A}_{IB}^T & \hat{A}_{BB} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_I \\ \hat{\mathbf{u}}_B \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{f}}_I \\ \hat{\mathbf{f}}_B \end{bmatrix}. \tag{3.62}$$

The Schur complement matrix $\hat{S} = (\hat{A}_{BB} - \hat{A}_{IB}^T \hat{A}_{II}^{-1} \hat{A}_{BB})$ in the model problem may then be employed as a preconditioner for S.

Remark 3.31. If $\hat{\Omega}$ is obtained by mapping Ω and if B maps into $\hat{B}$, we may seek $\hat{\Omega}_1$ and $\hat{\Omega}_2$ that are rectangular. Furthermore, the coefficients $\hat{a}(x)$ and $\hat{c}(x)$ may be chosen so that (3.61) is *separable*. If a uniform triangulation is employed on $\hat{\Omega}$, then a FFT solver can be constructed for $\hat{S}$.

Remark 3.32. If $\hat{\Omega}$ is a *small subregion* of Ω satisfying $B \subset \hat{\Omega} \subset \Omega$, then we may define $\hat{\Omega}_1 = \Omega_1 \cap \hat{\Omega}$, $\hat{\Omega}_2 = \Omega_2 \cap \hat{\Omega}$ and $\hat{B} = B$. In this case, we may choose $\hat{a}(\cdot) = a(\cdot)$ and $\hat{c}(\cdot) = c(\cdot)$, and system (3.62) will have a coefficient matrix which is a small submatrix of A. To construct a preconditioner $\hat{S}$ for S, we may substitute $\hat{\mathbf{f}}_I = \mathbf{0}$, $\hat{\mathbf{f}}_B = \mathbf{r}_B$ into (3.62) and define the action of the inverse of a preconditioner as $\hat{S}^{-1}\mathbf{r}_B \equiv \hat{\mathbf{u}}_B$, see [NE3, MA37, SM].

Next, we elaborate on the preconditioner outlined in Remark 3.31 for a *two dimensional* domain Ω. We shall assume that the interface B can be mapped onto a line segment $\hat{B}$, and choose $\hat{\Omega}$, $\hat{\Omega}_1$ and $\hat{\Omega}_2$ to be *rectangular* regions. The grid on $\hat{\Omega}$ will be chosen to be uniform, and the coefficients $\hat{a}(x)$ and $\hat{c}(x)$ will be chosen to be constant in each subdomain. In this case, matrix $\hat{S}$ may be *explicitly diagonalized* by a discrete sine transform matrix Q, with $\hat{S} = QDQ^T$ for a diagonal matrix D, see (3.57) and (3.58). If k denotes the number of unknowns on B (and hence on $\hat{B}$), we define the discrete sine transform matrix Q of size k as:

$$Q_{ij} \equiv \sqrt{2/(k+1)}\, \sin\left(i\, j\, \pi/(k+1)\right), \quad 1 \le i, j \le k. \tag{3.63}$$

We employ a model Schur complement preconditioner $\hat{S} = QDQ^T$ for S, for different choices of diagonal matrices D, see [DR, GO3, BJ9, CH2, BR11]. Heuristically, matrix $\hat{S}$ should be approximately spectrally equivalent to S, when the triangulation of $\hat{\Omega}$ restricted to $\hat{B}$ has the same connectivity as the original triangulation restricted to B.

Next, we list different choices of diagonal entries D_{ii} for $1 \le i \le k$, in two dimensions, yielding different preconditioners:

$$\begin{cases} D_{ii} = \left(a^{(1)} + a^{(2)}\right) (\sigma_i)^{1/2}, & \text{[DR]} \\ D_{ii} = \left(a^{(1)} + a^{(2)}\right) \left(\sigma_i + \sigma_i^2/4\right)^{1/2}, & \text{[GO3]} \\ D_{ii} = \left(a^{(1)} \dfrac{1+\gamma_i^{m_1+1}}{1-\gamma_i^{m_1+1}} + a^{(2)} \dfrac{1+\gamma_i^{m_2+1}}{1-\gamma_i^{m_2+1}}\right) \left(\sigma_i + \dfrac{1}{4}\sigma_i^2\right)^{1/2}, & \text{[BJ9, CH2]} \\ D_{ii} = \dfrac{1}{2} \left(a^{(1)} + a^{(2)}\right) \left(\sigma_i - \sigma_i^2/6\right)^{1/2}, & \text{[BR11, BJ9].} \end{cases} \tag{3.64}$$

Here, the parameters σ_i and γ_i are defined by:

$$\begin{cases} \sigma_i \equiv 4 \sin^2\left(\dfrac{i\,\pi}{2(k+1)}\right), & \text{for } 1 \le i \le k \\ \gamma_i \equiv \dfrac{1 + \frac{1}{2}\sigma_i - \sqrt{\sigma_i + \frac{1}{4}\sigma_i^2}}{1 + \frac{1}{2}\sigma_i + \sqrt{\sigma_i + \frac{1}{4}\sigma_i^2}}, & \text{for } 1 \le i \le k. \end{cases} \tag{3.65}$$

The scalars $a^{(1)}$ and $a^{(2)}$ denote values of $a(x)$ at some interior point in Ω_1 and Ω_2, respectively, when $a(x)$ is a scalar function. When $a(x)$ is a matrix function, $a^{(1)}$ and $a^{(2)}$ will be eigenvalues of $a(x)$ at chosen interior points in Ω_1 and Ω_2. The parameters m_1 and m_2 in the preconditioner of [BJ9, CH2] are integers chosen so that $(m_i+1)\,h$ and $(k+1)h$ represents the approximate length and width of subdomain Ω_i. The resulting preconditioner $\hat{S} = QDQ^T$ for S, is summarized next, where Q is the discrete sine transform.

Algorithm 3.4.4 *(FST Based Fractional Norm Preconditioner)*
Input: $\mathbf{r}_B$, D

```
1. Evaluate using the fast sine transform: $\mathbf{y}_B = Q\,\mathbf{r}_B$
2. Compute in linear complexity: $\mathbf{x}_B = D^{-1}\mathbf{y}_B$
3. Evaluate using the fast sine transform: $\mathbf{w}_B = Q\,\mathbf{x}_B$
```

Output: $\hat{S}^{-1}\mathbf{r}_B \equiv \mathbf{w}_B$.

Since the cost of applying a discrete sine transform is typically $O(k\log(k))$, the combined cost for solving the linear system $\hat{S}\,\mathbf{w}_B = \mathbf{r}_B$ will be $O(k\log(k))$. Since the discrete sine transform matrix Q is symmetric and orthogonal, it will hold that $Q^{-1} = Q$, so that the transform applied twice should yield the identity. However, in FFT packages [VA4] the discrete sine transform may be scaled differently, so the user may need to rescale the output.

Remark 3.33. The choice of diagonal matrix D in [BJ9, CH2] can be formally obtained as follows for a two dimensional domain $\hat{\Omega}$ with k interior nodes on B. Let $(m_i+1)\,h$ the approximate length of subdomain $\hat{\Omega}_i$ and let the coefficients of the model problem be *isotropic* with $\hat{a}^{(i)}$ constant in $\hat{\Omega}_i$. Then, for $h_{x_1} = h_{x_2} = h$, the eigenvalues D_{ii} of the Schur complement matrix $\hat{S}$ in the model problem simplifies to:

$$D_{ii} = \left(a^{(1)} \frac{1 + \gamma_i^{m_1+1}}{1 - \gamma_i^{m_1+1}} + a^{(2)} \frac{1 + \gamma_i^{m_2+1}}{1 - \gamma_i^{m_2+1}} \right) \left(\sigma_i + \frac{1}{4}\sigma_i^2 \right)^{1/2},$$

for $1 \leq i \leq k$, where σ_i and γ_i are as defined in (3.65). This follows by algebraic simplification of (3.57). When the parameters m_1 and m_2 are large above, the expression for the eigenvalues D_{ii} can be approximated as:

$$D_{ii} \to \left(a^{(1)} + a^{(2)} \right) \left(\sigma_i + \sigma_i^2/4 \right)^{1/2},$$

for $1 \leq i \leq k$, since $0 < \gamma_i < 1$. This heuristically motivates the preconditioner of [GO3]. The preconditioner of [DR] can be formally obtained from the preconditioner of [GO3] by replacing the terms $\left(\sigma_i + \sigma_i^2/4\right)^{1/2}$ by the terms $(\sigma_i)^{1/2}$. By construction, both preconditioners will be spectrally equivalent since $\left(\sigma_i + \sigma_i^2/4\right)^{1/2} = (\sigma_i)^{1/2}\left(1 + \sigma_i/4\right)^{1/2}$ and $1 < \left(1 + \sigma_i/4\right)^{1/2} < \sqrt{2}$ for $0 < \sigma_i < 4$.

Remark 3.34. Similar preconditioners can be constructed in *three* dimensions provided the grid on the interface B can be mapped into a two dimensional rectangular grid. Matrix Q will then be a two dimensional fast sine transform. When the grid on B is not rectangular, preconditioners approximating fractional Sobolev norms can be constructed using *multilevel* methodology, provided the grid has a multilevel structure, see [BR17] and Chap. 7.1.

Fractional Sobolev Norm Based Preconditioners. This approach is motivated by a norm equivalence, see (3.28) and (3.32), between the energy associated with a two subdomain Schur complement S and a fractional Sobolev norm energy. Such norm equivalences hold for harmonic and discrete harmonic functions, and is proved using elliptic regularity theory. For two subdomain decompositions, the norm equivalences (3.32) and (3.28) reduce to:

$$c \left\| u_h \right\|^2_{H^{1/2}_{00}(B)} \leq \mathbf{u}_B^T S \mathbf{u}_B \leq C \left\| u_h \right\|^2_{H^{1/2}_{00}(B)}, \tag{3.66}$$

for $0 < c < C$ independent of h, since the fractional Sobolev norm $\|u_h\|_{1/2,\partial\Omega_i}$ can be shown to be equivalent to $\|u_h\|_{H^{1/2}_{00}(B)}$ when u_h is zero on $\partial\Omega_i \backslash B$, see [LI4]. This norm equivalence suggests that a preconditioner M can be constructed for S, by representing the discrete fractional Sobolev energy as:

$$\left\| u_h \right\|^2_{H^{1/2}_{00}(B)} = \mathbf{u}_B^T M \mathbf{u}_B.$$

A matrix M satisfying this property can be constructed by employing the theory of Hilbert *interpolation* spaces [BA3, LI4, BE16] as outlined below.

Given two Hilbert spaces satisfying $\mathcal{H}_0 \supset \mathcal{H}_1$ where the latter space has a stronger norm $\|u\|_{\mathcal{H}_0} \leq C \|u\|_{\mathcal{H}_1}$ for all $u \in \mathcal{H}_1$, a family of interpolation spaces $\mathcal{H}^\alpha$ can be constructed for $0 \le \alpha \le 1$ with $\mathcal{H}^0 = \mathcal{H}_0$ and $\mathcal{H}^1 = \mathcal{H}_1$, with associated inner products defined as outlined below [BA3, LI4, BE16].

- Let $\mathcal{H}_0$ denote an Hilbert space with inner product $(.,.)_0$ and let $\mathcal{H}_1 \subset \mathcal{H}_0$ denote a subspace with a stronger inner product $(,.,)_1$:
$$(u,u)_0 \le C\,(u,u)_1\,, \qquad \forall u \in \mathcal{H}_1.$$
- Let T denote a *self adjoint* coercive operator satisfying:
$$(Tu,v)_0 = (u,v)_1\,, \qquad \forall u,v \in \mathcal{H}_1,$$
which corresponds to a Riesz representation map.
- Let T have the following spectral decomposition:
$$T = \sum_{i=1}^{\infty} \lambda_i P_i,$$
where $0 < \lambda_1 < \lambda_2 < \cdots$ are eigenvalues of T and P_i are $(.,.)_0$ orthogonal projections onto the eigenspace of T corresponding to eigenvalue λ_i.

Then, for $0 \le \alpha \le 1$ we may formally define a fractional operator T^α as:

$$T^\alpha \equiv \sum_{i=1}^{\infty} \lambda_i^\alpha P_i, \qquad 0 \le \alpha \le 1.$$

Then, for each $0 \le \alpha \le 1$ the interpolation space $\mathcal{H}^\alpha$ is formally defined as the domain of the fractional operator T^α, so that $\mathcal{H}^0 = \mathcal{H}_0$ and $\mathcal{H}^1 = \mathcal{H}_1$:

$$\mathcal{H}^\alpha \equiv \left\{ u \in \mathcal{H}_0 : (T^\alpha u, u)_0 < \infty \right\},$$

where the inner product on $\mathcal{H}^\alpha$ is consistently defined by:

$$(u,v)_\alpha \equiv (T^\alpha u, v)_0 = \textstyle\sum_{i=1}^{\infty} \lambda_i^\alpha (P_i u, v)_0.$$

This procedure defines interpolation spaces $\mathcal{H}^\alpha$ satisfying $\mathcal{H}_1 \subset \mathcal{H}^\alpha \subset \mathcal{H}_0$.

In elliptic regularity theory, the fractional index Sobolev space $H^{1/2}_{00}(B)$ is often constructed as an interpolation space $\mathcal{H}^{1/2}$ obtained by interpolating $\mathcal{H}_0 = L^2(B)$ and $\mathcal{H}_1 = H^1_0(B)$. The space $H^{1/2}_{00}(B)$ will correspond to the domain of the operator $T^{\frac{1}{2}}$ with associated fractional norm defined by:

$$\|u\|^2_{H^{1/2}_{00}(B)} \equiv \left(T^{\frac{1}{2}} u, u \right)_0 = \textstyle\sum_{i=1}^{\infty} \lambda_i^{\frac{1}{2}} (P_i u, u)_0, \qquad \forall u \in H^{1/2}_{00}(B).$$

The operator T corresponds to a Laplace-Beltrami operator $-\Delta_B$ defined on B with homogeneous boundary conditions on ∂B. Formally, the fractional powers of T may be computed by employing the eigenfunction expansion of T and replacing the eigenvalues of T by their fractional powers. These fractional operators T^α, however, will not remain differential operators for $0 < \alpha < 1$, and are examples of *pseudodifferential operators*.

In the finite dimensional case, we may employ fractional powers of matrices to represent fractional operators. To obtain a matrix representation of $\|u_h\|^2_{H^{1/2}_{00}(B)}$ on the finite element space $V_h(B)$ of finite element functions restricted to B, we seek a symmetric positive definite matrix T_h satisfying:

$$\begin{cases} \left(T_h^0 u_h, u_h\right) = \|u_h\|^2_{L^2(B)}, & \text{for } u_h \in V_h(B) \cap L^2(B) \\ \left(T_h^1 u_h, u_h\right) = \|u_h\|^2_{H_0^1(B)}, & \text{for } u_h \in V_h(B) \cap H_0^1(B), \end{cases}$$

where $(\cdot,\cdot)$ denotes the $L^2(B)$ inner product. Let G_h denote the mass (Gram) matrix associated with the finite element space $V_h \cap H_0^1(B)$ with standard nodal basis, and let A_h denote the finite element discretization of the Laplace-Beltrami operator with trivial boundary conditions imposed on ∂B. Then, by construction it will hold that:

$$\begin{cases} (T_h^\alpha u_h, u_h) = \mathbf{u}_B^T A_h \mathbf{u}_B, & \text{for } \alpha = 1 \\ (T_h^\alpha u_h, u_h) = \mathbf{u}_B^T G_h \mathbf{u}_B, & \text{for } \alpha = 0, \end{cases}$$

where $\mathbf{u}_B$ denotes the nodal vector corresponding to the finite element function $u_h(x)$ restricted to B. Formally, a matrix representation of fractional operators associated with T_h may be constructed as:

$$T_h^\alpha = G_h^{\frac{1}{2}} \left(G_h^{-\frac{1}{2}} A_h G_h^{-\frac{1}{2}} \right)^\alpha G_h^{\frac{1}{2}}, \qquad \text{for} \quad 1 \le \alpha \le 1.$$

This yields:

$$T_h^{\frac{1}{2}} = G_h^{\frac{1}{2}} \left(G_h^{-\frac{1}{2}} A_h G_h^{-\frac{1}{2}} \right)^{\frac{1}{2}} G_h^{\frac{1}{2}}.$$

When matrices A_h and G_h can be simultaneously diagonalized by the discrete sine transform Q, then $T_h^{\frac{1}{2}}$ can be efficiently computed and its associated linear system can be solved efficiently.

To construct an explicit representation of matrix $T_h^{1/2}$ we assume that the interface B corresponds to the line segment $(0,1)$. Then, the Laplace-Beltrami operator $-\Delta_B$ defined on B, with zero boundary conditions on ∂B is:

$$-\Delta_B u(x) \equiv -\frac{d^2 u}{dx^2}, \qquad \text{for} \quad u(0) = 0 \text{ and } u(1) = 0.$$

If the grid size is $h = 1/(k+1)$, and nodal vector $(\mathbf{u}_B)_i = u_h(ih)$ corresponds to the finite element function u_h, then the finite element discretization A_h of the Laplace-Beltrami operator and the Gram matrix G_h will be of size k:

$$A_h = \frac{1}{h} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \quad \text{and} \quad G_h = \frac{h}{6} \begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix}.$$

Matrices A_h and G_h can be simultaneously diagonalized by the one dimensional discrete sine transform matrix Q with entries:

$$Q_{ij} = \sqrt{2/(k+1)}\,\sin\left(i\,j\,\pi/(k+1)\right), \quad \text{for } 1 \le i,\, j \le k.$$

The eigenvalues of matrices A_h and G_h corresponding to eigenvector $\mathbf{q}_j$ is:

$$\begin{cases} \lambda_j(A_h) = 4(k+1)\sin^2(\frac{j\,\pi}{2(k+1)}), & \text{for } 1 \le j \le k \\ \lambda_j(G_h) = \frac{1}{3(k+1)}\left(3 - 2\sin^2(\frac{j\,\pi}{2(k+1)})\right) & \text{for } 1 \le j \le k. \end{cases}$$

The fractional power $T_h^{\frac{1}{2}}$ can be represented explicitly as $T_h^{\frac{1}{2}} = QDQ^T$ where:

$$\begin{cases} D_{jj} = \lambda_j(G_h)^{\frac{1}{2}}\,\lambda_j(A_h)^{\frac{1}{2}} \\ \quad = \left(\frac{1}{3(k+1)}\left(3 - 2\sin^2(\frac{j\,\pi}{2(k+1)})\right)\right)^{\frac{1}{2}}\left(4\,(k+1)\,\sin^2(\frac{j\,\pi}{2(k+1)})\right)^{\frac{1}{2}} \\ \quad = \frac{2}{\sqrt{3}}\left(3\,\sin^2(\frac{j\,\pi}{2(k+1)}) - 2\sin^4(\frac{j\,\pi}{2(k+1)})\right)^{\frac{1}{2}} \;=\; \left(\sigma_j - \frac{1}{6}\sigma_j^2\right)^{\frac{1}{2}} \end{cases}$$

for $\sigma_i = 4\sin^2(\frac{i\,\pi}{2(k+1)})$ where $1 \le j \le k$. This choice of D yields the preconditioner $M = Q^T DQ$ of [BR11, BJ9] in (3.64), with $a^{(1)} = a^{(2)} = 1$. It may be implemented as in Alg. 3.4.4.

Remark 3.35. Analogous FFT based preconditioners can be constructed for two subdomain Schur complements in *three* dimensions, provided that the grid on the interface B can be mapped onto a uniform rectangular grid [CH2, CH13]. In this case, the Schur complements S may be formulated and applied *heuristically*, by analogy with the two dimensional case.

The following result is proved in Chap. 3.9.

Lemma 3.36. *For any 2nd order, coercive self adjoint elliptic operator, the subdomain Schur complement preconditioner and the fractional Sobolev norm based preconditioner $\hat{S}$ will be spectrally equivalent to S as $h \to 0$.*

The convergence rate, however, may depend on the aspect ratios of the subdomains, and also on the coefficients.

3.4.3 Preconditioners Based on Algebraic Approximation of S

The Schur complement matrix S arising in a two subdomain decomposition is typically a *dense* matrix. This can be verified heuristically by computing its entries and plotting its magnitude, or by using expression (3.26) and noting that the discrete Green's function A_{II}^{-1} is a dense matrix as $h \to 0^+$ (due to the global domain of dependence of elliptic equations). As a result, traditional algebraic preconditioners based on *ILU* factorization [GO4, BE2, SA2, AX] will offer no advantages over direct solution. Furthermore, such factorizations cannot be employed when matrix S is *not assembled*, as is the case in iterative substructuring methods. Instead, in this subsection we shall describe two alternative algebraic preconditioners for the Schur complement matrix, one based on sparse approximation of the Schur complement using the *probing* technique, and the other based on *incomplete factorization* of the subdomain matrices $A_{II}^{(i)}$. Both preconditioners may be applied without assembly of S.

The first algebraic preconditioner we shall consider is based on the construction of a *sparse* matrix approximation of S using a *probing* technique [CH13, KE7, CH9]. A sparse approximation of S can be heuristically motivated by a *decay property* in the entries S_{ij} of a two subdomain Schur complement matrix, with increasing distance between the nodes x_i and x_j. This decay property can be observed when S is assembled explicitly, and arises from the decay in the entries $(A_{II}^{(l)^{-1}})_{rs}$ of the discrete Green's function associated with the elliptic equation on the subdomains, with increasing distance between the nodes x_r and x_s. This suggests that a sparse approximation M of S may be effective as a preconditioner, provided the nonzero entries of M approximate the dominant entries of S. In the following, we shall describe the probing technique for determining a sparse approximation M of S.

For $\Omega \subset \mathbb{R}^2$, if the nodes x_i on B are ordered consecutively along B, then the entries of the Schur complement matrix S typically decay along diagonal bands. This motivates choosing a band matrix M, say of band width d, to approximate S. Nonzero entries of the band matrix M can be determined by choosing *probe* vectors $\mathbf{p}_l$, say for $1 \le l \le (2d+1)$, and requiring that the matrix vector products of S with each probe vector $\mathbf{p}_l$ matches the matrix vector product of M with the same probe vector:

$$M\mathbf{p}_l = S\mathbf{p}_l, \qquad \text{for} \quad 1 \le l \le (2d+1).$$

If matrix S is of size k, these requirements yield $k\,(2d+1)$ equations for the unknown entries of M. A careful choice of the probe vectors based on the decay in the entries of S can increase the accuracy of the probe approximation M, and also simplify the linear system for the nonzero entries of M. The resulting *probing* technique [CH13, KE7, CH9] does not require the explicit assembly of matrix S, but does require the computation of the matrix-vector products of S with the chosen probe vectors.

Below, we illustrate a specific choice of probe vectors to construct a *tridiagonal* approximation M of S. In this case $2d+1 = 3$, and three probe vectors $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_3$ will be sufficient. Choose:

$$\begin{cases} \mathbf{p}_1 = (1,0,0,1,0,0,\ldots)^T \\ \mathbf{p}_2 = (0,1,0,0,1,0,\ldots)^T \\ \mathbf{p}_3 = (0,0,1,0,0,1,\ldots)^T. \end{cases}$$

Equating $M\mathbf{p}_k = S\mathbf{p}_k$ for $k = 1, 2, 3$ yields:

$$\begin{bmatrix} m_{11} & m_{12} & & & \\ m_{21} & m_{22} & m_{23} & & \\ & m_{32} & m_{33} & m_{34} & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & 0 \\ m_{21} & m_{22} & m_{23} \\ m_{34} & m_{32} & m_{33} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} S\mathbf{p}_1 & S\mathbf{p}_2 & S\mathbf{p}_3 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

All the nonzero entries of the tridiagonal matrix M can be computed explicitly using the above equations. An algorithm for constructing a tridiagonal matrix M of size k is summarized below. For an integer i, we employ the notation:

$$mod(i, 3) \equiv \begin{cases} 1 & \text{if } i = 3k + 1, \text{ for some integer } k \\ 2 & \text{if } i = 3k + 2, \text{ for some integer } k \\ 3 & \text{if } i = 3k, \text{ for some integer } k. \end{cases}$$

Thus, $mod(i, 3)$ denotes the remainder in the division of i by 3.

Algorithm 3.4.5 *(Probe Tridiagonal Approximation of S)*
Input: $S\mathbf{p}_1$, $S\mathbf{p}_2$, $S\mathbf{p}_3$

```
For $i = 1, \cdots, k$ do:
    Let $j = mod(i, 3)$.
    $m_{ii} = (S\mathbf{p}_j)_i$
    If $i < k$ define:
        $\begin{cases} m_{i,i+1} \equiv (S\mathbf{p}_{j+1})_i \\ m_{i+1,i} \equiv (S\mathbf{p}_j)_{i+1} \end{cases}$
    Endif
Endfor
```

Output: Tridiagonal matrix M.

Remark 3.37. As input, this algorithm requires three matrix vector products of the form $S\mathbf{p}_j$ for $j = 1, 2, 3$. These products can be computed without the assembly of S, using the identity (3.20) with $S^{(i)} = A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}$, or based on the identity $S = A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}$. The computational cost of constructing a tridiagonal approximation M of S will essentially be proportional to the cost of computing three matrix-vector products with S.

Remark 3.38. If the Schur complement matrix S is tridiagonal, it is easily verified that the entries M_{ij} of the reconstructed matrix will match S_{ij}. More generally, however, the reconstructed entries M_{ij} will only be approximations of the corresponding entries S_{ij}, due to the nonzero entries of S outside the tridiagonal band. However, if the entries of S decay rapidly outside the tridiagonal band, then this approximation may be reasonably accurate.

Remark 3.39. The reconstructed matrix M in the above algorithm may not be symmetric, i.e., $M_{ij} \neq M_{ji}$. However, given a nonsymmetric tridiagonal matrix M, a symmetric tridiagonal approximation $\tilde{M}$ may be obtained as:

$$\begin{cases} \tilde{M}_{ij} = \max\{M_{ij}, M_{ji}\}, & \text{if } i \neq j \\ \tilde{M}_{ii} = M_{ii}, & \text{if } i = j. \end{cases}$$

Alternatively, a different probing technique [KE7] involving only *two* probe vectors can be employed to construct a symmetric approximation M of S.

The following result concerns a tridiagonal probe approximation based on three probe vectors.

Lemma 3.40. *If S is an M-matrix, then the tridiagonal probe approximation M of S will also be an M-matrix. Furthermore, its symmetrization $\tilde{M}$ will also be an M-matrix. For a model Laplacian on a rectangular grid with periodic boundary conditions on two boundary edges, the condition number of the tridiagonal probe approximation will satisfy cond$(M, S) \leq C\, h^{-1/2}$ in comparison to cond$(S) \leq C\, h^{-1}$.*

Proof. See [CH9]. □

The tridiagonal probing procedure described above can be easily generalized to to band matrices with larger bandwidths. To generalize to other sparsity patterns, however, requires some care. Suppose G denotes the adjacency matrix representing the sparsity pattern desired for M. To construct an approximation M of S with the same sparsity pattern as G, the first step would be to determine a *coloring* or partitioning of the nodes so that nodes of the same color are not adjacent in G. Thus, if node i is adjacent to nodes j and k in G, then nodes j and k cannot be of the same color. Given such a coloring of the nodes, into d colors, define d probe vectors $\mathbf{p}_1, \ldots, \mathbf{p}_d$ so that $\mathbf{p}_j$ is one at all indices corresponding to the j'th color and zero on all other nodes. A reconstruction algorithm may be derived for M using the symmetry of M. Once a sparse approximation M of S has been constructed, it may be necessary to further approximate M by its ILU factorization, to enable efficient solvability of the preconditioner. We omit further details.

We conclude our discussion on algebraic approximation of S, with another approximation based on an *incomplete* factorization of the matrices $A_{II}^{(i)}$. Such approximations will be of interest primarily for multisubdomain decompositions [CA33]. In the two subdomain case, the method employs an incomplete factorization of the subdomain stiffness matrices $A_{II}^{(i)} \approx \tilde{L}_I^{(i)} \tilde{L}_I^{(i)^T}$ for $i = 1, 2$ to compute a low cost *dense* approximation M of S:

$$\begin{cases} S = A_{BB} - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} A_{IB}^{(1)} - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} A_{IB}^{(2)} \\ \quad \approx A_{BB} - A_{IB}^{(1)^T} \tilde{L}_{II}^{(1)^{-T}} \tilde{L}_{II}^{(1)^{-1}} A_{IB}^{(1)} - A_{IB}^{(2)^T} \tilde{L}_{II}^{(2)^{-T}} \tilde{L}_{II}^{(2)^{-1}} A_{IB}^{(2)} \\ \quad \equiv M. \end{cases} \tag{3.67}$$

If the matrix S is of size k, then the cost of constructing M will typically be proportional to $\mathcal{O}(k^2)$. The approximation M will typically be dense. However, sufficiently small entries of M may be truncated to zero using a threshold parameter $\eta > 0$:

$$\tilde{M}_{ij} \equiv \begin{cases} 0, & \text{if } |M_{ij}| \le \eta \left(|M_{ii}| + |M_{jj}|\right) \\ M_{ij}, & \text{otherwise.} \end{cases} \tag{3.68}$$

The use of incomplete factorization to construct a dense approximation M of S, followed by threshold truncation $\tilde{M}$ of M will yield a sparse approximation $\tilde{M}$ of S, which can then be used as a preconditioner [CA33] for S.

Lemma 3.41. *If A is an M-matrix, then the approximation M in (3.67) will also be an M-matrix. Furthermore, if threshold truncation as in (3.68) is applied to the resulting dense matrix M, the truncated approximation $\tilde{M}$ will also be an M-matrix.*

Proof. See [CA33]. □

3.5 Preconditioners in Two Dimensions

The Schur complement matrix S associated with the discretization of elliptic equation (3.1) is typically more difficult to precondition for *multisubdomain* decompositions and in higher dimensions. This difficulty can be attributed to the increasingly complex geometry of the interface B for a multisubdomain decomposition, and to the properties of the Steklov-Poincaré map on B. In the multisubdomain case, the Schur complement matrix will have *zero* block entries corresponding to nodes on disjoint subdomains $\partial\Omega_i \cap \partial\Omega_j = \emptyset$. Furthermore, the entries in the nonzero blocks will decay in magnitude with increasing distance between the nodes. As the size h_0 of each subdomain decreases, the condition number of the multisubdomain Schur complement matrix increases from $\mathcal{O}(h^{-1})$ to $\mathcal{O}(h^{-1}h_0^{-1})$, see [BR24]. However, Schwarz subspace preconditioners employing suitable overlap between blocks of S can be effective (see Chap. 3.7 on Neumann-Neumann preconditioners, in particular).

In this section, we shall describe the block Jacobi, BPS and vertex space preconditioners for a multisubdomain Schur complement matrix S associated with (3.5) on a domain $\Omega \subset \mathbb{R}^2$. Each of the preconditioners we describe will have the structure of an additive *Schwarz subspace* preconditioner from Chap. 2.3, for the space $V = \mathbb{R}^{n_B}$ of nodal vectors on B, endowed with the inner product generated by S. With the exception of the coarse space $V_0 \subset V$, the other subspaces $V_i \subset V$ required to define the Schwarz subspace algorithm for S, will be defined based on a partition of the interface B into subregions $\mathcal{G}_i \subset B$, referred to as *globs*. These globs may be extended to define overlapping or non-overlapping segments on B, but to implement such preconditioners, the submatrices of S must be *approximated* without assembling S. Importantly,

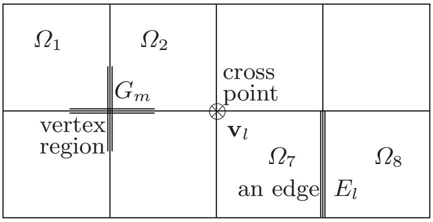


Fig. 3.4. A partition of Ω into 8 subdomains

as $h_0 \to 0^+$, if a coarse space V_0 is included, global transfer of information will be facilitated between the subdomains, and this will reduce the dependence of the condition number on h_0.

Let $\Omega_1, \ldots, \Omega_p$ form a nonoverlapping box type decomposition of $\Omega \subset \mathbb{R}^2$ as in Fig. 3.4, with subdomains of diameter h_0. Consider a finite element discretization of elliptic equation (3.1) with Dirichlet boundary $\mathcal{B}_D = \partial\Omega$. We shall employ the notation:

$$\begin{cases} B^{(i)} & \equiv \partial\Omega_i \backslash \mathcal{B}_D, & \text{for } 1 \le i \le p \\ B & \equiv \cup_{i=1}^p B^{(i)} & \\ B_{ij} & \equiv \text{int}\left(B^{(i)} \cap B^{(j)}\right), & \text{for } 1 \le i,j \le p. \end{cases}$$

Here $\text{int}(B^{(i)} \cap B^{(j)})$ refers to the *interior* of $B^{(i)} \cap B^{(j)}$. Each *connected* and nonempty boundary segment B_{ij} will be referred to as an *edge*. The distinct edges will be enumerated as $E_1, \cdots, E_q$ so that each E_l corresponds uniquely to a nonempty connected segment B_{ij}. Endpoints in B of open segments B_{ij} will be referred to as *vertices* or *cross-points*, see Fig. 3.4, and the collection of all vertices will be denoted $\mathcal{V}$:

$$\mathcal{V} = B \backslash (E_1 \cup \cdots \cup E_q).$$

The term *glob* will refer to subregions of the interface which partition B, see [MA14]. For $\Omega \subset \mathbb{R}^2$, edges and cross-points will be globs.

The interface B arising in the decomposition of a two dimensional domain can be partitioned based on edges and cross-points as follows:

$$B = E_1 \cup \cdots \cup E_q \cup \mathcal{V}.$$

If the indices of nodes on B are grouped and ordered based on the globs $E_1, \ldots, E_q, \mathcal{V}$, with some chosen ordering within each edge E_l and cross-point set $\mathcal{V}$, then the Schur complement matrix can be block partitioned as:

$$S = \begin{bmatrix} S_{E_1E_1} & \cdots & S_{E_1E_q} & S_{E_1\mathcal{V}} \\ \vdots & & \vdots & \vdots \\ S_{E_1E_q}^T & \cdots & S_{E_qE_q} & S_{E_q\mathcal{V}} \\ S_{E_1\mathcal{V}}^T & \cdots & S_{E_q\mathcal{V}}^T & S_{\mathcal{V}\mathcal{V}} \end{bmatrix}. \tag{3.69}$$

Here $S_{E_l E_r}$, $S_{E_l \mathcal{V}}$ and $S_{\mathcal{V}\mathcal{V}}$ denote submatrices of S corresponding to indices in the respective globs.

Since the Schur complement matrix S is not typically assembled in iterative substructuring methodology, the different submatrices of S in (3.69) will also not be assembled explicitly. However, the action of submatrix $S_{E_l E_l}$ on a subvector can be computed explicitly without assembly of $S_{E_l E_l}$ when edge $E_l = B^{(i)} \cap B^{(j)}$. This is because in this case submatrix $S_{E_l E_l}$ will correspond to a *two* subdomain Schur complement matrix, arising from the decomposition of the subregion $\Omega_i \cup \Omega_j \cup E_l$ into the subdomains Ω_i, Ω_j and interface E_l. This observation yields the formal expression:

$$S_{E_l E_l} = A_{E_l E_l} - A_{IE_l}^{(i)^T} A_{II}^{(i)^{-1}} A_{IE_l}^{(i)} - A_{IE_l}^{(j)^T} A_{II}^{(j)^{-1}} A_{IE_l}^{(j)}, \tag{3.70}$$

for submatrix $S_{E_l E_l}$. This may be applied to yield:

$$S_{E_l E_l}^{-1} \mathbf{r}_{E_l} = \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & 0 & A_{IE_l}^{(i)} \\ 0 & A_{II}^{(j)} & A_{IE_l}^{(j)} \\ A_{IE_l}^{(i)^T} & A_{IE_l}^{(j)^T} & A_{E_l E_l} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{r}_{E_l} \end{bmatrix}, \tag{3.71}$$

so that the action of $S_{E_l E_l}^{-1}$ on a subvector can be computed at the cost of solving the preceding linear system. More generally, by applying property (3.26) of Schur complement matrices, it can be noted that when edges E_l and E_k belong to a common subdomain boundary $B^{(i)}$, the block submatrices $S_{E_l E_k}$ will be nonzero and dense. Otherwise, the submatrices $S_{E_l E_k}$ will be zero. The block submatrices $S_{E_l \mathcal{V}}$ and $S_{\mathcal{V}\mathcal{V}}$ will typically have nonzero entries since there will be nodes in E_l adjacent to nodes in $\mathcal{V}$. We now describe preconditioners.

3.5.1 Block Jacobi Preconditioner

In two dimensions, a block Jacobi preconditioner for S can be defined based on the partition of B into the globs $E_1, \ldots, E_q$ and $\mathcal{V}$. In matrix form, such a preconditioner will correspond to the block diagonal of matrix (3.69):

$$M = \begin{bmatrix} S_{E_1 E_1} & & & 0 \\ & \ddots & & \\ & & S_{E_q E_q} & \\ 0 & & & S_{\mathcal{V}\mathcal{V}} \end{bmatrix}.$$

The action of the inverse of the block Jacobi preconditioner satisfies:

$$M^{-1} \equiv \sum_{i=1}^{p} \mathcal{R}_{E_i}^T S_{E_i E_i}^{-1} \mathcal{R}_{E_i} + \mathcal{R}_{\mathcal{V}}^T S_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}}, \tag{3.72}$$

using the interface restriction and extension matrices $\mathcal{R}_G$ and $\mathcal{R}_G^T$ defined in (3.19) between nodes on B and nodes on $G = E_l$ or $G = \mathcal{V}$. Since the Schur complement matrix S will not be assembled, the diagonal blocks $S_{E_l E_l} = \mathcal{R}_{E_l} S \mathcal{R}_{E_l}^T$ and $S_{\mathcal{V}\mathcal{V}} = \mathcal{R}_{\mathcal{V}} S \mathcal{R}_{\mathcal{V}}^T$ of S must typically be approximated, or alternatively, the action of the inverses of the submatrices $S_{E_l E_l}^{-1}$ and $S_{\mathcal{V}\mathcal{V}}^{-1}$ must be approximated. We outline how such approximations can be obtained.

Approximation of $S_{E_l E_l}$. If there are n_{E_l} nodes on E_l then $S_{E_l E_l}$ will be of size n_{E_l}. Since block submatrix $S_{E_l E_l}$ corresponds to a two subdomain Schur complement matrix by (3.70) when edge $E_l = B^{(i)} \cap B^{(j)}$, the action of $S_{E_l E_l}^{-1}$ on a vector can be computed exactly using (3.71). This does not require assembly of $S_{E_l E_l}$. Alternate approximations $M_{E_l E_l}$ of $S_{E_l E_l}$ can be obtained by employing any two subdomain preconditioner for $S_{E_l E_l}$. Choices of such preconditioners include Dirichlet-Neumann, Neumann-Neumann, fractional Sobolev norm, FFT based or algebraic approximation based preconditioners. Such preconditioners must be scaled based on the coefficient $a(x)$ within the subdomains Ω_i and Ω_j.

Approximation of $S_{\mathcal{V}\mathcal{V}}$. If there are $n_{\mathcal{V}}$ vertices in $\mathcal{V}$ then $S_{\mathcal{V}\mathcal{V}}$ will be of size $n_{\mathcal{V}}$. The block submatrix $S_{\mathcal{V}\mathcal{V}}$ can typically be approximated by a diagonal matrix based on the following heuristics. When Ω is a rectangular domain, and the subdomains are rectangular boxes, and a five point stencil is employed, it can easily be verified that matrix $S_{\mathcal{V}\mathcal{V}}$ will be identical to the submatrix $A_{\mathcal{V}\mathcal{V}}$ of stiffness matrix A. This is a consequence of the property that for five point stencils, the interior solution in a rectangular subdomain will not depend on the nodal value on corner vertices. This observation, heuristically suggests replacing $S_{\mathcal{V}\mathcal{V}}$ by $M_{\mathcal{V}\mathcal{V}} = A_{\mathcal{V}\mathcal{V}}$. The latter is easily seen to be diagonal.

Due to its block diagonal structure, the block Jacobi preconditioner M ignores coupling between distinct edges and between the edges and the vertex set $\mathcal{V}$. As a result, the block Jacobi preconditioner does not globally exchange information between the different subdomains, and this results in a non-optimal convergence rate as $h_0 \to 0$. We assume that the grid in quasi-uniform.

Theorem 3.42. *If M is the block Jacobi preconditioner and the subdomains are of diameter h_0, then there exists $C > 0$ independent of h_0 and h:*

$$\mathrm{cond}(M, S) \le C h_0^{-2} \left(1 + \log^2(h_0/h)\right).$$

Proof. See [BR12, DR14, DR10]. □

3.5.2 BPS Preconditioner

As with the block Jacobi preconditioner, the BPS preconditioner [BR12] also has the structure of a matrix additive Schwarz preconditioner for S. Formally, this preconditioner can be obtained by replacing the *local* residual correction term $\mathcal{R}_{\mathcal{V}}^T S_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}}$ on the vertices $\mathcal{V}$ in the block Jacobi preconditioner (3.72) by a *global* coarse space residual correction term of the form $\mathcal{R}_0^T S_0^{-1} \mathcal{R}_0$. We shall define $\mathcal{R}_0$ in (3.74), however, given $\mathcal{R}_0$, matrix $S_0 \equiv \mathcal{R}_0 S \mathcal{R}_0^T$, and the action of the inverse of the BPS preconditioner will have the form:

$$M^{-1} \equiv \sum_{i=1}^{q} \mathcal{R}_{E_l}^T S_{E_l E_l}^{-1} \mathcal{R}_{E_l} + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0. \tag{3.73}$$

Unlike the restriction matrices $\mathcal{R}_{E_l}$ and $\mathcal{R}_{\mathcal{V}}$ onto E_l and $\mathcal{V}$, respectively, which have zero-one entries, the matrix $\mathcal{R}_0$ whose row space defines the coarse space, will not be a matrix with zero-one entries. As a result, the matrix $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$ will not be a submatrix of S. When the region of *support* of Range$(\mathcal{R}_0^T)$ covers Ω, the residual correction term $\mathcal{R}_0^T S_0^{-1} \mathcal{R}_0$ in the BPS preconditioner will transfer information *globally* between the different subdomains. Heuristically, this can help reduce the dependence of the condition number of the preconditioned Schur complement matrix on h_0.

In applications, the coarse space restriction matrix $\mathcal{R}_0$ is usually defined when the subdomains $\Omega_1, \ldots, \Omega_p$ form a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω with elements of size h_0 and nodes corresponding to the vertices in $\mathcal{V}$. In this case, we enumerate the vertices in $\mathcal{V}$ as $v_1, \ldots, v_{n_0}$, and denote by $\phi_1^{h_0}(x), \ldots, \phi_{n_0}^{h_0}(x)$ the coarse grid nodal basis functions associated with these vertices. Suppose the nodes on B are enumerated as $x_1, \cdots, x_{n_B}$, where n_B denotes the number of nodes on B. Then, the coarse space restriction matrix $\mathcal{R}_0$ is defined as the following $n_0 \times n_B$ matrix:

$$\mathcal{R}_0 \equiv \begin{bmatrix} \phi_1^{h_0}(x_1) & \cdots & \phi_1^{h_0}(x_{n_B}) \\ \vdots & & \vdots \\ \phi_{n_0}^{h_0}(x_1) & \cdots & \phi_{n_0}^{h_0}(x_{n_B}). \end{bmatrix}. \tag{3.74}$$

Its transpose $\mathcal{R}_0^T$ of size $n_B \times n_0$ is an interpolation onto nodal values on B. As with the block Jacobi preconditioner, suitable approximations of the matrices $S_{E_l E_l} = \mathcal{R}_{E_l} S \mathcal{R}_{E_l}^T$ and $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$ must be employed in the BPS preconditioner (3.73), since the Schur complement matrix S is not assembled. Below, we indicate various such approximations [BR12].

Approximation of $\boldsymbol{S_{E_l E_l}}$. The submatrix $S_{E_l E_l}$ in (3.73) can be replaced by any suitable two subdomain Schur complement preconditioner $M_{E_l E_l}$ just as for the block Jacobi preconditioner (3.72). In the original BPS algorithm, $S_{E_l E_l}$ was approximated by a preconditioner of the form $(a^{(i)} + a^{(j)}) Q_l D_l Q_l^T$ where Q_l was a discrete sine transform of size n_{E_l}, and D_l was a suitably chosen diagonal matrix from (3.64), with $a^{(k)}$ corresponding to an evaluation of coefficient $a(x)$ at some point in Ω_k.

Approximation of $\boldsymbol{S_0}$. The matrix $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$ associated with the coarse space is typically approximated by a coarse grid stiffness matrix A_0 obtained by discretizing the underlying elliptic equation (3.1) on the coarse grid. Below, we heuristically indicate why such an approximation can be employed. Consider a Poisson problem on a rectangular domain Ω partitioned into subdomains Ω_i which form a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω of size h_0. Let $\phi_l^{h_0}(x)$ denote the coarse grid finite element nodal basis centered at vertex v_l. Then, the entries $(A_0)_{ij}$ of the coarse space stiffness matrix A_0 will satisfy:

$$(A_0)_{ij} = \mathcal{A}\left(\phi_i^{h_0}, \phi_j^{h_0}\right).$$

Let $\mathbf{u}_B$ and $\mathbf{w}_B$ denote the nodal vectors representing the coarse space nodal basis functions $\phi_i^{h_0}(x)$ and $\phi_j^{h_0}(x)$ on B. Then, the vector representation of $\phi_i^{h_0}(x)$ and $\phi_j^{h_0}(x)$ on Ω will be given by $\left((E\mathbf{u}_B)^T, \mathbf{u}_B^T\right)^T$ and $\left((E\mathbf{w}_B)^T, \mathbf{w}_B^T\right)^T$ where $E \equiv -A_{II}^{-1}A_{IB}$ denotes the discrete harmonic extension map from B into the interior $\cup_{i=1}^p \Omega_i$ of the subdomains. This holds because each coarse grid function $\phi_l^{h_0}(x)$ is linear within each subdomain, and thus also harmonic (and discrete harmonic) within each subdomain. Consequently, by (3.27), it will hold that:

$$(S_0)_{ij} = \mathbf{w}_B^T S \mathbf{u}_B = \begin{bmatrix} E\mathbf{w}_B \\ \mathbf{w}_B \end{bmatrix}^T A \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} = \mathcal{A}(\phi_i^{h_0}, \phi_j^{h_0}) = (A_0)_{ij}\,.$$

This yields that $A_0 = S_0$ for this geometry and choice of coefficients $a(x)$. More generally, matrix A_0 may be employed as an *approximation* of S_0. The following result concerns the condition number of the BPS preconditioner.

Theorem 3.43. *Let $\mathcal{T}_h(\Omega)$ denote a quasiuniform triangulation of Ω and let the subdomains $\Omega_1, \ldots, \Omega_p$ form a coarse triangulation of Ω of size h_0. Then, there exists $C > 0$ independent of h_0 and h such that:*

$$\mathrm{cond}(M, S) \leq C\,\left(1 + \log^2(h_0/h)\right).$$

If $a(\cdot)$ is constant within each Ω_i, then C will also be independent of $a(\cdot)$.

Proof. See [BR12, DR14, DR10]. □

3.5.3 Vertex Space Preconditioner for S

From a theoretical viewpoint, the logarithmic growth factor $\left(1 + \log^2(h_0/h)\right)$ in the condition number of the BPS preconditioner arises because the BPS preconditioner does not approximate the coupling in S between different edges E_l of B. The vertex space preconditioner extends the BPS preconditioner by including local residual correction terms based on *overlapping* segments of B, see [SM3]. It includes a local correction term of the form $\mathcal{R}_{G_l}^T S_{G_l G_l}^{-1} \mathcal{R}_{G_l}$ involving nodal unknowns on regions $G_l \subset B$, referred to as *vertex regions.* For each vertex $\mathbf{v}_l \in \mathcal{V}$, a vertex region G_l is a star shaped *connected* subset of B that contains segments of length $\mathcal{O}(h_0)$ of all edges E_r emanating from vertex v_l. By construction, each local residual correction term approximates coupling in S between edges adjacent to that vertex.

Formally, the vertex space preconditioner is obtained by adding the terms $\mathcal{R}_{G_l}^T S_{G_l G_l}^{-1} \mathcal{R}_{G_l}$ to the BPS preconditioner (3.73), yielding:

$$M^{-1} = \sum_{l=1}^{q} \mathcal{R}_{E_l}^T S_{E_l E_l}^{-1} \mathcal{R}_{E_l} + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0 + \sum_{i=1}^{n_0} \mathcal{R}_{G_i}^T S_{G_i G_i}^{-1} \mathcal{R}_{G_i}. \tag{3.75}$$

The resulting preconditioner has the structure of a matrix additive Schwarz preconditioner for S based on the *overlapping* decomposition:

$$B = (E_1 \cup \ldots \cup E_q) \cup \mathcal{V} \cup (G_1 \cup \ldots \cup G_{n_0}),$$

of interface B, with an additional coarse space correction term. In practice, it will be convenient to construct each vertex region G_l as the intersection of interface B with a subdomain $\Omega_{v_l} \supset v_l$ of diameter $\mathcal{O}(h_0)$ centered at v_l, see [NE3, MA37, SM3]. By construction, each G_l will be a cross shaped or star shaped subregion of B, see Fig. 3.4, and restriction matrix $\mathcal{R}_{G_l}$ will map a nodal vector on B to its subvector corresponding to nodes in G_l. Matrix $\mathcal{R}_{G_l}$ will be of size $n_{G_l} \times n_B$ when there are n_{G_l} nodes on vertex region G_l, and have entries which are zero or one, as defined by (3.19). Consequently, each matrix $S_{G_l G_l} = \mathcal{R}_{G_l} S \mathcal{R}_{G_l}^T$ will be a submatrix of S of size n_{G_l} corresponding to indices of nodes in G_l. The vertex space preconditioner can be implemented like the BPS preconditioner. Since matrix S is generally not assembled, the matrices $S_{E_l E_l}$, $S_{G_i G_i}$ and S_0 must be appropriately approximated to implement the preconditioner. The matrices $S_{E_l E_l}$ and S_0 can be approximated as described for the BPS preconditioner, since these terms will be identical to those in (3.73). Below, we focus on the action of $S_{G_i G_i}^{-1}$.

Approximation of $\boldsymbol{S_{G_i G_i}}$. Let $\Omega_{v_i} \subset \Omega$ denote a subregion used to define the vertex region $G_i \equiv B \cap \Omega_{v_i}$. Partition the nodes of $\mathcal{T}_h(\Omega)$ in Ω_{v_i} into those in $D_i \equiv \Omega_{v_i} \backslash G_i$ and those in G_i. This will induce a block partitioning of the submatrix $A^{(\Omega_{v_i})}$ of stiffness matrix A corresponding to all nodes in Ω_{v_i}:

$$A^{(\Omega_{v_i})} = \begin{bmatrix} A_{D_i D_i} & A_{D_i G_i} \\ A_{D_i G_i}^T & A_{G_i G_i} \end{bmatrix}.$$

Using the above block partitioned matrix, one may approximate the action of $S_{G_i G_i}^{-1}$ on a vector $\mathbf{r}_{G_i}$ as follows:

$$S_{G_i G_i}^{-1} \mathbf{r}_{G_i} \approx \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{D_i D_i} & A_{D_i G_i} \\ A_{D_i G_i}^T & A_{G_i G_i} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \mathbf{r}_{G_i} \end{bmatrix}.$$

Alternatively, sparse approximations of $S_{G_i G_i}$ can be computed efficiently using the probing technique, by weighted sums of FFT based matrices, or by use of inexact factorizations see [CH12, CA33]. The following convergence bound will hold for the vertex space preconditioner.

Theorem 3.44. *If the diameter of vertex subregions is $\beta\, h_0$, then the condition number of the vertex space preconditioned system will satisfy:*

$$\mathrm{cond}(M, S) \leq C_0 (1 + \beta^{-1}),$$

where $C_0 > 0$ is independent of h_0, h and β, but may depend on the variation of $a(\cdot)$. There also exists a constant C_1 independent of h_0, h, and the jumps in $a(\cdot)$ (provided $a(x)$ is constant on each subdomain Ω_i).

$$\mathrm{cond}(M, S) \leq C_1 \, \left(1 + \log^2(h_0/h)\right).$$

Proof. See [SM, DR10]. □

Thus, in the presence of large jumps in the coefficient $a(\cdot)$, the bounds for the condition number of the vertex space algorithm can deteriorate to $\left(1 + \log^2(h_0/h)\right)$, which is the same growth as for the BPS preconditioner.

3.6 Preconditioners in Three Dimensions

The Schur complement matrix S for a three dimensional multi-subdomain decomposition is more difficult to precondition than in two dimensions. This difficulty arises due to the more complex geometry of the interface B in three dimensions. However, effective preconditioners can be constructed (see in particular, Chap. 3.7 on Neumann-Neumann preconditioners) by employing Schwarz subspace methods with more overlap between blocks of S.

Our discussion of three dimensional preconditioners will focus on several block Jacobi preconditioners for S, a vertex space preconditioner, and a parallel wirebasket preconditioner. We consider a decomposition of $\Omega \subset \mathbb{R}^3$ into p non-overlapping box type or tetrahedral subdomains $\Omega_1, \cdots, \Omega_p$. having diameter h_0. Typically, these subdomains will be assumed to form a quasi-uniform coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω. For the Dirichlet boundary value problem (3.1) with $\mathcal{B}_D = \partial\Omega$, we let $B^{(i)} \equiv \partial\Omega_i \backslash \mathcal{B}_D$ denote the non-Dirichlet segment of $\partial\Omega_i$, and define the interface as $B = B^{(1)} \cup \cdots \cup B^{(p)}$.

The different additive Schwarz matrix preconditioners we shall consider for S will be based on a decomposition of the interface B into the following subregions of B referred to as *globs* [MA14], They will typically be well defined for tetrahedral or box type subdomains:

$$\begin{cases} F_{ij} & \equiv \mathrm{int}\left(B^{(i)} \cap B^{(j)}\right) \\ W^{(i)} & \equiv B^{(i)} \cap (\cup_{j \neq i} \partial F_{ij}) \\ W & \equiv W^{(1)} \cup \cdots \cup W^{(p)}. \end{cases}$$

Here $F_{ij} = \mathrm{int}\left(B^{(i)} \cap B^{(j)}\right)$ denotes the *interior* of region $B^{(i)} \cap B^{(j)}$, and is referred to as a *face* of Ω_i when it is nonempty. By definition, each face will be two dimensional. The subregion $W^{(i)}$ of $B^{(i)}$ is referred to as a *local*

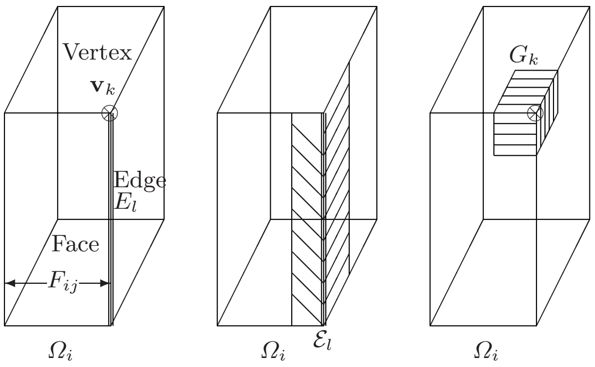


Fig. 3.5. Boundary segments and vertex regions for three dimensional subdomains

wirebasket of Ω_i and is the union of the boundaries ∂F_{ij} of all faces of Ω_i. Typically, each wirebasket will be connected and the union of several one dimensional segments. The union of all local wirebaskets is referred to as the global *wirebasket*. The above mentioned subregions are indicated in Fig. 3.5 for an individual subdomain.

In applications, we shall enumerate all the faces in B as $F_1, \ldots, F_q$ where q denotes the total number of faces. By definition, each face F_l will correspond uniquely to some nonempty intersection int $\left(B^{(i)} \cap B^{(j)}\right)$, and by construction, we may partition the interface B into the following globs:

$$B = F_1 \cup \cdots \cup F_q \cup W.$$

In practice, it will be convenient to decompose the wirebaskets into smaller globs. We define an *edge* as a maximal line segment of a local wirebasket, homeomorphic to an open interval. We define *vertices* as endpoints of edges.

Edges and vertices can be expressed formally as:

$$\begin{cases} E_{ijk} \equiv int\left(\overline{F}_{ij} \cap \overline{F}_{ik}\right) \\ \mathcal{V} \quad\;\; \equiv W \backslash \left(\cup_{i,j,k} E_{ijk}\right) \\ \qquad\; = \{v_l \,:\, v_l \in \mathcal{V}\}. \end{cases}$$

By definition each edge will be open, and we enumerate all the nonempty edges as $E_1, \ldots, E_r$ where r denotes the total number of such edges. Similarly, we enumerate all the vertices as $v_1, \ldots, v_{n_0}$, where n_0 will denote the total number of vertices in W. The collection of all vertices will be denoted $\mathcal{V}$, as in two dimensions.

Definition 3.45. *Given a glob $G \subset B$ containing n_G nodes of Ω_h, we let $\mathcal{R}_G \equiv R_G R_B^T$ denote a restriction matrix of size $n_B \times n_G$ which restricts a nodal vector on B to a subvector corresponding to nodes on G, as defined in (3.19). Its transpose $\mathcal{R}_G^T \equiv R_B R_G^T$ will extend a vector of nodal values on G to a vector of nodal values on B (extension by zero). The entries of these glob based restriction and extension matrices will be zeros or ones.*

In the three dimensional case, we will on occasion employ an additional restriction map, which we shall denote as $\mathcal{R}_{W^{(i)}W}$.

Definition 3.46. *Let $\mathcal{R}_{W^{(i)}W} \equiv R_{W^{(i)}} R_W^T$ denote the matrix which restricts a vector of nodal values on the global wirebasket W into a subvector of nodal values on the local wirebasket $W^{(i)}$. Its transpose $\mathcal{R}_{W^{(i)}W}^T \equiv R_W R_{W^{(i)}}^T$ will extend a vector of nodal values on $W^{(i)}$ to a vector of nodal values on W (extension by zero).*

3.6.1 Block Jacobi Preconditioner for S

We first describe a block Jacobi preconditioner based on the decomposition of B into the faces $F_1, \ldots, F_q$ and the wirebasket W:

$$B = F_1 \cup \cdots \cup F_q \cup W.$$

This nonoverlapping decomposition of B induces a block partition of S as:

$$S = \begin{bmatrix} S_{F_1F_1} & \cdots & S_{F_1F_q} & S_{F_1W} \\ \vdots & \ddots & \vdots & \vdots \\ S_{F_1F_q}^T & \cdots & S_{F_qF_q} & S_{F_qW} \\ S_{F_1W}^T & \cdots & S_{F_qW}^T & S_{WW} \end{bmatrix},$$

corresponding to indices of nodes within the chosen subregions of B. If n_{G_l} denotes the number of nodes on glob G_l, then $S_{G_iG_j}$ will denote a submatrix of S of size $n_{G_i} \times n_{G_j}$ corresponding to the nodes on glob G_i and G_j.

The block Jacobi preconditioner will be the block diagonal part of S:

$$M = \begin{bmatrix} S_{F_1F_1} & & & 0 \\ & \ddots & & \\ & & S_{F_qF_q} & \\ 0 & & & S_{WW} \end{bmatrix}.$$

In terms of restriction and extension matrices, the action of the inverse M^{-1} the block Jacobi preconditioner will be:

$$M^{-1} = \sum_{l=1}^{q} \mathcal{R}_{F_l}^T S_{F_lF_l}^{-1} \mathcal{R}_{F_l} + \mathcal{R}_W^T S_{WW}^{-1} \mathcal{R}_W, \tag{3.76}$$

where $S_{F_lF_l} = \mathcal{R}_{F_l} S \mathcal{R}_{F_l}^T$ and $S_{WW} = \mathcal{R}_W S \mathcal{R}_W^T$ are submatrices of S corresponding to indices in F_l and W. As with the other Schwarz preconditioners for S, in practice the submatrices $S_{F_lF_l}$ and S_{WW} of S must be replaced by suitable approximations since S will typically not be assembled. Various alternative approximations may be chosen for such approximations.

Approximation of $S_{F_l F_l}$. If $F_l = \text{int}\left(B^{(i)} \cap B^{(j)}\right)$, then $S_{F_l F_l}$ will correspond to a two subdomain Schur complement associated with the partition of $\Omega_i \cup F_l \cup \Omega_j$ into Ω_i and Ω_j. Consequently, the action $S_{F_l F_l}^{-1} \mathbf{r}_{F_l}$ may be computed *exactly* as follows:

$$S_{F_l F_l}^{-1} \mathbf{r}_{F_l} = \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & 0 & A_{IF_l}^{(i)} \\ 0 & A_{II}^{(j)} & A_{IF_l}^{(j)} \\ A_{IF_l}^{(i)^T} & A_{IF_l}^{(j)^T} & A_{F_l F_l} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{r}_{F_l} \end{bmatrix}, \tag{3.77}$$

where the above blocks are submatrices of A corresponding to indices in Ω_i, Ω_j and F_l. Alternatively, a Dirichlet-Neumann preconditioner may be employed, for instance based on subdomain Ω_i:

$$S_{F_l F_l}^{(i)^{-1}} \mathbf{r}_{F_l} \approx \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & A_{IF_l}^{(i)} \\ A_{IF_l}^{(i)^T} & A_{F_l F_l}^{(i)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_{F_l} \end{bmatrix}.$$

A Neumann-Neumann preconditioner will also approximate $S_{F_l F_l}$.

If the triangulation of F_l induced by Ω_h can be mapped bijectively into a rectangular grid, then we may employ an FFT based preconditioner of the form $S_{F_l F_l} \approx (a^{(i)} + a^{(j)}) Q D Q^T$ where Q is a two dimensional fast sine transform and D is a diagonal matrix approximating the eigenvalues of a reference Schur complement matrix $\hat{S}_{F_l F_l}$ associated with a three dimensional cubical domain partitioned into two strips [RE]. Here $a^{(l)}$ denotes the coefficient $a(x)$ evaluated at a sample point of Ω_l. Alternative preconditioners for $S_{F_l F_l}$ may be obtained using algebraic approximation of $S_{F_l F_l}$ based on generalization of the tridiagonal probing procedure [KE7, CH9] or ILU [CA33].

Approximation of S_{WW}. An approximation of $S_{WW} = \mathcal{R}_W S \mathcal{R}_W^T$ can be based on the following heuristic observation. When Ω is rectangular, and the subdomains are boxes, and a seven point stencil is used for the finite element discretization of (3.1), then matrix $S_{WW} = A_{WW}$. This can be verified by using the property that for seven point stencils the nodal values on the wirebasket will not influence the interior Dirichlet solution in a box subdomain. As a consequence, the piecewise discrete harmonic extension of nonzero nodal values on W and zero nodal values on $B \backslash W$ will be zero in the interior of the subdomains. The desired property that $S_{WW} = A_{WW}$ will now follow from (3.26). When the geometry of the subdomains is more general, A_{WW} may still be used as an approximation of S_{WW}. Replacing the submatrices $S_{F_l F_l}$ and S_{WW} by the preceding approximations will yield an approximate block Jacobi preconditioner for S.

Remark 3.47. Efficient sparse solvers may be employed to solve systems of the form $A_{WW} \mathbf{u}_W = \mathbf{r}_W$, since A_{WW} will typically be sparse. Indeed, for the seven point stencil, at most seven entries of the form $(A_{WW})_{ij}$ will be nonzero when $\mathbf{x}_i \in \mathcal{V}$, while at most three entries of the form $(A_{WW})_{ij}$ will be nonzero when $\mathbf{x}_i \in E_l$. However, the band width of A_{WW} will depend on the ordering of nodes within W. In practice, the wirebasket W can be further decomposed into the edges $E_1, \ldots, E_r$ and the vertex set $\mathcal{V}$, and the action of $\mathcal{R}_W^T S_{WW}^{-1} \mathcal{R}_W$ can be approximated by the following matrix additive Schwarz preconditioner:

$$\mathcal{R}_W^T S_{WW}^{-1} \mathcal{R}_W \approx \sum_{l=1}^{r} \mathcal{R}_{E_l}^T A_{E_l E_l}^{-1} \mathcal{R}_{E_l} + \mathcal{R}_{\mathcal{V}}^T A_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}}.$$

A variant of the block Jacobi preconditioner employs such an approximation. The following result concerns the convergence rate associated with (3.76).

Lemma 3.48. *The condition number of the Schur complement matrix preconditioned by the block Jacobi preconditioner (3.76) satisfies:*

$$\operatorname{cond}(M, S) \leq C h_0^{-2} \left(1 + \log(h_0/h)\right)^2,$$

for some $C > 0$ independent of h and h_0.

Proof. See [BR15, DR10]. □

As the preceding theorem indicates, the convergence rate of block Jacobi preconditioner (3.76) for S deteriorates as the subdomain sizes h_0 becomes small. This deterioration arises primarily because this block Jacobi preconditioner exchanges information only locally for the chosen diagonal blocks in the block partition of S. This convergence rate, however, can be improved by including some global transfer of information.

We next describe two variants of the block Jacobi preconditioner (3.76) incorporating coarse space correction [DR10]. To obtain the first variant, we substitute the approximation:

$$\mathcal{R}_W^T S_{WW}^{-1} \mathcal{R}_W \approx \sum_{l=1}^{r} \mathcal{R}_{E_l}^T S_{E_l E_l}^{-1} \mathcal{R}_{E_l} + \mathcal{R}_{\mathcal{V}}^T S_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}},$$

into (3.76) and replace the *local* correction term $\mathcal{R}_{\mathcal{V}}^{-1} S_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}}$ on the vertices $\mathcal{V}$ by a coarse space correction term $\mathcal{R}_0^{-1} S_0^{-1} \mathcal{R}_0$ to obtain the preconditioner:

$$M^{-1} = \sum_{l=1}^{q} \mathcal{R}_{F_l}^T S_{F_l F_l}^{-1} \mathcal{R}_{F_l} + \sum_{l=1}^{r} \mathcal{R}_{E_l}^T S_{E_l E_l}^{-1} \mathcal{R}_{E_l} + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0. \tag{3.78}$$

Here, the coarse space restriction matrix $\mathcal{R}_0$ is defined analogous to (3.74), with coarse grid nodal basis functions $\phi_i^{h_0}(x)$ corresponding to each vertex $\mathbf{v}_i \in \mathcal{V}$, and $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$. A second variant is obtained by adding the local correction term $\mathcal{R}_{\mathcal{V}}^{-1} S_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}}$ yielding:

$$M^{-1} = \sum_{l=1}^{q} \mathcal{R}_{F_l}^T S_{F_l F_l}^{-1} \mathcal{R}_{F_l} + \sum_{l=1}^{r} \mathcal{R}_{E_l}^T S_{E_l E_l}^{-1} \mathcal{R}_{E_l} + \mathcal{R}_{\mathcal{V}}^{-1} S_{\mathcal{V}\mathcal{V}}^{-1} \mathcal{R}_{\mathcal{V}} + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0. \tag{3.79}$$

The resulting preconditioners satisfy the following bounds.

Theorem 3.49. *The preconditioner M in (3.78) satisfies the bound:*

$$\operatorname{cond}(M, S) \le C_1 \frac{h_0}{h} (1 + \log(h_0/h))^2,$$

while the bound for the preconditioner M in (3.79) satisfies:

$$\operatorname{cond}(M, S) \le C_2 (1 + \log(h_0/h))^2,$$

where C_1 is independent of h_0, h and jumps in the coefficient $a(x)$, while C_2 is independent of h_0 and h, but may depend on the coefficient $a(x)$.

Proof. See [DR10]. □

As with the other matrix Schwarz preconditioners for S, the submatrices $S_{F_iF_i}$, $S_{E_lE_l}$, $S_{\mathcal{V}\mathcal{V}}$ and S_0 must be replaced by suitable approximations since S is not assembled in practice. As we have already described approximations of $S_{F_iF_i}$, we shall only focus on the other terms.

Approximation of $\boldsymbol{S_{E_lE_l}}$. To obtain a heuristic approximation of $S_{E_lE_l}$, we approximate $S_{WW} \approx A_{WW}$ as described earlier to obtain $S_{E_lE_l} \approx A_{E_lE_l}$. It is easily verified that the edge matrix $A_{E_lE_l}$ will be well conditioned and may effectively be replaced by a suitably scaled multiple of the identity matrix:

$$S_{E_lE_l} \approx h\, \sigma_{E_l} I_{E_l},$$

where σ_{E_l} represents the average of the coefficients $a(\cdot)$ in the subdomains adjacent to edge E_l. For finite difference schemes, the scaling factor for $S_{E_lE_l}$ must be proportional to h^{-2} instead of h.

Approximation of $\boldsymbol{S_{\mathcal{V}\mathcal{V}}}$. To obtain an approximation of $S_{\mathcal{V}\mathcal{V}}$, again we employ the approximation $S_{WW} \approx A_{WW}$ to obtain $S_{\mathcal{V}\mathcal{V}} \approx A_{\mathcal{V}\mathcal{V}}$. The submatrix $A_{\mathcal{V}\mathcal{V}}$ will also be diagonal, and may be approximated as follows:

$$(S_{\mathcal{V}\mathcal{V}})_{ii} \approx h\, \sigma_i,$$

for finite element discretizations, where σ_i denotes a suitably weighted average of the coefficients $a(\cdot)$ in subdomains adjacent to vertex $\mathbf{v}_i$. For finite difference discretizations, the scaling factor must be h^{-2} instead of h.

Approximation of $\boldsymbol{S_0}$. The coarse space matrix $S_0 = \mathcal{R}_0 A \mathcal{R}_0^T$ can be approximated by A_0 as in two dimensions.

Remark 3.50. For *smooth* coefficients, preconditioner (3.79) will yield better convergence than preconditioner (3.78), due to elimination of (h_0/h).

3.6.2 Vertex Space Preconditioner for S

The different variants of the block Jacobi preconditioner are nonoptimal. This arises due to the elimination of the off diagonal blocks in S. The vertex space preconditioner [SM3] for S, incorporates some of this coupling by including subspace correction terms on overlapping globs containing segments of faces adjacent to each vertex v_l and to each edge E_l, yielding improved bounds.

The three dimensional vertex space preconditioner is based on an *overlapping* extension of the following partition of B:

$$B = (F_1 \cup \cdots \cup F_q) \cup (E_1 \cup \cdots , E_r) \cup (v_1 \cup \cdots \cup v_{n_0}).$$

Each edge E_l is extended to a glob $\mathcal{E}_l$ which includes segments of all faces adjacent to this edge. Formally, a cylindrical subdomain $\Omega_{E_l} \supset E_l$ of width $\mathcal{O}(h_0)$ is employed to define:

$$\mathcal{E}_l \equiv \Omega_{E_l} \cap B, \qquad \text{for} \quad 1 \le l \le r,$$

see Fig. 3.5 for segments of $\mathcal{E}_l$ within a subdomain Ω_i. Similarly, each vertex v_k is extended to a glob G_k of width $\mathcal{O}(h_0)$ containing segments of all faces adjacent to vertex v_k. Formally, a domain $\Omega_{v_k} \supset v_k$ of size $\mathcal{O}(h_0)$ centered about vertex v_k is employed to define glob G_k:

$$G_k \equiv B \cap \Omega_{v_k}, \qquad \text{for} \quad 1 \le k \le n_0,$$

see [SM3, MA38]. A section of glob G_k restricted to subdomain Ω_i is illustrated in Fig. 3.5. The overlapping decomposition of B employed in the vertex space preconditioner can be expressed in terms of F_l, $\mathcal{E}_i$ and G_r:

$$B = (F_1 \cup \cdots \cup F_q) \cup (\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_r) \cup (G_1 \cup \cdots \cup G_{n_0}).$$

Additionally, a coarse space correction term based on a coarse space is employed. Corresponding to each glob F_l, $\mathcal{E}_i$ and G_k, we define the restriction maps $\mathcal{R}_{F_l}$, $\mathcal{R}_{\mathcal{E}_i}$ and $\mathcal{R}_{G_k}$ which restrict a vector of nodal values on B to the nodes on F_l, $\mathcal{E}_i$ and G_k, respectively. Such restriction maps are defined by (3.19) with zero-one entries so that $S_{F_lF_l} = \mathcal{R}_{F_l} S \mathcal{R}_{F_l}^T$, $S_{\mathcal{E}_i\mathcal{E}_i} = \mathcal{R}_{\mathcal{E}_i} S \mathcal{R}_{\mathcal{E}_i}^T$ and $S_{G_kG_k} = \mathcal{R}_{G_k} S \mathcal{R}_{G_k}^T$ are submatrices of S corresponding to indices of nodes on F_l, $\mathcal{E}_i$ and G_k. Additionally, $\mathcal{R}_0$ will denote a coarse space matrix defined by (3.74). The action M^{-1} of the vertex space preconditioner is then:

$$M^{-1} = \sum_{l=1}^{q} \mathcal{R}_{F_l}^T S_{F_lF_l}^{-1} \mathcal{R}_{F_l} + \sum_{i=1}^{r} \mathcal{R}_{\mathcal{E}_i}^T S_{\mathcal{E}_i\mathcal{E}_i}^{-1} \mathcal{R}_{\mathcal{E}_i} + \sum_{k=1}^{n_0} \mathcal{R}_{G_k}^T S_{G_kG_k}^{-1} \mathcal{R}_{G_k} + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0. \tag{3.80}$$

As with the other matrix Schwarz preconditioners for S, the matrices $S_{F_lF_l}$, $S_{\mathcal{E}_i\mathcal{E}_i}$ $S_{G_kG_k}$ and S_0 must be approximated without explicit construction of S. We outline below such approximations.

Approximation of $S_{F_lF_l}$. The action of $S_{F_lF_l}^{-1}$ on a vector can either be computed exactly or approximately, as described for block Jacobi preconditioners. We shall omit further discussion of it here.

Approximation of $S_{\mathcal{E}_l\mathcal{E}_l}$. The action of $S_{\mathcal{E}_l\mathcal{E}_l}^{-1}$ on a vector $\mathbf{r}_{\mathcal{E}_l}$ can be approximated as follows. Given the domain Ω_{E_l} such that $\mathcal{E}_l = B \cap \Omega_{E_l}$, partition the nodes in Ω_{E_l} into $\mathcal{D}_l \equiv \Omega_{E_l} \backslash \mathcal{E}_l$ and $\mathcal{E}_l$. Let $A^{(\Omega_{E_l})}$ denote the submatrix of A corresponding to indices of nodes in $\mathcal{D}_l$ and $\mathcal{E}_l$. Then, the action of $S_{\mathcal{E}_l \mathcal{E}_l}^{-1}$ may be approximated as:

$$S_{\mathcal{E}_l \mathcal{E}_l}^{-1} \mathbf{r}_{\mathcal{E}_l} \approx \begin{bmatrix} 0 \\ I \end{bmatrix} \begin{bmatrix} A_{\mathcal{D}_l \mathcal{D}_l} & A_{\mathcal{D}_l \mathcal{E}_l} \\ A_{\mathcal{D}_l \mathcal{E}_l}^T & A_{\mathcal{E}_l \mathcal{E}_l} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \mathbf{r}_{\mathcal{E}_l} \end{bmatrix}.$$

Alternative approximations of $S_{\mathcal{E}_k \mathcal{E}_k}$ can be constructed based on extensions of the probing technique or based on inexact Cholesky factorizations.

Approximation of $\boldsymbol{S_{G_k G_k}}$. The action of $S_{G_k G_k}^{-1}$ on a vector $\mathbf{r}_{G_k}$ can be approximated as follows. Let $\Omega_{\mathbf{v}_k}$ denote a domain of width $\mathcal{O}(h_0)$ such that $G_k = B \cap \Omega_{\mathbf{v}_k}$. Partition the nodes in $\Omega_{\mathbf{v}_k}$ based on $H_k \equiv \Omega_{\mathbf{v}_k} \backslash G_k$ and G_k. Let $A^{(\Omega_{\mathbf{v}_k})}$ denote the submatrix of corresponding to nodes in H_k and G_k. Then, the action of $S_{G_k G_k}^{-1}$ may be approximated as:

$$S_{G_k G_k}^{-1} \mathbf{r}_{G_k} \approx \begin{bmatrix} 0 \\ I \end{bmatrix} \begin{bmatrix} A_{H_k H_k} & A_{H_k G_k} \\ A_{H_k G_k}^T & A_{G_k G_k} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \mathbf{r}_{G_k} \end{bmatrix}.$$

Alternative matrix approximations of $S_{G_k G_k}$ may be constructed based on extensions of the probing technique or inexact Cholesky decomposition.

Approximation of $\boldsymbol{S_0}$. The coarse space matrix $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$ can be approximated by coarse grid stiffness matrix A_0 as in the two dimensional case. The rate of convergence of the *vertex space* preconditioner will be of optimal order provided the globs $\{\mathcal{E}_l\}$ and $\{G_k\}$ have sufficient overlap of size $\beta\, h_0$ when the coefficients $a(\cdot)$ is smooth.

Theorem 3.51. *There exists $C_1 > 0$ independent of h_0 and h, but depending on the coefficients $a(\cdot)$ such that:*

$$\operatorname{cond}(M, S) \;\le\; C_1 \, \left(1 + \log^2(\beta^{-1})\right).$$

If the coefficient $a(\cdot)$ is constant on each subdomain, but has large jumps across subdomains, then the above bound deteriorates to:

$$\operatorname{cond}(M, S) \le C_2(\beta)\, \frac{h_0}{h},$$

where $C_2 > 0$ is independent of h_0, h and $a(\cdot)$.

Proof. See [SM, DR10]. □

3.6.3 A Parallel Wirebasket Preconditioner for S

Wirebasket methods for the Schur complement S are preconditioners which employ special coarse spaces [BR14, BR15, DR3, MA12, SM2, DR10], based on the wirebasket region of the interface. These preconditioners are typically formulated to yield robust convergence in the presence of large jump discontinuities in the coefficient $a(x)$, with rates of convergence that compare favorably with those for the block Jacobi and vertex space preconditioners. Due to a weaker discrete Sobolev inequality holding for traditional coarse spaces in three dimensions, theoretical bounds for the latter two preconditioners deteriorate in the presence of large jump discontinuities in $a(x)$. With the use of an appropriately chosen wirebasket coarse space, improved bounds can be obtained. Like traditional coarse spaces, wirebasket coarse spaces help transfer information globally between different subdomains, but involve significantly more unknowns. Typically however, an efficient algebraic solver can be formulated to solve the resulting coarse problems.

The *parallel* wirebasket preconditioner [SM2] we describe has the form of a matrix additive Schwarz preconditioner for S. Like the preconditioner (3.76), it is based on a partition of the interface into faces and the wirebasket:

$$B = F_1 \cup \cdots \cup F_q \cup W,$$

However, unlike (3.76) which employs a *local* correction term $\mathcal{R}_W^T S_{WW}^{-1} \mathcal{R}_W$ corresponding to the nodes on the wirebasket region W, where $\mathcal{R}_W$ is a *pointwise* nodal restriction matrix with zero-one entries, the parallel wirebasket preconditioner employs a coarse space correction term of the form $\mathcal{I}_W^T\, S_{WB}^{-1}\, \mathcal{I}_W$ based on a *weighted* restriction matrix $\mathcal{I}_W$ whose rows span the *wirebasket coarse space*. Once $\mathcal{I}_W$ is defined, the wirebasket preconditioner is obtained by formally replacing the term $\mathcal{R}_W^T S_{WW}^{-1} \mathcal{R}_W$ in (3.76) by the wirebasket coarse space correction term $\mathcal{I}_W S_{WB}^{-1} \mathcal{I}_W^T$ where $S_{WB} \equiv \mathcal{I}_W S \mathcal{I}_W^T$:

$$M^{-1} = \sum_{i=1}^{q} \mathcal{R}_{F_i}^T S_{F_i F_i}^{-1} \mathcal{R}_{F_i} + \mathcal{I}_W^T S_{WB}^{-1} \mathcal{I}_W, \tag{3.81}$$

If n_W and n_B denote the number of nodes on the wirebasket region W and interface B, respectively, then $\mathcal{I}_W$ will be a matrix of size $n_W \times n_B$ and S_{WB} will be a symmetric positive definite matrix of size n_W. Once the coarse space given by Range $\left(\mathcal{I}_W^T\right)$ has been defined, a suitable matrix approximation $M_{WB} \approx S_{MB} \equiv \mathcal{I}_W\, S\, \mathcal{I}_W^T$ must also be specified, to ensure that linear systems of the form $M_{WB}\mathbf{u}_W = \mathbf{r}_W$ can be solved efficiently within the wirebasket preconditioner. We shall describe $\mathcal{I}_W^T$ and M_{WB} in the following.

We first define the extension map $\mathcal{I}_W^T$. Let $\partial F_l \subset W$ denote the boundary segment of face F_l and let $n_{\partial F_l}$ denote the number of nodes on ∂F_l. Then, the wirebasket extension map $\mathcal{I}_W^T$ is defined as the following $n_B \times n_W$ matrix:

$$\left(\mathcal{I}_W^T \mathbf{v}_W\right)_i = \begin{cases} \left(\mathbf{v}_W\right)_i, & \text{if } x_i \in W \\ \frac{1}{n_{\partial F_l}} \sum_{j: x_j \in \partial F_l} (\mathbf{v}_W)_j, & \text{if } x_i \in F_l, \end{cases} \tag{3.82}$$

where x_i is a node on W with index i in the local ordering of nodes on B. By definition, the extension $\left(\mathcal{I}_W^T \mathbf{v}_W\right)_i$ equals the *average* nodal value of $\mathbf{v}_W$ on ∂F_l when node $x_i \in F_l$. It can thus be verified that its transpose $\mathcal{I}_W$ satisfies:

$$(\mathcal{I}_W \mathbf{v}_B)_i = (\mathbf{v}_B)_i + \sum_{\{k: x_i \in \partial F_k\}} \sum_{\{j: x_j \in F_k\}} \frac{(\mathbf{v}_B)_j}{n_{\partial F_k}}, \tag{3.83}$$

which yields a weighted combination of the nodal values of $\mathbf{v}_B$ on B.

Remark 3.52. Since the Schur complement S is not assembled in iterative substructuring, the matrices $S_{F_l F_l} = \mathcal{R}_{F_l} S \mathcal{R}_{F_l}^T$ and $S_{WB} \equiv \mathcal{I}_W S \mathcal{I}_W^T$ must be approximated in practice. Symmetric positive definite approximations of the submatrices $S_{F_l F_l}$ of S have already been described in the section on block Jacobi preconditioners, and so will not be described further. A symmetric positive definite approximation M_{WB} of S_{WB} and an associated algebraic solver for linear systems of the form $M_{WB}\mathbf{v}_W = \mathbf{r}_W$ will be formulated in the remainder of this subsection.

To construct a heuristic approximation M_{WB} of S_{WB}, we consider the subassembly identity for the Schur complement matrix:

$$S = \sum_{i=1}^{p} \mathcal{R}_B^{(i)^T} S^{(i)} \mathcal{R}_B^{(i)},$$

Substituting this identity into $S_{WB} = \mathcal{I}_W S \mathcal{I}_W^T$ yields:

$$S_{WB} = \sum_{i=1}^{p} \mathcal{I}_W \mathcal{R}_B^{(i)^T} S^{(i)} \mathcal{R}_B^{(i)} \mathcal{I}_W^T. \tag{3.84}$$

Using definition (3.82), it can be verified that the extension (interpolation) map $\mathcal{I}_W^T$ acts *locally* on each subdomain boundary. Indeed, the nodal values of $\mathcal{I}_W^T \mathbf{v}_W$ on each subdomain boundary $B^{(i)}$ can be expressed solely in terms of the nodal values of $\mathbf{v}_W$ on the wirebasket $W^{(i)}$, yielding the following identity on each boundary $B^{(i)}$:

$$\mathcal{E}_{B^{(i)}W^{(i)}} \mathbf{v}_{W^{(i)}} = \mathcal{R}_B^{(i)} \mathcal{I}_W^T \mathbf{v}_W, \tag{3.85}$$

where $\mathcal{E}_{B^{(i)}W^{(i)}} \mathbf{v}_{W^{(i)}}$ is defined next.

$$\left(\mathcal{E}_{B^{(i)}W^{(i)}} \mathbf{v}_{W^{(i)}}\right)_k \equiv \begin{cases} \left(\mathbf{v}_{W^{(i)}}\right)_k, & \text{if } x_k \in W^{(i)} \\ \frac{1}{n_{\partial F_l}} \sum_{j: x_j \in \partial F_l} \left(\mathbf{v}_{W^{(i)}}\right)_j, & \text{if } x_k \in F_l \subset B^{(i)}. \end{cases}$$

Thus $\mathcal{E}_{B^{(i)}W^{(i)}} \mathcal{R}_{W^{(i)}W} = \mathcal{R}_{B^{(i)}} \mathcal{I}_W^T$. Substituting this into (3.84) yields:

$$S_{WB} = \sum_{i=1}^{p} \mathcal{R}_{W^{(i)}W}^T S_{WB}^{(i)} \mathcal{R}_{W^{(i)}W}, \tag{3.86}$$

where $S_{WB}^{(i)} \equiv \mathcal{E}_{B^{(i)}W^{(i)}}^T S^{(i)} \mathcal{E}_{B^{(i)}W^{(i)}}$. This expresses S_{WB} as a sum of local contributions. Given a local approximation $M_{WB}^{(i)}$ of $S_{WB}^{(i)}$, an approximation M_{WB} of S_{WB} can be constructed by replacing $S_{WB}^{(i)}$ by $M_{WB}^{(i)}$ in (3.86).

To construct an approximation $M_{WB}^{(i)}$ of $S_{WB}^{(i)}$ so that M_{WB} is spectrally equivalent to S_{WB} independent of the coefficient $a(\cdot)$, we will require that each $M_{WB}^{(i)}$ be spectrally equivalent to $S_{WB}^{(i)}$ independent of $a(\cdot)$. The following heuristic observations will be employed when $a(\cdot)$ is piecewise constant. *Firstly*, when $c(.) = 0$ in elliptic equation (3.1), and $a(x) \equiv a^{(i)}$ on each Ω_i, then the local Schur complement $S^{(i)}$ (and consequently $S_{WB}^{(i)}$) will scale in proportion to coefficient $a^{(i)}$. In particular, if Ω_i is immersed in Ω, i.e., $B^{(i)} = \partial\Omega_i$, then $S^{(i)}$ (and also $S_{WB}^{(i)}$) will be *singular*. *Secondly*, when $c(.) = 0$ and $a(x) \equiv a^{(i)}$ on Ω_i and Ω_i is immersed, let $\mathbf{z}_{W^{(i)}} \equiv (1, \ldots, 1)^T$ denote a vector of size $n_{W^{(i)}}$ corresponding to the number of nodes on $W^{(i)}$. Then, its extension $\mathcal{E}_{B^{(i)}W^{(i)}}\mathbf{z}_{W^{(i)}}$ of size $n_{B^{(i)}}$ will satisfy:

$$\mathcal{E}_{B^{(i)}W^{(i)}}\mathbf{z}_{W^{(i)}} = (1, \ldots, 1)^T,$$

where vector $(1, \ldots, 1)^T$ of size $n_{B^{(i)}}$ generates the null space of $S^{(i)}$. As a consequence, $S_{WB}^{(i)}$ will be singular when $S^{(i)}$ is singular, and $\mathbf{z}_{W^{(i)}}$ will span its null space. *Thirdly*, since $S_{WB}^{(i)}$ will scale in proportion to coefficient $a^{(i)}$, it will be necessary to choose $M_{WB}^{(i)}$ also proportional to $a^{(i)}$ to ensure spectral equivalence between $M_{WB}^{(i)}$ and $S_{WB}^{(i)}$ independent of $\{a^{(l)}\}$.

Employing these heuristic observations, we may seek to approximate $S_{WB}^{(i)}$ by a scalar multiple $D^{(i)} = \beta\, a^{(i)}\, I$ of the identity matrix of size $n_{W^{(i)}}$ for a scaling factor $\beta > 0$ to be specified. However, to ensure that $S_{WB}^{(i)}$ and $M_{WB}^{(i)}$ also both have the same *null spaces*, we shall post-multiply and pre-multiply matrix $D^{(i)}$ and define $M_{WB}^{(i)} = (I - P_i)^T D^{(i)} (I - P_i)$ where P_i is defined as:

$$P_i \equiv \frac{\mathbf{z}_{W^{(i)}}\mathbf{z}_{W^{(i)}}^T D^{(i)}}{\mathbf{z}_{W^{(i)}}^T D^{(i)} \mathbf{z}_{W^{(i)}}} = \frac{\mathbf{z}_{W^{(i)}}\mathbf{z}_{W^{(i)}}^T}{\mathbf{z}_{W^{(i)}}^T \mathbf{z}_{W^{(i)}}}, \tag{3.87}$$

corresponding to a $D^{(i)}$-orthogonal projection onto the null space span($\mathbf{z}_{W^{(i)}}$) of $S_{WB}^{(i)}$. This yields the choice of $M_{WB}^{(i)}$ as:

$$M_{WB}^{(i)} = (I - P_i)^T D^{(i)} (I - P_i) = \beta\, a^{(i)}\, (I - P_i).$$

Matrix $M_{WB}^{(i)}$ may also be equivalently characterized by the requirement:

$$\mathbf{v}_{W^{(i)}}^T M_{WB}^{(i)} \mathbf{v}_{W^{(i)}} = \min_{\omega_i} \left(\mathbf{v}_{W^{(i)}} - \omega_i \mathbf{z}_{W^{(i)}}\right)^T D^{(i)} \left(\mathbf{v}_{W^{(i)}} - \omega_i \mathbf{z}_{W^{(i)}}\right), \tag{3.88}$$

where ω_i is a parameter chosen to minimize the above expression. This can easily be verified. Theoretical analysis [SM2, DR10] suggests choosing the scaling factor as $\beta = h\,(1 + \log(h_0/h))$. Combining the preceding observations yields a global approximation $M_{WB} \approx S_{WB}$ based on the local approximations $M_{WB}^{(i)} \approx S_{WB}^{(i)}$ as:

$$\begin{cases} M_{WB} = \sum_{i=1}^{p} \mathcal{R}_{W^{(i)}W}^T M_{WB}^{(i)} \mathcal{R}_{W^{(i)}W} \\ \qquad\;\; = \sum_{i=1}^{p} \mathcal{R}_{W^{(i)}W}^T (I - P_i)^T D^{(i)} (I - P_i) \mathcal{R}_{W^{(i)}W}, \end{cases} \tag{3.89}$$

where $D^{(i)} = h\ (1 + \log(h_0/h))\ a^{(i)} I$, and P_i is defined by (3.87).

Remark 3.53. Matrix M_{WB} may also be equivalently characterized using (3.89) and (3.88) as satisfying:

$$\begin{array}{c} \mathbf{v}_W^T M_{WB} \mathbf{v}_W \\ = \min_{(\omega_1,\ldots,\omega_p)} \sum_{i=1}^{p} \left(\mathbf{v}_{W^{(i)}} - \omega_i \mathbf{z}_{W^{(i)}}\right)^T D^{(i)} \left(\mathbf{v}_{W^{(i)}} - \omega_i \mathbf{z}_{W^{(i)}}\right), \end{array} \tag{3.90}$$

where $\mathbf{v}_{W^{(i)}} = \mathcal{R}_{W^{(i)}W} \mathbf{v}_W$. This alternative expression will be useful in constructing an efficient solver for linear systems of the form $M_{WB} \mathbf{v}_W = \mathbf{r}_W$.

Remark 3.54. For elliptic systems such as the equations of linear elasticity, the null space of $S^{(i)}$ may have several linearly independent vectors. In this case $\mathbf{z}_{W^{(i)}}$ will need to be replaced by a matrix whose columns are restrictions to $W^{(i)}$ of a basis for the null space of $S^{(i)}$.

Remark 3.55. By construction, matrix M_{WB} is symmetric, and will also be positive semidefinite since $\mathbf{v}_W^T M_{WB} \mathbf{v}_W$ is a sum of nonnegative quadratic forms. A vector $\mathbf{v}_W$ will belong to the *null space* of M_{WB} only if:

$$M_{WB} \mathbf{v}_W = \mathbf{0} \quad \Leftrightarrow \quad \mathcal{R}_{W^{(i)}W} \mathbf{v}_W = \alpha_i\, \mathbf{z}_{W^{(i)}}, \;\; \text{for } 1 \le i \le p.$$

This can be verified to hold for nonzero α_i only if S_{WB} is singular. As a result, M_{WB} will be positive definite whenever S_{WB} is positive definite.

We now describe an algebraic solver for $M_{WB} \mathbf{u}_W = \mathbf{r}_W$. Since M_{WB} will be a symmetric and positive definite matrix, the solution $\mathbf{u}_W$ to the linear system $M_{WW} \mathbf{u}_W = \mathbf{r}_W$ will also solve the following minimization problem:

$$J(\mathbf{u}_W) = \min_{\mathbf{v}_W} J(\mathbf{v}_W), \tag{3.91}$$

where $J(\mathbf{u}_W) \equiv \frac{1}{2} \mathbf{v}_{\scriptscriptstyle W}^T M_{WB} \mathbf{v}_W - \mathbf{v}_W^T \mathbf{r}_W$ is its associated energy.

$$\begin{aligned} J(\mathbf{v}_{\scriptscriptstyle W}) &\equiv \tfrac{1}{2} \mathbf{v}_{\scriptscriptstyle W}^T M_{WB} \mathbf{v}_W - \mathbf{v}_W^T \mathbf{r}_W \\ &= \tfrac{1}{2} \sum_{i=1}^{p} \min_{\omega_i} \left(\mathcal{R}_{W^{(i)}W} \mathbf{v}_W - \omega_i \mathbf{z}_{W^{(i)}}\right)^T D^{(i)} \\ &\quad \left(\mathcal{R}_{W^{(i)}W} \mathbf{v}_W - \omega_i \mathbf{z}_{W^{(i)}}\right) - \mathbf{v}_W^T \mathbf{r}_W. \end{aligned}$$

The minimization of (3.91) will thus also be equivalent to:

$$\tilde{J}\left(\mathbf{u}_{\scriptscriptstyle W}, \omega_1^*, \ldots, \omega_p^*\right) = \min_{(\mathbf{v}_W, \omega_1, \cdots, \omega_p)} \tilde{J}\left(\mathbf{v}_W, \omega_1, \ldots, \omega_p\right),$$

where

$$\begin{array}{c} \tilde{J}\left(\mathbf{v}_W, \omega_1, \ldots, \omega_p\right) \\ \equiv \frac{1}{2} \sum_{i=1}^{p} \left(\mathbf{v}_{W^{(i)}} - \omega_i \mathbf{z}_{W^{(i)}}\right)^T D^{(i)} \left(\mathbf{v}_{W^{(i)}} - \omega_i \mathbf{z}_{W^{(i)}}\right) - \mathbf{v}_W^T \mathbf{r}_W. \end{array}$$

Applying the first order derivative conditions for a minimum (differentiating the above expression with respect to $\mathbf{v}_W$ and $\omega_1, \ldots, \omega_p$ and requiring it to equal zero) yields the following system of equations:

$$\begin{cases} \mathbf{z}_{W^{(i)}}^T D^{(i)} \left(\mathcal{R}_{W^{(i)}W} \mathbf{u}_W - \omega_i^* \mathbf{z}_{W^{(i)}} \right) & = 0, \quad \text{for} \quad 1 \le i \le p, \\ D_{WB} \mathbf{u}_W - \sum_{i=1}^{p} \omega_i^* \mathcal{R}_{W^{(i)}W}^T D^{(i)} \mathbf{z}_{W^{(i)}} & = \mathbf{r}_W, \end{cases} \tag{3.92}$$

where D_{WB} is the following *diagonal* matrix of size n_W:

$$D_{WB} \equiv \sum_{i=1}^{p} \mathcal{R}_{W^{(i)}W}^T D^{(i)} \mathcal{R}_{W^{(i)}W},$$

with *diagonal* entries:

$$(D_{WB})_{ii} = \sum_{\{k : \mathbf{v}_i \in B^{(k)}\}} a^{(k)}\, h\, \left(1 + \log(h_0/h)\right).$$

An efficient solver for $M_{WB}\mathbf{u}_W = \mathbf{r}_W$ can be formulated by solving (3.92).

For each choice of parameters $\omega_1^*, \ldots, \omega_p^*$, the vector unknown $\mathbf{u}_W$ can be determined by solving the second block row in (3.92):

$$\mathbf{u}_W = D_{WB}^{-1} \left(\mathbf{r}_W + \sum_{i=1}^{p} \omega_i^* \mathcal{R}_{W^{(i)}W}^T D^{(i)} \mathbf{z}_{W^{(i)}} \right).$$

A reduced system can thus be obtained for the parameters $\omega_1^*, \ldots, \omega_p^*$ by substituting the preceding expression for $\mathbf{u}_W$ into the first block row in (3.92):

$$\begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & & \vdots \\ K_{1p} & \cdots & K_{pp} \end{bmatrix} \begin{bmatrix} \omega_1^* \\ \vdots \\ \omega_p^* \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_p \end{bmatrix},$$

where the entries K_{ij} and g_i are defined as follows.

$$\begin{cases} K_{ij} \equiv -\mathbf{z}_{W^{(i)}}^T D^{(i)} \mathcal{R}_{W^{(i)}W} D_{WB}^{-1} \mathcal{R}_{W^{(j)}W}^T D^{(j)} \mathbf{z}_{W^{(j)}}, \quad \text{for } i \neq j \\ K_{ii} \equiv \left(\mathbf{z}_{W^{(i)}}^T D^{(i)} \mathbf{z}_{W^{(i)}} \right) - \mathbf{z}_{W^{(i)}}^T D^{(i)} \mathcal{R}_{W^{(i)}W} D_{WB}^{-1} \mathcal{R}_{W^{(i)}W}^T D^{(i)} \mathbf{z}_{W^{(j)}}, \\ \quad g_i \equiv \mathbf{z}_{W^{(i)}}^T D^{(i)} \mathcal{R}_{W^{(i)}W} D_{WB}^{-1} \mathbf{r}_W. \end{cases} \tag{3.93}$$

Matrix K can be verified to be symmetric and sparse, and the preceding linear system can be solved using any suitable sparse direct solver. We summarize the implementation of the parallel wirebasket preconditioner for S.

Algorithm 3.6.1 *(Wirebasket Preconditioner)*

$$M^{-1}\mathbf{r}_B \equiv \sum_{i=1}^{q} \mathcal{R}_{F_k}^T S_{F_k F_k}^{-1} \mathcal{R}_{F_k} \mathbf{r}_B + \mathcal{I}_W^T M_{WB}^{-1} \mathcal{I}_W \mathbf{r}_B.$$

The terms $S_{F_k F_k}^{-1} \mathcal{R}_{F_k} \mathbf{r}_B$ can be computed as described for the block Jacobi preconditioner. The solution to $M_{WB}\,\mathbf{u}_W = \mathcal{I}_W \mathbf{r}_B$ can be computed as follows. *Firstly*, using $\mathbf{r}_W \equiv \mathcal{I}_W \mathbf{r}_B$ solve for $\omega_1^*, \ldots, \omega_p^*$:

$$\begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & & \vdots \\ K_{1p} & \cdots & K_{pp} \end{bmatrix} \begin{bmatrix} \omega_1^* \\ \vdots \\ \omega_p^* \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_p \end{bmatrix},$$

where the entries K_{ij} and g_j are defined in (3.93). *Secondly*, solve for $\mathbf{u}_W$:

$$D_{WB}\,\mathbf{u}_W = \left(\mathbf{r}_W + \sum_{i=1}^{p} \omega_i \mathcal{R}_{W^{(i)}W}^T D^{(i)} \mathbf{z}_{W^{(i)}} \right).$$

This yields $\mathbf{u}_W$. The following result concerns the convergence rate of the preceding parallel wirebasket algorithm.

Theorem 3.56. *If the coefficient $a(\cdot)$ is constant within each subdomain, there exists $C > 0$ independent of h_0, h and $a(\cdot)$ such that*

$$\mathrm{cond}(M, S) \le C(1 + \log(h_0/h))^2.$$

Proof. See [SM2, DR10]. □

Remark 3.57. The heuristic approximation M_{WB} of S_{WB} assumed that the coefficient $c(x) = 0$. In practice the same matrix M_{WB} described above (based on the vectors $\mathbf{z}_{W^{(i)}}$) can be employed even when $c(x) \neq 0$ though $S_{WB}^{(i)}$ will not be singular in such a case. Indeed, omitting such terms will remove the mechanism for global transfer of information. Alternate wirebasket algorithms are described in [BR15, MA12, DR10], including an algorithm with condition number $(1 + \log(h_0/h))$.

3.7 Neumann-Neumann and Balancing Preconditioners

Neumann-Neumann and balancing domain decomposition methods are a widely used family of preconditioners for multisubdomain Schur complement matrices in *two* and *three* dimensions. From a computational viewpoint, these preconditioners solve a Neumann problem on each subdomain, and hence the name. Furthermore, such preconditioners have an algebraic form that may be applied to arbitrary subdomain geometries in two or three dimensions, without the requirement that the subdomains be boxes or tetrahedra.

Theoretical analysis indicates that these methods precondition effectively, yielding condition number bounds which grow polylogarithmic in the mesh parameters, independent of the jump discontinuities in the coefficient. Our discussion will focus on the family of Neumann-Neumann preconditioners [BO7, DE2, DE3, DR14, DR16, LE, DR18, LE5], and the balancing domain decomposition preconditioner [MA14, MA17]. We also outline an algebraic preconditioner [CA33] based on the Neumann-Neumann preconditioner.

From the viewpoint of Schwarz subspace methods, a Neumann-Neumann preconditioner has the structure of an additive Schwarz preconditioner for S, while the balancing domain decomposition preconditioner has the structure of a *hybrid* Schwarz preconditioner for S. Given a decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, both preconditioners decompose the interface B into the segments:

$$B = B^{(1)} \cup \cdots \cup B^{(p)}, \quad \text{where} \quad B^{(i)} \equiv \partial\Omega_i \backslash \mathcal{B}_D. \tag{3.94}$$

Both preconditioners employ the subdomain Schur complement matrix $S^{(i)}$ to approximate the unassembled submatrix $S_{B^{(i)}B^{(i)}} = \mathcal{R}_{B^{(i)}} S \mathcal{R}_{B^{(i)}}^T$ of S, corresponding to the nodes on $B^{(i)}$. Different coarse spaces facilitating global transfer of information are also employed in each preconditioner.

3.7.1 Neumann-Neumann Preconditioners

Multi-subdomain Neumann-Neumann preconditioners are extensions of the two subdomain Neumann-Neumann preconditioner from Chap. 3.4. It has the formal structure of an additive Schwarz subspace preconditioner for S, based on the decomposition of B into the overlapping boundary segments $B^{(1)}, \ldots, B^{(p)}$, with restriction and extension matrices $\mathcal{R}_{B^{(i)}}$ and $\mathcal{R}_{B^{(i)}}^T$ respectively, defined in (3.19). Since S is *not assembled*, $S_{B^{(i)}B^{(i)}} \equiv \mathcal{R}_{B^{(i)}} S \mathcal{R}_{B^{(i)}}^T$ submatrix of S is *approximated* by the subdomain Schur complement $S^{(i)}$. If no *coarse space* is employed, the preconditioner has the form:

$$M^{-1} = \sum_{i=1}^{p} \mathcal{R}_{B^{(i)}}^T S^{(i)^\dagger} \mathcal{R}_{B^{(i)}}, \tag{3.95}$$

where $S^{(i)^\dagger}$ denotes the Moore-Penrose pseudoinverse [GO4] of the local Schur complement matrix $S^{(i)}$, since $S^{(i)}$ can be *singular*, unlike $S_{B^{(i)}B^{(i)}}$.

Remark 3.58. In practical implementation, the local Schur complement $S^{(i)}$ need not be assembled. Instead the following may be noted. When matrix $S^{(i)}$ is *nonsingular*, then $S^{(i)^\dagger} = S^{(i)^{-1}}$. In this case, terms of the form $S^{(i)^\dagger}\mathbf{r}_{B^{(i)}}$ can be computed by solving the linear system $S^{(i)}\mathbf{w}_{B^{(i)}} = \mathbf{r}_{B^{(i)}}$ corresponding to a discrete Neumann problem on Ω_i:

$$S^{(i)^{-1}}\mathbf{r}_{B^{(i)}} = \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \mathbf{r}_{B^{(i)}} \end{bmatrix}.$$

However, when $c(x) = 0$ and Ω_i is *immersed* inside Ω (i.e., $B^{(i)} = \partial\Omega_i$), then matrices $A^{(i)}$ and $S^{(i)}$ will be *singular*. In this case, the null space of $A^{(i)}$ and $S^{(i)}$ will be spanned by vectors of the form $\mathbf{1} = (1, \ldots, 1)^T$ of appropriate sizes. As a result, the linear system $S^{(i)}\mathbf{w}_{B^{(i)}} = \mathbf{r}_{B^{(i)}}$ will be solvable only if $\mathbf{r}_{B^{(i)}}$ satisfies the compatability condition:

$$\mathbf{1}^T\mathbf{r}_{B^{(i)}} = 0.$$

When this compatibility condition is satisfied, a solution $\mathbf{w}_{B^{(i)}}$ will exist, though it will not be unique, as any scalar multiple of $\mathbf{1}$ may be added to it. When $S^{(i)}$ is singular, the action of $S^{(i)^\dagger}$ on a vector it typically approximated in Neumann-Neumann algorithms [DE3], as follows. If direct solvers are employed, then when the Cholesky factorization $L^{(i)}L^{(i)^T}$ of $A^{(i)}$ is computed on each subdomain, zero or "small" pivots can be set to a prescribed nonzero number $\epsilon > 0$, and this approximate factorization can be employed to formally compute $\tilde{\mathbf{w}}_{B^{(i)}} \approx S^{(i)^\dagger}\mathbf{r}_{B^{(i)}}$. If desired, this approximate solution $\tilde{\mathbf{w}}_{B^{(i)}}$ may then be projected onto the orthogonal complement of the null space:

$$\mathbf{w}_{B^{(i)}} \equiv \tilde{\mathbf{w}}_{B^{(i)}} - \left(\frac{\mathbf{1}^T\tilde{\mathbf{w}}_{B^{(i)}}}{\mathbf{1}^T\mathbf{1}}\right)\mathbf{1}.$$

Alternatively, a projected gradient method may be used to iteratively solve $S^{(i)}\mathbf{w}_{B^{(i)}} = \mathbf{r}_{B^{(i)}}$, yielding an approximate solution satisfying $\mathbf{1}^T\mathbf{w}_{B^{(i)}} = 0$. We summarize the algorithm below assuming nonsingular subproblems.

Algorithm 3.7.1 *(Neumann-Neumann Preconditioner-No Coarse Space)*
Given $\mathbf{r}_B$ the vector $M^{-1}\mathbf{r}_B$ is computed as follows.

1. *For $i = 1, \cdots, p$ in parallel solve:*

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(i)} \\ \mathbf{w}_B^{(i)} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{R}_{B^{(i)}}\mathbf{r}_B \end{bmatrix}.$$

2. *Endfor*

Output: $M^{-1}\mathbf{r}_B \equiv \sum_{i=1}^p \mathcal{R}_{B^{(i)}}^T\mathbf{w}_B^{(i)}$.

Remark 3.59. As noted earlier, if a local problem is *singular*, the local Cholesky factorization can be modified. However, the balancing domain decomposition preconditioner [MA14, MA17], described later, elegantly addresses the issue arising with singular local problems and its non-unique solution.

We shall next describe a Neumann-Neumann preconditioner employing an algebraic *partition of unity*. For convenience, we omit a coarse space correction term, though it may be added. To motivate this version of the preconditioner, note that because of overlap between adjacent boundaries $B^{(i)}$, the Neumann-Neumann preconditioner adds duplicates of the solution on the regions of overlap. Such duplication can be reduced by employing a discrete partition of unity on B subordinate to the subdomain boundaries $B^{(1)}, \ldots, B^{(p)}$. Accordingly, let $n_{B^{(l)}}$ denote the number of nodes on $B^{(l)}$ for $1 \leq l \leq p$ and let $\mathbf{x}_i^{(l)}$ for $1 \leq i \leq n_{B^{(l)}}$ denote an ordering of the nodes on $B^{(l)}$. For each $1 \leq l \leq p$ let $D^{(l)}$ denote a *diagonal matrix* of size $n_{B^{(l)}}$ with *nonnegative entries* so that a discrete partition (decomposition) of the identity matrix is obtained:

$$\sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T \, D^{(l)} \, \mathcal{R}_{B^{(l)}} = I. \tag{3.96}$$

Various choices of such diagonal matrices exist. The diagonal entries of $D^{(l)}$ is also commonly defined based on the coefficient $a(x)$:

$$\left(D^{(l)}\right)_{ii} = \frac{(a^{(l)})^{\rho}}{\sum_{j:\mathbf{x}_i^{(l)} \in B^{(j)}} (a^{(j)})^{\rho}} \tag{3.97}$$

where $0 \leq \rho \leq 1$ denotes some user chosen scaling factor and $a^{(l)}$ denotes some sample value of coefficient $a(x)$ in Ω_l. When $a(x) \equiv 1$, the above definition yields $\left(D^{(l)}\right)_{ii} = 1/\deg(x_i^{(l)})$, where $\deg(x_i^{(l)})$ denotes the degree of node $x_i^{(l)}$, i.e., the number of distinct subdomain boundaries $B^{(j)}$ to which node $x_i^{(l)}$ belongs to. Such a discrete partition of the identity on B can be employed to distribute an interface load $\mathbf{r}_B$ to the subdomain boundaries $\mathbf{r}_B = \sum_{i=1}^{p} \mathcal{R}_{B^{(i)}}^T D^{(i)} \mathcal{R}_{B^{(i)}} \mathbf{r}_B$ so that the load is not duplicated. The partition of unity Neumann-Neumann preconditioner can now be formulated as:

$$M^{-1}\mathbf{r}_B = \sum_{i=1}^{p} \mathcal{R}_{B^{(i)}}^T \, {D^{(i)}}^T \, {S^{(i)}}^{\dagger} \, D^{(i)} \, \mathcal{R}_{B^{(i)}} \mathbf{r}_B \tag{3.98}$$

where we have omitted a coarse space correction term. To ensure that the preconditioner is *symmetric*, each matrix $D^{(i)}$ has been employed twice. Preconditioner (3.98) corresponds to a matrix additive Schwarz preconditioner for S based on the subspaces $\mathrm{Range}(\mathcal{R}_{B^{(i)}}^T {D^{(i)}}^T)$ for $1 \leq i \leq p$ with the matrices $S^{(i)}$ approximating $D^{(i)} \mathcal{R}_{B^{(i)}} S \mathcal{R}_{B^{(i)}}^T {D^{(i)}}^T$.

The following bounds will hold for the standard and partition of unity versions of the Neumann-Neumann preconditioner without a coarse space.

Lemma 3.60. *If M denotes the preconditioner in (3.95) or in (3.98), then the following condition number bound will hold:*

$$\mathrm{cond}(M, S) \leq C \, h_0^{-2} \left(1 + \log(h_0/h)^2\right),$$

where $C > 0$ is independent of h and h_0.

Proof. See [DE3, DR18]. □

To improve the convergence rate of the preceding Neumann-Neumann algorithms as the subdomain size h_0 decreases, a *coarse space* correction term can be included, thereby providing some global exchange of information. Any coarse space from Chap. 2.1 may be employed, in principle. However, if the subdomains $\Omega_1, \ldots, \Omega_p$ correspond to elements in a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of size h_0, let $y_l^{(0)}$ denote the coarse nodes for $1 \le l \le n_0$, and let $\phi_l^{(h_0)}(x)$ denote the coarse space nodal basis satisfying $\phi_l^{(h_0)}(y_j^{(0)}) = \delta_{ij}$. If $x_1, \ldots, x_{n_B}$ denotes the nodes on B, then the coarse space matrix $\mathcal{R}_0^T$ is:

$$\mathcal{R}_0^T = \begin{bmatrix} \phi_1^{(h_0)}(x_1) & \cdots & \phi_{n_0}^{(h_0)}(x_1) \\ \vdots & & \vdots \\ \phi_1^{(h_0)}(x_{n_B}) & \cdots & \phi_{n_0}^{(h_0)}(x_{n_B}) \end{bmatrix}, \tag{3.99}$$

A coarse space version of the Neumann-Neumann preconditioner can now be obtained by including the correction term $\mathcal{R}_0^T S_0^{-1} \mathcal{R}_0$ with $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$:

$$M^{-1}\mathbf{r}_B = \sum_{i=1}^{p} \mathcal{R}_{B^{(i)}}^T D^{(i)^T} S^{(i)^\dagger} D^{(i)} \mathcal{R}_{B^{(i)}} \mathbf{r}_B + \mathcal{R}_0 S_0^{-1} \mathcal{R}_0 \mathbf{r}_B. \tag{3.100}$$

As with other matrix additive Schwarz preconditioners for S, the coarse matrix S_0 may be approximated by the coarse grid discretization A_0 of (3.1). The Neumann-Neumann preconditioner with coarse space correction can be implemented in parallel using (3.100) with the subdomain problems solved as in Alg. 3.7.1. For brevity, we shall not summarize the resulting algorithm.

Lemma 3.61. *If coefficient $a(x)$ satisfies $a(x) = a^{(i)}$ on each subdomain Ω_i, then the condition number of the partition of unity Neumann-Neumann preconditioner with coarse space correction will satisfy:*

$$\mathrm{cond}(M, S) \le C\, \left(1 + \log(h_0/h)^2\right),$$

where $C > 0$ is independent of h, h_0 and $\{a^{(l)}\}$.

Proof. See [DE3, DR18].

3.7.2 Balancing Domain Decomposition Preconditioner

The balancing domain decomposition preconditioner [MA14, MA17] for the Schur complement S, employs an algebraic procedure referred to as *balancing*, which ensures that each *singular* subdomain problem arising in the Neumann-Neumann preconditioner is solvable. Additionally, the procedure eliminates arbitrariness in the output of the Neumann-Neumann preconditioner, arising from non-unique subdomain solutions, and provides a natural *coarse space* which transfers information globally, see also [GL14, FA16].

We shall *heuristically* motivate the balancing procedure, before outlining its implementation. The methodology will be illustrated for balancing the discrete partition of unity version of the Neumann-Neumann preconditioner:

$$M^{-1}\mathbf{r}_B = \sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T D^{(l)} S^{(l)^\dagger} D^{(l)} \mathcal{R}_{B^{(l)}} \mathbf{r}_B. \tag{3.101}$$

When $c(x) = 0$ and Ω_l is *floating* in Ω, matrix $S^{(l)}$ will be *singular.* Let $\tilde{N}_l$ denote a matrix whose columns form a basis for $\text{Kernel}(S^{(l)})$, so that $\text{Range}(\tilde{N}_l) = \text{Kernel}(S^{(l)})$. If $n_B^{(l)}$ denotes the size of $S^{(l)}$ and $\tilde{d}_l$ the dimension of the null space of $S^{(l)}$, then $\tilde{N}_l$ will be a matrix of size $n_B^{(l)} \times \tilde{d}_l$. When the matrix $S^{(l)}$ is singular, the subdomain problem:

$$S^{(l)}\mathbf{w}_B^{(l)} = D^{(l)}\mathcal{R}_{B^{(l)}}\mathbf{r}_B, \tag{3.102}$$

will be solvable only if the following compatibility condition holds:

$$\tilde{N}_l^T D^{(l)}\mathcal{R}_{B^{(l)}}\mathbf{r}_B = \mathbf{0}. \tag{3.103}$$

When (3.103) holds, the general solution to (3.102) will be:

$$\mathbf{w}_B^{(l)} = \mathbf{v}_B^{(l)} + \tilde{N}_l\boldsymbol{\alpha}_l, \tag{3.104}$$

where $\mathbf{v}_B^{(l)}$ is a particular solution, and $\tilde{N}_l\boldsymbol{\alpha}_l$ represents a general term in the null space of $S^{(l)}$ for $\boldsymbol{\alpha}_l \in \mathbb{R}^{\tilde{d}_l}$. The balancing procedure will employ a more general matrix N_l of size $n_B^{(l)} \times d_l$ with $d_l \geq \tilde{d}_l$ such that:

$$\text{Kernel}(S^{(l)}) = \text{Range}(\tilde{N}_l) \subset \text{Range}(N_l).$$

For instance when $c(x) > 0$, matrix $S^{(l)}$ will be nonsingular, but it may be advantageous to choose N_l as the matrix whose columns span the null space of the local Schur complement associated with $c(x) = 0$. By construction, if $N_l^T D^{(l)}\mathcal{R}_{B^{(l)}}\mathbf{r}_B^{(l)} = \mathbf{0}$, then system (3.102) will be consistent (even if $N_l \neq \tilde{N}_l$).

Definition 3.62. *A vector $\mathbf{r}_B \in \mathbb{R}^{n_B}$ will be said to be balanced if:*

$$N_l^T D^{(l)}\mathcal{R}_{B^{(l)}}\mathbf{r}_B = \mathbf{0}, \qquad \textit{for } 1 \leq l \leq p. \tag{3.105}$$

In this case, each system $S^{(l)}\mathbf{w}_B^{(l)} = D^{(l)}\mathcal{R}_{B^{(l)}}\mathbf{r}_B$ will be solvable.

By the preceding definition, when vector $\mathbf{r}_B$ is balanced, each subproblem $S^{(l)}\mathbf{w}_B^{(l)} = D^{(l)}\mathcal{R}_{B^{(l)}}\mathbf{r}_B$ in (3.101) will be solvable. When $\mathbf{r}_B$ is not balanced, it may be modified by subtracting a correction term $P_0\,\mathbf{r}_B$ so that $(I - P_0)\,\mathbf{r}_B$ is balanced, where P_0 is an S-orthogonal *projection*, which will be described in the following. Equation (3.105) which describes a balanced vector can be compactly represented using a matrix C of size $n_B \times d$ for $d = (d_1 + \cdots + d_p)$, where the columns of C consists of the columns of $\mathcal{R}_{B^{(l)}}^T D^{(l)^T} N_l$ for $1 \le l \le p$:

$$C = \left[\mathcal{R}_{B^{(1)}}^T D^{(1)^T} N_1 \quad \cdots \quad \mathcal{R}_{B^{(p)}}^T D^{(p)^T} N_p \right].$$

Then, equation (3.105) for a balanced vector $\mathbf{r}_B$ becomes:

$$C^T \mathbf{r}_B = \mathbf{0}.$$

When $C^T \mathbf{r}_B \neq \mathbf{0}$, a correction term $(S\, C\, \boldsymbol{\alpha})$ may be sought for $\alpha \in \mathbb{R}^d$:

$$C^T \left(\mathbf{r}_B - S\, C\, \boldsymbol{\alpha} \right) = \mathbf{0},$$

so that $(\mathbf{r}_B - S\, C\, \boldsymbol{\alpha})$ is balanced. This yields the following linear system of equations for determining $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T \right)^T \in \mathbb{R}^d$:

$$\left(C^T S C \right) \boldsymbol{\alpha} = C^T \mathbf{r}_B. \tag{3.106}$$

When C is of full rank, this system will be uniquely solvable by positive definiteness of S. The correction term $S\, C\, \boldsymbol{\alpha}$ may then be represented as:

$$P_0\, \mathbf{r}_B \equiv S\, C\, \boldsymbol{\alpha} = S\, C\, \left(C^T S C \right)^{-1} C^T \mathbf{r}_B,$$

where $P_0\, \mathbf{r}_B$ can be easily verified to be an S-orthogonal projection of $\mathbf{r}_B$ onto the column space of C (with $P_0\, P_0 = P_0$ and $P_0\, S = S\, P_0^T$).

Motivated by the preceding, the balancing domain decomposition preconditioner M employs the structure of a *hybrid* Schwarz preconditioner:

$$M^{-1} S = P_0 + (I - P_0) \left(\sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T D^{(l)} S^{(l)^\dagger} D^{(l)} \mathcal{R}_{B^{(l)}}\, \mathbf{r}_B \right) (I - P_0). \tag{3.107}$$

The first application of $(I - P_0)$ ensures that the residual is balanced so that when the partition of unity Neumann-Neumann preconditioner is applied, the subproblems are solvable (but with non-unique solutions). To ensure symmetry, the output of the Neumann-Neumann preconditioner is subsequently balanced by another application of the $(I - P_0)$ in a post-processing step. Since this output will lie in the subspace Kernel(C^T) of balanced vectors, the term P_0 is employed to compute the projection of the solution onto the coarse space $V_0 = \text{Kernel}(C^T)^\perp$, which is the S-orthogonal complement of the space Kernel(C^T) of balanced vectors.

Computing the action $M^{-1}\mathbf{r}_B$ of the inverse of the hybrid Schwarz preconditioner M in (3.107) involves three steps. In the first step, solve:

$$(C^T S C) \boldsymbol{\alpha} = C^T \mathbf{r}_B.$$

If $\mathbf{r}_B = S\,\mathbf{u}_B$, this yields $C\,\boldsymbol{\alpha} = P_0\,\mathbf{u}_B$. Using $S\,C\,\boldsymbol{\alpha}$, a balanced residual $\tilde{\mathbf{r}}_B$ is constructed from $\mathbf{r}_B$ by subtraction of the term $S\,C\,\boldsymbol{\alpha}$:

$$\tilde{\mathbf{r}}_B = \mathbf{r}_B - S\,C\,\boldsymbol{\alpha}.$$

In the second step, the partition of unity Neumann-Neumann preconditioner is formally applied to the balanced residual $\tilde{\mathbf{r}}_B$:

$$\mathbf{v}_B = \sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T D^{(l)} S^{(l)^\dagger} D^{(l)} \mathcal{R}_{B^{(l)}}\, \tilde{\mathbf{r}}_B.$$

In the third step, to obtain $\tilde{\mathbf{v}}_B = (I - P_0)\,\mathbf{v}_B$ requires solving the system:

$$(C^T SC)\boldsymbol{\beta} = C^T \mathbf{v}_B,$$

and defining $\tilde{\mathbf{v}}_B = (\mathbf{v}_B - S\,C\,\boldsymbol{\beta})$. Then $M^{-1}\mathbf{r}_B \equiv (S\,C\,\boldsymbol{\alpha} + \mathbf{v}_B - S\,C\,\boldsymbol{\beta})$.

Remark 3.63. System (3.106) has a block structure which can be obtained by substituting the block structure $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T\right)^T$ and the block structure of C into (3.106) to yield the following block partitioned linear system:

$$\begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & & \vdots \\ K_{1p}^T & \cdots & K_{pp} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \end{bmatrix} = \begin{bmatrix} N_1^T D^{(1)} \mathcal{R}_{B^{(1)}} \mathbf{r}_B \\ \vdots \\ N_p^T D_p \mathcal{R}_{B^{(p)}} \mathbf{r}_B \end{bmatrix}, \qquad (3.108)$$

involving $(d_1 + \cdots + d_p)$ unknowns corresponding to the subvectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$. Here, the block submatrices K_{ij} will be $d_i \times d_j$ matrices defined by:

$$K_{ij} \equiv N_i^T D^{(i)} \mathcal{R}_{B^{(i)}} S \mathcal{R}_{B^{(j)}}^T D^{(j)^T} N_j, \qquad \text{for} \quad 1 \le i,\, j \le p, \qquad (3.109)$$

and $\boldsymbol{\alpha}_i \in \mathbb{R}^{d_i}$. If $d_i = 0$ for any index i, then the corresponding block rows and columns of K and $\boldsymbol{\alpha}$ should be omitted.

Remark 3.64. In most applications, K will be symmetric and positive definite. However, when C is not of full rank, matrix K can be singular. In this case, the columns of C will be linearly dependent with:

$$\sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T D^{(l)} N_l \boldsymbol{\gamma}_l = \mathbf{0},$$

for some choice of coefficient vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$. To avoid a singular matrix K, some care must be exercised when extending each matrix $\tilde{N}_l$ to N_l.

Below, we summarize the action of the inverse of the balancing domain decomposition preconditioner.

```
Algorithm 3.7.2 (Balancing Domain Decomposition Preconditioner)
Input: $\mathbf{r}_B$.
 1. Solve:
      $\begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & & \vdots \\ K_{1p}^T & \cdots & K_{pp} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \end{bmatrix} = \begin{bmatrix} N_1^T D^{(1)} \mathcal{R}_{B^{(1)}} \mathbf{r}_B \\ \vdots \\ N_p^T D^{(p)} \mathcal{R}_{B^{(p)}} \mathbf{r}_B \end{bmatrix}.$
 2. Define:
      $\begin{cases} \mathbf{w}_B^* \equiv \sum_{j=1}^p \mathcal{R}_{B^{(j)}}^T D^{(j)^T} N_j \boldsymbol{\alpha}_j \\ \mathbf{r}_B^* \equiv \mathbf{r}_B - S \mathbf{w}_B^*. \end{cases}$
 3. For $i = 1, \cdots, p$ in parallel solve:
      $S^{(i)} \mathbf{w}_{B^{(i)}} = D^{(i)} \mathcal{R}_{B^{(i)}} \mathbf{r}_B^*.$
 4. Endfor
 5. Compute:
      $\begin{cases} \mathbf{w}_B = \sum_{j=1}^p \mathcal{R}_{B^{(j)}}^T D^{(j)^T} \mathbf{w}_{B^{(j)}} \\ \mathbf{t}_B = \mathbf{r}_B^* - S \mathbf{w}_B. \end{cases}$
 6. Solve:
      $\begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & & \vdots \\ K_{1p}^T & \cdots & K_{pp} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_p \end{bmatrix} = \begin{bmatrix} N_1^T D^{(1)} \mathcal{R}_{B^{(1)}} \mathbf{t}_B \\ \vdots \\ N_p^T D^{(p)} \mathcal{R}_{B^{(p)}} \mathbf{t}_B \end{bmatrix}.$
 7. Define:
      $\mathbf{v}_B^* \equiv \sum_{j=1}^p \mathcal{R}_{B^{(j)}}^T D^{(j)^T} N_j \boldsymbol{\beta}_j.$
Output: $M^{-1} \mathbf{r}_B \equiv \mathbf{w}_B^* + \mathbf{w}_B + \mathbf{v}_B^*.$
```

Remark 3.65. If the input $\mathbf{r}_B$ to the preconditioner is balanced, then step 1 can be omitted in the preconditioner, yielding $\mathbf{w}_B^* = \mathbf{0}$. In this case, the output $M^{-1}\mathbf{r}_B = \mathbf{w}_B + \mathbf{v}_B^*$ will also be balanced. Motivated by this, in practice steps 1 and 2 are employed in a pre-processing stage to ensure that the initial residual is balanced. Then, steps 1 and 2 can be omitted in all subsequent applications of M^{-1} in the CG algorithm. Each iteration will require one matrix multiplication with S and one multiplication by M^{-1}. Thus, the computational cost of each iteration will be proportional to the cost of two subdomain solves on each subdomain and the cost of balancing (which requires the solution of a coarse problem P_0). The following convergence bound will hold for the balanced domain decomposition preconditioner.

Theorem 3.66. *Suppose that $c(x) = 0$ and that coefficient $a(x) = a^{(i)}$ on each subdomain Ω_i. Then, if each $N_l = \tilde{N}_l$, there will be a constant C independent of h_0, h and the $\{a^{(i)}\}$ such that:*

$$\mathrm{cond}(M, S) \le C \, \left(1 + \log(h_0/h)\right)^2,$$

where M denotes the balancing domain decomposition preconditioner.

Proof. See [MA14, MA17, DR18]. □

Remark 3.67. If $c(x) > 0$, then each subdomain problem will be *nonsingular*. In this case, the *coarse space* $V_0 = \mathrm{Kernel}(C^T)^{\perp}$ will be trivial, and the convergence rate of the balancing domain decomposition preconditioner will deteriorate. However, this can be remedied by choosing a nontrivial matrix $N_l \neq \tilde{N}_l$ on each subdomain, such that $\mathrm{Kernel}(N_l)$ corresponds to the null space of $S^{(l)}$ when $c(x) = 0$ (typically with $N_l = \mathrm{Span}(\mathbf{1})$).

3.7.3 An Algebraic Preconditioner

We conclude this section by outlining an algebraic preconditioner of [CA33]. It *approximates* the following additive Schwarz preconditioner for S, based on the segments $B^{(1)}, \ldots, B^{(p)}$ of B:

$$M^{-1} = \sum_{i=1}^{p} \mathcal{R}_{B^{(i)}}^T S_{B^{(i)}B^{(i)}}^{-1} \mathcal{R}_{B^{(i)}} + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0.$$

Here $\mathcal{R}_{B^{(i)}}$ denotes a restriction matrix with zero-one entries corresponding to nodes on $B^{(i)}$, and $\mathcal{R}_0$ denotes the coarse space weighted restriction matrix, with $S_{B^{(i)}B^{(i)}} = \mathcal{R}_{B^{(i)}} S \mathcal{R}_{B^{(i)}}^T$ and $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$. An exact application of the preceding preconditioner requires assembly of the submatrices $S_{B^{(i)}B^{(i)}}$ and the coarse matrix S_0. However, an approximation $\tilde{S}_{B^{(i)}B^{(i)}} \approx S_{B^{(i)}B^{(i)}}$ can be constructed based on the **ILU** factorization $A_{II}^{(i)} \approx \tilde{L}_{(i)} \tilde{L}_{(i)}^T$ of each subdomain stiffness matrix $A^{(i)}$, with ${A^{(i)}}^{-1} \approx \tilde{L}_{(i)}^{-T} \tilde{L}_{(i)}^{-1}$:

$$\tilde{S}_{B^{(i)}B^{(i)}} \equiv \sum_{l=1}^{p} \mathcal{R}_{B^{(i)}} \mathcal{R}_{B^{(l)}}^T \left(A_{BB}^{(l)} - {A_{IB}^{(l)}}^T \tilde{L}_{(l)}^{-T} \tilde{L}_{(l)}^{-1} A_{IB}^{(l)} \right) \mathcal{R}_{B^{(l)}} \mathcal{R}_{B^{(i)}}^T.$$

Efficient algorithms for assembling such approximations are described in [CA33]. Unlike the subdomain stiffness matrices $S^{(i)}$, the algebraic approximations $\tilde{S}_{B^{(i)}B^{(i)}}$ of $S_{B^{(i)}B^{(i)}}$ will not be singular. Matrix $\tilde{S}_{B^{(i)}B^{(i)}}$ will be dense, and can be truncated to a sparse matrix, and its incomplete factorization can be found. The coarse matrix S_0 may be approximated by a coarse grid discretization A_0 of (3.1). Numerical studies indicate attractive convergence properties for such preconditioners [CA33].

3.8 Implementational Issues

Schur complement algorithms are generally more difficult to implement than Schwarz methods, since more geometric information is required about the subdomains and their boundaries (Neumann-Neumann and balancing preconditioners may be exceptions). However, an effectively preconditioned Schur complement algorithm can converge at almost optimal rates with respect to h, h_0 and jumps in the coefficient $a(.)$, just as Schwarz algorithms, where the implementation, storage and communication costs, may be reduced due to the lack of overlap between the subdomains.

In this section, we remark on implementational issues in applications of Schur complement algorithms to solve a discretization of (3.1). They include, choice of subdomains, general boundary conditions, preconditioning S or A, local solvers, parallel libraries, and remarks on discontinuous coefficient problems, anisotropic problems, and time stepped problems. The condition number bounds of several Schur complement preconditioners are summarized in Table 3.1, when the coefficient $a(.)$ is constant within each subdomain. Estimates are presented for the case when the jumps in $a(\cdot)$ are *mild*, and when the jumps are *large*. $C(a)$ denotes a parameter independent of h_0 and h but dependent on the coefficient $a(\cdot)$, while C is independent of h_0, h and $a(\cdot)$. For the vertex space algorithm $C(\beta)$ depends on the overlap factor β.

3.8.1 Choice of Subdomains

Various factors influence the choice of a decomposition $\Omega_1, \ldots, \Omega_p$ of Ω. These include, the geometry of the domain, location of the essential and natural boundary, regularity of the solution, availability of fast local solvers, and heterogeneity of the coefficients. For instance, when $a(.)$ has large jumps, the subdomains should ideally be aligned with the discontinuities in $a(.)$, to reduce the variation of $a(.)$ within each subdomain. For anisotropic coefficients, strip like subdomains may be chosen so that the elliptic equation is coupled more strongly within the strips. When a natural decomposition is not obvious, an automated strategy, see Chap. 5.1, may be employed to minimize the communication between the subdomains, and to balance the loads [BE14, FO2, SI2, FA9, BA20, PO3, PO2].

Table 3.1. Condition number bounds for Schur complement preconditioners

Algorithm	Mild Coeff.	Disc Coeff.
2D BPS	$C\left(1+\log^2(h_0/h)\right)$	$C\left(1+\log^2(h_0/h)\right)$
2D Vertex Space	$C(a)\left(1+\log^2(\beta^{-1})\right)$	$C(\beta)\left(1+\log^2(h_0/h)\right)$
3D Vertex Space	$C(a)\left(1+\log^2(\beta^{-1})\right)$	$C(\beta)(h_0/h)$
3D Wirebasket	$C\left(1+\log^2(h_0/h)\right)$	$C\left(1+\log^2(h_0/h)\right)$
Neumann & Balancing	$C\left(1+\log^2(h_0/h)\right)$	$C\left(1+\log^2(h_0/h)\right)$

3.8.2 General Boundary Conditions

Our discussion of Schur complement preconditioners has focused primarily on Dirichlet problems, i.e., for $\mathcal{B}_D = \partial\Omega$. When more general boundary conditions are imposed, the natural boundary $\mathcal{B}_N \neq \emptyset$. and the solution will be unknown not only in Ω, but also in $\mathcal{B}_N$. In this case, the triangulation must ideally be chosen so that its elements are aligned with $\mathcal{B}_N$. Then, given a decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, the nodal unknowns can in principle be block partitioned in *two* alternate ways, yielding two different Schur complement systems. In the following, we shall indicate both *block partitionings*, and remark on the construction of Schur complement preconditioners for a discretization of (3.1) with stiffness matrix A and load vector $\mathbf{f}$.

In both of the above cases, a discretization of (3.1) can be block partitioned as in (3.5), using the block vectors $\mathbf{u}_I$ and $\mathbf{u}_B$ of unknowns, yielding a Schur complement $S = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$. However, A_{II}, A_{IB} and A_{BB} will have different sizes for each partition.

First Case. In the first block partitioning, each $\mathbf{u}_I^{(l)}$ will denote a vector of unknowns in $\Omega_l \cup (\partial\Omega_l \cap \mathcal{B}_N)$, while $\mathbf{u}_B^{(l)}$ will denote unknowns on $(\partial\Omega_l \cap \Omega)$. Thus, the unknowns on $\mathcal{B}_N \cap \partial\Omega_l$ will be included in $\mathbf{u}_I^{(l)}$ though they do not strictly lie in the interior of the subdomain, while $B = \cup_{l=1}^p (\partial\Omega_l \cap \Omega)$ will not include the natural boundary $\mathcal{B}_N$. We then define $\mathbf{u}_I = (\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T})^T$, and let $\mathbf{u}_B$ denote the vector of nodal values on B. In this case, Schur complement preconditioners can be constructed as for a Dirichlet problem, since the interface B will be identical to the interface for a Dirichlet problem, and it can be decomposed into globs or overlapping segments, as before. However, the subdomain matrix $A_{II}^{(l)}$ will involve natural boundary conditions on $(\partial\Omega_l \cap \mathcal{B}_N)$. Care must be exercised in defining a coarse space when $\mathcal{B}_N \neq \emptyset$, since the coarse space must be a subspace of $V_h \cap H_D^1(\Omega)$.

Second Case. In the second block partitioning, each $\mathbf{u}_I^{(l)}$ will denote unknowns in Ω_l and $\mathbf{u}_I = (\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T})^T$. We shall define the "interface" as $B = \cup_{l=1}^p \partial\Omega_l \cap (\Omega \cup \mathcal{B}_N)$ and let $\mathbf{u}_B$ denote the unknowns on B. Since B will include the natural boundary $\mathcal{B}_N$, it may be difficult to decompose it into standard globs if $\mathcal{B}_N$ has an irregular shape. This may complicate the formulation of glob based preconditioners (such as block Jacobi, vertex space and wirebasket preconditioners), and it may also be difficult to formulate a traditional coarse space. However, given a decomposition of B into globs or overlapping segments, Schwarz subspace preconditioners can be formulated for S, and the subdomain matrix $A_{II}^{(l)}$ will only involve interior nodal unknowns in Ω_l. Neumann-Neumann and balancing methods apply in both cases.

Remark 3.68. If $\mathcal{B}_N \neq \emptyset$ and $\mathcal{B}_D \neq \emptyset$, then stiffness matrix A and the Schur complement matrix S, will be *nonsingular*. However, if $\mathcal{B}_N = \partial\Omega$ and coefficient $c(x) = 0$, then stiffness matrix A and the Schur complement matrix S

will be *singular*. In this case, the coarse space matrix $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$ will also be singular. As a result, the Schur complement system $S\mathbf{u}_B = \tilde{\mathbf{f}}_B$ and the coarse problem $S_0 \mathbf{w}_0 = \mathcal{R}_0 \mathbf{r}_B$ will be solvable only if $\mathbf{1}^T \tilde{\mathbf{f}}_B = 0$ and $\mathbf{1}^T \mathcal{R}_0 \mathbf{r}_B = 0$, respectively, for $\mathbf{1} = (1, \ldots, 1)^T$. To obtain a unique solution, each iterate should be normalized to have zero mean value. For instance, if $\mathbf{w}_B \in \mathbb{R}^{n_B}$ denotes the output of the preconditioned system in the k'th iterate, then modify it to have mean value zero:

$$\mathbf{w}_B \leftarrow \mathbf{w}_B - \left(\frac{\mathbf{1}^T \mathbf{w}_B}{\mathbf{1}^T \mathbf{1}} \right) \mathbf{1}.$$

Such normalizations will not be needed when $c(x) \neq 0$.

3.8.3 Preconditioning S or A

Given subdomains $\Omega_1, \ldots, \Omega_p$ of Ω, the solution to (3.5) may in principle be sought in two alternate ways. In the first approach, the Schur complement system may be solved for $\mathbf{u}_B$ using Alg. 3.2.1 and a CG algorithm with an appropriate preconditioner for S. Once $\mathbf{u}_B$ has been determined, $\mathbf{u}_I$ can be obtained at the cost of one subdomain solve. This approach will require matrix-vector products with S computed exactly (to machine precision), and so require solving systems of the form $A_{II}^{(l)} \mathbf{w}_I^{(l)} = \mathbf{r}_I^{(l)}$ exactly (to machine precision) each iteration. A sparse direct solver may be used for $A_{II}^{(l)}$.

In the second approach, the global stiffness matrix A is solved by a preconditioned CG algorithm, where the action of the inverse of the preconditioner $\tilde{A}$ for A has the following block matrix structure:

$$\begin{aligned} \tilde{A}^{-1} &= \begin{bmatrix} I & -\tilde{A}_{II}^{-1} A_{IB} \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{A}_{II}^{-1} & 0 \\ 0 & \tilde{S}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\, A_{IB}^T \tilde{A}_{II}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} I & -\tilde{A}_{II}^{-1} A_{IB} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \tilde{S}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\, A_{IB}^T & I \end{bmatrix} \begin{bmatrix} \tilde{A}_{II}^{-1} & 0 \\ 0 & I \end{bmatrix}, \end{aligned} \tag{3.110}$$

where $\tilde{S}$ denotes a preconditioner for S and $\tilde{A}_{II}$ a preconditioner for A_{II}. Computing the solution to $\tilde{A}\mathbf{u} = \mathbf{f}$ formally requires computing the action of $\tilde{A}_{II}^{-1}$ twice, and $\tilde{S}^{-1}$ once. The advantage is that an exact solver is not required for $A_{II}^{(l)}$, but the disadvantage is that the inexact solver must be applied twice.

Remark 3.69. The second approach has not be studied extensively. If a preconditioner $\tilde{A}$ is employed for A, it is important that the submatrices $\tilde{A}_{II}^{(i)}$ and $\tilde{A}_{IB}$ be *scaled* similar to $A_{II}^{(i)}$ and A_{IB}, respectively, or the convergence rate can deteriorate significantly to $O(h^{-2})$ even if $\mathrm{cond}(\tilde{A}_{II}^{(i)}, A_{II}^{(i)}) = 1$, see [BO4].

3.8.4 Local Solvers, Parallelization and Libraries

Typically, sparse direct solvers are employed for solving the subdomain problems arising in a Schur complement algorithm. In some applications, however, FFT based solvers and iterative solvers are used for subdomain problems. In the Schur complement method, the action of A_{II}^{-1} and the action of preconditioners typically involve parallel tasks, which require synchronization between the processors assigned to different subdomains. Importantly, the PETSc library contains parallel codes implementing most Schur complement algorithms, see Chap. 2.4 for additional comments on local solvers, parallelization, and the MPI and PETSc libraries.

3.8.5 Remarks on Discontinuous Coefficient Problems

When $a(.)$ has *large* jump discontinuities, care must be exercised in the choice of a subdomain decomposition and a coarse problem, or the rate of convergence of a Schur preconditioned algorithm can deteriorate. Ideally, the *subdomains* must align with the discontinuities of $a(.)$, i.e., if Γ denotes the curve or surface along which the coefficient $a(.)$ is discontinuous, then $\Gamma \subset B = \cup_{i=1}^{p} \partial \Omega_i$. If an initial decomposition of Ω yields subdomains on which $a(.)$ is smooth, then larger subdomains may be further decomposed to improve load balancing. Choosing subdomains with reduced variation in $a(.)$ also yields better conditioned local problems.

Another consideration is the choice of a *coarse space*. Theoretical bounds for Schur complement preconditioners, are better when a coarse space is included. For instance, on a two dimensional domain, typical bounds are $\mathcal{O}\left(1+\log(h_0/h)\right)^2$ when a traditional coarse space is employed, provided the coefficient $a(.)$ is constant within each subdomain. For a three dimensional domain, such bounds can deteriorate to $\mathcal{O}\left((h_0/h)(1+\log(h_0/h))^2\right)$ when a traditional coarse based on a coarse triangulation is employed, but improve to $\mathcal{O}\left(1+\log(h_0/h)\right)^2$ when a *piecewise constant* coarse space is employed (see Remark 3.70 below). Other coarse spaces include wirebasket and partition of unity spaces, see [BR15, MA12, WI6, DR10, SA11, SA12]. For a Schur complement preconditioner with optimal order complexity, see [NE5].

Remark 3.70. The "piecewise constant coarse space" $V_{0,P}$ is defined as follows. Let n_B and $n_{B^{(i)}}$ denote the number of nodes on B and $B^{(i)}$, respectively. Let N_i denote a matrix of size $n_{B^{(i)}}$ whose columns form a basis for the null space of the local Schur complement matrix $S^{(i)}$ when $c(x) = 0$. For 2nd order scalar elliptic equations $N_i = (1, \ldots, 1)^T$. Let $D^{(i)}$ be a diagonal matrix of size $n_{B^{(i)}}$ with nonnegative entries defined by (3.97). Then, $V_{0,P} \equiv \text{Range}(\mathcal{R}_0^T)$ where:

$$\mathcal{R}_0^T = \left[\mathcal{R}_{B^{(1)}}^T D^{(1)^T} N_1 \ \cdots \ \mathcal{R}_{B^{(p)}}^T D^{(p)^T} N_p \right].$$

Such a coarse space will be defined even when the subdomains do not form a coarse triangulation of Ω, see [CO8, MA14, SA7, SA8, MA15].

3.8.6 Remarks on Anisotropic Problems

To motivate Schur complement algorithms for anisotropic problems, consider the following model equation:

$$\begin{cases} -\alpha_1\, u_{x_1x_1} - \alpha_2\, u_{x_2x_2} = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{3.111}$$

posed on a domain $\Omega \equiv (-L_1, L_2) \times (0,1) \subset \mathbb{R}^2$ with parameters $\alpha_1 > 0$ and $\alpha_2 > 0$ which determine the degree of anisotropy in the equation. This problem will be strongly anisotropic when $(\alpha_1/\alpha_2) \ll 1$ or $(\alpha_1/\alpha_2) \gg 1$. When this holds, elliptic equation (3.111) may be of singular perturbation type with boundary layers in the solution [KE5, LA5]. However, we shall assume that the boundary layer need not be captured, and instead heuristically motivate issues for consideration when formulating a Schur complement preconditioner.

Consider a discretization of the above equation on a uniform grid and suppose that Ω is partitioned into *vertical strip* subdomains. Suppose that the unknowns are ordered consecutively along each vertical line $x_1 = c$, with increasing indices as x_2 increases and as x_1 increases. Then, the coefficient matrix A will have a block tridiagonal structure, as in Chap. 3.3, and its eigendecomposition may be obtained exactly. The following special limiting cases may be noted.

When $\boldsymbol{\alpha_1 = 1}$ and $\boldsymbol{\alpha_2 \to 0^+}$. If $\alpha_1 = 1$ and $\alpha_2 \to 0^+$, then the linear system will be strongly coupled along the x_1-axis, but weakly coupled along the x_2 axis. As a result, each diagonal block of S will formally approach a scalar multiple of the identity (and will be well conditioned), but S will still have a block tridiagonal structure in this limiting case. In particular, if the off diagonal blocks in S are neglected when a preconditioner is formulated, it will result in deteriorated convergence rates.

When $\boldsymbol{\alpha_1 \to 0^+}$ and $\boldsymbol{\alpha_2 = 1}$. If $\alpha_2 = 1$ and $\alpha_1 \to 0^+$, then the linear system will be strongly coupled along the x_2-axis, but weakly coupled along the x_1 axis. Formally, A_{IB} will be proportional to α_1, and A_{II} will remain nonsingular as $\alpha_1 \to 0^+$, yielding that $S = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}) \to A_{BB}$ as $\alpha_1 \to 0^+$, and the off diagonal blocks of S will formally approach zero. The diagonal blocks of S approach a discretization of $-(\partial^2/\partial_{x_2}^2)$, yielding an ill-conditioned matrix. This suggests A_{BB} as a heuristic preconditioner for S.

The limiting cases above indicate that the square root of the discrete Laplace-Beltrami operator on the interface B will generally *not* be an effective preconditioner for S in the strongly anisotropic case. The traditional norm equivalence between the subdomain Schur complement energy $\mathbf{u}^{(i)^T} S^{(i)} \mathbf{u}^{(i)}$ and the fractional Sobolev energy $|u_h|_{1/2,\partial\Omega_i}^2$ on the subdomain boundary:

$$c_1\, |u_h|_{1/2,\partial\Omega_i}^2 \le \mathbf{u}^{(i)^T} S^{(i)} \mathbf{u}^{(i)} \le c_2\, |u_h|_{1/2,\partial\Omega_i}^2,$$

will deteriorate for an anisotropic problem, with the ratio (c_2/c_1) increasing in proportion to the anisotropy in $a(.)$.

However, a preconditioner based on subdomain Schur complements (such as the Neumann-Neumann or balancing preconditioner), or one based on algebraic approximation may be employed. Heuristically, for the former, each subdomain problem may have similar anisotropic limits, while for the latter, an algebraic approximation may be constructed to have the same anisotropic limits. Depending on the alignment of the sides of the subdomains relative to direction of weak coupling, a coarse space may be required. For instance, if $a(.)$ is a constant (or mildly varying) but strongly anisotropic matrix function on a domain Ω (not necessarily rectangular). Then, *strip subdomains* may be chosen so that the equation is strongly coupled within each strip, with sides perpendicular to the direction in which the equation is weakly coupled. Then, the Schur complement matrix will have a block tridiagonal structure, and by analogy with the model problem as $\alpha_1 \to 0^+$, S will formally approach A_{BB} in the limit. Matrix A_{BB} may then be employed as a *heuristic* algebraic preconditioner for S (without coarse space correction). However, if the strips were chosen with its sides perpendicular to an axis of strong coupling, as when $\alpha_2 \to 0^+$, then a coarse space will be required.

In three dimensions, the coefficient matrix $a(x)$ will have three eigenvalues for each $x \in \Omega$ and the elliptic equation will be strongly anisotropic if either one or two eigenvalues of $a(x)$ are very small relative to the others. When $a(.)$ is a constant matrix having only one relatively small eigenvalue, then the elliptic equation will be strongly coupled on planes perpendicular to the eigenvector of $a(.)$ corresponding to the smallest eigenvalue. When $a(.)$ is a constant matrix having two relatively small eigenvalues, then the elliptic equation will be strongly coupled along rays (lines) parallel to the eigenvector associated with the largest eigenvalue. Heuristically, strip subdomains may still be employed, provided its sides are perpendicular to the eigenvector associated with the smallest eigenvalue of $a(.)$.

3.8.7 Remarks on Time Stepped Problems

In time stepped problems, the condition number of an *unpreconditioned* Schur complement matrix improves with decreasing time step. However, care must be exercised, if a preconditioner is employed, We consider an implicit scheme in time and a finite difference discretization in space for the parabolic equation:

$$\begin{cases} u_t + L\,u = f, & \text{in } \Omega \times (0,T) \\ \qquad\quad u = 0, & \text{on } \partial\Omega \times (0,T) \\ \quad u(x,0) = u_0(x), & \text{in } \Omega, \end{cases}$$

where $Lu \equiv -\nabla \cdot (a\nabla u)$. This will yield a linear system $(I + \tau\, A)\,\mathbf{u} = \tilde{\mathbf{f}}$, at each time step, where $0 < \tau$ denotes the time step and $(I + \tau\, A)$ corresponds to a finite difference discretization of the elliptic operator $(I + \tau\, L)$.

Given a nonoverlapping decomposition $\Omega_1, \ldots, \Omega_p$ of Ω with interface B, block partition $\mathbf{u} = (\mathbf{u}_I^T, \mathbf{u}_B^T)^T$ and $\tilde{\mathbf{f}} = (\tilde{\mathbf{f}}_I^T, \tilde{\mathbf{f}}_B^T)^T$, based on the subdomain interiors $\cup_{i=1}^{p} \Omega_i$ and interface B. The time stepped system $(I + \tau\, A)\, \mathbf{u} = \tilde{\mathbf{f}}$, will have the following block structure:

$$\begin{bmatrix} I + \tau A_{II} & \tau A_{IB} \\ \tau A_{IB}^T & I + \tau A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_I \\ \tilde{\mathbf{f}}_B \end{bmatrix}. \tag{3.112}$$

The Schur complement system will be:

$$S(\tau)\, \mathbf{u}_B = (\tilde{\mathbf{f}}_B - \tau A_{IB}^T (I + \tau A_{II})^{-1} \tilde{\mathbf{f}}_I),$$

where the Schur complement matrix $S(\tau)$ satisfies:

$$S(\tau) = \Big(I + \tau A_{BB} - \tau^2\, A_{IB}^T\, (I + \tau\, A_{II})^{-1}\, A_{IB} \Big).$$

Due to τ and h dependent terms, a preconditioner $M(\tau)$ must ideally adapt to both parameters uniformly, such as would be heuristically expected for the Neumann-Neumann and balancing preconditioners. For such preconditioners, by *heuristic* analogy with Schwarz algorithms, we expect that a coarse space may not be required if some time step constraint of the form $\tau \,\leq\, c\, h_0^2$ holds.

Remark 3.71. In the strip or two subdomain case, FFT based preconditioners $M(\tau)$ can be constructed to adapt to the τ and h dependent terms. However, using a *fixed* FFT based based preconditioner M for $S(\tau)$, such as the square root of the discrete Laplace-Beltrami matrix for a two subdomain decomposition, will not perform uniformly, since formally, $S(\tau) \to I$ as $\tau \to 0^+$, for a fixed h. The entries of A_{BB}, A_{IB} and A_{II} grow as $\mathcal{O}(h^{-2})$ as $h \to 0^+$.

In time stepped problems, it may also be of interest to formulate a stable *one iteration* algorithm which computes the discrete solution at each time step to within the local truncation error, see [DA4, DA5, DR5, LA3, LA4, ZH5]. Below, we outline a *heuristic* approach based on a subassembly identity for the time stepped Schur complement $S(\tau)$ in terms of $S^{(l)}(\tau)$:

$$S(\tau) = \sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T S^{(l)}(\tau) \mathcal{R}_{B^{(l)}},$$

where each $S^{(l)}(\tau) = (I^{(l)} + \tau A_{BB}^{(l)}) - \tau^2 A_{IB}^{(l)^T} (I + \tau A_{II}^{(l)})^{-1} A_{IB}^{(l)}$ is a subdomain Schur complement, and $I = \sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T I^{(l)} \mathcal{R}_{B^{(l)}}$ forms an algebraic partition of the identity. Given such a decomposition, we may split $S(\tau)$ as:

$$S(\tau) = I + \tau \sum_{l=1}^{p} \mathcal{R}_{B^{(l)}}^T \left(A_{BB}^{(l)} - \tau\, A_{IB}^{(l)^T} (I + \tau A_{II}^{(l)})^{-1} A_{IB}^{(l)} \right) \mathcal{R}_{B^{(l)}},$$

and apply a generalized ADI (alternating directions implicit) method to construct an approximate solution [DR5, LA3, LA4, VA, VA2], see Chap. 9. For an alternative scheme, see [ZH5]. Time step constraints may apply.

3.9 Theoretical Results

In this section, we describe theoretical methods for estimating the condition number of selected Schur complement preconditioners. We focus primarily on the dependence of the condition numbers on the mesh parameter h and subdomain size h_0, and in some cases on the jumps in the coefficients $a(\cdot)$. To obtain such bounds, we will employ the abstract Schwarz convergence theory described in Chap. 2.5, and employ theoretical properties of elliptic equations and Sobolev norms to estimate the dependence of partition parameters on the mesh size h, subdomain size h_0 and jumps in the coefficient $a(.)$.

Our discussion will be organized as follows. In Chap. 3.9.1, we introduce scaled Sobolev norms, Poincaré-Freidrich's inequalities, and trace and extension theorems. We use these background results to derive an equivalence between the energy associated with the Schur complement matrix and a scaled sum of fractional Sobolev norm energies. Chap. 3.9.2 describes *discrete Sobolev inequalities* for finite element spaces and uses them to prove a result referred to as the *glob theorem* (our proof will hold only in two dimensions), useful in estimating partition parameters for glob based algorithms. In Chap. 3.9.3, we describe theoretical properties of the traditional and piecewise constant coarse spaces. In Chap. 3.9.4, we estimate the condition number of several *two subdomain* preconditioners. In Chap. 3.9.5, we estimate the condition number of multisubdomain block Jacobi, BPS and vertex space preconditioners. We omit theoretical discussion of wirebasket preconditioners. In Chap. 3.9.6, we describe estimates for the condition number of the *balancing domain decomposition preconditioner.*

3.9.1 Background Results

We will consider a finite element discretization of elliptic equation (3.1) on a quasiuniform triangulation $\mathcal{T}_h(\Omega)$ of Ω. The domain Ω will be assumed to be partitioned into nonoverlapping subdomains $\Omega_1, \ldots, \Omega_p$ which forms a quasiuniform triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω of diameter h_0. The coefficient $c(.)$ in (3.1) will be assumed to be zero, while coefficient $a(.)$ will be assumed to be constant on each subdomain:

$$\begin{cases} a(x) = \rho_i, & \text{for } x \in \Omega_i, \text{ for } 1 \le i \le p \\ c(x) = 0, & \text{in } \Omega. \end{cases}$$

We will denote the finite element space defined on Ω as $V_h(\Omega)$ and by $V_h(D)$ the space of finite element functions restricted to D, for any subregion $D \subset \Omega$ (including $D \subset B$). In most applications, we assume that the finite element space consists of continuous piecewise linear finite elements.

The following *scaled* norms and seminorms will be employed throughout this section [NE, LI4, DR2, GR8, BR15, DR14, DR10, MA17]:

$$\begin{cases} |u|_{1,\Omega_i}^2 & \equiv \int_{\Omega_i} |\nabla u|^2 dx \\ \|u\|_{1,\Omega_i}^2 & \equiv \int_{\Omega_i} |\nabla u|^2 dx + \frac{1}{h_0^2} \int_{\Omega_i} |u|^2 dx \\ |u|_{1/2,B^{(i)}}^2 & \equiv \int_{B^{(i)}} \int_{B^{(i)}} \frac{|u(x)-u(y)|^2}{|x-y|^d} dx dy, \qquad \Omega_i \subset \mathbb{R}^d \\ \|u\|_{1/2,B^{(i)}}^2 & \equiv \int_{B^{(i)}} \int_{B^{(i)}} \frac{|u(x)-u(y)|^2}{|x-y|^d} dx dy + \frac{1}{h_0} \int_{B^{(i)}} |u|^2 dx. \end{cases} \tag{3.113}$$

By construction, the above norms and seminorms will scale similarly under dilations of the underlying domain in $\mathbb{R}^d$. As a consequence, we may map a subdomain Ω_i of width h_0 to a reference domain of width 1, apply trace or extension theorems on the reference domain, and map the results back to the original domain and obtain estimates of the norms in the trace and extension theorems independent of the width h_0 of the subdomain. We will thus assume heuristically that the bounds in the trace, extension, and Poincaré-Freidrich's type inequalities are independent of h_0 when scaled norms are employed.

We will frequently encounter norms of the form $\|v\|_{1/2,\partial\Omega_i}^2$ when the function $v(\cdot) \in H^{1/2}(\partial\Omega_i)$ is zero outside some subregion $D_i \subset \partial\Omega_i$. In such cases, the norm $\|v\|_{1/2,\partial\Omega_i}$ will be *stronger* than $\|v\|_{1/2,D_i}$ and will be denoted $\|v\|_{H_{00}^{1/2}(D_i)}^2$ as formalized below.

Definition 3.72. *Let $D_i \subset \partial\Omega_i$. We define an extension by zero map $\mathcal{E}_0$ as:*

$$\mathcal{E}_0\, v = \begin{cases} v & \text{on } D_i \\ 0 & \text{in } \partial\Omega_i \backslash D_i, \end{cases}$$

and define a Sobolev space $H_{00}^{1/2}(D_i)$ and its norm by:

$$\begin{cases} H_{00}^{1/2}(D_i) & \equiv \left\{ v \in H^{1/2}(D_i) : \mathcal{E}_0 v \in H^{1/2}(\partial\Omega_i) \right\} \\ \|v\|_{H_{00}^{1/2}(D_i)} & \equiv \|\mathcal{E}_0 v\|_{1/2,\partial\Omega_i}. \end{cases} \tag{3.114}$$

Substitution of the above definition into the integral form of the fractional Sobolev norm on $H^{1/2}(\partial\Omega_i)$ yields:

$$\|v\|_{H_{00}^{1/2}(D_i)}^2 \equiv \int_{D_i} \int_{D_i} \frac{|v(x)-v(y)|^2}{|x-y|^d} dx dy + 2 \int_{D_i} \int_{\partial\Omega_i \backslash D_i} \frac{|v(x)|^2}{|x-y|^d} dy dx + \frac{\|u\|_{0,D_i}^2}{h_0}.$$

When $\Omega_i \subset \mathbb{R}^2$, this is easily verified to be *equivalent* to:

$$\|v\|_{H_{00}^{1/2}(D_i)}^2 \equiv \int_{D_i} \int_{D_i} \frac{|v(x)-v(y)|^2}{|x-y|^d} dx dy + \frac{1}{h_0} \int_{D_i} \frac{|u(x)|^2}{\operatorname{dist}(x, \partial\Omega_i \backslash D_i)} dx.$$

Here $\operatorname{dist}(x, \partial\Omega_i \backslash D_i)$ denotes the distance of x to $\partial\Omega_i \backslash D_i$. Importantly, the fractional Sobolev space $H_{00}^{1/2}(D_i)$ may also be defined *equivalently* as an *interpolation space* of the form $[L^2(D_i), H_0^1(D_i)]_{1/2}$ using interpolation between embedded Hilbert spaces, see [LI4, BA3, BE16]. This equivalence enables an alternate *formal* representation of fractional Sobolev spaces and their norm using fractional powers of eigenvalues in the spectral expansion of a Laplace-Beltrami operator associated with the underlying spaces, as described below.

Lemma 3.73. *Suppose the following assumptions hold.*

1. *Let* $H_0^1(D_i) = \left\{ v : \mathcal{E}_0 v \in H^1(\partial\Omega_i) \right\}$ *and let* $-\Delta_{D_i}$ *formally denote a self adjoint coercive operator which generates the Dirichlet form:*

$$(-\Delta_{D_i} u, u)_{L^2(D_i)} \equiv \|u\|_{1,D_i}^2, \quad \forall u \in H_0^1(D_i) \subset L^2(D_i),$$

as guaranteed by the Riesz representation theorem. Let:

$$-\Delta_{D_i} = \sum_{l=1}^{\infty} \lambda_l P_l,$$

denote its spectral representation where each P_l *denotes an* $L^2(D_i)$-*orthogonal projection onto the null space of* $-\Delta_{D_i}$ *associated with eigenvalue* $\lambda_l > 0$.

2. *Formally define the fractional power* $\left(-\Delta_{D_i}\right)^{1/2}$ *of operator* $-\Delta_{D_i}$ *as:*

$$\left(-\Delta_{D_i}\right)^{1/2} = \sum_{l=1}^{\infty} \lambda_l^{1/2} P_l.$$

Then, the fractional Sobolev space $H_{00}^{1/2}(D_i)$ *defined by (3.114) will satisfy:*

$$H_0^{1/2}(D_i) = \left\{ v \in L^2(D_i) : -\Delta_{D_i}^{1/2} v \in L^2(D_i) \right\},$$

while its fractional Sobolev norm will satisfy:

$$c \sum_{l=1}^{\infty} \lambda_l^{1/2} \|P_l v\|_{0,D_i}^2 \leq \|v\|_{H_{00}^{1/2}(D_i)}^2 \leq C \sum_{l=1}^{\infty} \lambda_l^{1/2} \|P_l v\|_{0,D_i}^2,$$

for some $0 < c < C$.

Proof. See [LI4, BA3, BE16]. □

We next describe a result referred to as the *Poincaré-Freidrich's* inequality, which establishes a bound for the $L^2(\Omega_i)$ norm of a function in terms of one of its Sobolev seminorms, provided the function either has zero mean value on the underlying domain, or is zero on a segment of the boundary.

Lemma 3.74 (Poincaré-Freidrich's). *The following bounds will hold.*

1. *If $v \in H^1(\Omega_i)$ satisfies $\int_{\Omega_i} v dx = 0$ then:*
$$\begin{cases} \|v\|^2_{0,\Omega_i} \le C |v|^2_{1,\Omega_i} \\ \|v\|^2_{1,\Omega_i} \le (1+C)\,|v|^2_{1,\Omega_i}, \end{cases}$$
for some $C > 0$ independent of v and h_0.
2. *If $v \in H^1(\Omega_i)$ satisfies $v = 0$ on $D_i \subset \partial\Omega_i$ where measure$(D_i) > 0$, then:*
$$\begin{cases} \|v\|^2_{0,\Omega_i} \le C |v|^2_{1,\Omega_i} \\ \|v\|^2_{1,\Omega_i} \le (1+C)\,|v|^2_{1,\Omega_i}, \end{cases}$$
for some $C > 0$ independent of v and h_0.
3. *If $g \in H^{1/2}(B^{(i)})$ satisfies $\int_{B^{(i)}} g ds = 0$, then:*
$$\begin{cases} |g\|^2_{0,B^{(i)}} & \le C\,|g|^2_{1/2,B^{(i)}} \\ |g\|^2_{1/2,B^{(i)}} & \le (1+C)\,|g|^2_{1/2,B^{(i)}}, \end{cases}$$
for some $C > 0$ independent of g and h_0.
4. *If $g \in H^{1/2}(B^{(i)})$ satisfies $g = 0$ on $D_i \subset \partial\Omega_i$ where measure$(D_i) > 0$, then:*
$$\begin{cases} \|g\|^2_{0,B^{(i)}} & \le C\,|g|^2_{1/2,B^{(i)}} \\ \|g\|^2_{1/2,B^{(i)}} & \le (1+C)\,|g|^2_{1/2,B^{(i)}}, \end{cases}$$
for some $C > 0$ independent of g and h_0.

Proof. See [NE]. □

For the choice of scaled Sobolev norms and seminorms defined in (3.113), the parameter C will be independent of h_0. Additionally, since the seminorms are invariant under shifts by constants, the first and third Poincaré-Freidrich's inequalities may equivalently be stated in the *quotient* space $H^1(\Omega_i)/\mathbb{R}$ or $H^{1/2}(B^{(i)})/\mathbb{R}$, respectively. The next result we describe is referred to as a *trace* theorem, and states that when $\Omega_i \subset \mathbb{R}^d$ for $d = 2, 3$, functions in $H^1(\Omega)$ will have boundary values (or trace) of some regularity (smoothness).

Theorem 3.75 (Trace Theorem). *If $v \in H^1(\Omega_i)$, then its restriction to the boundary $\partial\Omega_i$ will be well defined, with $v \in H^{1/2}(\partial\Omega_i) \subset L^2(\partial\Omega_i)$, and:*
$$\|v\|_{1/2,\partial\Omega_i} \le C \|v\|_{1,\Omega_i},$$
where $C > 0$ is independent of v and h_0.

Proof. See [NE, LI4, GR8]. The parameter C will be independent of h_0 because of the scaled norms employed. □

The linear mapping of $v \in H^1(\Omega_i)$ to its boundary value $v \in H^{1/2}(\partial\Omega_i)$ is not only bounded, it is *surjective*, see [NE, LI4, GR8]. As a consequence, by the closed graph theorem this mapping will have a bounded right inverse. This result is stated below, and referred to as an *extension* theorem.

Theorem 3.76 (Extension Theorem). *There exists a bounded linear map $\mathcal{E} : H^{1/2}(\partial\Omega_i) \rightarrow H^1(\Omega_i)$ such that for each $g \in H^{1/2}(\partial\Omega_i)$:*

$$\mathcal{E}\, g = g \qquad \text{on } \partial\Omega_i,$$

satisfying the following bound:

$$\|\mathcal{E}g\|_{1,\Omega_i} \leq C\, \|g\|_{1/2,\partial\Omega_i}$$

for $C > 0$ independent of g and h_0.

Proof. See [ST7, LI4, GR8]. The independence of C from h_0 is a consequence of the scaled norms employed. □

As we will be working with finite element functions, we will require a discrete version of the preceding extension theorem in which the extended function is a finite element function. We refer to such a result as a *discrete extension theorem.*

Lemma 3.77. *Let Ω_i be a polygonal domain of size h_0 triangulated by a grid $\mathcal{T}_h(\Omega_i)$ quasiuniform of size h. Then there exists a bounded linear map:*

$$\mathcal{E}_h : V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i) \rightarrow V_h(\Omega_i) \cap H^1(\Omega_i),$$

such that for $g_h \in V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i)$

$$\mathcal{E}_h g_h = g_h, \qquad \text{on } \partial\Omega_i,$$

with the following bound holding:

$$\|\mathcal{E}_h g_h\|_{1,\Omega_i} \leq C\|g_h\|_{1/2,\partial\Omega_i},$$

where $C > 0$ is independent of g_h, h and h_0.

Proof. We will outline a proof when $\Omega_i \subset \mathbb{R}^2$, in which case the solution to Laplace's equation has sufficiently regular solutions. More general results are described in [AS4, WI, BJ9, BR11, NE6]. To construct a finite element extension $\mathcal{E}_h g_h \in V_h(\Omega_i)$, given $g_h \in V_h(\partial\Omega_i)$, we will first extend g_h to the interior of the subdomain as a harmonic function $\mathcal{H}g_h$:

$$\begin{cases} -\Delta(\mathcal{H}g_h) = 0, & \text{in } \Omega_i \\ \quad\;\; \mathcal{H}g_h = g_h, & \text{on } \partial\Omega_i. \end{cases}$$

Applying the continuous extension theorem and using the weak formulation of Laplace's equation on Ω_i, it can easily be shown that $\mathcal{H}g_h \in H^1(\Omega_i)$.

Furthermore, $\mathcal{H}g_h$ will satisfy the *a priori* bound:

$$\|\mathcal{H}g_h\|_{1,\Omega_i} \leq C\, \|g_h\|_{1/2,\partial\Omega_i}$$

where $C > 0$ is independent of g_h and h. The harmonic extension $\mathcal{H}g_h$ will, however, not be a finite element function. So define $\mathcal{E}_h g_h$ as the interpolant $I_h \mathcal{H} g_h$ of $\mathcal{H}g_h$ onto the finite element space $V^h(\Omega_i)$:

$$\mathcal{E}_h g_h \equiv I_h \mathcal{H} g_h.$$

Since $\mathcal{H}g_h$ is a harmonic function, it will be continuous in the interior and so the interpolant $I_h\mathcal{H}g_h$ will be well defined on the interior nodes of Ω_i. By construction, the interpolant $I_h\mathcal{H}g_h$ is well defined on the boundary $\partial\Omega_i$ since $\mathcal{H}g_h = g_h$ is continuous and piecewise polynomial on $\partial\Omega_i$. Thus, $I_h\mathcal{H}g_h$ will be well defined in $V_h(\Omega)$. We now verify that the discrete extension map $\mathcal{E}_h$ is bounded. Since g_h is continuous and piecewise polynomial it will hold that $g_h \in H^1(\partial\Omega_i)$. Consequently, the harmonic extension $\mathcal{H}g_h$ will be $H^{1+\epsilon}(\Omega_i)$ regular on the polygonal domain [GR8] and satisfy the following *a priori* bound, see [NE, GI, GR8, EV]:

$$|\mathcal{H}g_h|_{1+\epsilon,\Omega_i} \leq C\,|g_h|_{1/2+\epsilon,\partial\Omega_i}.$$

Applying standard error bounds [CI2, JO2] for the interpolation map yields:

$$\begin{cases} |I_h\mathcal{H}g_h - \mathcal{H}g_h|_{1,\Omega_i} \leq C\,h^{\epsilon}\,|\mathcal{H}g_h|_{1+\epsilon,\Omega_i} \\ \qquad\qquad\qquad\qquad\quad\;\; \leq C\,h^{\epsilon}\,|g_h|_{1/2+\epsilon,\partial\Omega_i}. \end{cases}$$

Substituting an inverse inequality [CI2, JO2] of the form:

$$|g_h|_{1/2+\epsilon,\partial\Omega_i} \leq C\,h^{-\epsilon}\,|g_h|_{1/2,\partial\Omega_i},$$

into the preceding yields:

$$|I_h\mathcal{H}g_h - \mathcal{H}g_h|_{1,\Omega_i} \leq C\,h^{\epsilon}\,h^{-\epsilon}\,|g_h|_{1/2,\partial\Omega_i}.$$

Applying the triangle inequality and employing the preceding bounds yields:

$$|I_h\mathcal{H}g_h|_{1,\Omega_i} \leq |I_h\mathcal{H}g_h - \mathcal{H}g_h|_{1,\Omega_i} + |\mathcal{H}g_h|_{1,\Omega_i} \leq C\,|g_h|_{1/2,\partial\Omega_i}$$

where C is independent of h_0 and h. By construction we obtain $I_h\mathcal{H}g_h = \alpha$ when $g_h(x) = \alpha \in \mathbb{R}$, so that using a quotient space and applying Poincaré-Freidrich's inequality yields:

$$\|I_h\mathcal{H}g_h\|_{1,\Omega_i} \leq C\,\|g_h\|_{1/2,\partial\Omega_i}.$$

Since ρ_i is not involved in this construction, C will be independent of ρ_i. □

We shall next state and prove a basic *norm equivalence* between the energy associated with the Schur complement matrix on a subdomain and a weighted fractional Sobolev norm energy on the boundary of the subdomain.

Lemma 3.78. *Suppose the following assumptions hold.*

1. *Let $a(\cdot)$ in (3.1) satisfy $a(x) = \rho_i$ on Ω_i for $1 \le i \le p$ and $c(\cdot) \equiv 0$ on Ω.*
2. *Let $g_i \in V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i)$ satisfy:*

$$g_i = 0 \quad \text{on } \mathcal{B}_D \cap \partial\Omega_i.$$

3. *Let $u_i \in V_h(\Omega_i) \cap H^1(\Omega_i)$ satisfy:*

$$\begin{cases} \mathcal{A}_i(u_i, v) = 0, \ \forall v \in V_h(\Omega_i) \cap H_0^1(\Omega_i) \\ \qquad u_i = g_i, \ \text{ on } \partial\Omega_i, \end{cases}$$

where

$$\mathcal{A}_i(u, v) \equiv \rho_i \int_{\Omega_i} \nabla u \cdot \nabla v dx.$$

Then the following norm equivalence will hold:

$$c\,\rho_i\,|g_i|^2_{1/2,\partial\Omega_i} \le \mathcal{A}_i(u_i, u_i) \le C\,\rho_i\,|g_i|^2_{1/2,\partial\Omega_i},$$

for $0 < c < C$ independent of h, h_0, ρ_i and g_i.

Proof. We will describe the proof for the case $\partial\Omega_i \cap \mathcal{B}_D = \emptyset$. The proof when $\partial\Omega_i \cap \mathcal{B}_D \neq \emptyset$, will be analogous provided the boundary norm $\|g_i\|^2_{1/2,\partial\Omega_i}$ is replaced by $\|g_i\|^2_{H_{00}^{1/2}(B^{(i)})}$ and provided the appropriate version of the Poincaré-Freidrich's inequality is employed. We will employ the notation $u_i = \mathcal{H}_i^h g_i \in V_h(\Omega_i) \cap H^1(\Omega_i)$ to denote a discrete harmonic function with boundary values $g_i \in V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i)$. Since $c(x) = 0$, it will follow that if γ_i is a constant then $\mathcal{H}_i^h \gamma_i = \gamma_i$, and that $\mathcal{H}_i^h(g_i - \gamma_i) = u_i - \gamma_i$.

To prove the lower bound, given data $g_i \in V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i)$ let α_i denote the mean value of $u_i = \mathcal{H}_i^h g_i$ on Ω_i. Apply the invariance of seminorms under shifts by constants and the trace theorem to obtain:

$$\begin{aligned} |g_i|^2_{1/2,\partial\Omega_i} &= |g_i - \alpha_i|^2_{1/2,\partial\Omega_i} \\ &\le C_1\,\|u_i - \alpha_i\|^2_{1,\Omega_i} \\ &\le C_2\,|u_i - \alpha_i|^2_{1,\Omega_i} \\ &= C_2\,|u_i|^2_{1,\Omega_i} \\ &= \left(\tfrac{C_2}{\rho_i}\right) \mathcal{A}_i\left(u_i, u_i\right), \end{aligned}$$

where the third line above follows by Poincaré-Freidrich's inequality since α_i corresponds to the mean value of u_i on Ω_i. Here C_1, $C_2 > 0$ denote generic constants independent of h, h_0 and ρ_i.

To prove the upper bound, represent the extension $u_i = \mathcal{H}_i^h g_i$ in the form:

$$\mathcal{H}_i^h g_i = \mathcal{E}_i g_i + w_i,$$

where $\mathcal{E}_i^h g_i \in V_h(\Omega_i) \cap H^1(\Omega_i)$ is an extension of g_i satisfying the following:

$$\begin{aligned} \mathcal{E}_i^h g_i &= g_i, \quad \text{on } \partial\Omega_i \\ \|\mathcal{E}_i^h g_i\|_{1,\Omega_i} &\leq C\, \|g_i\|_{1/2,\partial\Omega_i} \end{aligned} \tag{3.115}$$

as given by the discrete extension theorem (Lemma 3.77), and w_i is defined by $w_i \equiv \left(\mathcal{H}_i^h g_i - \mathcal{E}_i^h g_i\right) \in V_h(\Omega_i) \cap H_0^1(\Omega_i)$. We substitute the above representation into the equation satisfied by $\mathcal{H}_i^h g_i$:

$$\mathcal{A}_i(\mathcal{E}_i g_i + w_i, v) = 0 \quad \forall v \in V_h(\Omega_i) \cap H_0^1(\Omega_i),$$

and choose $v = w_i \in V_h(\Omega_i) \cap H_0^1(\Omega_i)$ to obtain:

$$\rho_i |w_i|_{1,\Omega_i}^2 = \mathcal{A}_i(w_i, w_i) = -\mathcal{A}_i(\mathcal{E}_i^h g_i, w_i) \leq \rho_i |\mathcal{E}_i^h g_i|_{1,\Omega_i} |w_i|_{1,\Omega_i}.$$

It thus follows that:

$$\begin{aligned} |w_i|_{1,\Omega_i} &\leq |\mathcal{E}_i^h g_i|_{1,\Omega_i} \\ &\leq C_1\, \|g_i\|_{1/2,\partial\Omega_i}. \end{aligned}$$

Applying the triangle inequality to $u_i = \mathcal{E}_i^h g_i + w_i$, and using the preceding bound and equation (3.115) yields:

$$|u_i|_{1,\Omega_i} \leq C_2 \|g_i\|_{1/2,\partial\Omega_i}.$$

The same bound will hold if g_i is replaced by $g_i - \alpha_i$ for any constant α_i:

$$|u_i - \gamma_i|_{1,\Omega_i} \leq C_2\, \|g_i - \gamma_i\|_{1/2,\partial\Omega_i},$$

where $\mathcal{H}_i^h(g_i - \gamma_i) = u_i - \gamma_i$. If we choose γ_i as the mean value of g_i on $\partial\Omega_i$, then $g_i - \gamma_i$ will have zero mean value on $\partial\Omega_i$, and an application of the Poincaré-Freidrich's inequality will yield:

$$\begin{aligned} |u_i - \gamma_i|_{1,\Omega_i} &\leq C_2\, \|g_i - \gamma_i\|_{1/2,\partial\Omega_i} \\ &\leq C_3\, |g_i - \gamma_i|_{1/2,\partial\Omega_i} \\ &= C_3\, |g_i|_{1/2,\partial\Omega_i}. \end{aligned}$$

It will thus hold that:

$$\begin{aligned} \mathcal{A}_i(u_i, u_i) &= \rho_i\, |u_i|_{1,\Omega_i}^2 \\ &= \rho_i\, |u_i - \gamma_i|_{1,\Omega_i}^2 \\ &\leq C_3\, \rho_i\, |g_i|_{1/2,\partial\Omega_i}^2, \end{aligned}$$

which is the desired upper bound. □

Remark 3.79. The parameters C_i in the preceding estimates will be independent of h and ρ_i, by construction. In addition, they will be independent of h_0 due to the scale invariance of the seminorms. In general, C_i may depend on other geometrical properties of Ω_i, such as its aspect ratio.

Applying the preceding equivalence on each subdomain and summing over all the subdomains yields a global equivalence between the Schur complement energy and a weighted sum of the subdomain fractional Sobolev energies.

Theorem 3.80. *Suppose the following assumptions hold.*

1. *Let $\Omega_1, \ldots, \Omega_p$ form a quasiuniform triangulation of Ω of width h_0.*
2. *Given a vector $\mathbf{u}_B$ of nodal values on interface B, define $\mathbf{u} = \left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T$ where $\mathbf{u}_I = -A_{II}^{-1} A_{IB} \mathbf{u}_B$. Let u_h denote the discrete harmonic finite element function corresponding the nodal vector $\mathbf{u}$.*
3. *Let the coefficient $a(x) = \rho_i$ on Ω_i and $c(x) = 0$ on Ω with:*

$$\mathcal{A}_i(u, v) \equiv \rho_i \int_{\Omega_i} \nabla u \cdot \nabla v dx.$$

Then, the following estimate will hold:

$$c \sum_{i=1}^{p} \rho_i |u_h|^2_{1/2, \partial\Omega_i} \leq \mathbf{u}_B^T S \mathbf{u}_B \; = \mathcal{A}(u_h, u_h) \; \leq C \sum_{i=1}^{p} \rho_i |u_h|^2_{1/2, \partial\Omega_i},$$

for $0 < c < C$ independent of h, h_0 and ρ_i.

Proof. Since u_h is piecewise discrete harmonic by assumption, it will satisfy:

$$\mathbf{u}_B^T S \mathbf{u}_B = \mathbf{u}^T A \mathbf{u} = \mathcal{A}(u_h, u_h) \equiv \sum_{i=1}^{p} \mathcal{A}_i(u_h, u_h).$$

The result now follows by an application of the preceding lemma on each subdomain, and summing over all subdomains using that $g_i = u_h$ on $\partial\Omega_i$. □

Remark 3.81. In view of the preceding result, a preconditioner M for S must ideally be chosen so that its interface energy $\mathbf{u}_B^T M \mathbf{u}_B$ approximates the above weighted sum of fractional Sobolev energies on its subdomain boundaries.

3.9.2 Discrete Sobolev Inequalities

We next describe a discrete Sobolev inequality [BR12] which holds for finite element functions on $\Omega \subset \mathbb{R}^2$, see also [DR2, DR10, MA14, MA17],

Lemma 3.82. *Let $V_h(\Omega_i)$ denote a finite element space defined on a domain $\Omega_i \subset \mathbb{R}^2$ of diameter h_0 triangulated by a quasiuniform grid of size h. Then the following bound will hold for the maximum norm on $V_h(\Omega_i) \subset H^1(\Omega_i)$:*

$$\|u_h\|^2_{\infty, \Omega_i} \leq C \left(1 + \log(h_0/h)\right) \left(h_0^{-2} \|u_h\|^2_{0, \Omega_i} + |u_h|^2_{1, \Omega_i}\right), \quad \forall u_h \in V_h(\Omega_i),$$

where $C > 0$ is independent of h_0 and h.

Proof. We follow the proof in [BR12]. Let $x_* \in \overline{\Omega}_i$ denote a point where the finite element function u_h attains it maximum modulus $|u_h(x_*)| = \|u_h\|_{\infty,\Omega_i}$. Let $\mathcal{C} \subset \overline{\Omega}_i$ denote a cone of radius R and angle α at vertex x_*. Introduce polar coordinates (r, θ) within the cone so that $(0, 0)$ corresponds to x_* and so that the cone is specified in polar coordinates by $0 \leq r \leq R$ and $0 \leq \theta \leq \alpha$. Apply the fundamental theorem of calculus along a ray within the cone:

$$u_h(0,0) = u_h(R,\theta) + \int_0^R \frac{\partial u_h}{\partial r}(r,\theta)\, dr.$$

Split the integral using the intervals $0 \leq r \leq \epsilon\, h$ and $\epsilon\, h \leq r \leq R$ for some $0 < \epsilon \ll 1$, take absolute values of all terms, and employ the inverse inequality $\|\frac{du_h}{dr}\|_{\infty,\Omega_i} \leq \|u_h\|_{\infty,\Omega_i} h^{-1}$ within the interval $0 \leq r \leq \epsilon\, h$ (which holds trivially for piecewise linear finite elements) to obtain:

$$\begin{aligned} |u_h(0,0)| &\leq |u_h(R,\theta)| + \textstyle\int_0^{\epsilon\, h} |\frac{\partial u_h(r,\theta)}{\partial r} dr| + \int_{\epsilon\, h}^R |\frac{\partial u_h(r,\theta)}{\partial r} dr| \\ &\leq |u_h(R,\theta)| + \epsilon\, h\, \|u_h\|_{\infty,\Omega_i} h^{-1} + \textstyle\int_{\epsilon\, h}^R |\frac{\partial u_h(r,\theta)}{\partial r} dr| \\ &= |u_h(R,\theta)| + \epsilon\, \|u_h\|_{\infty,\Omega_i} + \textstyle\int_{\epsilon\, h}^R |\frac{\partial u_h(r,\theta)}{\partial r} dr| \end{aligned}$$

Since $|u_h(0,0)| = \|u_h\|_{\infty,\Omega_i}$, bringing back the term $\epsilon\, \|u_h\|_{\infty,\Omega_i}$ yields:

$$(1-\epsilon)\, \|u_h\|_{\infty,\Omega_i} \leq |u_h(R,\theta)| + \textstyle\int_{\epsilon\, h}^R |\frac{\partial u_h(r,\theta)}{\partial r} dr|.$$

Integrating the above expression as θ ranges in $(0, \alpha)$ yields:

$$\begin{aligned} (1-\epsilon)\, \alpha \|u_h\|_{\infty,\Omega_i} &\leq \textstyle\int_0^\alpha |u_h(R,\theta)|\, d\theta + \int_0^\alpha \int_{\epsilon\, h}^R |\frac{\partial u_h(r,\theta)}{\partial r}| dr\, d\theta \\ &= \textstyle\int_0^\alpha |u_h(R,\theta)|\, d\theta + \int_0^\alpha \int_{\epsilon\, h}^R \frac{1}{r} |\frac{\partial u_h(r,\theta)}{\partial r}| r\, dr\, d\theta. \end{aligned}$$

Squaring both sides, applying the triangle inequality and the Cauchy-Schwartz inequality to the terms on the right side yields:

$$\begin{aligned} \|u_h\|^2_{\infty,\Omega_i} \leq\ & \tfrac{2}{\alpha^2(1-\epsilon)^2} \left((\textstyle\int_0^\alpha |u_h(R,\theta)|^2\, d\theta)(\int_0^\alpha d\theta) \right) \\ &+ \tfrac{2}{\alpha^2(1-\epsilon)^2} \Big((\textstyle\int_0^\alpha \int_{\epsilon\, h}^R |\frac{\partial u_h(r,\theta)}{\partial r}|^2 r\, dr\, d\theta)(\int_0^\alpha \int_{\epsilon\, h}^R \frac{1}{r^2} r\, dr\, d\theta) \Big). \end{aligned}$$

Simplifying the expression yields the bound:

$$\|u_h\|^2_{\infty,\Omega_i} \leq \tfrac{2\alpha}{\alpha^2(1-\epsilon)^2} \Big(\textstyle\int_0^\alpha |u_h(R,\theta)|^2 d\theta + \log(R/\epsilon h) \int_0^\alpha \int_{\epsilon h}^R |\frac{\partial u_h(r,\theta)}{\partial r}|^2 r\, dr\, d\theta \Big).$$

Since $(\frac{\partial u_h}{\partial r})^2 \leq (\frac{\partial u_h}{\partial r})^2 + \frac{1}{r^2}(\frac{\partial u_h}{\partial \theta})^2 = |\nabla u_h|^2$, we obtain:

$$\begin{aligned} \|u_h\|^2_{\infty,\Omega_i} &\leq \tfrac{2}{\alpha(1-\epsilon)^2} \left(\textstyle\int_0^\alpha |u_h(R,\theta)|^2\, d\theta + (\log(1/\epsilon) + \log(R/h)\, |u_h|^2_{1,\mathcal{C}} \right) \\ &\leq \tfrac{2}{\alpha(1-\epsilon)^2} \left(\textstyle\int_0^\alpha |u_h(R,\theta)|^2\, d\theta + C(1 + \log(h_0/h)\, |u_h|^2_{1,\Omega_i} \right). \end{aligned}$$

Multiplying both sides by $R\,dR$ and integrating over $0 \le R \le \beta\,h_0$ (assuming that the cone $\mathcal{C}$ can be extended within Ω_i to have diameter $\beta\,h_0$, for some $0 \ll \beta \le 1$) yields the estimate:

$$\begin{aligned}
&\tfrac{\beta^2 h_0^2}{2} \|u_h\|_{\infty,\Omega_i}^2 \\
&\le \tfrac{2}{\alpha(1-\epsilon)^2} \left(\int_0^{\beta h_0} \int_0^{\alpha} |u_h(R,\theta)|^2 R\,dR\,d\theta + \tfrac{C\beta^2 h_0^2}{2} (1 + \log(h_0/h))\, |u_h|_{1,\Omega_i}^2 \right) \\
&\le \tfrac{2}{\alpha(1-\epsilon)^2} \left(\|u_h\|_{0,\Omega_i}^2 + \tfrac{C\beta^2 h_0^2}{2} (1 + \log(h_0/h))\, |u_h|_{1,\Omega_i}^2 \right).
\end{aligned}$$

Dividing both sides by the factor $\beta^2 h_0^2/2$ yields the desired result. □

As a corollary of the preceding result, we obtain a discrete Sobolev inequality holding on the *boundary* $\partial\Omega_i$ of a two dimensional domain.

Lemma 3.83. *Let $\Omega_i \subset \mathbb{R}^2$ be of diameter h_0 and triangulated by a quasiuniform grid of size h. Then, the following bound will hold:*

$$\|v_h\|_{\infty,\partial\Omega_i}^2 \le C\,(1 + \log(h_0/h)) \left(|v_h|_{1/2,\partial\Omega_i}^2 + h_0 \|v_h\|_{0,\partial\Omega_i}^2 \right) \qquad (3.116)$$

for $v_h \in V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i)$, where $C > 0$ is independent of h_0 and h.

Proof. Given $v_h \in V_h(\partial\Omega_i) \cap H^{1/2}(\partial\Omega_i)$ let $\mathcal{H}_i^h v_h \in V_h(\Omega_i) \cap H^1(\Omega_i)$ denote the discrete harmonic extension of v_h into Ω_i, satisfying:

$$\|\mathcal{H}_i^h v_h\|_{1,\Omega_i}^2 \le C \|v_h\|_{1/2,\partial\Omega_i}^2,$$

for $C > 0$ independent of h_0 and h. Applying the preceding lemma to $\mathcal{H}_i^h v_h$ and using the boundedness of $\mathcal{H}_i^h$ yields:

$$\begin{cases}
\|v_h\|_{\infty,\partial\Omega_i}^2 \le \|\mathcal{H}_i^h v_h\|_{\infty,\Omega_i}^2 \\
\qquad\qquad\quad \le C\,(1 + \log(h_0/h))\, \|\mathcal{H}_i^h v_h\|_{1,\Omega_i}^2 \\
\qquad\qquad\quad \le C\,(1 + \log(h_0/h))\, \|v_h\|_{1/2,\partial\Omega_i}^2,
\end{cases}$$

for $C > 0$ independent of h_0 and h. □

We shall now present an alternate proof of the preceding discrete Sobolev inequality based on Fourier series [BR29]. This proof will use the property that the boundary $\partial\Omega_i$ of a simply connected polygonal domain $\Omega_i \subset \mathbb{R}^2$ will be Lipschitz homeomorphic to the unit circle S^1 (i.e., there will be a one to one correspondence between $\partial\Omega_i$ and the unit circle S^1, under a Lipschitz continuous parameterization). Given such a parameterization $x(\theta)$ of the boundary $\partial\Omega_i$ by a 2π periodic function $x(\theta)$ with arclength measure $ds(x(\theta)) = |x'(\theta)|\,d\theta$ defined along the curve, we may represent any function $u(\cdot) \in L^2(\partial\Omega_i)$ by a Fourier series expansion of the form:

$$u(x(\theta)) = \sum_{k=-\infty}^{\infty} c_k e^{ik\theta}.$$

When Ω_i is shape regular, the following equivalences will hold [BR29]:

$$\begin{cases} \|u\|^2_{L^2(\partial\Omega_i)} &= \frac{h_0}{2\pi}\sum_{k=-\infty}^{\infty} 2\pi |c_k|^2, \\ |u|^2_{H^\beta(\partial\Omega_i)} &= \frac{h_0^{1-2\beta}}{2\pi}\sum_{k=-\infty}^{\infty} 2\pi\,|k|^{2\beta}\,|c_k|^2, \text{ for } 0<\beta<1, \end{cases}$$

where $h_0 = |\partial\Omega_i|$ denotes the length of $\partial\Omega_i$. The alternate proof of the *discrete* Sobolev inequality will be obtained based on the following *continuous* Sobolev inequality for 2π periodic functions.

Lemma 3.84. *Let $v(x) \in H^{\frac{1+\epsilon}{2}}(0,2\pi)$ denote a real periodic function on $[0,2\pi]$ with Fourier expansion:*

$$v(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}.$$

Then, the following bound will hold:

$$\|v\|^2_{L^\infty(0,2\pi)} \le C\left(\|v\|^2_{L^2(0,2\pi)} + \epsilon^{-1}\|v\|^2_{H^{\frac{1+\epsilon}{2}}(0,2\pi)}\right), \tag{3.117}$$

for $0<\epsilon<1$ and C independent of ϵ.

Proof. To prove the bound take absolute values of the Fourier expansion and apply the Cauchy-Schwartz inequality to obtain:

$$\begin{aligned}\|v\|^2_{L^\infty(0,2\pi)} &\le \left(|c_0| + \sum_{k=-\infty,\,k\neq 0}^{\infty}|c_k|\right)^2 \\ &= \left(|c_0| + \sum_{k=-\infty,\,k\neq 0}^{\infty}(|k|^{\frac{1+\epsilon}{2}}|c_k|)\,|k|^{-\frac{1+\epsilon}{2}}\right) \\ &\le 2\pi|c_0|^2 + 2\pi\left(\sum_{k=-\infty,\,k\neq 0}^{\infty}|k|^{1+\epsilon}|c_k|^2\right)\left(\sum_{k=-\infty,\,k\neq 0}^{\infty}|k|^{-1-\epsilon}\right) \\ &\le \|v\|^2_{L^2(0,2\pi)} + |v|^2_{H^{\frac{1+\epsilon}{2}}(0,2\pi)}\left(\sum_{k=-\infty,\,k\neq 0}^{\infty}|k|^{-1-\epsilon}\right).\end{aligned}$$

Using the integral test, we may bound:

$$\sum_{k=-\infty,\,k\neq 0}^{\infty}|k|^{-1-\epsilon} \le 2\left(1+\int_1^\infty \frac{dx}{x^{1+\epsilon}}\right) = 2\left(1+\frac{1}{\epsilon}\right) \le 4\epsilon^{-1},$$

for $0<\epsilon<1$. Substituting this into the preceding bound yields:

$$\|v\|^2_{L^\infty(0,2\pi)} \le \|v\|^2_{L^2(0,2\pi)} + 4\,\epsilon^{-1}\,|v|^2_{H^{\frac{1+\epsilon}{2}}(0,2\pi)}$$

which is the desired result. □

The discrete Sobolev inequality (3.116) can now be obtained from (3.117) by choosing ϵ appropriately and using an inverse inequality for finite elements.

Lemma 3.85. *Let $v_h \in V_h(\partial\Omega_i) \cap H^{\frac{1+\epsilon}{2}}(\partial\Omega_i)$ be 2π-periodic. Then, the following bound will hold:*

$$\|v_h\|^2_{L^\infty(\partial\Omega_i)} \le C\ (1+\log(h_0/h))\, \|v_h\|^2_{H^{1/2}(\partial\Omega_i)}$$

for $C>0$ independent of h and h_0.

Proof. We follow the proof in [BR29]. Apply the preceding continuous Sobolev inequality to the 2π-periodic representation of v_h, and employ norm equivalences to obtain the bound:

$$\|v_h\|^2_{\infty,\partial\Omega_i} \le C\, h_0^{-1} \|v_h\|^2_{0,\partial\Omega_i} + C\,\epsilon^{-1}\, h_0^{\epsilon} |v_h|^2_{H^{\frac{1+\epsilon}{2}}(\partial\Omega_i)}.$$

Substitute the following inverse inequality:

$$|v_h|^2_{H^{\frac{1+\epsilon}{2}}(\partial\Omega_i)} \le C\, h^{-\epsilon} |v_h|^2_{H^{\frac{1}{2}}(\partial\Omega_i)}, \qquad \forall v_h \in V_h(\partial\Omega_i)$$

in the preceding bound, with $C>0$ independent of h, to obtain:

$$\|v_h\|^2_{\infty} \le C\, h_0^{-1} \|v_h\|^2_{0,\partial\Omega_i} + C\,\epsilon^{-1}\, h_0^{\epsilon}\, h^{-\epsilon}\, |v_h|^2_{H^{\frac{1}{2}}(\partial\Omega_i)}.$$

Importantly, the parameter $\epsilon>0$ may be chosen small enough so that

$$\epsilon^{-1}\,(h_0/h)^{\epsilon} \le (1+\log(1/h)).$$

This will hold provided:

$$\epsilon \equiv \begin{cases} \frac{1}{4}, & \text{if } (h/h_0) \ge e^{-4} \\ \frac{-1}{log(h_0/h)}, & \text{if } (h/h_0) < e^{-4}, \end{cases}$$

and can be verified by an application of the derivative test for a maximum in the parameter ϵ. The desired result follows immediately. □

We now apply the discrete Sobolev inequalities to derive results useful for estimating the condition number of Schur complement preconditioners.

Lemma 3.86. *Suppose the following assumptions hold.*

1. *Let $D_i \subset \partial\Omega_i$ denote a connected subset of length $d_i \le h_0$.*
2. *Let $w_h \in V_h(\partial\Omega_i)$ satisfy:*

$$w_h(x) = 0 \qquad \textit{for } x \in \partial\Omega_i \backslash D_i.$$

Then, the following results will hold:

$$|w_h|^2_{1/2,\partial\Omega_i} = |w_h|^2_{H^{1/2}_{00}(D_i)} \le C\,(1+\log(h_0/h))\ \|w_h\|^2_{\infty,D_i} + |w_h|^2_{1/2,D_i} \tag{3.118}$$

for $C>0$ independent of h_0 and h.

Proof. Since $w_h(x)$ is zero outside D_i, we may employ the equivalent integral expression for the fractional Sobolev seminorm:

$$|w_h|^2_{1/2,\partial\Omega_i} \leq |w_h|^2_{1/2,D_i} + 2\int_{D_i} \frac{|w_h(x)|^2}{\operatorname{dist}(x,\partial\Omega_i\backslash D_i)}\, ds(x), \tag{3.119}$$

where $ds(x)$ denotes the arclength measure along $\partial\Omega_i$ for $s \in (0, d_i)$ and $\operatorname{dist}(x,\partial\Omega_i\backslash D_i)$ denotes the arclength distance between x and $\partial\Omega_i\backslash D_i$. Since the arclength distance satisfies:

$$\operatorname{dist}(x,\partial\Omega_i\backslash D_i) = \min\{s, d_i - s\},$$

the above integral can be split as:

$$\int_{D_i} \frac{|w_h(x)|^2}{\operatorname{dist}(x,\partial\Omega_i\backslash D_i)}\, ds(x) = \int_0^{d_i/2} \frac{|w_h(x(s))|^2}{s} ds + \int_{d_i/2}^{d_i} \frac{|w_h(x(s))|^2}{d_i - s} ds. \tag{3.120}$$

Since $w_h(x)$ is zero when $s = 0$ and linear for $0 \leq s \leq h$, The first integral may further be split over the intervals $[0, h]$ and $[h, d_i/2]$. For $0 \leq s \leq h$, we may bound $|w_h(x(s))| \leq \|w_h\|_{\infty,D_i}(s/h)$ since $w_h(x(s))$ is linear on the interval and $w_h(x(0)) = 0$. For $h \leq s \leq d_i/2$, we may bound $|w_h(x(s))| \leq \|w_h\|_{\infty,D_i}$. Substituting these yields:

$$\begin{aligned}
\int_0^{d_i/2} \tfrac{|w_h(x(s))|^2}{s} ds &= \int_0^h \tfrac{|w_h(x(s))|^2}{s} ds + \int_h^{d_i/2} \tfrac{|w_h(x(s))|^2}{s} ds \\
&\leq \|w_h\|^2_{\infty,D_i} \int_0^h \tfrac{s^2}{h^2\, s} ds + \|w_h\|^2_{\infty,D_i} \int_h^{d_i/2} \tfrac{1}{s} ds \\
&= \|w_h\|^2_{\infty,D_i} \tfrac{h^2}{2h^2} + \|w_h\|^2_{\infty,D_i} \log(d_i/2h) \\
&\leq C\, \|w_h\|^2_{\infty,D_i} \left(1 + \log(h_0/h)\right).
\end{aligned}$$

We may similarly bound:

$$\int_{d_i/2}^{d_i} \frac{|w_h(x(s))|^2}{d_i - s} ds \leq C\, \|w_h\|^2_{\infty,D_i} \left(1 + \log(h_0/h)\right).$$

Combining bounds and substituting them into (3.120) yields the result. □

We now describe estimates of finite element decompositions based on globs. Recall that a *glob* is either an edge or a vertex of B when $\Omega \subset \mathbb{R}^2$, and a face, an edge or a vertex of B when $\Omega \subset \mathbb{R}^3$. Let $\mathcal{G}$ denote all globs of B.

Definition 3.87. *If $G \subset B$ is a glob, define the map $I_G : V_h(B) \to V_h(B)$ which assigns zero nodal values at all nodes in B outside G:*

$$I_G v_h(x) \equiv \begin{cases} v_h(x), & \text{for nodes } x \in G \\ 0, & \text{for nodes } x \notin G \end{cases} \quad \text{for} \quad v_h \in V_h(B).$$

In particular, the following decomposition of the identity will hold:

$$I = \sum_{G\in\mathcal{G}} I_G.$$

We associate the following parameters with a subdomain decomposition.

Definition 3.88. *Let $L > 0$ denote the maximum number of globs on any shared interface of the form $\partial\Omega_i \cap \Omega_j$:*

$$L \equiv \max_{i,j} |\{G : G \subset \partial\Omega_i \cap \partial\Omega_j\}| . \tag{3.121}$$

Let $K > 0$ denote the maximum number of neighboring subdomains:

$$K \equiv \max_{i} |\{j : \partial\Omega_i \cap \partial\Omega_j \neq \emptyset\}| . \tag{3.122}$$

For typical subdomain decompositions arising from a coarse triangulation, the parameters K and L will be bounded independent of h, h_0 and ρ_i.

We shall now outline an important theoretical result referred to as a *glob theorem*. The glob theorem provides a bound for the $H^{1/2}(\partial\Omega_i)$ seminorm of the finite element interpolation map I_G. It will be useful for estimating partition parameters in abstract Schwarz algorithms. The following preliminary result establishes a bound for I_G when G is a *vertex glob* in two dimensions.

Lemma 3.89. *Let $\Omega \subset \mathbb{R}^2$ and suppose the following assumptions hold.*

1. *Let $G \in \partial\Omega_i$ denote a vertex glob, and let $\psi_G^h(x) \in V_h(B)$ denote a finite element nodal basis function centered at vertex G on B:*

$$\psi_G^h(x_j) = \begin{cases} 1, & \text{if } x_j = G \\ 0, & \text{if } x_j \neq G, \end{cases}$$

where each x_j denotes a node in B.

2. *Given $w_h \in V_h(B)$ let $I_G w_h \in V_h(B)$ denote the finite element function:*

$$I_G w_h(x) \equiv w_h(G)\, \psi_G^h(x).$$

Then, the following bound will hold:

$$|I_G w_h|^2_{1/2,\partial\Omega_i} \leq C\, (1 + \log(h_0/h))^2\, \|w_h\|^2_{1/2,\partial\Omega_i} \qquad \forall w_h \in V_h(B)$$

for some $C > 0$ independent of h_0 and h.

Proof. See [MA17]. Since $I_G w_h(x) = w_h(G)\psi_G^h(x)$, by linearity we obtain:

$$\begin{cases} |I_G w_h|_{1/2,\partial\Omega_i} = |w_h(G)|\, |\psi_G^h|_{1/2,\partial\Omega_i} \\ \qquad\qquad\qquad\quad \leq \|w_h\|_{\infty,\partial\Omega_i}\, |\psi_G^h|_{1/2,\partial\Omega_i}. \end{cases}$$

Let $B_G \subset \partial\Omega_i$ denote the union of elements adjacent to G on which $\psi_G^h(x)$ has support. Apply Lemma 3.86 to $\psi_G^h(x)$ which has support on B_G to obtain:

$$|\psi_G^h|^2_{1/2,\partial\Omega_i} \leq C\, (1 + \log(h_0/h))\, \|\psi_G^h\|^2_{\infty,B_G} + |\psi_G^h|^2_{1/2,B_G}, \tag{3.123}$$

for $C > 0$ independent of h_0 and h. We estimate $|\psi_G^h|^2_{1/2,B_G}$ by substituting that $|\psi_G^h(x) - \psi_G^G(y)| \le \frac{|x-y|}{h}$ for $x, y \in B_G$ to obtain:

$$|\psi_G^h|^2_{1/2,B_G} = \int_{-h}^{h}\int_{-h}^{h} \frac{|\psi_G^h(x) - \psi_G^h(y)|^2}{|x-y|^2} ds(x)\, ds(y) \le 4.$$

Substituting the preceding bound and using $\|\psi_G^h\|_{\infty,B_G} = 1$ in (3.123) yields:

$$\begin{aligned}
|I_G w_h|^2_{1/2,\partial\Omega_i} &\le \|w_h\|^2_{\infty,\partial\Omega_i} \left(C\,(1+\log(h_0/h))\; \|\psi_G^h\|^2_{\infty,B_G} + |\psi_G^h|^2_{1/2,B_G} \right) \\
&\le \|w_h\|^2_{\infty,\partial\Omega_i} \; (C\,(1+\log(h_0/h)) + 4) \\
&\le C\,(1+\log(h_0/h))\; \|w_h\|^2_{\infty,\partial\Omega_i} \\
&\le C\,(1+\log(h_0/h))\; (1+\log(h_0/h))\, \|w_h\|^2_{1/2,\partial\Omega_i},
\end{aligned}$$

where we employed the discrete Sobolev inequality in the last step. □

Remark 3.90. A bound of the form $\|I_G v_h\|_{0,\partial\Omega_i} \le C\, \|v_h\|_{0,\partial\Omega_i}$ will hold trivially since the mass matrix on B is spectrally equivalent to an identity matrix. Combining such a bound with Lemma 3.89 yields an estimate of the form:

$$\|I_G w_h\|^2_{1/2,\partial\Omega_i} \le C\; (1+\log(h_0/h))^2\, \|w_h\|^2_{1/2,\partial\Omega_i},\; \forall w_h \in V_h(B)$$

for $C > 0$ independent of h_0 and h.

The next result bounds the $H^{1/2}(\partial\Omega_i)$ seminorm of I_G when G corresponds to an *edge glob* in $\partial\Omega_i \subset B$ on a two dimensional domain Ω_i.

Lemma 3.91. *Let $\Omega \subset \mathbb{R}^2$ and suppose the following assumptions hold.*

1. *Let $G \in \partial\Omega_i$ denote a edge glob, and given $v_h \in V_h(B)$ let $I_G v_h \in V_h(B)$ denote the finite element function defined by:*

$$I_G v_h(x_j) \equiv \begin{cases} v_h(x_j), & \text{if } x_j \in G \\ 0, & \text{if } x_j \in \partial\Omega_i \backslash G, \end{cases}$$

where x_j denotes nodes on $\partial\Omega_i$.

2. *Let B_G denote the union of all elements of $\partial\Omega_i$ intersecting the glob G.*

Then, the following bound will hold:

$$|I_G v_h|^2_{1/2,\partial\Omega_i} \le C\,(1+\log(h_0/h))^2\, \|v_h\|^2_{1/2,\partial\Omega_i},$$

for some $C > 0$ independent of h_0 and h.

Proof. See [MA17]. Given an *edge* glob $G \subset \partial\Omega_i$, let $G_L, G_R \in \partial\Omega_i$ denote its endpoints, corresponding to vertex globs. By construction, the finite element function $w_h(x)$ will be zero at these endpoints G_L and G_R and outside the glob, and may alternatively be expressed as:

$$w_h(x) \equiv I_G v_h(x) = \begin{cases} v_h(x) - I_{G_L} v_h(x) - I_{G_R} v_h(x) & \text{for } x \in B_G \\ 0 & \text{for } x \in \partial\Omega_i \backslash B_G. \end{cases}$$

Applying bound (3.118) to $w_h(x)$ on B_G yields:

$$|w_h|^2_{1/2,\partial\Omega_i} \leq C\,(1 + \log(h_0/h))\, \|w_h\|^2_{\infty,B_G} + |w_h|^2_{1/2,B_G}$$

for $C > 0$ independent of h_0 and h. Substituting that $\|w_h\|_{\infty,B_G} = \|v_h\|_{\infty,B_G}$ (which holds by construction), and estimating the latter term by the discrete Sobolev inequality yields:

$$\begin{cases} |w_h|^2_{1/2,\partial\Omega_i} \leq C\,(1 + \log(h_0/h))\, \|v_h\|^2_{\infty,B_G} + |w_h|^2_{1/2,B_G} \\ \qquad\qquad\quad \leq C\,(1 + \log(h_0/h))^2\, \|v_h\|^2_{1/2,\partial\Omega_i} + |w_h|^2_{1/2,B_G}. \end{cases}$$

Since $w_h(x) = v_h(x) - I_{G_L} v_h(x) - I_{G_R} v_h(x)$ on B_G, we may apply the generalized triangle inequality to estimate the seminorm $|w_h|^2_{1/2,B_G}$ as follows:

$$\begin{cases} |w_h|^2_{1/2,B_G} \leq 3\left(|v_h|^2_{1/2,B_G} + |I_{G_L} v_h|^2_{1/2,B_G} + |I_{G_R} v_h|^2_{1/2,B_G}\right). \\ \qquad\qquad\quad \leq C\,\left(|v_h|^2_{1/2,B_G} + (1 + \log(h_0/h))^2\, |v_h|^2_{1/2,\partial\Omega_i}\right), \end{cases}$$

where the latter expression was obtained using $|I_{G_L} v_h|_{1/2,B_G} \leq |I_{G_L} v_h|_{1/2,\partial\Omega_i}$ and employing bounds for the vertex glob interpolants. Similarly for the term $|I_{G_R} v_h|_{1/2,B_G}$. Combining the above estimate with the trivial bound $|v_h|^2_{1/2,B_G} \leq |v_h|^2_{1/2,\partial\Omega_i}$, we obtain:

$$|w_h|^2_{1/2,\partial\Omega_i} \leq C\,(1 + \log(h_0/h))^2\, \|v_h\|^2_{1/2,\partial\Omega_i}$$

which is the desired estimate. □

Remark 3.92. As for vertex globs, a bound $\|I_G v_h\|_{0,\partial\Omega_i} \leq C\, \|v_h\|_{0,\partial\Omega_i}$ will also hold trivially for edge globs since the mass matrix on B is spectrally equivalent to an identity matrix. Combining such a bound with the preceding lemma will yield an estimate of the form:

$$\|I_G w_h\|^2_{1/2,\partial\Omega_i} \leq C\, (1 + \log(h_0/h))^2\, \|w_h\|^2_{1/2,\partial\Omega_i},\ \forall w_h \in V_h(B)$$

for $C > 0$ independent of h_0 and h.

Combining the preceding results yields the two dimensional glob theorem.

Lemma 3.93. *Let $\Omega \subset \mathbb{R}^2$ and suppose the following assumptions hold.*

1. *Let $V_h(\Omega)$ be a finite element space on a quasiuniform triangulation.*
2. *Let I_G denote the glob interpolation map for vertex or edge globs $G \subset \partial\Omega_i$.*

Then, the following bound will hold for $v_h \in V_h(B)$:

$$\|I_G v_h\|^2_{1/2,\partial\Omega_i} \leq C\,(1 + \log(h_0/h))^2\, \|v_h\|^2_{1/2,\partial\Omega_i}, \qquad \forall v_h \in V_h(B).$$

Proof. The proof follows by combining the *seminorm* bounds for $|I_G v_h|^2_{1/2,\partial\Omega_i}$ in the preceding lemmas, with estimates for I_G in the $L^2(\partial\Omega_i)$ norm:

$$\|I_G v_h\|^2_{0,\partial\Omega_i} \le C\|v_h\|^2_{0,\partial\Omega_i},$$

which will hold for some $C > 0$ independent of h_0 and h for any glob G because of the spectral equivalence between the mass matrix and a scaled identity matrix on $\partial\Omega_i$. □

We now state the general glob theorem [MA17] in two or three dimensions.

Theorem 3.94 (Glob Theorem). *Suppose the following assumptions hold.*

1. *Let $\mathcal{T}_h(\Omega)$ be a quasiuniform triangulation of $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$.*
2. *Let $\Omega_1, \ldots, \Omega_p$ form a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω of size h_0.*
3. *Let $G \subset \partial\Omega_i$ be a glob within B and let $v_h \in V_h(B)$.*

Then, the following results will hold.

1. *There exists $C > 0$ independent of h, h_0 and ρ_i such that:*
$$\|I_G v_h\|^2_{1/2,\partial\Omega_i} \le C\left(1 + \log(h_0/h)\right)^2 \|v_h\|^2_{1/2,\partial\Omega_i}.$$
2. *There exists $C > 0$ independent of h, h_0 and ρ_i such that:*
$$\|I_G v_h\|^2_{1/2,B} \le C \sum_{j:\partial\Omega_j \cap G \neq \emptyset} \left(1 + \log(h_0/h)\right)^2 \|u_h\|^2_{1/2,\partial\Omega_j}.$$

Proof. See [BR12, BR15, DR17, DR10, MA17]. □

3.9.3 Properties of Coarse Spaces

We shall now summarize theoretical properties of two types of coarse spaces employed in Schur complement algorithms, the *traditional* coarse space defined based on an underlying coarse triangulation of the domain, and the *piecewise constant* coarse space employed in the balancing domain decomposition preconditioner, based on a decomposition of the interface into globs. We shall omit discussion of wirebasket coarse spaces, see [BR15, SM2].

Definition 3.95. *When the subdomains $\Omega_1, \ldots, \Omega_p$ of size h_0 form a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω, the traditional coarse space $V_{0,T}(B) \subset V_h(B)$ corresponds to the restriction to B of the finite element space defined on the coarse triangulation. If $y_1, \ldots, y_{n_0}$ denote the coarse vertices with associated coarse space nodal basis functions $\psi_1^{h_0}(x), \ldots, \psi_{n_0}^{h_0}(x)$ which satisfy $\psi_i^{h_0}(y_j) = \delta_{ij}$ (where δ_{ij} is the Kronecker delta), then the coarse space interpolation map $I_{0,T} : V_h(B) \to V_{0,T}(B) \subset V_h(B)$ is defined by:*

$$I_{0,T} v_h(x) = \sum_{i=1}^{n_0} v_h(y_i)\, \psi_i^{h_0}(x).$$

The traditional interpolation map $I_{0,T}$ is also denoted I_{h_0}, I_0, π_0 or π_{h_0}.

In the following, we summarize known bounds for the coarse grid interpolation map $I_{0,T}$ onto the standard coarse space $V_{0,T}(B) = V_{h_0}(B) \subset V_h(B)$.

Lemma 3.96. *Let $\mathcal{T}_h(\Omega)$ be a quasiuniform triangulation of Ω of size h. Let $\Omega_1, \ldots, \Omega_p$ form a quasiuniform coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω of size h_0. Then, for $1 \le i \le p$ the following results will hold.*

1. *The following bound will hold locally on each $\partial\Omega_i$ for $v_h \in V_h(B)$:*
$$|I_{0,T} v_h|^2_{1/2,\partial\Omega_i} \le \begin{cases} C\ (1 + \log(h_0/h))\, \|v_h\|^2_{1/2,\partial\Omega_i}, & \textit{if } \Omega \subset \mathbb{R}^2 \\ C\ (h_0/h)\, \|v_h\|^2_{1/2,\partial\Omega_i}, & \textit{if } \Omega \subset \mathbb{R}^3 \end{cases} \qquad (3.124)$$
for $C > 0$ independent of h, h_0 and ρ_i.
2. *The interpolation error will satisfy:*
$$|v_h - I_{0,T} v_h|^2_{0,\partial\Omega_i} \le C h_0 |v_h|^2_{1/2,\partial\Omega_i}, \qquad \forall v_h \in V_h(B) \qquad (3.125)$$
for $C > 0$ independent of h, h_0 and ρ_i.

Proof. For the general proof, see [BR15, DR10]. We shall only outline the proof of boundedness of $I_{0,T}$ in two dimensions. Employ the equivalence between $|I_{0,T} v_h|^2_{1/2,\partial\Omega_i}$ and $|\mathcal{H}_h I_{0,T} v_h|^2_{1,\Omega_i}$ when $\mathcal{H}_h$ is the discrete harmonic extension map into the subdomains. Since $\mathcal{H}_h I_{0,T} v_h$ will be linear on each triangular subdomain Ω_i, the term $|\mathcal{H}_h I_{0,T} v_h|^2_{1,\Omega_i}$ will involve the difference quotients:

$$\sum_{l,j} a^{(i)}_{lj}\, (v_h(x_l) - v_h(x_j))^2$$

without a h_0 scaling factor, in two dimensions.

The term $|\mathcal{H}_h I_{0,T} v_h|^2_{1,\Omega_i}$ can thus be estimated by $C\,|v_h|^2_{\infty,\Omega_i}$ which in turn can be estimated by the discrete Sobolev inequality as bounded by $C(1 + \log(h_0/h))\|v_h\|^2_{1/2,\partial\Omega_i}$. This yields the desired bound for $\Omega \subset \mathbb{R}^2$. The interpolation error:

$$|v_h - I_{0,T} v_h|^2_{0,\partial\Omega_i} \le C h_0 |v_h|^2_{1/2,\partial\Omega_i}$$

is standard [ST14, CI2, JO2]. □

Since each of the bounds in Lemma 3.96 are local, we may multiply them by a factor ρ_i on each subdomain, and sum over all subdomains to obtain global estimates involving weighted terms, as indicated below.

Lemma 3.97. *Under the same assumptions as Lemma 3.96, the following bound will hold for $v_h \in V_h(B)$:*

$$\sum_{i=1}^{p} \rho_i |I_{0,T} v_h|^2_{1/2,\partial\Omega_i} \le \begin{cases} C\,(1 + \log(h_0/h)) \sum_{i=1}^{p} \rho_i \|v_h\|^2_{1/2,\partial\Omega_i}, & \textit{if } \Omega \subset \mathbb{R}^2 \\ C\,(h_0/h) \sum_{i=1}^{p} \rho_i \|v_h\|^2_{1/2,\partial\Omega_i}, & \textit{if } \Omega \subset \mathbb{R}^3 \end{cases}$$

for $C > 0$ independent of h, h_0 and ρ_i.

Proof. Multiply the local bounds in Lemma 3.96 by the factor ρ_i and sum over all the subdomains. □

We next consider the *piecewise constant* coarse space $V_{0,P}(B)$ used in the balancing domain decomposition preconditioner, and describe analogous estimates. The space $V_{0,P}(B)$ defined by (3.126) is referred to as *piecewise constant* since the finite element functions within this space have constant values on nodes within each glob of B.

Definition 3.98. *Given a glob $G \in \mathcal{G}$, we let $0 \leq d_j(G) \leq 1$ denote non-negative partition of unity parameters which satisfy:*

$$\begin{cases} d_j(G) = 0, \;\; \textit{if } G \cap \partial\Omega_j = \emptyset \\ \sum_{\{j : G \subset \partial\Omega_j\}} d_j(G) = 1. \end{cases}$$

Remark 3.99. The quantities $d_j(G)$ are typically defined by:

$$d_j(G) = \frac{\rho_j^t}{\sum_{\{l : G \subset \partial\Omega_l\}} \rho_l^t}, \qquad \text{with } t \geq \frac{1}{2}.$$

For $1 \leq j \leq p$ we define a map $Q_j : V_h(B) \to \mathbb{R}$ by:

$$Q_j u = \frac{\int_{\partial\Omega_i} u \, ds}{\int_{\partial\Omega_i} ds}.$$

The piecewise constant coarse space, denoted $V_{0,P}(B) \subset V_h(B)$, is formally defined next as the *range* of an associated interpolation map $I_{0,P}$.

Definition 3.100. *We define an interpolation map $I_{0,P} : V_h(B) \to V_h(B)$:*

$$I_{0,P} v_h = \sum_{G \subset \mathcal{G}} I_G \sum_{\{j : G \subset \partial\Omega_i\}} d_j(G) \, Q_j v_h,$$

where Q_j and $d_j(G)$ are as defined in the preceding. The piecewise constant coarse space is then defined as the range of the interpolation map $I_{0,P}$:

$$V_{0,P}(B) \equiv \text{Range}(I_{0,P}). \tag{3.126}$$

The interpolation map $I_{0,P}$ is also denoted $\mathcal{Q}_0$ elsewhere in these notes.

We shall now turn to theoretical estimates of the interpolation map $I_{0,P}$. Unlike the traditional interpolation map $I_{0,T}$ the map $I_{0,P}$ is *not local*, and its values on a glob G depend on the mean value of the function on the boundaries of adjacent subdomains. Fortunately, by carefully choosing the partition of unity parameters $d_j(G)$, global norm bounds can be obtained which do not depend on $\{\rho_i\}$ and furthermore do not deteriorate in three dimensions.

Lemma 3.101. *Suppose the following assumptions hold.*

1. *Let $\Omega_1, \ldots, \Omega_p$ form a quasiuniform triangulation of Ω of size h_0.*
2. *Let $I_{0,P}$ denote the operator defined earlier based on the globs $G \in \mathcal{G}$:*

$$I_{0,P} v_h \equiv \sum_{G \in \mathcal{G}} I_G \sum_{\{j: G \subset \partial\Omega_j\}} d_j(G) (Q_j v_h), \tag{3.127}$$

where

$$\begin{cases} I = \sum_{G \in \mathcal{G}} I_G \\ 1 = \sum_{\{j: G \subset \partial\Omega_j\}} d_j(G). \end{cases} \tag{3.128}$$

Then, the following bounds will hold for $v_h \in V_h(B)$.

1. *If glob $G \subset \partial\Omega_i$, then:*

$$\begin{cases} \|I_G (I - I_{0,P}) v_h\|^2_{1/2, \partial\Omega_i} \\ \leq C \left(1 + \log(h_0/h)\right)^2 \sum_{\{j: G \subset \partial\Omega_j\}} d_j(G)^2 |v_h|^2_{1/2, \partial\Omega_j}. \end{cases} \tag{3.129}$$

2. *If the partition parameters $d_j(G)$ based on the globs are defined by:*

$$d_j(G) = \frac{\rho_j^t}{\sum_{\{l: G \subset \partial\Omega_l\}} \rho_l^t}, \qquad \text{for } G \subset \mathcal{G}$$

for $t \geq \frac{1}{2}$, then:

$$\begin{cases} \sum_{i=1}^p \rho_i \, |(I - I_{0,P}) v_h|^2_{1/2, \partial\Omega_i} \\ \leq C \, L^2 \, K^2 \, \left(1 + \log(h_0/h)\right)^2 \, \sum_{i=1}^p \rho_i |v_h|^2_{1/2, \partial\Omega_i}. \end{cases} \tag{3.130}$$

In both of the above, $C > 0$ is independent of h, h_0 and $\{\rho_j\}$.

Proof. We follow the proof in [MA14, MA17, MA15]. Substituting (3.127) and (3.128) in the expression for $I_G(v_h - I_{0,P} v_h)$ and using that $I_{G_1} I_{G_2} = 0$ whenever G_1 and G_2 are distinct globs, we obtain:

$$I_G (v_h - I_{0,P} v_h) = \sum_{\{j: G \subset \partial\Omega_j\}} d_j(G) I_G (I - Q_j) v_h.$$

Applying the generalized triangle inequality to the above expression yields:

$$|I_G (v_h - I_{0,P} v_h)|^2_{1/2, \partial\Omega_i} \leq L \sum_{\{j: G \subset \partial\Omega_j\}} d_j(G)^2 \|I_G (I - Q_j) v_h\|^2_{1/2, \partial\Omega_i}.$$

Since $G \subset \partial\Omega_i$ and $G \subset \partial\Omega_j$ the following norms will be equivalent:

$$c_1 \|I_G (I - Q_j) v_h\|^2_{1/2, \partial\Omega_i} \leq \|I_G (I - Q_j) v_h\|^2_{1/2, \partial\Omega_j} \leq c_2 \|I_G (I - Q_j) v_h\|^2_{1/2, \partial\Omega_i},$$

with $0 < c_1 < c_2$ independent of h, h_0 and $\{\rho_l\}$. This will hold because of the compact support of $I_G w$ so that $\|I_G w\|^2_{1/2,\partial\Omega_j}$ and $\|I_G w\|^2_{H^{1/2}_{00}(G)}$ will be equivalent by definition of $H^{1/2}_{00}(G)$. The latter will in turn be equivalent to $\|I_G w\|^2_{1/2,\partial\Omega_i}$. Applying this norm equivalence yields:

$$|I_G(v_h - I_{0,P}v_h)|^2_{1/2,\partial\Omega_i} \le c_2\, L \sum_{\{j:G\subset\partial\Omega_j\}} d_j(G)^2 \|I_G(I-Q_j)v_h\|^2_{1/2,\partial\Omega_j}.$$

Applying the glob theorem to the above yields:

$$\begin{cases} \quad |I_G(v_h - I_{0,P}v_h)|^2_{1/2,\partial\Omega_i} \\ \le q(h/h_0) \, \sum_{\{j:G\subset\partial\Omega_j\}} d_j(G)^2 \|(I-Q_j)v_h\|^2_{1/2,\partial\Omega_j} \\ = q(h/h_0) \, \sum_{\{j:G\subset\partial\Omega_j\}} d_j(G)^2 \Big(|(I-Q_j)v_h|^2_{1/2,\partial\Omega_j} + \frac{1}{h_0}\|(I-Q_j)v_h\|^2_{0,\partial\Omega_j} \Big) \end{cases}$$

where $q(h/h_0) \equiv C\, c_2\, L\, \left(1+\log(h_0/h)\right)^2$. Using a quotient space argument [CI2] (mapping $\partial\Omega_j$ to a reference domain, using that Q_j preserves constants and employing the scaling of seminorms under dilation) we obtain:

$$\|(I-Q_j)v_h\|^2_{0,\partial\Omega_j} \le C\, h_0\, |v_h|^2_{1/2,\partial\Omega_j},$$

for $c_3 > 0$ independent of h, h_0 and $\{\rho_l\}$. Since the seminorms are invariant under shifts by constants, we may replace $|(I-Q_j)v_h|^2_{1/2,\partial\Omega_j}$ by $|v_h|^2_{1/2,\partial\Omega_j}$:

$$|I_G(v_h - I_{0,P}v_h)|^2_{1/2,\partial\Omega_i} \le q(h/h_0) \sum_{\{j:G\subset\partial\Omega_j\}} d_j(G)^2\, |v_h|^2_{1/2,\partial\Omega_j}.$$

where $q(h/h_0) \equiv C\, c_2\, L\, \left(1+\log(h_0/h)\right)^2$. This yields (3.129).

To obtain (3.130), we multiply (3.129) by the factor ρ_i and rearrange terms:

$$\begin{cases} \qquad\qquad \rho_i\, |I_G(v_h - I_{0,P}v_h)|^2_{1/2,\partial\Omega_i} \\ \le C\, c_2\, L\, \left(1+\log(h_0/h)\right)^2 \sum_{\{j:G\subset\partial\Omega_j\}} \rho_i\, d_j(G)^2\, |v_h|^2_{1/2,\partial\Omega_j} \\ = C\, c_2\, L\, \left(1+\log(h_0/h)\right)^2 \sum_{\{j:G\subset\partial\Omega_j\}} \frac{\rho_i d_j(G)^2}{\rho_j} \rho_j\, |v_h|^2_{1/2,\partial\Omega_j}. \end{cases} \tag{3.131}$$

When $G \subset (\partial\Omega_i \cap \partial\Omega_j)$ the following bound may be obtained for $d_j(G)$:

$$d_j(G)^2 = \frac{\rho_j^{2t}}{\left(\sum_{\{l:G\subset\partial\Omega_l\}} \rho_l^t\right)^2} \le \frac{\rho_j^{2t}}{\left(\rho_i^t + \rho_j^t\right)^2} \le \frac{\rho_j^{2t}}{\rho_i^{2t} + \rho_j^{2t}},$$

which yields the following estimate for $\left(\rho_i d_j(G)^2/\rho_j\right)$:

$$\frac{\rho_i d_j(G)^2}{\rho_j} \le \frac{\rho_i \rho_j^{2t}}{\rho_j \rho_i^{2t} + \rho_j^{1+2t}} = \frac{(\rho_j/\rho_i)^{2t-1}}{1 + (\rho_j/\rho_i)^{2t}}.$$

Since the factor (ρ_j/ρ_i) is positive, the preceding expression will be uniformly bounded when $2t \geq 1$, with an upper bound of one. Substituting this upper bound into (3.131) yields:

$$\begin{cases} \qquad\qquad \rho_i \, |I_G(v_h - I_{0,P} v_h)|^2_{1/2,\partial\Omega_i} \\ \leq C\, c_2\, L\, \left(1 + \log(h_0/h)\right)^2 \sum_{\{j: G \subset \partial\Omega_j\}} \rho_j |v_h|^2_{1/2,\partial\Omega_j}. \end{cases}$$

To estimate $|v_h - I_{0,P} v_h|^2_{1/2,\partial\Omega_i}$ we employ the property of I_G on $\partial\Omega_i$:

$$v_h - I_{0,P} v_h = \sum_{\{G \subset \mathcal{G}\}} I_G(v_h - I_{0,P} v_h), \qquad \text{on } \partial\Omega_i.$$

This yields the bound:

$$\begin{cases} \rho_i \, |v_h - I_{0,P} v_h|^2_{1/2,\partial\Omega_i} \\ \leq \rho_i \, K \sum_{\{G \subset \partial\Omega_i\}} |I_G(v_h - I_{0,P} v_h)|^2_{1/2,\partial\Omega_i} \\ \leq C\, c_2\, K \sum_{\{G \subset \partial\Omega_i\}} L\, \left(1 + \log(h_0/h)\right)^2 \sum_{\{j: G \subset \partial\Omega_j\}} \rho_j |v_h|^2_{1/2,\partial\Omega_j}. \end{cases}$$

Summing over all subdomain boundaries $\partial\Omega_i$ yields:

$$\sum_{i=1}^{p} \rho_i \, |v_h - I_{0,P} v_h|^2_{1/2,\partial\Omega_i} \leq C\, c_2\, K^2\, L^2\, \left(1 + \log(h_0/h)\right)^2 \sum_{j=1}^{p} \rho_j |v_h|^2_{1/2,\partial\Omega_j},$$

which is the desired bound (3.130). □

As an immediate corollary, we obtain the following bounds for $I_{0,P}\, v_h$.

Lemma 3.102. *Let the assumptions in Lemma 3.101 hold. Then:*

$$\sum_{i=1}^{p} \rho_i \, |I_{0,P} v_h|^2_{1/2,\partial\Omega_i} \leq C\, L^2\, K^2\, \left(1 + \log(h_0/h)\right)^2 \sum_{i=1}^{p} \rho_i |v_h|^2_{1/2,\partial\Omega_i}, \quad (3.132)$$

for $C > 0$ independent of h, h_0 and $\{\rho_j\}$.

Proof. Follows immediately by an application of Lemma 3.101 and the triangle inequality to $I_{0,P}\, v_h = v_h - (I - I_{0,P}) v_h$. □

Remark 3.103. The reader is referred to [BR15, DR10, SM2] for theoretical estimates of the wirebasket interpolation map $I_{0,W}$.

3.9.4 Two Subdomain Preconditioners for S

As an application of the preceding theoretical results, we estimate the condition number of the two subdomain Dirichlet-Neumann [BJ9, BR11, FU, MA29], Neumann-Neumann [BO7], and the fractional Sobolev norm preconditioners [DR, GO3, BJ9, CH2, BR11], for the Schur complement. Such estimates can be obtained by applying Lemma 3.78.

Lemma 3.104. *Suppose that $\mathcal{T}_h(\Omega)$ is a quasiuniform triangulation of Ω and that neither Ω_1 nor Ω_2 is immersed. Let the coefficient $a(x) = \rho_i$ in Ω_i for $i = 1, 2$ and $c(x) = 0$ on Ω. Then, the following bound will hold for $v_h \in V_h(\Omega)$ with associated nodal vector $\mathbf{v}_B$ on $B = \partial\Omega_1 \cap \partial\Omega_2$:*

$$c_i\, \rho_i\, |v_h|^2_{H^{1/2}_{00}(B)} \leq \mathbf{v}_B^T S^{(i)} \mathbf{v}_B \leq C_i\, \rho_i\, |v_h|^2_{H^{1/2}_{00}(B)},$$

for $0 < c_i < C_i$ independent of h and $\{\rho_1, \rho_2\}$.

Proof. Follows from Lemma 3.78 since $|v_h|^2_{1/2,\partial\Omega_i}$ is norm equivalent to $|v_h|^2_{H^{1/2}_{00}(B)}$ by definition of $H^{1/2}_{00}(B)$. □

Remark 3.105. Suppose F represents the fractional Sobolev norm energy:

$$\mathbf{v}_B^T F \mathbf{v}_B = |v_h|^2_{H^{1/2}_{00}}, \quad \forall v_h \in V_h(B),$$

then the preceding lemma yields that for $\mathbf{v}_B \neq \mathbf{0}$:

$$c_i\, \rho_i \leq \frac{\mathbf{v}_B^T S^{(i)} \mathbf{v}_B}{\mathbf{v}_B^T F \mathbf{v}_B} \leq C_i\, \rho_i, \qquad \forall \mathbf{v}_B \neq \mathbf{0},$$

where $S^{(i)}$ denotes the subdomain Schur complement matrix.

We have the following condition number estimates.

Lemma 3.106. *Suppose the assumptions from Lemma 3.104 hold.*

1. *Let M denote any of the preconditioners [DR, GO3, BJ9, CH2, BR11] from (3.64), then M will be spectrally equivalent to F and satisfy:*

$$\mathrm{cond}(M, S) \leq \beta_1,$$

for some $\beta_1 > 0$ independent of h and $\{\rho_1, \rho_2\}$.

2. *If $M = S^{(i)}$, i.e., M is a Dirichlet-Neumann preconditioner, then:*

$$\mathrm{cond}(M, S) \leq \beta_2,$$

for some $\beta_2 > 0$ independent of h and $\{\rho_1, \rho_2\}$.

3. *If $M^{-1} = \alpha\, S^{(1)^{-1}} + (1-\alpha) S^{(2)^{-1}}$ for some $0 < \alpha < 1$, i.e., M corresponds to a Neumann-Neumann preconditioner, then:*

$$\mathrm{cond}(M, S) \leq \beta_3,$$

for some $\beta_3 > 0$ independent of h and $\{\rho_1, \rho_2\}$.

Proof. By Lemma 3.104, we obtain that $S = (S^{(1)} + S^{(2)}) \asymp (\rho_1 + \rho_2)\, F$. The desired result follows immediately. □

3.9.5 Multi-Subdomain Preconditioners for S

We now estimate the condition number of several *multisubdomain* Schwarz subspace preconditioners for the Schur complement matrix S. To study such convergence, we shall employ the Schwarz subspace framework from Chap. 2.5.2 with the linear space $V = V_h(B)$, endowed with the inner product $\mathcal{A}(.,.) \equiv \mathcal{S}(.,.)$ defined later. Subspaces $V_i \subset V$ will be chosen as subspaces of the form $V_h(G) \subset V_h(B)$ based on globs $G \subset \mathcal{G}$, and a coarse space $V_0 \subset V_h(B)$. Estimates for the parameters K_0 and K_1 from Chap. 2.5.2 can then be obtained by applying the glob theorem and other theoretical tools described in this section. Our estimates will be applicable when the coefficients $\{\rho_i\}$ have large variation across subdomains. However, when the variation in the coefficients is mild, improved bounds may be obtained in some cases by employing other tools, and the reader is referred to [BR12, DR17, BR15, DR10, TO10]. We also omit wirebasket preconditioners.

The inner produce $\mathcal{S}(\cdot,\cdot) : V_h(B) \times V_h(B) \rightarrow \mathbb{R}$, that we shall employ in $V_h(B)$ will be generated by the Schur complement matrix S. Given finite element functions $u_h, v_h \in V_h(B)$ defined on B with associated nodal vectors $\mathbf{u}_B$, $\mathbf{v}_B$, we define the bilinear form $\mathcal{S}(.,.)$ as:

$$\mathcal{S}(u_h, v_h) \equiv \mathbf{u}_B^T S \mathbf{v}_B = \begin{bmatrix} E\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} E\mathbf{v}_B \\ \mathbf{v}_B \end{bmatrix},$$

where $E \equiv -A_{II}^{-1} A_{IB}$. By Thm. 3.80, the following equivalence will hold:

$$c \sum_{i=1}^{p} \rho_i |u_h|_{1/2,\partial\Omega_i}^2 \leq \mathcal{S}(u_h, u_h) \leq C \sum_{i=1}^{p} \rho_i |u_h|_{1/2,\partial\Omega_i}^2, \qquad \forall u_h \in V_h(B).$$

For notational convenience, we shall use the notation:

$$\|u_h\|_{1/2,B}^2 \equiv \sum_{l=1}^{p} \rho_l |u_h|_{1/2,\partial\Omega_l}^2, \tag{3.133}$$

so that $c\ \|u_h\|_{1/2,B}^2 \leq \mathcal{S}(u_h, u_h) \leq C\,\|u_h\|_{1/2,B}^2$. For convenience, the globs in $\mathcal{G}$ shall be enumerated as $G_1, \ldots, G_n$ for some n. The additive Schwarz subspace preconditioners we shall consider will be based on the subspaces $V_h(G_i) \equiv \mathrm{Range}(I_{G_i}) \subset V_h(B)$ corresponding to globs $G_i \in \mathcal{G}$, and a coarse space $V_0(B) \subset V_h(B)$. We shall consider only the traditional coarse space $V_{0,T}(B) \subset V_h(B)$ and the piecewise constant space $V_{0,P} \subset V_h(B)$. The local bilinear forms $\tilde{\mathcal{A}}_i(\cdot,\cdot)$ in the abstract Schwarz framework of Chap. 2.5.2 will be denoted $\mathcal{S}_i(\cdot,\cdot)$ on $V_h(G_i)$ as defined below:

$$\mathcal{S}_i(u_h, v_h) \equiv \mathcal{S}(u_h, v_h), \quad \forall u_h, v_h \in V_h(G_i).$$

Similarly, for the coarse space.

To simplify our discussion, we shall assume that exact solvers are employed for the submatrices so that the parameters $\omega_0 = \omega_1 = 1$ and $K_0 = C_0$. If a coarse space is not employed, we define an $n \times n$ matrix $\epsilon = (\epsilon_{ij})$ of strengthened Cauchy-Schwartz parameters such that:

$$\mathcal{S}(v_i, v_j) \le \epsilon_{ij}\, \mathcal{S}(v_i, v_i)^{1/2}\, \mathcal{S}(v_j, v_j)^{1/2} \quad \forall\, v_i \in V_h(G_i) \text{ and } \forall\, v_j \in V_h(G_j).$$

It is easily verified that the spectral radius $\rho(\epsilon)$ of matrix ϵ is bounded by $K\,L$. When a coarse space is employed, matrix ϵ will be of size $(n+1)$ and its spectral radius will be bounded by $(K\,L + 1)$ regardless of the choice of coarse space. By the abstract theory of Chap. 2.5.2, the condition number of additive Schwarz subspace preconditioner for S will satisfy:

$$\mathrm{cond}(M, S) \le \begin{cases} C_0\, K\, L, & \text{No Coarse Space} \\ C_0\, (K\, L + 1), & \text{With Coarse Space.} \end{cases}$$

Since K and L are typically independent of h, h_0 and $\{\rho_i\}$, we only need to focus on the partition parameter C_0. The next result yields an estimate for C_0 when there is no coarse space.

Lemma 3.107. *Let $G_1, \ldots, G_n$ denote an enumeration of all the distinct globs in $\mathcal{G}$ so that the following decomposition of identity property holds:*

$$I = \sum_{i=1}^{n} I_{G_i},$$

where $I : V_h(B) \to V_h(B)$, and $I_{G_i} : V_h(B) \to V_h(G_i)$ for $1 \le i \le n$. Then, given $v_h \in V_h(B)$ there exists $v_i \in V_h(G_i)$ for $1 \le i \le n$ satisfying

$$v_h = v_1 + \cdots + v_n$$

and

$$\sum_{i=1}^{p} \mathcal{S}(v_i, v_i) \le C_0\, \mathcal{S}(v_h, v_h),$$

where

$$C_0 \le C\, L\, \left(1 + \log(h_0/h)\right)^2\, h_0^{-2} \left(\frac{\rho_{\max}}{\rho_{\min}}\right), \quad \forall v_h \in V_h(B),$$

with C independent of h, h_0 and $\{\rho_i\}$.

Proof. See [TO10, MA17]. We shall estimate the partition parameter C_0 in the weighted boundary norm (3.133) instead of $\mathcal{S}(\cdot,\cdot)$ since both are equivalent. Given $v_h \in V_h(B)$ define $v_i = I_{G_i} v_h \in V_h(G_i)$ for $1 \le i \le n$. Due to the decomposition of unity property for the I_{G_i}, it will hold that:

$$v_1 + \cdots + v_n = v_h.$$

If $G_l \subset \partial\Omega_i$, then by the glob theorem, we obtain that:

$$|I_{G_l} v_h|^2_{1/2,\partial\Omega_i} \le C\,(1+\log(h_0/h))^2\,\|v_h\|^2_{1/2,\partial\Omega_i}. \tag{3.134}$$

Define $\mathcal{H}_h v_h$ as the discrete harmonic extension of the the interface value v_h into the subdomain interiors Ω_i for $1 \le i \le p$

$$\mathcal{A}(\mathcal{H}_h v_h, v_i) = 0, \qquad \forall v_i \in V_h(\Omega_i) \cap H^1_0(\Omega_i),$$

where

$$\mathcal{A}(u,v) \equiv \sum_{i=1}^p \rho_i \int_{\Omega_i} \nabla u \cdot \nabla v\, dx.$$

Then, an application of the trace theorem to $\mathcal{H}_h v_h$ on $\partial\Omega_i$ yields:

$$\begin{cases} \|v_h\|^2_{1/2,\partial\Omega_i} \le C\,\|\mathcal{H}_h v_h\|^2_{1,\Omega_i} \\ \qquad\qquad\quad = C\,\left(|\mathcal{H}_h v_h|^2_{1,\Omega_i} + \frac{1}{h_0^2}\|\mathcal{H}_h v_h\|^2_{0,\Omega_i}\right). \end{cases}$$

Substituting the preceding bound into (3.134), multiplying by the factor ρ_i, and summing over all adjacent subdomains yields the following:

$$\begin{aligned} \|I_{G_l} v_h\|^2_{1/2,B} &= \textstyle\sum_{i=1}^p \rho_i |I_{G_l} v_h|^2_{1/2,\partial\Omega_i} \\ &\le C\,(1+\log(h_0/h))^2 \textstyle\sum_{i:G_l\subset\partial\Omega_i} \rho_i \left(|\mathcal{H}_h v_h|^2_{1,\Omega_i} + \frac{1}{h_0^2}\|\mathcal{H}_h v_h\|^2_{0,\Omega_i}\right) \\ &\le C\,(1+\log(h_0/h))^2\, \rho_{\max} \textstyle\sum_{i:G_l\subset\partial\Omega_i} \left(|\mathcal{H}_h v_h|^2_{1,\Omega_i} + \frac{1}{h_0^2}\|\mathcal{H}_h v_h\|^2_{0,\Omega_i}\right). \end{aligned}$$

Summing over all globs yields the estimate:

$$\begin{aligned} \textstyle\sum_{l=1}^n \|I_{G_l} v_h\|^2_{1/2,B} &= \textstyle\sum_{l=1}^n \sum_{i=1}^p \rho_i |I_{G_l} v_h|^2_{1/2,\partial\Omega_i} \\ &\le C\,(1+\log(h_0/h))^2 \textstyle\sum_{l=1}^n \sum_{i:G_l\subset\partial\Omega_i} \rho_i \left(|\mathcal{H}_h v_h|^2_{1,\Omega_i} + \frac{1}{h_0^2}\|\mathcal{H}_h v_h\|^2_{0,\Omega_i}\right) \\ &\le C\,(1+\log(h_0/h))^2\, \rho_{\max}\, L \textstyle\sum_{i=1}^p \left(|\mathcal{H}_h v_h|^2_{1,\Omega_i} + \frac{1}{h_0^2}\|\mathcal{H}_h v_h\|^2_{0,\Omega_i}\right) \\ &= C\,(1+\log(h_0/h))^2\, \rho_{\max}\, L \left(|\mathcal{H}_h v_h|^2_{1,\Omega} + \frac{1}{h_0^2}\|\mathcal{H}_h v_h\|^2_{0,\Omega}\right). \end{aligned}$$

Since $\mathcal{H}_h v_h$ is zero on $\mathcal{B}_D$, we apply Poincaré-Freidrich's inequality:

$$\|\mathcal{H}_h v_h\|^2_{0,\Omega} \le C\,|\mathcal{H}_h v_h|^2_{1,\Omega},$$

and substitute it in the preceding bound to obtain:

$$\begin{cases} \sum_{l=1}^n \|I_{G_l} v_h\|^2_{1/2,B} = \sum_{l=1}^n \sum_{i=1}^p \rho_i |I_{G_l} v_h|^2_{1/2,\partial\Omega_i} \\ \le C\,(1+\log(h_0/h))^2\, \rho_{\max}\, L\,(1+\frac{1}{h_0^2}) \sum_{i=1}^p \rho_i |\mathcal{H}_h v_h|^2_{1,\Omega_i}. \end{cases} \tag{3.135}$$

Since $\mathcal{H}_h v_h$ is piecewise discrete harmonic, Thm. 3.80 yields the equivalence:

$$c\,\|v_h\|^2_{1/2,B} \le \mathcal{S}(v_h, v_h) \;=\; \sum_{i=1}^{p} \rho_i |\mathcal{H}_h v_h|^2_{1,\Omega_i} \le C\,\|v_h\|^2_{1/2,B},$$

for c, C independent of h, h_0 and $\{\rho_i\}$. Substituting this into (3.135) yields:

$$\sum_{l=1}^{n} \mathcal{S}(I_{G_l} v_h, I_{G_l} v_h) \;\le\; C\,L\,\left(1 + \log(h_0/h)\right)^2 \frac{\rho_{\max}}{\rho_{\min}} \left(1 + \frac{1}{h_0^2}\right) \mathcal{S}(v_h, v_h)$$

which is the desired result. □

Remark 3.108. As an immediate corollary, we obtain the following condition number estimate for the *block Jacobi* Schur complement preconditioner in two or three dimensions:

$$\text{cond}(M, S) \le C\,L\,\left(1 + \log(h_0/h)\right)^2 \left(\frac{\rho_{\max}}{\rho_{\min}}\right) \left(1 + \frac{1}{h_0^2}\right),$$

for some $C > 0$ independent of h, h_0 and $\{\rho_i\}$. This upper bound may be unduly pessimistic when the factor $(\rho_{\max}/\rho_{\min})$ is large. However, if a suitable coarse space V_0 is employed, these bounds can be improved significantly.

Our next estimate is for the Schur complement additive Schwarz preconditioner when a coarse space is included [TO10, MA17].

Lemma 3.109. *Let $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$. Let the following conditions hold.*

1. *Let $G_1, \ldots, G_n$ denote an enumeration of all the distinct globs in $\mathcal{G}$ so that the following decomposition of identity property holds:*

$$I = \sum_{i=1}^{n} I_{G_i},$$

where $I : V_h(B) \to V_h(B)$, and $I_{G_i} : V_h(B) \to V_h(G_i)$ for $1 \le i \le n$.

2. *Let coarse space $V_{0,T}(B) \subset V_h(B)$ or $V_{0,P}(B) \subset V_h(B)$ be employed.*
3. *Let the partition parameters $d_j(G)$ be defined for $t \ge \frac{1}{2}$ by:*

$$d_j(G) = \frac{\rho_j^t}{\sum_{\{l: G \subset \partial\Omega_l\}} \rho_l^t}.$$

Then, given $v_h \in V_h(B)$ there exists $v_0 \in V_0$ and $v_i \in V_h(G_i)$ for $1 \le i \le n$ with $v_h = v_0 + v_1 + \cdots + v_n$ satisfying:

$$\sum_{i=0}^{p} \mathcal{S}(v_i, v_i) \le C_0\, \mathcal{S}(v_h, v_h),$$

where

$$C_0 \le \begin{cases} C\,\left(1 + \log(h_0/h)\right)^2, & \text{if } V_0 = V_{0,P} \text{ for } d = 2, 3 \\ C\,\left(1 + \log(h_0/h)\right)^2, & \text{if } V_0 = V_{0,T} \text{ and } d = 2 \\ C\,(h_0/h), & \text{if } V_0 = V_{0,T} \text{ and } d = 3, \end{cases}$$

with C independent of h, h_0 and $\{\rho_i\}$.

Proof. We shall only outline the proof for the choice $V_0 = V_{0,P}(B)$. The choice $V_0 = V_{0,T}(B)$ will be analogous, with differences arising from the bound for $I_{0,T} : V_h(B) \to V_{0,T}$ depending on whether $d = 2$ or $d = 3$. Given $v_h \in V_h(B)$, define $v_0 \equiv I_{0,P}\, v_h$ where $I_{0,P}$ is the interpolation onto $V_{0,P}(B)$. For $1 \le i \le n$ define $v_i \equiv I_{G_i}\,(v_h - I_0 v_h)$. By construction, it will hold that:

$$v_h = v_0 + v_1 + \cdots + v_n.$$

Bound (3.132) from the preceding section yields that:

$$\|I_{0,P} v_h\|^2_{1/2,B} \le C\,(1 + \log(h_0/h))^2\, \|v_h\|^2_{1/2,B}.$$

To estimate $v_l = I_{G_l}(v_h - I_{0,P} v_h)$ when $G_l \subset \partial\Omega_i$, bound (3.129) yields:

$$\|I_{G_l}(I - I_{0,P}) v_h\|^2_{1/2,\partial\Omega_i} \le C\,(1 + \log(h_0/h))^2 \sum_{\{j:G_l \subset \partial\Omega_j\}} d_j(G_l)^2 |v_h|^2_{1/2,\partial\Omega_j},$$

where $C > 0$ is independent of h, h_0 and $\{\rho_j\}$. Multiply the above expression by ρ_i and sum over all subdomains containing G_l to obtain:

$$\begin{aligned}
&\textstyle\sum_{\{i:G_l\subset\partial\Omega_i\}} \rho_i \|I_{G_l}(I - I_{0,P}) v_h\|^2_{1/2,\partial\Omega_i} \\
&\le C\,(1 + \log(h_0/h))^2 \textstyle\sum_{\{i:G_l\subset\partial\Omega_i\}} \sum_{\{j:G_l\subset\partial\Omega_j\}} \rho_i d_j(G_l)^2 |v_h|^2_{1/2,\partial\Omega_j} \\
&= C\,(1 + \log(h_0/h))^2 \textstyle\sum_{\{i:G_l\subset\partial\Omega_i\}} \sum_{\{j:G_l\subset\partial\Omega_j\}} \frac{\rho_i d_j(G_l)^2}{\rho_j} \rho_j |v_h|^2_{1/2,\partial\Omega_j}.
\end{aligned} \tag{3.136}$$

When $G_l \subset (\partial\Omega_i \cap \partial\Omega_j)$, the following bound can be obtained, as before:

$$\frac{\rho_i d_j(G_l)^2}{\rho_j} \le \frac{\rho_i \rho_j^{2t}}{\rho_j\left(\rho_i^{2t} + \rho_j^{2t}\right)} \le 1, \qquad \text{for } t \ge 1/2.$$

Substitution of the above into (3.136) yields the bound:

$$\begin{aligned}
&\|I_{G_l}(I - I_{0,P}) v_h\|^2_{1/2,B} \\
&= \textstyle\sum_{\{i:G_l\subset\partial\Omega_i\}} \rho_i \|I_{G_l}(I - I_{0,P}) v_h\|^2_{1/2,\partial\Omega_i} \\
&\le C\,(1 + \log(h_0/h))^2 \textstyle\sum_{\{i:G_l\subset\partial\Omega_i\}} \sum_{\{j:G_l\subset\partial\Omega_j\}} \rho_j |v_h|^2_{1/2,\partial\Omega_j}.
\end{aligned}$$

Summing over all globs G_l yields:

$$\begin{aligned}
&\textstyle\sum_{l=1}^{n} \|I_{G_l}(I - I_{0,P}) v_h\|^2_{1/2,\{\rho_i\},B} \\
&\le C\,(1 + \log(h_0/h))^2 \textstyle\sum_{l=1}^{n} \sum_{\{i:G_l\subset\partial\Omega_i\}} \sum_{\{j:G_l\subset\partial\Omega_j\}} \rho_j |v_h|^2_{1/2,\partial\Omega_j} \\
&\le C\,(1 + \log(h_0/h))^2\, L^2 \textstyle\sum_{j=1}^{p} \rho_j |v_h|^2_{1/2,\partial\Omega_j} \\
&= C\,(1 + \log(h_0/h))^2\, L^2 \|v_h\|^2_{1/2,B}.
\end{aligned}$$

Combining the above bound with (3.132) yields:

$$\begin{cases} \|I_{0,P}v_h\|_{1/2,B}^2 + \sum_{l=1}^n \|I_{G_l}(I - I_{0,P})v_h\|_{1/2,\{\rho_i\},B}^2 \\ \qquad \le C\,(1+\log(h_0/h))^2\,\left(1+L^2\right)\|v_h\|_{1/2,B}^2. \end{cases}$$

Since $v_0 = I_{0,P}v_h$ and $v_l = I_{G_l}(I - I_{0,P})v_h$ for $1 \le l \le n$, the desired result follows by equivalence between $\mathcal{S}(w_h, w_h)$ and $\|w_h\|_{1/2,B}^2$. □

As a corollary, we estimate the condition number of the additive Schwarz preconditioner for S based on $V_h(G_1), \ldots, V_h(G_n)$, and $V_{0,T}$ or $V_{0,P}$.

Lemma 3.110. *Let the assumptions in Lemma 3.109 hold. Then:*

$$\mathrm{cond}(M,S) \le \begin{cases} C\,K\,L\,(1+\log(h_0/h))^2\,, & \text{if } V_0 = V_{0,P} \text{ and } d = 2,3 \\ C\,K\,L\,(1+\log(h_0/h))^2\,, & \text{if } V_0 = V_{0,T} \text{ and } d = 2 \\ C\,K\,L\,(h_0/h)\,, & \text{if } V_0 = V_{0,T} \text{ and } d = 3. \end{cases}$$

Proof. Since *cond*$(M,S) \le K_0K_1$, we combine the upper bound $K_1 \le M\,L$ with preceding bounds for $K_0 = C_0$ to obtain the desired result. □

The preceding lemma may be applied to estimate the condition number of several Schur complement preconditioners in two and three dimensions.

- The BPS preconditioner in *two dimensions* is an additive Schwarz subspace preconditioner based on the edge and vertex globs, and a coarse space $V_{0,T}$ or $V_{0,P}$. The preceding result yields logarithmic bounds.
- The vertex space preconditioner in two or three dimensions, based on the vertex, edge and face globs (in three dimensions) or their extensions, is also an additive Schwarz subspace preconditioner. If coarse space $V_{0,T}$ is employed, then the bound $C\,(1+\log(h_0/h))^2$ will hold in two dimensions, while $C\,(1+(h_0/h))$ will hold in three dimensions. If coarse space $V_{0,P}$ is employed, then the bound $C\,(1+\log(h_0/h))^2$ will hold in two and three dimensions. Improved bounds independent of h_0 and h can be proved, depending only on the amount β of overlap, when the coefficient $a(x)$ is smooth, see [SM, DR17, DR10].
- The Schwarz subspace preconditioner for S based on the overlapping subregions $\partial\Omega_1, \ldots, \partial\Omega_p$ of B and either coarse space $V_{0,T}$ or $V_{0,P}$ will yield similar bounds as the vertex space preconditioner.

Readers are referred to [BJ8, BJ9, BR11, BR12, BR13, BR14, BR15, DR14] and [DE3, DR10, WI6, MA17, XU10, KL8, TO10] for additional theory.

3.9.6 Balancing Domain Decomposition Preconditioner

We conclude our discussion on bounds for Schur complement preconditioners, by estimating the condition number of the balancing domain decomposition preconditioner using an algebraic framework introduced in [MA14, MA17]. We refer the reader to [DR14, DE3, DR18, TO10] for general convergence estimates on Neumann-Neumann preconditioners.

We shall employ the following notation in our discussion. The number of nodes on B will be denoted n and the number of nodes on $B^{(i)} = \partial\Omega_i \backslash \mathcal{B}_D$ will be denoted n_i. The Euclidean inner product on $\mathbb{R}^n$ will be denoted:

$$(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{v}, \quad \forall\, \mathbf{u},\, \mathbf{v} \in \mathbb{R}^n.$$

We also employ the inner product generated by the Schur complement S:

$$S(\mathbf{u}, \mathbf{v}) \equiv (S\mathbf{u}, \mathbf{v}) = \mathbf{u}^T S \mathbf{v}, \qquad \forall\, \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

On each non-Dirichlet boundary segment $B^{(i)}$, we define a *semi-norm*:

$$|\mathbf{w}_i|^2_{S^{(i)}} \equiv \left(S^{(i)} \mathbf{w}_i, \mathbf{w}_i\right), \qquad \text{for } \mathbf{w}_i \in \mathbb{R}^{n_i}.$$

The following Cauchy-Schwartz inequality will hold:

$$\left(S^{(i)} \mathbf{u}_i, \mathbf{v}_i\right) \leq \left(S^{(i)} \mathbf{u}_i, \mathbf{u}_i\right)^{1/2} \left(S^{(i)} \mathbf{v}_i, \mathbf{v}_i\right)^{1/2}, \quad \forall \mathbf{u}_i,\, \mathbf{v}_i \in \mathbb{R}^{n_i},$$

even when the Schur complement matrices $S^{(i)}$ is singular. Indeed, such a Cauchy-Schwartz inequality follows from the Euclidean Cauchy-Schwartz inequality since the fractional powers $(S^{(i)})^\alpha$ are well defined for $\alpha \geq 0$ because $S^{(i)}$ is symmetric positive semidefinite:

$$\begin{aligned} \left(S^{(i)} \mathbf{u}_i, \mathbf{v}_i\right) &= \left((S^{(i)})^{1/2} \mathbf{u}_i, (S^{(i)})^{1/2} \mathbf{v}_i\right) \\ &\leq \left((S^{(i)})^{1/2} \mathbf{u}_i, (S^{(i)})^{1/2} \mathbf{u}_i\right)^{1/2} \left((S^{(i)})^{1/2} \mathbf{v}_i, (S^{(i)})^{1/2} \mathbf{v}_i\right)^{1/2} \\ &= \left(S^{(i)} \mathbf{u}_i, \mathbf{u}_i\right)^{1/2} \left(S^{(i)} \mathbf{v}_i, \mathbf{v}_i\right)^{1/2} \end{aligned}$$

for all $\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^{n_i}$. When the subdomain stiffness matrix $S^{(i)}$ is singular, we shall denote by Z_i (identical to N_i in Chap. 3.7) an $n_i \times d_i$ matrix:

$$\text{Kernel}(S^{(i)}) \subset \text{Range}(Z_i),$$

whose column space (i.e., range) contains the null space of $S^{(i)}$.

For each subdomain, let $\mathcal{R}_i$ (same as $\mathcal{R}_{B^{(i)}}$ in Chap. 3.7) denote the $n_i \times n$ matrix which restricts a nodal vector on B to its subvector corresponding to nodes on $B^{(i)}$. For each subdomain, let D_i denote a diagonal matrix of size n_i with positive diagonal entries such that the following identity holds:

$$I = \sum_{i=1}^{p} \mathcal{R}_i^T D_i \mathcal{R}_i,$$

which we refer to as a decomposition of unity. Define N (identical to matrix C in Chap. 3.7) as the following $n \times d$ matrix, where $d \equiv (d_1 + \cdots + d_p)$:

$$N \equiv \left[\mathcal{R}_1^T D_1^T Z_1 \cdots \mathcal{R}_p^T D_p^T Z_p \right].$$

We define T as the following $n \times n$ matrix:

$$T = \sum_{i=1}^{p} \mathcal{R}_i^T D_i^T S^{(i)^\dagger} D_i \mathcal{R}_i,$$

where $S^{(i)^\dagger}$ denotes the Moore-Penrose pseudoinverse of matrix $S^{(i)}$, see [ST13, GO4]. We define $\tilde{P}_0$ as the following $n \times n$ symmetric matrix:

$$\tilde{P}_0 = N \left(N^T S N \right)^{-1} N^T \quad \text{and} \quad P_0 = \tilde{P}_0 S.$$

By definition, $P_0 = \tilde{P}_0 S$ corresponds to the S-orthogonal projection onto Range(N). Consequently, the following properties will hold:

$$\begin{array}{ll} P_0 P_0 & = P_0 \\ P_0(I - P_0) & = 0 \\ S(P_0\mathbf{u}, \mathbf{v}) & = S(\mathbf{u}, \mathbf{v}), \quad \forall \mathbf{v} \in \text{Range}(N), \;\; \mathbf{u} \in \mathrm{I\!R}^n \\ S(P_0\mathbf{u}, \mathbf{v}) & = S(\mathbf{u}, P_0\mathbf{v}), \forall \mathbf{u}, \, \mathbf{v} \in \mathrm{I\!R}^n \\ S(P_0\mathbf{u}, P_0\mathbf{u}) & \le S(\mathbf{u}, \mathbf{u}), \quad \forall \mathbf{u}. \end{array}$$

Employing the above notation, we express the matrix form of the balancing domain decomposition preconditioner for S.

Lemma 3.111. *The following properties will hold.*

1. *The inverse M^{-1} of the balanced domain decomposition preconditioner is:*

$$M^{-1} = \tilde{P}_0 + (I - \tilde{P}_0 S) T (I - S \tilde{P}_0).$$

2. *The preconditioned Schur complement matrix $M^{-1}S$ will have the form:*

$$\begin{aligned} M^{-1}S &= \tilde{P}_0 S + (I - \tilde{P}_0 S) T S (I - \tilde{P}_0 S) \\ &= P_0 + (I - P_0) T S (I - P_0), \end{aligned} \tag{3.137}$$

where $M^{-1}S$ will be symmetric in the S-inner product.

Proof. Follows from the hybrid Schwarz description of the balancing domain decomposition preconditioner, in Chap. 3.7. □

Since the preconditioned matrix $M^{-1}S$ is symmetric in the S-inner product, its condition number can be estimated as:

$$\mathrm{cond}(M,S) = \frac{\lambda_M}{\lambda_m}, \tag{3.138}$$

where λ_m and λ_M denote the minimum and maximum values of the generalized Rayleigh quotient associated with $M^{-1}S$ in the S-inner product:

$$\lambda_m \le \frac{S\left(M^{-1}S\mathbf{u},\mathbf{u}\right)}{S\left(\mathbf{u},\mathbf{u}\right)} \le \lambda_M, \quad \forall \mathbf{u} \in \mathbb{R}^n \backslash \{\mathbf{0}\}. \tag{3.139}$$

Estimation of the parameters λ_m, λ_M may be simplified using the S-orthogonality of the decomposition $\mathbf{u} = P_0\mathbf{u} + (I - P_0)\mathbf{u}$, as the following result shows.

Lemma 3.112. *Suppose the following condition holds:*

$$\gamma_m \le \frac{S\left(TS(I-P_0)\mathbf{u},(I-P_0)\mathbf{u}\right)}{S\left((I-P_0)\mathbf{u},(I-P_0)\mathbf{u}\right)} \le \gamma_M, \quad \forall \mathbf{u} \in \mathbb{R}^n \backslash \{\mathbf{0}\}, \tag{3.140}$$

for parameters $0 < \gamma_m \le \gamma_M$ *Then, the following bound will hold:*

$$\mathrm{cond}(M,S) \le \frac{\max\{1,\gamma_M\}}{\min\{1,\gamma_m\}}. \tag{3.141}$$

Proof. We will derive bounds for $\mathrm{cond}(M,S)$ by estimating the extreme values of the generalized Rayleigh quotient of $M^{-1}S$ in the S-inner product, as described in (3.138) and (3.139). By substituting (3.137) into $S\left(M^{-1}S\mathbf{u},\mathbf{u}\right)$, we obtain the following equivalent expression:

$$\begin{aligned} S\left(M^{-1}S\mathbf{u},\mathbf{u}\right) &= S\left(P_0\mathbf{u} + (I-P_0)TS(I-P_0)\mathbf{u},\mathbf{u}\right) \\ &= S\left(P_0\mathbf{u},\mathbf{u}\right) + S((I-P_0)TS(I-P_0)\mathbf{u},\mathbf{u}) \\ &= S\left(P_0\mathbf{u},P_0\mathbf{u}\right) + S(TS(I-P_0)\mathbf{u},(I-P_0)\mathbf{u})\,. \end{aligned} \tag{3.142}$$

Employing the Pythagorean theorem:

$$S(\mathbf{u},\mathbf{u}) = S(P_0\mathbf{u},P_0\mathbf{u}) + S((I-P_0)\mathbf{u},(I-P_0)\mathbf{u})$$

and substituting the bounds in (3.140) into (3.142) yields the estimates:

$$\min\{1,\gamma_m\} \le \frac{S\left(P_0\mathbf{u},P_0\mathbf{u}\right) + S\left(TS(I-P_0)\mathbf{u},(I-P_0)\mathbf{u}\right)}{S\left(P_0\mathbf{u},P_0\mathbf{u}\right) + S\left((I-P_0)\mathbf{u},(I-P_0)\mathbf{u}\right)} \le \max\{1,\gamma_M\},$$

for $\mathbf{u} \neq \mathbf{0}$. □

Next an alternative expression is derived for $S\left(TS(I-P_0)\mathbf{u},\mathbf{u}\right)$. It is then proved that $\gamma_m = 1$. Following that, Lemma 3.114 proves a bound for γ_M. Readers are referred to [MA17] for additional details.

Lemma 3.113. *Given* $\mathbf{u} \in \mathrm{I\!R}^n$ *define* $\mathbf{u}_i \in \mathrm{I\!R}^{n_i}$ *as follows:*

$$\mathbf{u}_i \equiv S^{(i)^\dagger} D_j \mathcal{R}_i (I - P_0)\mathbf{u} \qquad \textit{for} \quad 1 \le i \le p,$$

and define

$$|\mathbf{u}_i|^2_{S^{(i)}} \equiv \left(S^{(i)} \mathbf{u}_i, \mathbf{u}_i \right), \qquad \textit{for} \quad 1 \le i \le p.$$

Then, the following identity will hold:

$$S\left(TS(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u}\right) = \sum_{i=1}^{p} |\mathbf{u}_i|^2_{S^{(i)}}. \tag{3.143}$$

Furthermore, the lower bound $\gamma_m = 1$ *will hold, i.e.,*

$$S\left((I-P_0)\mathbf{u}, (I-P_0)\mathbf{u}\right) \le S\left(TS(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u}\right), \quad \forall \mathbf{u} \in \mathrm{I\!R}^n. \tag{3.144}$$

Proof. To derive (3.143), express $S\left(TS(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u}\right)$ in the Euclidean inner product, substitute $T = \sum_{i=1}^p \mathcal{R}_i^T D_i^T S^{(i)^\dagger} D_i \mathcal{R}_i$ and simplify as follows:

$$\begin{aligned}
&(TS(I-P_0)\mathbf{u}, S(I-P_0)\mathbf{u}) \\
&= \left(\textstyle\sum_{i=1}^p \mathcal{R}_i^T D_i^T S^{(i)^\dagger} D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, S(I-P_0)\mathbf{u} \right) \\
&= \textstyle\sum_{i=1}^p \left(S^{(i)^\dagger} D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, D_i \mathcal{R}_i S(I-P_0)\mathbf{u} \right) \\
&= \textstyle\sum_{i=1}^p \left(S^{(i)} S^{(i)^\dagger} D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, S^{(i)^\dagger} D_i \mathcal{R}_i S(I-P_0)\mathbf{u} \right) \\
&= \textstyle\sum_{i=1}^p \left(S^{(i)} \mathbf{u}_i, \mathbf{u}_i \right) \\
&= \textstyle\sum_{i=1}^p |\mathbf{u}_i|^2_{S^{(i)}}.
\end{aligned}$$

To derive a lower bound for $S\left(TS(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u}\right)$, insert $I = \sum_{i=1}^p \mathcal{R}_i^T D_i \mathcal{R}_i$ in $S\left((I-P_0)\mathbf{u}, (I-P_0)\right)$ and expand to obtain:

$$\begin{aligned}
S\left((I-P_0)\mathbf{u}, (I-P_0)\right) &= (S(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u}) \\
&= \left(\textstyle\sum_{i=1}^p \mathcal{R}_i^T D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u} \right) \\
&= \textstyle\sum_{i=1}^p \left(\mathcal{R}_i^T D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u} \right) \\
&= \textstyle\sum_{i=1}^p (D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, \mathcal{R}_i (I-P_0)\mathbf{u}).
\end{aligned} \tag{3.145}$$

By definition of P_0, the vector $S(I-P_0)\mathbf{u}$ will be balanced. Therefore it will hold that $D_i \mathcal{R}_i S(I-P_0)\mathbf{u} \perp \text{Kernel}(S^{(i)})$, so a property of the pseudoinverse yields $D_i \mathcal{R}_i S(I-P_0)\mathbf{u} = S^{(i)} S^{(i)^\dagger} D_i \mathcal{R}_i S(I-P_0)\mathbf{u}$, for $1 \le i \le p$. Substituting this into (3.145) and expressing the result in terms of $\mathbf{u}_i$, and applying the Cauchy-Schwartz inequality yields:

$$\begin{aligned}
&S\left((I-P_0)\mathbf{u},(I-P_0)\right)\\
&\quad=\sum_{i=1}^{p}\left(D_i\mathcal{R}_iS(I-P_0)\mathbf{u},\mathcal{R}_i(I-P_0)\mathbf{u}\right)\\
&\quad=\sum_{i=1}^{p}\left(S^{(i)}S^{(i)^{\dagger}}D_i\mathcal{R}_iS(I-P_0)\mathbf{u},\mathcal{R}_i(I-P_0)\mathbf{u}\right)\\
&\quad=\sum_{i=1}^{p}\left(S^{(i)}\mathbf{u}_i,\mathcal{R}_i(I-P_0)\mathbf{u}\right)\\
&\quad\leq\sum_{i=1}^{p}\left(S^{(i)}\mathbf{u}_i,\mathbf{u}_i\right)^{1/2}\left(S^{(i)}\mathcal{R}_i(I-P_0)\mathbf{u},\mathcal{R}_i(I-P_0)\mathbf{u}\right)^{1/2}\\
&\quad\leq\left(\sum_{i=1}^{p}(S^{(i)}\mathbf{u}_i,\mathbf{u}_i)\right)^{1/2}\left(\sum_{i=1}^{p}(S^{(i)}\mathcal{R}_i(I-P_0)\mathbf{u},\mathcal{R}_i(I-P_0)\mathbf{u})\right)^{1/2}\\
&\quad=\left(\sum_{i=1}^{p}|\mathbf{u}_i|_{S^{(i)}}^2\right)^{1/2}\left((\sum_{i=1}^{p}\mathcal{R}_i^TS^{(i)}\mathcal{R}_i(I-P_0)\mathbf{u},(I-P_0)\mathbf{u})\right)^{1/2}\\
&\quad=\left(\sum_{i=1}^{p}|\mathbf{u}_i|_{S^{(i)}}^2\right)^{1/2}\left(S(I-P_0)\mathbf{u},(I-P_0)\mathbf{u}\right)^{1/2}.
\end{aligned}$$

Canceling common terms and squaring the resulting expression yields:

$$S\left((I-P_0)\mathbf{u},(I-P_0)\right)\,\leq\,\sum_{i=1}^{p}|\mathbf{u}_i|_{S^{(i)}}^2\,=\,S\left(TS(I-P_0)\mathbf{u},(I-P_0)\mathbf{u}\right)$$

where bound (3.143) was employed. This yields the bound $\gamma_m=1$. □

Lemma 3.114. *Suppose the following assumptions hold.*

1. *Let* $\mathcal{K}$ *denote the following set:*

$$\mathcal{K}\equiv\left\{(\mathbf{u}_1,\ldots,\mathbf{u}_p):\ \mathbf{u}_i\in\mathbb{R}^{n_i},\ \mathbf{u}_i\perp\text{Kernel}(S^{(i)}),\ S^{(i)}\mathbf{u}_i\perp\text{Range}(Z_i)\right\}.$$

2. *Let* $C>0$ *be as defined below:*

$$C=\sup_{(\mathbf{u}_1,\ldots,\mathbf{u}_p)\in\mathcal{K}\backslash\mathbf{0}}\frac{\sum_{i=1}^{p}|R_i\sum_{j=1}^{p}R_j^TD_i^T\mathbf{u}_j|_{S^{(i)}}^2}{\sum_{i=1}^{p}|\mathbf{u}_i|_{S^{(i)}}^2}. \tag{3.146}$$

Then, the following estimate will hold:

$$\gamma_M\leq C. \tag{3.147}$$

Proof. γ_M corresponds to the maximum of the generalized Rayleigh quotient associated with TS on the subspace Range$(I-P_0)$ in the inner product $S(.,.)$. To estimate γ_M, expand $(TS(I-P_0)\mathbf{u},S(I-P_0)\mathbf{u})$ employing (3.143). Then substitute that $S=\sum_{j=1}^{p}\mathcal{R}_j^TS^{(j)}\mathcal{R}_j$, and simplify the resulting expression:

$$\begin{aligned}
&(TS(I-P_0)\mathbf{u}, S(I-P_0)\mathbf{u}) = \textstyle\sum_{i=1}^{p} \left(S^{(i)}\mathbf{u}_i, \mathbf{u}_i\right) \\
&= \textstyle\sum_{i=1}^{p} \left(S^{(i)} S^{(i)^{\dagger}} D_i \mathcal{R}_i S(I-P_0)\mathbf{u}, \mathbf{u}_i\right) \\
&= \textstyle\sum_{i=1}^{p} \left(S(I-P_0)\mathbf{u}, \mathcal{R}_i^T D_i^T \mathbf{u}_i\right) \\
&= \textstyle\sum_{i=1}^{p} \left((\sum_{j=1}^{p} \mathcal{R}_j^T S^{(j)} \mathcal{R}_j)(I-P_0)\mathbf{u}, \mathcal{R}_i^T D_i^T \mathbf{u}_i\right) \\
&= \textstyle\sum_{i=1}^{p} \sum_{j=1}^{p} \left(S^{(j)} \mathcal{R}_j (I-P_0)\mathbf{u}, \mathcal{R}_j \mathcal{R}_i^T D_i^T \mathbf{u}_i\right) \\
&= \textstyle\sum_{j=1}^{p} \left(S^{(j)} \mathcal{R}_j (I-P_0)\mathbf{u}, \mathcal{R}_j (\sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i)\right) \\
&\leq \textstyle\sum_{j=1}^{p} \left(S^{(j)} \mathcal{R}_j (I-P_0)\mathbf{u}, \mathcal{R}_j (I-P_0)\mathbf{u}\right)^{1/2} \\
&\quad \left(S^{(j)} \mathcal{R}_j (\textstyle\sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i), \mathcal{R}_j (\sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i)\right)^{1/2} \\
&\leq \left(\textstyle\sum_{j=1}^{p} (S^{(j)} \mathcal{R}_j (I-P_0)\mathbf{u}, \mathcal{R}_j (I-P_0)\mathbf{u})\right)^{1/2} \\
&\quad \left(\textstyle\sum_{j=1}^{p} (S^{(j)} \mathcal{R}_j (\sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i), \mathcal{R}_j (\sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i))\right)^{1/2} \\
&= (S(I-P_0)\mathbf{u}, (I-P_0)\mathbf{u})^{1/2} \left(\textstyle\sum_{j=1}^{p} |\mathcal{R}_j \sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i|_{S^{(j)}}^2\right)^{1/2} \\
&\leq \left(\textstyle\sum_{i=1}^{p} |\mathbf{u}_i|_{S^{(i)}}^2\right)^{1/2} \left(\textstyle\sum_{j=1}^{p} |\mathcal{R}_j \sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i|_{S^{(j)}}^2\right)^{1/2},
\end{aligned}$$

where (3.144) was applied to obtain the last line. Canceling the common terms and squaring the resulting expression, yields:

$$(TS(I-P_0)\mathbf{u}, S(I-P_0)\mathbf{u}) \leq \sum_{j=1}^{p} |\sum_{i=1}^{p} \mathcal{R}_i^T D_i^T \mathbf{u}_i|_{S^{(j)}}^2.$$

Applying (3.146) yields $(TS(I-P_0)\mathbf{u}, S(I-P_0)\mathbf{u}) \leq C\,S((I-P_0)\mathbf{u}, (I-P_0)\mathbf{u})$, which yields an upper bound with $\gamma_M \leq C$. See [MA17]. □

By combining bound (3.141) with bounds (3.144) and (3.147), we obtain the condition number estimate $\mathrm{cond}(M, S) \leq C$, where C is defined in (3.146). Next, we shall estimate C for a finite element discretization. The following notation will be employed. Let $x_1, \ldots, x_n$ denote the nodes on B. For each glob G, let I_G denote the following $n \times n$ diagonal matrix:

$$(I_G)_{ii} = \begin{cases} 1, & \text{if } x_i \in G \\ 0, & \text{if } x_i \notin G. \end{cases}$$

By construction, it will hold that:

$$\sum_{\{G \subset \mathcal{G}\}} I_G = I. \tag{3.148}$$

On each $\partial\Omega_i$, let $y_j^{(i)}$ for $1 \leq j \leq n_i$ denote the nodes on $B^{(i)}$ in the local ordering. Given glob $G \in (\partial\Omega_i \cap \partial\Omega_j)$ define I_G^{ji} as the matrix of size $n_i \times n_j$

$$I_G^{ji} \equiv \mathcal{R}_j I_G \mathcal{R}_i^T.$$

Expressing $\mathcal{R}_j \mathcal{R}_i^T = \mathcal{R}_j I \mathcal{R}_i^T$ and substituting for I using (3.148) yields:

$$\mathcal{R}_j \mathcal{R}_i^T = \sum_{\{G \subset (\partial\Omega_i \cap \partial\Omega_j)\}} I_G^{ji}.$$

Diagonal matrix D_i of size n_i has the following representation:

$$D_i \equiv \sum_{\{G \subset \partial\Omega_i\}} d_i(G) I_G^{ii},$$

where the scalars $d_i(G)$ are defined by:

$$d_i(G) \equiv \frac{\rho_i^t}{\sum_{\{l: G \subset \partial\Omega_l\}} \rho_l^t}, \qquad \text{for } t \geq \frac{1}{2}.$$

When $G \subset (\partial\Omega_i \cap \partial\Omega_j)$, then the following will hold:

$$d_i(G) \leq \frac{\rho_i^t}{\rho_i^t + \rho_j^t}.$$

Additionally:

$$\begin{aligned} \mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i &= \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} \mathcal{R}_j I_G \mathcal{R}_i^T D_i \mathbf{u}_i \\ &= \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} I_G^{ji} D_i \mathbf{u}_i \\ &= \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} I_G^{ji} d_i(G) I_G^{ii} \mathbf{u}_i \\ &= \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} d_i(G) I_G^{ji} \mathbf{u}_i. \end{aligned} \tag{3.149}$$

Lemma 3.115. *Suppose the following assumptions hold.*

1. *Let $u_i \in V_h(\partial\Omega_i)$ denote a finite element function with associated nodal vector $\mathbf{u}_i \in \mathbb{R}^{n_i}$ on $B^{(i)}$.*
2. *Let $R > 0$ be the bound in the following discrete harmonic extension:*

$$\frac{1}{\rho_j} |I_G^{ji} \mathbf{u}_i|^2_{S^{(j)}} \leq R \frac{1}{\rho_j} |\mathbf{u}_i|^2_{S^{(i)}}, \tag{3.150}$$

for $\mathbf{u}_i \perp \text{Kernel}(S^{(i)})$ *and* $S^{(i)} \mathbf{u}_i \perp \text{Range}(Z_i)$.

For K and L defined by (3.122) and (3.121), the following estimate will hold:

$$\sup_{(\mathbf{u}_1, \ldots, \mathbf{u}_p) \neq \mathbf{0}} \frac{\sum_{i=1}^p |R_i \sum_{j=1}^p R_j^T D_i^T \mathbf{u}_j|^2_{S^{(i)}}}{\sum_{i=1}^p |\mathbf{u}_i|^2_{S^{(i)}}} \leq K^2 \, L^2 \, R,$$

yielding $\text{cond}(M, S) \leq K^2 \, L^2 \, R$.

Proof. We follow the proof in [MA14, MA17]. Apply the generalized triangle inequality to estimate the term $|\mathcal{R}_j \sum_{i=1}^p \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}}^2$ and use assumption (3.122) so that at most K terms of the form $\mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i$ will be nonzero:

$$\begin{aligned} |\textstyle\sum_{i=1}^p \mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}}^2 &\le \left(\textstyle\sum_{i=1}^p |\mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}}\right)^2 \\ &\le K \textstyle\sum_{i=1}^p |\mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}}^2. \end{aligned}$$

Summing over the indices j yields:

$$\sum_{j=1}^p |\sum_{i=1}^p \mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}}^2 \le \; K^2 \sum_{i=1}^p \max_j |\mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}}^2. \qquad (3.151)$$

By property (3.149) it holds that:

$$\mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i = \sum_{\{G : G \subset \partial\Omega_i \cap \partial\Omega_j\}} d_i(G) I_G^{ji} \mathbf{u}_i.$$

Applying the triangle inequality and employing assumption (3.150) yields:

$$\begin{aligned} |\mathcal{R}_j \mathcal{R}_i^T D_i \mathbf{u}_i|_{S^{(j)}} &\le \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} d_i(G) |I_G^{ji} \mathbf{u}_i|_{S^{(j)}} \\ &\le \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} \frac{\rho_i^t}{\rho_i^t + \rho_j^t} |I_G^{ji} \mathbf{u}_i|_{S^{(j)}} \\ &\le \textstyle\sum_{\{G \subset \partial\Omega_i \cap \partial\Omega_j\}} \frac{\rho_i^{t-\frac{1}{2}} \rho_j^{\frac{1}{2}}}{\rho_i^t + \rho_j^t} R^{1/2} |\mathbf{u}_i|_{S^{(i)}} \\ &\le L\, R^{1/2} \sup_{\rho > 0} \frac{\rho^{1/2}}{1 + \rho^t} |\mathbf{u}_i|_{S^{(i)}} \\ &\le L\, R^{1/2} \, |\mathbf{u}_i|_{S^{(i)}}. \end{aligned}$$

Substituting the above in (3.151) yields the desired result. □

The condition number of the balancing domain decomposition system now follows immediately from the preceding result.

Theorem 3.116. *If the assumptions in Lemma 3.115 hold, then the balancing domain decomposition preconditioned system will satisfy:*

$$\mathrm{cond}(M, S) \; \le \; C\, K^2\, L^2\, \left(1 + \log(h_0/h)\right)^2,$$

when $t \ge \frac{1}{2}$, *for some* $C > 0$ *independent of* h_0, h *and coefficients* $\{\rho_j\}$.

Proof. By Lemma 3.115, the condition number of the balancing domain decomposition preconditioned system will satisfy the bound:

$$\mathrm{cond}(M, S) \le K^2\, L^2\, R,$$

where the parameters K, L and R are as defined earlier. The parameters K and L are generally independent of h, h_0 and $\{\rho_j\}$, depending only on the spatial dimension and the shape regularity properties of the subdomain decomposition. Thus, we only need to estimate parameter R.

Accordingly, let u_i denote a finite element function on $\partial\Omega_i$ with associated vector of nodal values $\mathbf{u}_i \in \mathbb{R}^{n_i}$ satisfying $\mathbf{u}_i \perp \text{Kernel}(S^{(i)})$. If matrix $S^{(i)}$ is singular, then $(1,\ldots,1)^T \in \text{Kernel}(S^{(i)})$ and so a Poincaré-Freidrich's inequality of the following form will hold for u_i

$$|u_i|^2_{1/2,\partial\Omega_i} \leq C\,\frac{1}{\rho_i}|\mathbf{u}_i|^2_{S^{(i)}},$$

for some $C > 0$ independent of h_0, h and $\{\rho_j\}$. A similar Poincaré-Freidrich's inequality will hold for u_i if $S^{(i)}$ is not singular (since $c(x) = 0$ and due to zero Dirichlet values on a segment of $\partial\Omega_i$). In either case, $I_G^{ji}\mathbf{v}_i$ will correspond to the finite element function $I_G u_i$ restricted to $\partial\Omega_j$, with $\partial\Omega_i \cap \partial\Omega_j$ as the support. Applying the equivalence between the scaled Schur complement energy $\frac{1}{\rho_j}|I_G^{ji}\mathbf{u}_i|^2_{S^{(j)}}$ and the fractional Sobolev boundary energy $|I_G u_i|^2_{1/2,\partial\Omega_j}$, the equivalence between $|I_G u_i|^2_{1/2,\partial\Omega_j}$ and $|I_G u_i|^2_{1/2,\partial\Omega_i}$ (due to support of $I_G u_i$ on $\partial\Omega_i \cap \partial\Omega_j$), and subsequently applying the glob theorem, we arrive at the following estimates:

$$\begin{aligned}
\tfrac{1}{\rho_j}|I_G^{ji}\mathbf{u}_i|^2_{S^{(j)}} &= |I_G u_i|^2_{1/2,\partial\Omega_j} \\
&\leq C|I_G u_i|^2_{1/2,\partial\Omega_i} \\
&\leq C\left(1 + \log(h_0/h)\right)^2 |u_i|^2_{1/2,\partial\Omega_i} \\
&\leq C\left(1 + \log(h_0/h)\right)^2 \tfrac{1}{\rho_i}|\mathbf{u}_i|^2_{S^{(i)}},
\end{aligned}$$

where we have employed the Poincaré-Freidrich's inequality in the last line, due to the constraint $\mathbf{u}_i \perp \text{Kernel}(S^{(i)})$. Thus $R \leq C\left(1 + \log(h_0/h)\right)^2$ for some $C > 0$ independent of h_0, h and $\{\rho_i\}$. □

For additional details, the reader is referred to [MA14, MA17, KL8].

4

Lagrange Multiplier Based Substructuring: FETI Method

In this chapter, we describe the FETI method (the Finite Element Tearing and Interconnecting method) [FA2, FA16, FA15, MA25, FA14, KL8]. It is a *Lagrange multiplier* based *iterative substructuring* method for solving a finite element discretization of a self adjoint and coercive elliptic equation, based on a *non-overlapping* decomposition of its domain. In traditional substructuring, each subdomain solution is parameterized by its Dirichlet value on the boundary of the subdomain. The global solution is sought by solving a reduced Schur complement system for determining the unknown Dirichlet boundary values of each subdomain solution. By contrast, in Lagrange multiplier substructuring, each subdomain solution is parameterized by a Lagrange multiplier flux variable which represents the Neumann data of each subdomain solution on the subdomain boundary. The global solution is then sought by determining the unknown Lagrange multiplier flux variable, by solving a saddle point problem, resulting in a highly parallel algorithm with Neumann subproblems. Applications include elasticity, shell and plate problems [FA2, FA16, FA15].

Our discussion is organized as follows. Chap. 4.1 describes the *constrained minimization* problem underlying the FETI method. Given a *non-overlapping* decomposition, the FETI method employs an extended energy functional associated with the self adjoint and coercive elliptic. It is obtained by weakening the continuity of the displacements across the subdomain boundaries. The FETI method then minimizes this extended energy, subject to the constraint that the local displacements be continuous across the subdomains. Chap. 4.2 describes the Lagrange multiplier formulation associated with this constrained minimization problem. The Lagrange multiplier variables correspond to flux or Neumann data on the subdomain boundaries. Chap. 4.3 describes a projected gradient algorithm for determining the Lagrange multiplier flux variables in the FETI method. Several preconditioners are outlined. Chap. 4.4 describes the FETI-DP and BDDC variants of the FETI algorithm. Both methods are based on the PCG method with a special *coarse* space and with local problems that impose constraints on the *globs*. They yield identical convergence rates and provide advantages in parallelizability.

4.1 Constrained Minimization Formulation

The FETI method is based on a constrained minimization formulation of an elliptic equation. Given a non-overlapping decomposition of a domain, local solutions are sought on each subdomain which minimize an extended global energy, subject to the constraint that the local solutions match across the subdomain boundaries. In this section, we describe the constrained minimization formulation of a self adjoint and coercive elliptic equation, based on a non-overlapping decomposition. We also describe the finite element discretization of the elliptic equation and its constrained minimization formulation.

4.1.1 Constrained Minimization Problem: Continuous Case

Consider the following self adjoint and coercive elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad u(x) = 0, & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

The weak formulation of (4.1) seeks $u \in H_0^1(\Omega)$ satisfying:

$$\begin{cases} \mathcal{A}(u,v) = F(v), \qquad \forall v \in H_0^1(\Omega), \ \text{ where} \\ \mathcal{A}(u,v) \equiv \int_\Omega (a\,\nabla u \cdot \nabla v + c\,u\,v)\,dx \\ \quad F(v) \equiv \int_\Omega f\,v\,dx. \end{cases} \tag{4.2}$$

The minimization formulation of (4.1) seeks $u \in H_0^1(\Omega)$:

$$J(u) = \min_{v \in H_0^1(\Omega)} J(v),$$

where $J(v) \equiv \frac{1}{2}\mathcal{A}(v,v) - F(v)$. Given a non-overlapping decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, we define its internal boundary segments $B^{(l)} = \partial\Omega_l \cap \Omega$ and external boundary segments $B_{[l]} = \partial\Omega_l \cap \partial\Omega$, and common interfaces $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$. We also define the following *subdomain* forms and spaces:

$$\begin{cases} \mathcal{A}_{\Omega_l}(u_l, v_l) \equiv \int_{\Omega_l} (a\,\nabla u_l \cdot \nabla v_l + c\,u_l\,v_l)\,dx, \quad \forall u_l,\, v_l \in H^1_{B_{[l]}}(\Omega_l) \\ \quad F_{\Omega_l}(v_l) \equiv \int_{\Omega_l} f\,v_l\,dx, \quad \forall v_l \in H^1_{B_{[l]}}(\Omega_l) \ \text{ where} \\ \ H^1_{B_{[l]}}(\Omega_l) \equiv \{v \in H^1(\Omega_l) \,:\, v = 0 \ \text{ on } \ B_{[l]}\}. \end{cases}$$

Given a collection of subdomain functions $v_{\mathcal{E}} = (v_1, \ldots, v_p)$ where each local function $v_l(\cdot) \in H^1_{B_{[l]}}(\Omega_l)$, we define an *extended energy* functional $J_{\mathcal{E}}(\cdot)$ as:

$$J_{\mathcal{E}}(v_{\mathcal{E}}) = \sum_{l=1}^{p} \left(\frac{1}{2}\,\mathcal{A}_{\Omega_l}(v_l, v_l) - F_{\Omega_l}(v_l) \right). \tag{4.3}$$

By construction, if $v \in H_0^1(\Omega)$ and $v_l(\cdot) \equiv v(\cdot)$ on Ω_l for $1 \le l \le p$, then it can be verified that $J(v) = J_{\mathcal{E}}(v_{\mathcal{E}})$. Generally, $v_l \neq v_j$ need not match across common interfaces $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$, yet $J_{\mathcal{E}}(v_{\mathcal{E}})$ is well defined.

Define $\mathcal{I}^*(l) \equiv \{j : B_{lj} \neq \emptyset\}$. Note that $j \in \mathcal{I}^*(l)$ if and only if $l \in \mathcal{I}^*(j)$. For $1 \leq l \leq p$, choose $\mathcal{I}(l) \subset \mathcal{I}^*(l)$ as a subindex set, such that if $B_{lj} \neq \emptyset$, then either $j \in \mathcal{I}(l)$ or $l \in \mathcal{I}(j)$, but *not* both. Additionally, define $\mathcal{I}_*(l) \subset \mathcal{I}(l)$ as the subindex set of interface segments of dimension $(d-1)$ when $\Omega \subset \mathbb{R}^d$. Heuristically, we define the following constraint set of local functions:

$$\mathcal{V}_0 \equiv \{v_{\mathcal{E}} : v_l = v_j \quad \text{on} \quad B_{lj} \quad \text{if} \quad j \in \mathcal{I}(l),\ 1 \leq l \leq p\}.$$

Then $\mathcal{V}_0$ will consist of local functions which match across subdomains. Heuristically, we expect that minimizing $J_{\mathcal{E}}(v_{\mathcal{E}})$ in $\mathcal{V}_0$ will yield:

$$J_{\mathcal{E}}(u_{\mathcal{E}}) = \min_{v_{\mathcal{E}} \in \mathcal{V}_0} J_{\mathcal{E}}(v_{\mathcal{E}}) \tag{4.4}$$

where $u_{\mathcal{E}} = (u_1, \ldots, u_p)$ will satisfy $u_l = u$ on Ω_l for $1 \leq l \leq p$, for the desired solution $u(.)$. The FETI method employs a Lagrange multiplier formulation of a discrete version of this problem, and iteratively solves the resulting saddle point system using a preconditioned projected gradient method.

4.1.2 Constrained Minimization Problem: Discrete Case

Let $\mathcal{T}_h(\Omega)$ denote a quasiuniform triangulation of Ω with n nodes in Ω. Let V_h denote a space of finite element functions on the triangulation $\mathcal{T}_h(\Omega)$ of Ω. A finite element discretization of (4.1) will seek $u_h \in V_h \cap H_0^1(\Omega)$ such that:

$$\mathcal{A}(u_h, v_h) = F(v_h), \qquad \forall v_h \in V_h \cap H_0^1(\Omega).$$

If $\{\phi_1, \ldots, \phi_n\}$ denotes a nodal basis for $V_h \cap H_0^1(\Omega)$, then, the resulting discretization will yield the linear system:

$$A\,\mathbf{u} = \mathbf{f}, \tag{4.5}$$

where A is symmetric positive definite with entries $A_{ij} = \mathcal{A}(\phi_i, \phi_j)$, and $\mathbf{u}$ denotes the displacement vector with $u_h(x) = \sum_{i=1}^{n} (\mathbf{u})_i \phi_i(x)$ and $\mathbf{f}$ denotes the load vector, with $(\mathbf{f})_i = F(\phi_i)$, for a chosen ordering of the nodes.

Given a nonoverlapping decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, as in Fig. 4.1, we shall block partition the nodal unknowns on each subdomain as follows. Nodes in Ω_i will be regarded as “interior” nodes in Ω_i, while nodes on $B^{(i)}$

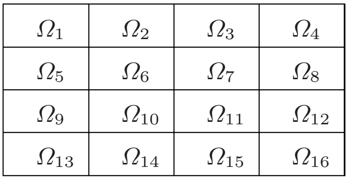


Fig. 4.1. A non-overlapping decomposition

as "subdomain boundary" nodes. The common interface will be denoted as $B = \cup_{i=1}^{p} B^{(i)}$, and the number of nodes in Ω_i and $B^{(i)}$ will be denoted as $n_I^{(i)}$ and $n_B^{(i)}$, respectively, with $n_i \equiv (n_I^{(i)} + n_B^{(i)})$. We shall denote by $\mathbf{u}_I^{(i)} \in \mathbb{R}^{n_I^{(i)}}$ and $\mathbf{u}_B^{(i)} \in \mathbb{R}^{n_B^{(i)}}$ vectors of finite element nodal values on Ω_i and $B^{(i)}$, respectively, for the chosen local ordering of the nodes. The local displacements, local stiffness matrices and load vectors will be denoted as:

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{u}_I^{(i)} \\ \mathbf{u}_B^{(i)} \end{bmatrix}, \quad A^{(i)} = \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix}, \quad \mathbf{f}_i = \begin{bmatrix} \mathbf{f}_I^{(i)} \\ \mathbf{f}_B^{(i)} \end{bmatrix}, \quad \text{for } 1 \le i \le p, \tag{4.6}$$

given a local ordering of the nodes. We shall denote by R_i the restriction map which maps a nodal vector $\mathbf{u} \in \mathbb{R}^n$ of nodal values on Ω onto its subvector $\mathbf{u}_i = R_i \mathbf{u}$ of size n_i of nodal values on $\Omega_i \cup B^{(i)}$. Its transpose R_i^T will extend by *zero* a nodal vector on $\Omega_i \cup B^{(i)}$ to the rest of Ω. Decomposing $\mathcal{A}(.,.)$ and $F(.)$ based on the subdomains, will yield the subassembly identity (3.10):

$$\begin{cases} A = \sum_{i=1}^{p} R_i^T A^{(i)} R_i \\ \mathbf{f} = \sum_{i=1}^{p} R_i^T \mathbf{f}_i, \end{cases} \tag{4.7}$$

relating the local and global stiffness matrices and load vectors.

When coefficient $c(x) = 0$ in (4.1) and Ω_i is *floating*, i.e., $\overline{\Omega_i} \subset \Omega$, then the local stiffness matrix $A^{(i)}$ will be singular with $\mathbf{1} \equiv (1, \ldots, 1)^T$ spanning Kernel $\left(A^{(i)}\right)$. For discretizations of more general elliptic equations, such as the equations of linear elasticity, Kernel $\left(A^{(i)}\right)$ may have dimension d_i up to six (for $\Omega \subset \mathbb{R}^3$). When matrix $A^{(i)}$ is singular, we shall let $Z^{(i)}$ denote an $n_i \times d_i$ matrix whose columns form a basis for the kernel of $A^{(i)}$:

$$\text{Range}(Z^{(i)}) = \text{Kernel}(A^{(i)}). \tag{4.8}$$

When $A^{(i)}$ is nonsingular, we define $Z^{(i)} = \mathbf{0}$ and set $d_i = 0$. The FETI algorithm solves (4.5) by a constrained minimization reformulation of (4.5). The next result describes a minimization problem equivalent to (4.5).

Lemma 4.1. *Suppose $A = A^T > 0$ and let $\mathbf{u}$ solve the linear system (4.5). Then $\mathbf{u}$ will minimize the associated energy functional:*

$$\mathbf{J}(\mathbf{u}) = \min_{\mathbf{v} \in \mathbb{R}^n} \mathbf{J}(\mathbf{v}), \quad \text{where } \mathbf{J}(\mathbf{v}) \equiv \frac{1}{2} \mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}, \quad \textit{for } \mathbf{v} \in \mathbb{R}^n. \tag{4.9}$$

Proof. At the critical point of $\mathbf{J}(\cdot)$, we obtain $\mathbf{0} = \nabla \mathbf{J}(\mathbf{u}) = A\mathbf{u} - \mathbf{f}$. Since $A = A^T > 0$, the critical point $\mathbf{u}$ will correspond to a minimum. □

The constrained minimization problem employed in the FETI method is obtained by *weakening* the requirement that the subdomain finite element functions be continuous across the interface B, and by subsequently enforcing continuity across B as a *constraint.* In terms of nodal vectors, each *local* displacement vector $\mathbf{v}_i = (\mathbf{v}_I^{(i)^T}, \mathbf{v}_B^{(i)^T})^T$ in an extended *global displacement* $\mathbf{v}_\mathcal{E} \equiv (\mathbf{v}_1^T, \ldots, \mathbf{v}_p^T)^T$ of size $n_\mathcal{E} = (n_1 + \cdots + n_p)$, need not match with adjacent displacements on $B^{(i)} \cap B^{(j)}$. To determine an extended displacement $\mathbf{v}_\mathcal{E}$ whose components *match* on the interface B, *constraints* are imposed on $\mathbf{v}_\mathcal{E}$ and an extended energy functional is minimized subject to these constraints.

The FETI method also employs extended loads $\mathbf{f}_\mathcal{E} \equiv (\mathbf{f}_1^T, \ldots, \mathbf{f}_p^T)^T$, where $\mathbf{f}_i = (\mathbf{f}_I^{(i)^T}, \mathbf{f}_B^{(i)^T})^T$ denote local loads, and an extended block diagonal stiffness matrix $A_{\mathcal{E}\mathcal{E}} \equiv \text{blockdiag}\left(A^{(1)}, \ldots, A^{(p)}\right)$ of size $n_\mathcal{E}$, based on the local stiffness matrices $A^{(i)}$. By construction, the extended stiffness matrices, displacement and load vectors will have the following block structure:

$$A_{\mathcal{E}\mathcal{E}} \equiv \begin{bmatrix} A^{(1)} & & 0 \\ & \ddots & \\ 0 & & A^{(p)} \end{bmatrix}, \quad \mathbf{v}_\mathcal{E} \equiv \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_p \end{bmatrix}, \quad \mathbf{f}_\mathcal{E} \equiv \begin{bmatrix} \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_p \end{bmatrix}. \tag{4.10}$$

Given matrices $Z^{(i)}$ of size $n_i \times \min\{1, d_i\}$ whose columns span the null space of $A^{(i)}$ with $\text{Range}(Z^{(i)}) = \text{Kernel}(A^{(i)})$, a block matrix Z of size $n_\mathcal{E} \times d$:

$$Z \equiv \begin{bmatrix} Z^{(1)} & & 0 \\ & \ddots & \\ 0 & & Z^{(p)} \end{bmatrix} \tag{4.11}$$

will also be employed, where $d = \min\{1, d_1\} + \cdots + \min\{1, d_p\}$.

In the following, we introduce the *extended* energy functional $\mathbf{J}_\mathcal{E}(\mathbf{w}_\mathcal{E})$ that corresponds to the sum of the local displacement energies.

Lemma 4.2. *Suppose the following assumptions hold for* $\mathbf{v} \in \mathbb{R}^n$*:*

1. *Define* $\mathbf{w}_i = R_i \mathbf{v} \in \mathbb{R}^{n_i}$ *and* $\mathbf{w}_\mathcal{E} = \left(\mathbf{w}_1^T, \ldots, \mathbf{w}_p^T\right)^T \in \mathbb{R}^{n_\mathcal{E}}$.
2. *Given local load vectors* $\mathbf{f}_i = \left(\mathbf{f}_I^{(i)^T}, \mathbf{f}_B^{(i)^T}\right)^T \in \mathbb{R}^{n_i}$ *define:*

$$\mathbf{f} = \sum_{i=1}^{p} R_i^T \mathbf{f}_i \in \mathbb{R}^n \quad \textit{and} \quad \mathbf{f}_\mathcal{E} = \left(\mathbf{f}_1^T, \ldots, \mathbf{f}_p^T\right)^T \in \mathbb{R}^{n_\mathcal{E}}.$$

3. *Let* $\mathbf{J}_\mathcal{E}(\mathbf{w}_\mathcal{E})$ *denote the following extended energy functional:*

$$\mathbf{J}_\mathcal{E}(\mathbf{w}_\mathcal{E}) \equiv \frac{1}{2}\mathbf{w}_\mathcal{E}^T A_{\mathcal{E}\mathcal{E}} \mathbf{w}_\mathcal{E} - \mathbf{w}_\mathcal{E}^T \mathbf{f}_\mathcal{E}. \tag{4.12}$$

Then, it will hold that $\mathbf{J}(\mathbf{v}) = \mathbf{J}_\mathcal{E}(\mathbf{w}_\mathcal{E})$, *for* $\mathbf{J}(\mathbf{v})$ *defined by (4.9).*

Proof. The subassembly identity (4.7) for the stiffness matrix yields:

$$\mathbf{v}^T A \mathbf{v} = \sum_{i=1}^{p} \mathbf{v}^T R_i^T A^{(i)} R_i \mathbf{v} = \mathbf{w}_\mathcal{E}^T A_{\mathcal{E}\mathcal{E}} \mathbf{w}_\mathcal{E},$$

since $\mathbf{w}_i = R_i\mathbf{v}$. The subassembly identity for load vectors yields:

$$\mathbf{v}^T\mathbf{f} = \mathbf{v}^T\left(\sum_{i=1}^{p} R_i^T\mathbf{f}_i\right) = \sum_{i=1}^{p}(R_i\mathbf{v})^T\mathbf{f}_i = \sum_{i=1}^{p}\mathbf{w}_i^T\mathbf{f}_i = \mathbf{w}_{\mathcal{E}}^T\mathbf{f}.$$

Substituting these into $\mathbf{J}(\mathbf{v})$ and $\mathbf{J}_{\mathcal{E}}(\mathbf{w}_{\mathcal{E}})$ yields the desired result. □

Remark 4.3. When $\mathbf{f} = \sum_{i=1}^{p} R_i^T\mathbf{f}_i$, the above equivalence between $\mathbf{J}(\mathbf{v})$ and $\mathbf{J}_{\mathcal{E}}(\mathbf{w}_{\mathcal{E}})$ will hold *only* when the constraints $\mathbf{w}_i = R_i\mathbf{v}$ for $1 \le i \le p$ are satisfied. By construction, the parametric representation $\mathbf{w}_i = R_i\mathbf{v}$ in terms of $\mathbf{v} \in \mathbb{R}^n$ ensures that the nodal values of $\mathbf{w}_B^{(i)}$ match with those of $\mathbf{w}_B^{(j)}$ for nodes on $B^{(i)} \cap B^{(j)}$, corresponding to nodal values of $\mathbf{v}$ at the specific nodes.

A constrained minimization formulation of (4.5) can now be obtained, provided a matrix M can be constructed, such that $\mathbf{w}_i = R_i\mathbf{v}$ for $1 \le i \le p$ if and only if $M\mathbf{w}_{\mathcal{E}} = \mathbf{0}$ for $\mathbf{w}_{\mathcal{E}} = \left(\mathbf{w}_1^T, \ldots, \mathbf{w}_p^T\right)^T$:

$$\mathcal{V}_0 \equiv \left\{\left((R_1\mathbf{v})^T, \ldots, (R_p\mathbf{v})^T\right)^T : \mathbf{v} \in \mathbb{R}^n\right\} = \{\mathbf{w}_{\mathcal{E}} \in \mathbb{R}^{n_{\mathcal{E}}} : M\mathbf{w}_{\mathcal{E}} = \mathbf{0}\}. \tag{4.13}$$

Here M will be a matrix of size $m \times n_{\mathcal{E}}$. When matrix M can be constructed, the minimization problem (4.9) can be expressed as a constrained minimization of the extended energy functional $\mathbf{J}_{\mathcal{E}}(\mathbf{w}_{\mathcal{E}})$ within the constraint set $\mathcal{V}_0$.

Lemma 4.4. *Suppose the following assumptions hold.*

1. *Let* $\mathbf{f} = \sum_{i=1}^{p} R_i^T\mathbf{f}_i$ *and* $\mathbf{f}_{\mathcal{E}} = \left(\mathbf{f}_1^T, \ldots, \mathbf{f}_p^T\right)^T$.
2. *Let* $\mathbf{u}$ *denote the minimum of (4.9).*
3. *Let* $\mathcal{V}_0$ *and matrix* M *of size* $m \times n_{\mathcal{E}}$ *be as in (4.13).*
4. *Let* $\mathbf{w}_{\mathcal{E}} = \left(\mathbf{w}_1^T, \ldots, \mathbf{w}_p^T\right)^T$ *denote the constrained minimum:*

$$\mathbf{J}_{\mathcal{E}}(\mathbf{w}_{\mathcal{E}}) = \min_{\mathbf{v}_{\mathcal{E}} \in \mathcal{V}_0} \mathbf{J}_{\mathcal{E}}(\mathbf{v}_{\mathcal{E}}). \tag{4.14}$$

Then, the following results will hold:

$$\mathbf{w}_i = R_i\mathbf{u}, \qquad \textit{for } i = 1, \ldots, p.$$

Proof. By definition of subspace $\mathcal{V}_0$, the following parametric representation $\mathbf{w}_i = R_i\mathbf{v}$ will hold for $1 \le i \le p$ and for some $\mathbf{v} \in \mathbb{R}^n$. An application of the preceding lemma will yield the desired result. □

Construction of Matrix *M*. We shall now describe how to construct a matrix M so that the representation $\mathcal{V}_0 = \text{Kernel}(M)$ holds in (4.13). The matrix M will be chosen so that the equation $M\mathbf{w}_{\mathcal{E}} = \mathbf{0}$ enforces each admissible pair of local displacement vectors $\mathbf{w}_B^{(i)}$ and $\mathbf{w}_B^{(j)}$ to match on the nodes in $B^{(i)} \cap B^{(j)}$. We let n_B denote the number of nodes on $B = \cup_{i=1}^{p} B^{(i)}$.

Definition 4.5. *Given nodes $x_1, \ldots, x_{n_B}$ on interface B, we define:*

$$\begin{cases} \mathcal{W}\left(x_i\right) \equiv \{j : x_i \in \partial\Omega_j\} \\ \text{degree}\left(x_i\right) \equiv \left|\mathcal{W}\left(x_i\right)\right| \\ \text{index}\left(x_l, B^{(j)}\right) \equiv \textit{ local index of } x_l \text{ in } B^{(j)}. \end{cases} \tag{4.15}$$

Here $\mathcal{W}\left(x_i\right)$ denotes the indices of all subdomains whose boundaries contain x_i, and the degree of a node x_i denotes the number of distinct subdomain boundaries to which it belongs.

There is much arbitrariness in the choice of matrix M. Each row of matrix M must be chosen to enforce a constraint which matches two nodal values. Each node $x_i \in B$ will belong to degree (x_i) distinct subdomain boundaries. In principle, we may require matching of nodal values of $\mathbf{v}_l$ and $\mathbf{v}_j$ at node x_i for each pair of indices $l, j \in \mathcal{W}\left(x_i\right)$. This can be done by requiring that the difference between the nodal value of $\mathbf{v}_l$ and $\mathbf{v}_j$ be *zero* at node x_i, for each pair of indices $l, j \in \mathcal{W}\left(x_i\right)$. However, this will typically yield redundant equations when degree $(x_i) \geq 3$. In practice, it will be sufficient to select a subset of linearly dependent constraints so that all such matching conditions can be derived from the selected few constraints. We describe two alternate choices of matrix M (not necessarily full rank), having the block structure:

$$M = \left[\, M^{(1)} \cdots \ M^{(p)} \,\right], \tag{4.16}$$

so that $M\mathbf{v}_{\mathcal{E}} = M^{(1)}\mathbf{v}_1 + \cdots + M^{(p)}\mathbf{v}_p$ where $M^{(i)}$ is of size $m \times n_i$. Since each $\mathbf{v}_i = \left(\mathbf{v}_I^{(i)^T}, \mathbf{v}_B^{(i)^T}\right)^T$ corresponds to interior and boundary nodal values, each $M^{(i)}$ may further be partitioned as:

$$M^{(i)} = [M_I^{(i)} \;\; M_B^{(i)}] = [0 \;\; M_B^{(i)}]. \tag{4.17}$$

The submatrix $M_I^{(i)}$ will be *zero* since the matching of *boundary* values does not involve *interior* nodal values. There is arbitrariness in the choice of entries of M. The matrices we shall construct will have their entries M_{ij} chosen from $\{-1, 0, +1\}$, selected based on the following observations. Corresponding to each node $x_i \in B$, there will be $\frac{1}{2}$ degree(x_i) (degree$(x_i) - 1$) distinct pairs of subdomains which contain node x_i. For each $l, j \in \mathcal{W}(x_i)$ we will require that the difference of the entries of $\mathbf{v}_l$ and $\mathbf{v}_j$ be zero at x_i.

Specifically, if $l, j \in \mathcal{W}(x_i)$ let $\tilde{l}_i = \text{index}(x_i, B^{(l)})$ and $\tilde{j}_i = \text{index}(x_i, B^{(j)})$. Then the continuity of $\mathbf{v}_l$ and $\mathbf{v}_j$ at node x_i can be enforced as follows:

$$\left(\mathbf{v}_B^{(l)}\right)_{\tilde{l}_i} - \left(\mathbf{v}_B^{(j)}\right)_{\tilde{j}_i} = 0, \qquad \text{if } \; l, j \in \mathcal{W}(x_i). \tag{4.18}$$

This will yield entries of M to be from $\{-1, 0, +1\}$. By convention, we shall require $l < j$, and in the following, describe two different choices of matrices M depending on how many index pairs l, j are selected from $\mathcal{W}(x_i)$.

Choice 1. For each node x_i, arrange all the indices in $\mathcal{W}(x_i)$ in increasing order. For each consecutive pair of such indices, impose one constraint, yielding a total of degree$(x_i) - 1$ constraints corresponding to node x_i. In this case, the constraints will not be redundant, and the total number m of constraints:

$$m = \sum_{i=1}^{n_B} (\text{degree}(x_i) - 1) \, .$$

By construction, all such constraints will be linearly independent, and matrix M will be of *full rank* (with rank equal to m). The actual entries of matrix M will depend on the ordering of the constraints used. If $l < j$ are consecutive indices in $\mathcal{W}(x_i)$, let $k(i, l, j)$ denote the numbering (between 1 and m) assigned to the constraint involving node x_i and subvectors $\mathbf{v}_l$ and $\mathbf{v}_j$. Then, for each such node x_i and consecutive indices $l < j$ from $\mathcal{W}(x_i)$ define the entries of M as:

$$\begin{cases} \left(M_B^{(l)}\right)_{k,r} = 1, & \text{if } r = \tilde{l}_i \\ \left(M_B^{(l)}\right)_{k,r} = 0, & \text{if } r \neq \tilde{l}_i \\ \left(M_B^{(j)}\right)_{k,r} = -1, & \text{if } r = \tilde{j}_i \\ \left(M_B^{(l)}\right)_{k,r} = 0, & \text{if } r \neq \tilde{j}_i . \end{cases} \tag{4.19}$$

All other entries in the k'th row of M are defined to be zero.

Choice 2. An alternative choice of matrix M may be obtained as follows. For each node x_i, impose one constraint corresponding to each distinct pair $l < j$ of indices in $\mathcal{W}(x_i)$. Since there are degree(x_i) such indices, there will be $\frac{1}{2}$ degree(x_i) (degree$(x_i) - 1$) such constraints, so that:

$$m = \sum_{i=1}^{n_B} \frac{1}{2} \, \text{degree}(x_i) \; (\text{degree}(x_i) - 1) \, .$$

In this case, several of the constraints will be *redundant* if degree$(x_i) \geq 3$. Consequently, matrix M will *not* be of full rank if degree$(x_i) \geq 3$ for at least one node x_i. The entries of matrix M can be defined as in (4.19), noting that $l, j \in \mathcal{W}(x_i)$ need not be consecutive indices.

Choice 1 for M is easier to analyze than choice 2, due to it being of full rank. However, choice 2 is preferable for parallel implementation [FA14]. In both cases, however, the constraint set $\mathcal{V}_0$ will satisfy:

$$\mathcal{V}_0 = \left\{ \left((R_1 \mathbf{v})^T, \ldots, (R_p \mathbf{v})^T \right)^T : \mathbf{v} \in \mathbb{R}^n \right\} = \text{Kernel}\,(M) \, ,$$

as may be verified by the reader.

Remark 4.6. For a two subdomain decomposition, all nodes on B will have degree two. Consequently, choices 1 and 2 will coincide. In this case, matrix M will have the following block structure with $M_B^{(1)} = I$ and $M_B^{(2)} = -I$:

$$M = \begin{bmatrix} 0 & I & 0 & -I \end{bmatrix} ,$$

provided all nodes on B are ordered identically in both subdomains.

4.2 Lagrange Multiplier Formulation

To determine the solution to the constrained minimization problem (4.14), the FETI method reformulates (4.14) as a saddle point problem (saddle point or Lagrange multiplier methodology is described in [CI4, GI3] and Chap. 10). It introduces new variables, referred to as Lagrange multipliers, one for each constraint. Further, it associates a function, referred to as the Lagrangian function, whose saddle point (a critical point which is neither a local maximum nor a local minimum) yields the constrained minimum from its components. At the saddle point of the Lagrangian function, its gradient with respect to the original and Lagrange multiplier variables will be zero, and the resulting system of equations can be solved to determine the constrained minimum. In the following result, we describe the saddle point system associated with the constrained minimization problem (4.14).

Lemma 4.7. *Suppose the following assumptions hold.*

1. *Let* $\mathbf{u}_{\mathcal{E}} = \left(\mathbf{u}_1^T, \cdots, \mathbf{u}_p^T\right)^T \in \mathbb{R}^{n_{\mathcal{E}}}$ *denote the solution of:*
$$\mathbf{J}_{\mathcal{E}}(\mathbf{u}_{\mathcal{E}}) = \min_{\mathbf{w}_{\mathcal{E}} \in \mathcal{V}_0} \mathbf{J}_{\mathcal{E}}(\mathbf{w}_{\mathcal{E}}) \tag{4.20}$$
where
$$\mathcal{V}_0 \equiv \{\mathbf{w}_{\mathcal{E}} \in \mathbb{R}^{n_{\mathcal{E}}} : M\mathbf{w}_{\mathcal{E}} = \mathbf{0}\}. \tag{4.21}$$
2. *Let* M *be a matrix of size* $m \times n_{\mathcal{E}}$ *of full rank* m.

Then, there will exist a vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ *such that:*

$$\begin{bmatrix} A_{\mathcal{E}\mathcal{E}} & M^T \\ M & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\mathcal{E}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{\mathcal{E}} \\ \mathbf{0} \end{bmatrix}. \tag{4.22}$$

Proof. To verify the first block row of (4.22), for each choice of nonzero vector $\mathbf{v}_{\mathcal{E}} \in \mathcal{V}_0$ consider the line $\mathbf{x}(t) = \mathbf{u}_{\mathcal{E}} + t\,\mathbf{v}_{\mathcal{E}} \in \mathcal{V}_0$ for $t \in \mathbb{R}$. By construction, it passes through $\mathbf{u}_{\mathcal{E}}$ when $t = 0$ with:

$$\left.\frac{d\mathbf{x}(t)}{dt}\right|_{t=0} = \mathbf{v}_{\mathcal{E}}.$$

Since $\mathbf{u}_{\mathcal{E}}$ corresponds to the minimum of $\mathbf{J}_{\mathcal{E}}(\cdot)$ in $\mathcal{V}_0$, and since $\mathbf{x}(t) \subset \mathcal{V}_0$ with $\mathbf{x}(0) = \mathbf{u}_{\mathcal{E}}$ the function $\mathbf{J}_{\mathcal{E}}(\mathbf{x}(t))$ will attain a minimum along the line at $t = 0$. Applying the derivative test yields:

$$\begin{aligned} \left.\tfrac{d\mathbf{J}_{\mathcal{E}}(\mathbf{x}(t))}{dt}\right|_{t=0} = 0, \;\; \forall \mathbf{v}_{\mathcal{E}} \in \mathcal{V}_0 &\Leftrightarrow \nabla \mathbf{J}_{\mathcal{E}}(\mathbf{u}_{\mathcal{E}}) \cdot \mathbf{v}_{\mathcal{E}} = 0, \;\; \forall \mathbf{v}_{\mathcal{E}} \in \mathcal{V}_0 \\ &\Leftrightarrow \nabla \mathbf{J}_{\mathcal{E}}(\mathbf{u}_{\mathcal{E}}) \perp \mathcal{V}_0 \\ &\Leftrightarrow \nabla \mathbf{J}_{\mathcal{E}}(\mathbf{u}_{\mathcal{E}}) \in \text{Kernel}(M)^{\perp} \\ &\Leftrightarrow \nabla \mathbf{J}_{\mathcal{E}}(\mathbf{u}_{\mathcal{E}}) \in \text{Range}(M^T). \end{aligned}$$

We may represent any vector in Range(M^T) in the form $-M^T\boldsymbol{\lambda}$ for $\boldsymbol{\lambda} \in \mathbb{R}^m$.

Choosing $-M^T\boldsymbol{\lambda}$ (the negative sign here is for convenience), we obtain:

$$A_{\mathcal{E}\mathcal{E}}\mathbf{u}_{\mathcal{E}} - \mathbf{f}_{\mathcal{E}} = \nabla \mathbf{J}_{\mathcal{E}}\left(\mathbf{u}_{\mathcal{E}}\right) = -M^T\boldsymbol{\lambda}, \qquad \text{for some } \boldsymbol{\lambda} \in \mathbb{R}^m,$$

which yields the first block row of (4.22). To verify the second block row of (4.22), note that since $\mathbf{u}_{\mathcal{E}} \in \mathcal{V}_0$, we obtain $M\,\mathbf{u}_{\mathcal{E}} = \mathbf{0}$. □

Remark 4.8. Each λ_i in $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$ is referred to as a Lagrange multiplier. There will be m Lagrange multipliers, one corresponding to each row of M which enforces one of the m constraints. Since each λ_i is a dual variable to the Dirichlet data (see Chap. 11.2), it will represent an inter-subdomain flux.

Remark 4.9. To ensure solvability of (4.22), it is sufficient to require that M is an $m \times n_{\mathcal{E}}$ matrix of full rank m, and to require that matrix $A_{\mathcal{E}\mathcal{E}}^T = A_{\mathcal{E}\mathcal{E}} \geq 0$ be *coercive* on the null space $\mathcal{V}_0$ of M. This latter requirement can be equivalently stated as $\text{Kernel}(M) \cap \text{Kernel}(A_{\mathcal{E}\mathcal{E}}) = \{\mathbf{0}\}$. When M is not of full rank, $\boldsymbol{\lambda}$ will not be uniquely determined.

Given $\boldsymbol{\mu} \in \mathbb{R}^m$ of Lagrange multipliers, we associate a *Lagrangian function* $\mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\mu})$ with the constrained minimization problem (4.20):

$$\begin{cases} \mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\mu}) \equiv \mathbf{J}_{\mathcal{E}}\left(\mathbf{v}_{\mathcal{E}}\right) + \boldsymbol{\mu}^T M \mathbf{v}_{\mathcal{E}} \\ \qquad\qquad\quad = \frac{1}{2}\mathbf{v}_{\mathcal{E}}^T A_{\mathcal{E}\mathcal{E}} \mathbf{v}_{\mathcal{E}} - \mathbf{v}_{\mathcal{E}}^T \mathbf{f}_{\mathcal{E}} + \boldsymbol{\mu}^T M \mathbf{v}_{\mathcal{E}}. \end{cases} \tag{4.23}$$

By construction, the derivative test for the critical point of $\mathcal{L}(\cdot, \cdot)$ yields (4.22). We shall associate the following *dual* function with the Lagrangian function.

Definition 4.10. *For $\boldsymbol{\mu} \in \mathbb{R}^m$ define the dual function $D(\boldsymbol{\mu})$:*

$$D(\boldsymbol{\mu}) \equiv \inf_{\mathbf{v}_{\mathcal{E}}} \mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\mu}).$$

Remark 4.11. Since matrix $A_{\mathcal{E}\mathcal{E}}$ may be singular, the above infimum could be $-\infty$ if $(\mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\mu}) \notin \text{Range}(A_{\mathcal{E}\mathcal{E}})$. Recall that Z of rank d satisfies:

$$\text{Range}(Z) = \text{Kernel}(A_{\mathcal{E}\mathcal{E}}).$$

Using Z, we may define the class $\mathcal{G}$ of *admissible* Lagrange multipliers as:

$$\mathcal{G} \equiv \{\boldsymbol{\mu} : Z^T(\mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\mu}) = \mathbf{0}\}.$$

By definition, if $\boldsymbol{\mu} \in \mathcal{G}$, then $D(\boldsymbol{\mu}) > -\infty$.

Definition 4.12. *For $\mathbf{v}_{\mathcal{E}} \in \mathbb{R}^{n_{\mathcal{E}}}$ define a function $E(\mathbf{v}_{\mathcal{E}})$:*

$$E(\mathbf{v}_{\mathcal{E}}) \equiv \sup_{\boldsymbol{\mu}} \mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\mu}).$$

Remark 4.13. It is easily verified that since $\mathcal{L}(\cdot,\cdot)$ is linear in $\boldsymbol{\mu}$:

$$E(\mathbf{v}_{\mathcal{E}}) = \begin{cases} +\infty, & \text{if } M\mathbf{v}_{\mathcal{E}} \neq \mathbf{0} \\ \mathbf{J}_{\mathcal{E}}\left(\mathbf{v}_{\mathcal{E}}\right), & \text{if } M\mathbf{v}_{\mathcal{E}} = \mathbf{0}. \end{cases}$$

So we define a class of *admissible* displacements $\mathbf{v}_{\mathcal{E}}$ as $\mathcal{V}_0$:

$$\mathcal{V}_0 \equiv \left\{\mathbf{v}_{\mathcal{E}} : M\mathbf{v}_{\mathcal{E}} = \mathbf{0}\right\}.$$

By definition, if $\mathbf{v}_{\mathcal{E}} \in \mathcal{V}_0$ then we will have $E(\mathbf{v}_{\mathcal{E}}) = \mathbf{J}_{\mathcal{E}}\left(\mathbf{v}_{\mathcal{E}}\right) < \infty$.

The term “saddle point” is motivated by the following property.

Definition 4.14. *We say that* $(\mathbf{u}_{\mathcal{E}}, \boldsymbol{\lambda})$ *is a saddle point of the Lagrangian functional* $\mathcal{L}(.,.)$ *if the following conditions are satisfied:*

$$\mathcal{L}(\mathbf{u}_{\mathcal{E}}, \boldsymbol{\mu}) \le E(\mathbf{u}_{\mathcal{E}}) = \mathcal{L}(\mathbf{u}_{\mathcal{E}}, \boldsymbol{\lambda}) = D(\boldsymbol{\lambda}) \le \mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\lambda}), \quad \forall \mathbf{v}_{\mathcal{E}},\, \boldsymbol{\mu}.$$

Remark 4.15. Thus, the saddle point $(\mathbf{u}_{\mathcal{E}}, \boldsymbol{\lambda})$ corresponds to a *minimum* of $\mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\lambda})$ as $\mathbf{v}_{\mathcal{E}}$ is varied, and to a *maximum* of $\mathcal{L}(\mathbf{u}_{\mathcal{E}}, \boldsymbol{\mu})$ as $\boldsymbol{\mu}$ is varied. As mentioned before, the first order derivative test (differentiation with respect to $\mathbf{v}_{\mathcal{E}}$ and $\boldsymbol{\mu}$) for a critical point of $\mathcal{L}(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\mu})$ at $(\mathbf{u}_{\mathcal{E}}, \boldsymbol{\lambda})$ yields system (4.22).

In the next section, we describe an algorithm for determining $\mathbf{u}_{\mathcal{E}}$ and $\boldsymbol{\lambda}$.

4.3 Projected Gradient Algorithm

In this section, following [FA14] we describe an iterative algorithm for obtaining the solution $\mathbf{u}_{\mathcal{E}}$ and $\boldsymbol{\lambda}$ to saddle point system (4.22). Since matrix $A_{\mathcal{E}\mathcal{E}}$ may be *singular*, traditional saddle point iterative algorithms from Chap. 10 need to be modified, and we discuss these modifications [FA15, FA14]. We assume that if $A_{\mathcal{E}\mathcal{E}}$ is singular, that Z has rank d. We define $G \equiv MZ$ as a matrix of size $m \times d$. Due to the block structure of matrices M and Z, we obtain:

$$G = MZ = \left[\, M^{(1)}Z^{(1)} \quad \cdots \quad M^{(p)}Z^{(p)} \,\right]. \tag{4.24}$$

When local stiffness matrix $A^{(i)}$ is nonsingular, $Z^{(i)} = \mathbf{0}$ and $M^{(i)}Z^{(i)} = \mathbf{0}$. The next result describes a system for determining $\boldsymbol{\lambda}$, and subsequently $\mathbf{u}_{\mathcal{E}}$.

Lemma 4.16. *Suppose the following assumptions hold.*

1. *Let* $\left(\mathbf{u}_{\mathcal{E}}^T, \boldsymbol{\lambda}^T\right)^T$ *denote the solution to the saddle point system (4.22):*

$$\begin{cases} A_{\mathcal{E}\mathcal{E}}\mathbf{u}_{\mathcal{E}} + M^T\boldsymbol{\lambda} = \mathbf{f}_{\mathcal{E}} \\ \qquad\qquad\; M\mathbf{u}_{\mathcal{E}} = \mathbf{0}. \end{cases} \tag{4.25}$$

Then, the following results will hold for $G = M\,Z$ defined by (4.24):

1. *The Lagrange multiplier $\boldsymbol{\lambda}$ will solve the following reduced system:*
$$\begin{cases} P_0\,K\,\boldsymbol{\lambda} = P_0\,\mathbf{e} \\ \quad G^T\,\boldsymbol{\lambda} = \mathbf{g}, \end{cases} \tag{4.26}$$
where $P_0 \equiv I - G(G^TG)^{\dagger}G^T$, $K \equiv MA_{\mathcal{E}\mathcal{E}}^{\dagger}M^T$, $\mathbf{e} \equiv MA_{\mathcal{E}\mathcal{E}}^{\dagger}\mathbf{f}_{\mathcal{E}}$, $\mathbf{g} \equiv Z^T\mathbf{f}_{\mathcal{E}}$, and $A_{\mathcal{E}\mathcal{E}}^{\dagger}$ and $(G^TG)^{\dagger}$ denote Moore-Penrose pseudoinverses.
2. *Given $\boldsymbol{\lambda}$, the displacement $\mathbf{u}_{\mathcal{E}}$ can be determined as follows:*
$$\begin{cases} \mathbf{u}_{\mathcal{E}} = A_{\mathcal{E}\mathcal{E}}^{\dagger}\left(\mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda}\right) + Z\boldsymbol{\alpha}, \qquad \textit{where} \\ \ \boldsymbol{\alpha} = (G^TG)^{\dagger}G^T\left(K\boldsymbol{\lambda} - MA_{\mathcal{E}\mathcal{E}}^{\dagger}\mathbf{f}_{\mathcal{E}}\right). \end{cases} \tag{4.27}$$

Proof. Since $A_{\mathcal{E}\mathcal{E}}$ is singular, the first block row in (4.25) yields:

$$\begin{cases} A_{\mathcal{E}\mathcal{E}}\mathbf{u}_{\mathcal{E}} = \mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda} \iff \mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda} \in \text{Range}(A_{\mathcal{E}\mathcal{E}}) \\ \qquad\qquad\qquad\qquad\quad\ \iff \mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda} \perp \text{Kernel}(A_{\mathcal{E}\mathcal{E}}) \\ \qquad\qquad\qquad\qquad\quad\ \iff Z^T\left(\mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda}\right) = \mathbf{0} \\ \qquad\qquad\qquad\qquad\quad\ \iff G^T\boldsymbol{\lambda} = \mathbf{g}, \end{cases}$$

where $\mathbf{g} \equiv Z^T\mathbf{f}_{\mathcal{E}}$. When the compatability condition $G^T\boldsymbol{\lambda} = \mathbf{g}$ is satisfied, the general solution to the *singular* system $A_{\mathcal{E}\mathcal{E}}\mathbf{u}_{\mathcal{E}} = \mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda}$ will be:

$$\mathbf{u}_{\mathcal{E}} = A_{\mathcal{E}\mathcal{E}}^{\dagger}\left(\mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda}\right) + Z\boldsymbol{\alpha}.$$

Here $\boldsymbol{\alpha} \in \mathbb{R}^d$ is arbitrary, since matrix Z has rank d, and $A_{\mathcal{E}\mathcal{E}}^{\dagger}$ is the Moore-Penrose pseudoinverse of $A_{\mathcal{E}\mathcal{E}}$, see [ST13, GO4]. Applying the constraint $M\mathbf{u}_{\mathcal{E}} = \mathbf{0}$ to the above expression for $\mathbf{u}_{\mathcal{E}}$ yields:

$$MA_{\mathcal{E}\mathcal{E}}^{\dagger}\left(\mathbf{f}_{\mathcal{E}} - M^T\boldsymbol{\lambda}\right) + MZ\boldsymbol{\alpha} = \mathbf{0}.$$

This corresponds to $K\,\boldsymbol{\lambda} - G\boldsymbol{\alpha} = \mathbf{e}$, for $K \equiv MA_{\mathcal{E}\mathcal{E}}^{\dagger}M^T$ and $\mathbf{e} \equiv MA_{\mathcal{E}\mathcal{E}}^{\dagger}\mathbf{f}_{\mathcal{E}}$. Combining the compatability condition with the preceding yields the system:

$$\begin{cases} K\,\boldsymbol{\lambda} - G\boldsymbol{\alpha} = \mathbf{e} \\ \qquad G^T\,\boldsymbol{\lambda} = \mathbf{g}, \end{cases} \tag{4.28}$$

which constitutes $m+d$ equations for the $m+d$ unknown entries of $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$. The term $G\boldsymbol{\alpha}$ in the first block equation in (4.28) can be eliminated by applying $P_0 = I - G(G^TG)^{\dagger}G^T$, which corresponds to the Euclidean orthogonal projection onto $\text{Range}(G)^{\perp}$:

$$\begin{cases} P_0\,K\boldsymbol{\lambda} = P_0\,\mathbf{e} \\ \quad G^T\boldsymbol{\lambda} = \mathbf{g}. \end{cases}$$

Since $d = \text{Rank}(Z)$, it follows that $\text{Rank}(G) = d$ and P_0 is an orthogonal projection onto a space of dimension $m - d$. This effectively constitutes

$m = (m-d)+d$ equations for the unknown $\boldsymbol{\lambda} \in \mathbb{R}^m$. Once $\boldsymbol{\lambda}$ is determined by solving the above problem, the unknown coefficient vector $\boldsymbol{\alpha}$ can be determined using (4.28) as $\boldsymbol{\alpha} \equiv (G^T G)^{\dagger} \left(G^T K \boldsymbol{\lambda} - G^T \mathbf{e} \right)$. □

Remark 4.17. When matrix $A_{\mathcal{E}\mathcal{E}}$ is *nonsingular*, matrix Z will have zero rank and vector $\boldsymbol{\alpha}$ can be omitted. In this case, $K = M A_{\mathcal{E}\mathcal{E}}^{-1} M^T$, $\mathbf{e} = M A_{\mathcal{E}\mathcal{E}}^{-1} \mathbf{f}_{\mathcal{E}}$, $G = 0$, $P_0 = I$, and $\mathbf{g} = \mathbf{0}$. Furthermore, the reduced system (4.26) will correspond to the stationarity condition for a maximum of the dual function $D(\boldsymbol{\mu})$ associated with the Lagrange multiplier variables.

4.3.1 Projected Gradient Algorithm to Solve (4.26)

Since the solution to (4.25) can be obtained using (4.27) once $\boldsymbol{\lambda}$ is determined, the FETI method seeks the Lagrange multiplier variables $\boldsymbol{\lambda} \in \mathbb{R}^m$ by solving:

$$\begin{cases} P_0 K \boldsymbol{\lambda} = P_0 \mathbf{e} \\ \quad G^T \boldsymbol{\lambda} = \mathbf{g}. \end{cases} \tag{4.29}$$

In Lemma 4.19 it is shown that this system is symmetric and positive definite within a certain *subspace* $\mathcal{G}_*$ of $\mathbb{R}^m$, and consequently, it will be solvable by a conjugate gradient method in that subspace. However, the FETI method solves a modified linear system equivalent to (4.29), to include *global transfer* of information within the algorithm, as outlined below.

Let C denote an $m \times q$ matrix having rank q where $q < m$. Employing matrix C we modify system (4.29) as follows:

$$\begin{cases} \quad P_0 K \boldsymbol{\lambda} = P_0 \mathbf{e} \\ C^T P_0 K \boldsymbol{\lambda} = C^T P_0 \mathbf{e} \\ \quad\quad G^T \boldsymbol{\lambda} = \mathbf{g}. \end{cases} \tag{4.30}$$

The first and third block equations in (4.30) are identical to the first and second block equations in (4.29), while the second block equation in (4.30) is redundant, corresponding to linear combinations of the first block in (4.29) with weights based on matrix C. Typically, either matrix $G = 0$ or $C = 0$, however, both will be included for generality [FA14].

Remark 4.18. If Z (and hence $G = MZ$) has rank d, then the orthogonal projection matrix P_0 will have rank $(m-d)$. Thus, the coefficient matrix in the first block equation in (4.30) will have rank $(m-d)$, third block equation will have rank d, while the second block equation will be redundant consisting of q linear combinations of rows of the first block equation.

We now motivate a projected gradient algorithm to solve (4.30). Suppose $\boldsymbol{\lambda}_* \in \mathbb{R}^m$ can be found satisfying the 2nd and 3rd block equations in (4.30):

$$\begin{cases} C^T P_0 K \boldsymbol{\lambda}_* = C^T P_0 \mathbf{e} \\ \quad\quad G^T \boldsymbol{\lambda}_* = \mathbf{g}. \end{cases} \tag{4.31}$$

Then, we may seek the solution to (4.30) as $\boldsymbol{\lambda} = \boldsymbol{\lambda}_* + \tilde{\boldsymbol{\lambda}}$ provided $\tilde{\boldsymbol{\lambda}}$ solves:

$$\begin{cases} P_0 K\, \tilde{\boldsymbol{\lambda}} = P_0\, (\mathbf{e} - K\, \boldsymbol{\lambda}_*) \\ C^T P_0 K\, \tilde{\boldsymbol{\lambda}} = \mathbf{0} \\ G^T \tilde{\boldsymbol{\lambda}} = \mathbf{0}. \end{cases} \tag{4.32}$$

If we seek the correction $\tilde{\boldsymbol{\lambda}}$ within the *subspace* $\mathcal{G}_0 \subset \mathrm{IR}^m$ defined by:

$$\mathcal{G}_0 \equiv \left\{ \boldsymbol{\mu} \in \mathrm{IR}^m : C^T P_0 K\, \boldsymbol{\mu} = \mathbf{0},\ G^T \boldsymbol{\mu} = \mathbf{0} \right\}, \tag{4.33}$$

then, the second and third block equations in (4.30) will automatically hold. Importantly, by Lemma 4.19 below, matrix $P_0 K$ will be symmetric and positive definite in subspace $\mathcal{G}_0$ equipped with the Euclidean inner product $(\cdot, \cdot)$:

$$\begin{cases} \left(P_0\, K \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}} \right) = \left(\tilde{\boldsymbol{\lambda}}, P_0\, K\, \tilde{\boldsymbol{\mu}} \right), & \forall \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\lambda}} \in \mathcal{G}_0 \\ (P_0\, K\, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}) \geq c\, (\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}), & \forall \tilde{\boldsymbol{\mu}} \in \mathcal{G}_0, \end{cases} \tag{4.34}$$

for some $c > 0$. Consequently, a projected conjugate gradient iterative method may be applied to determine $\tilde{\boldsymbol{\lambda}}$ within $\mathcal{G}_0$ so that:

$$\left(P_0\, K \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}} \right) = (P_0\, (\mathbf{e} - K\, \boldsymbol{\lambda}_*), \tilde{\boldsymbol{\mu}}), \quad \forall \tilde{\boldsymbol{\mu}} \in \mathcal{G}_0. \tag{4.35}$$

To determine $\boldsymbol{\lambda}_* \in \mathrm{IR}^m$ such that (4.31) holds, seek it as $\boldsymbol{\lambda}_* = G\boldsymbol{\beta}_* + C\boldsymbol{\gamma}_*$ where the coefficient vectors $\boldsymbol{\beta}_* \in \mathrm{IR}^d$ and $\boldsymbol{\gamma}_* \in \mathrm{IR}^q$ are to be determined. By applying the constraints (4.31) to $G\boldsymbol{\beta}_* + C\boldsymbol{\gamma}_*$, we obtain the following block equations for $\boldsymbol{\beta}_*$ and $\boldsymbol{\gamma}_*$:

$$\begin{cases} C^T P_0 K G \boldsymbol{\beta}_* + C^T P_0 K C \boldsymbol{\gamma}_* = C^T P_0\, \mathbf{e} \\ G^T G\, \boldsymbol{\beta}_* + G^T C\, \boldsymbol{\gamma}_* = \mathbf{g}. \end{cases} \tag{4.36}$$

Rather than solve this system involving $d + q$ unknowns, it will be advantageous to combine the computation of $P_0 K \boldsymbol{\lambda}_*$ into the above system. Accordingly, represent $P_0 K \boldsymbol{\lambda}_*$ as $K\boldsymbol{\lambda}_* + G\boldsymbol{\delta}_*$ where $\boldsymbol{\delta}_* \in \mathrm{IR}^d$ denotes an unknown coefficient vector to be selected so that $K\boldsymbol{\lambda}_* + G\boldsymbol{\delta}_* \in \mathrm{Range}\,(G)^{\perp}$:

$$G^T \left(K\boldsymbol{\lambda}_* + G\boldsymbol{\delta}_* \right) = \mathbf{0}. \tag{4.37}$$

Substituting $P_0 K \boldsymbol{\lambda}_* = K\boldsymbol{\lambda}_* + G\boldsymbol{\delta}_*$ into (4.36) and applying the constraint (4.37) yields the following block system for $\boldsymbol{\beta}_*$, $\boldsymbol{\gamma}_*$ and $\boldsymbol{\delta}_*$:

$$\begin{cases} G^T K \left(G\boldsymbol{\alpha}_* + C\boldsymbol{\beta}_* \right) + G^T G \boldsymbol{\mu}_* = G^T P_0\, \mathbf{e} \\ C^T K \left(G\boldsymbol{\alpha}_* + C\boldsymbol{\beta}_* \right) + C^T G \boldsymbol{\mu}_* = C^T P_0\, \mathbf{e} \\ G^T \left(G\boldsymbol{\alpha}_* + C\boldsymbol{\beta}_* \right) = G^T \mathbf{g}. \end{cases}$$

This system has the following block matrix form:

$$\begin{bmatrix} G^T K G & G^T K C & G^T G \\ C^T K G & C^T K C & C^T G \\ G^T G & G^T C & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_* \\ \boldsymbol{\gamma}_* \\ \boldsymbol{\delta}_* \end{bmatrix} = \begin{bmatrix} G^T P_0 \, \mathbf{e} \\ C^T P_0 \, \mathbf{e} \\ G^T \mathbf{g} \end{bmatrix}. \tag{4.38}$$

Once system (4.38) is solved, the solution $\boldsymbol{\lambda}_*$ to problem (4.31) is:

$$\boldsymbol{\lambda}_* = G\boldsymbol{\alpha}_* + C\boldsymbol{\beta}_*.$$

The solution $\boldsymbol{\lambda}$ of (4.30) may now be expressed as $\boldsymbol{\lambda} = \boldsymbol{\lambda}_* + \tilde{\boldsymbol{\lambda}}$ where $\tilde{\boldsymbol{\lambda}}$ solves (4.35). We next verify that $P_0 \, K$ is symmetric positive definite in $\mathcal{G}_*$.

Lemma 4.19. *Suppose the following assumptions hold.*

1. *Let M be of full rank, $K = M A_{\mathcal{E}\mathcal{E}}^{\dagger} M^T$, $G = MZ$, $P_0 = I - G \left(G^T G \right)^{\dagger} G^T$ and* Range(Z) = Kernel$(A_{\mathcal{E}\mathcal{E}})$.
2. *Let $\sigma_*(A_{\mathcal{E}\mathcal{E}})$ be the smallest nonzero singular value of matrix $A_{\mathcal{E}\mathcal{E}}$:*

$$\mathbf{w}_{\mathcal{E}}^T A_{\mathcal{E}\mathcal{E}} \mathbf{w}_{\mathcal{E}} \geq \sigma_*(A_{\mathcal{E}\mathcal{E}}) \, \mathbf{w}_{\mathcal{E}}^T \mathbf{w}_{\mathcal{E}}, \qquad \forall \mathbf{w}_{\mathcal{E}} \ \text{such that } Z^T \mathbf{w}_{\mathcal{E}} = \mathbf{0}.$$

3. *Define*

$$\mathcal{G}_* \equiv \left\{ \boldsymbol{\mu} \in \mathrm{IR}^m : G^T \boldsymbol{\lambda} = \mathbf{0} \right\}.$$

4. *Let $\sigma_*(M)$ be the smallest singular value of M.*

Also, let $(\cdot, \cdot)$ denote the Euclidean inner product.
Then, the following results will hold.

1. *The matrix $P_0 \, K$ will be symmetric in the subspace $\mathcal{G}_*$ with:*

$$(P_0 \, K \boldsymbol{\lambda}, \boldsymbol{\mu}) = (\boldsymbol{\lambda}, P_0 \, K \boldsymbol{\mu}), \qquad \forall \boldsymbol{\lambda}, \, \boldsymbol{\mu} \in \mathcal{G}_*.$$

2. *Matrix $P_0 \, K$ will be positive definite, satisfying:*

$$(P_0 \, K \boldsymbol{\mu}, \boldsymbol{\mu}) \geq \sigma_*(A_{\mathcal{E}\mathcal{E}}) \, \sigma_*(M) \; (\boldsymbol{\mu}, \boldsymbol{\mu}), \qquad \text{if} \;\; G^T \boldsymbol{\mu} = \mathbf{0}.$$

Proof. To show that $P_0 \, K$ is symmetric in $\mathcal{G}_*$, choose $\boldsymbol{\mu}$, $\boldsymbol{\lambda} \in \mathcal{G}_*$. Then, by definition $P_0 \boldsymbol{\lambda} = \boldsymbol{\lambda}$ and $P_0 \boldsymbol{\mu} = \boldsymbol{\mu}$. Since $P_0^T = P_0$ and $K^T = K$, we obtain:

$$\begin{cases} (P_0 \, K \boldsymbol{\lambda}, \boldsymbol{\mu}) = (K \boldsymbol{\lambda}, P_0 \boldsymbol{\mu}) \\ \qquad\qquad\quad = (K \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \qquad\qquad\quad = (\boldsymbol{\lambda}, K \boldsymbol{\mu}) \\ \qquad\qquad\quad = (P_0 \boldsymbol{\lambda}, K \boldsymbol{\mu}) \\ \qquad\qquad\quad = (\boldsymbol{\lambda}, P_0 K \boldsymbol{\mu}). \end{cases}$$

To verify positive definiteness, suppose that $\boldsymbol{\mu} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$ and that $G^T \boldsymbol{\mu} = \mathbf{0}$, i.e., $Z^T \left(M^T \boldsymbol{\mu} \right) = \mathbf{0}$. Then, applying the hypothesis yields:

$$\begin{cases} (P_0 K \boldsymbol{\mu}, \boldsymbol{\mu}) = (K \boldsymbol{\mu}, \boldsymbol{\mu}) \\ \qquad\qquad\quad = \left(M A_{\mathcal{E}\mathcal{E}}^{\dagger} M^T \boldsymbol{\mu}, \boldsymbol{\mu} \right) \\ \qquad\qquad\quad = \left(A_{\mathcal{E}\mathcal{E}}^{\dagger} M^T \boldsymbol{\mu}, M^T \boldsymbol{\mu} \right) \\ \qquad\qquad\quad \geq \sigma_*(A_{\mathcal{E}\mathcal{E}}) \left(M^T \boldsymbol{\mu}, M^T \boldsymbol{\mu} \right) \\ \qquad\qquad\quad \geq \sigma_*(A_{\mathcal{E}\mathcal{E}})\, \sigma_*(M)\, (\boldsymbol{\mu}, \boldsymbol{\mu}), \end{cases}$$

since $Z^T \left(M^T \boldsymbol{\mu} \right) = \mathbf{0}$ and since $\boldsymbol{\mu} \neq \mathbf{0}$. □

Lemma 4.19 shows that $P_0 K$ is symmetric and positive definite in $\mathcal{G}_*$, and hence in $\mathcal{G}_0 \subset \mathcal{G}_*$, see (4.33). As a result, the PCG method may be employed to determine $\tilde{\boldsymbol{\lambda}} \in \mathcal{G}_0$. Care must be exercised, however, to ensure that all iterates and residuals in the conjugate gradient algorithm remain within the subspace $\mathcal{G}_0$. To do this, a projection matrix Q (possibly oblique, satisfying $Q^2 = Q$) will be employed to project residuals or preconditioned updates onto the subspace $\mathcal{G}_0$ each iteration. In the following, we derive an expression for such a (possibly oblique) projection matrix Q. Given $\boldsymbol{\lambda} \in \mathbb{R}^m$, we shall seek its projection $Q\boldsymbol{\lambda} \in \mathcal{G}_0$ in the form:

$$Q\boldsymbol{\lambda} \equiv \boldsymbol{\lambda} + G\boldsymbol{\beta} + C\boldsymbol{\gamma}, \tag{4.39}$$

where the coefficient vectors $\boldsymbol{\beta} \in \mathbb{R}^d$, $\boldsymbol{\gamma} \in \mathbb{R}^q$ are chosen to satisfy:

$$\begin{cases} G^T K \left(G\boldsymbol{\beta} + C\boldsymbol{\gamma} \right) + G^T G \boldsymbol{\delta} = -G^T K \boldsymbol{\lambda} \\ C^T K \left(G\boldsymbol{\beta} + C\boldsymbol{\gamma} \right) + C^T G \boldsymbol{\delta} = -C^T K \boldsymbol{\lambda} \\ \qquad\qquad G^T \left(G\boldsymbol{\beta} + C\boldsymbol{\gamma} \right) = -G^T \boldsymbol{\lambda}. \end{cases}$$

In the above, $\boldsymbol{\delta} \in \mathbb{R}^d$ was introduced to represent:

$$P_0\, K \left(G\boldsymbol{\beta} + C\boldsymbol{\gamma} \right) = K \left(G\boldsymbol{\beta} + C\boldsymbol{\gamma} \right) + G\boldsymbol{\delta}.$$

The resulting projection Q will thus have matrix representation:

$$Q \equiv I - \begin{bmatrix} G^T \\ C^T \\ 0 \end{bmatrix}^T \begin{bmatrix} G^T K G & G^T K C & G^T G \\ C^T K G & C^T K C & C^T G \\ G^T G & G^T C & 0 \end{bmatrix}^{\dagger} \begin{bmatrix} G^T K \\ C^T K \\ G^T \end{bmatrix}. \tag{4.40}$$

A pseudoinverse was employed in the above since in the cases of interest, either $C = 0$ or $G = 0$, and this coefficient matrix will become singular. By construction $Q\boldsymbol{\lambda} \in \mathcal{G}_0$. In the following, we describe the projection matrix (4.40) in the two special cases of interest.

Form of $\boldsymbol{Q}$ when $\boldsymbol{c(x) = 0}$. If $c(x) = 0$ in (4.1), then the subdomain stiffness matrix $A^{(i)}$ will be singular when Ω_i is a *floating* subdomain. In this case Z, and hence $G = MZ$, will be nontrivial, and typically C is chosen to be 0 (or equivalently omitted). Importantly, due to the block diagonal terms $M^{(i)} Z^{(i)}$ in matrix G, the projection onto subspace $\mathcal{G}_0$ will provide *global transfer of information.* When $P_0\, K$ is suitably preconditioned, the convergence rate may deteriorate only mildly with h. The nonhomogeneous term $\boldsymbol{\lambda}_*$ can be sought as $\boldsymbol{\lambda}_* = G\boldsymbol{\beta}_*$ with:

$$G^T G\boldsymbol{\beta}_* = G^T \mathbf{g},$$

so that:

$$\boldsymbol{\lambda}_* = G(G^T G)^{-1} G^T \mathbf{g}.$$

In this case, the operator $Q = I - G(G^T G)^{-1} G^T$ reduces to P_0 and will be an orthogonal projection in the Euclidean inner product.

Form of $\boldsymbol{Q}$ when $\boldsymbol{c(x) \geq c_0 > 0}$. If the coefficient $c(x) \geq c_0 > 0$ in (4.1), then the local stiffness matrices $A^{(i)}$, and hence $A_{\mathcal{E}\mathcal{E}}$, will be nonsingular. In this case $G = 0$ and $P_0 = I$. While this may be viewed as an advantage, it results in an algorithm *without* any built in mechanism for global transfer of information. Such transfer may be included in a suitably constructed preconditioner. However, it will be advantageous to include it by selecting a nontrivial matrix $C \equiv M\tilde{Z}$ where $\tilde{Z}$ is an $n_{\mathcal{E}} \times d$ matrix whose columns form a basis for Kernel$(\tilde{A}_{\mathcal{E}\mathcal{E}})$ where $\tilde{A}_{\mathcal{E}\mathcal{E}} = \text{blockdiag}(\tilde{A}^{(1)}, \ldots, \tilde{A}^{(p)})$ denotes the extended stiffness matrix arising from discretization of the elliptic operator in (4.1) with $c(x) = 0$. For this choice of matrix C, and a suitable preconditioner, the FETI algorithm will typically have convergence rates deteriorating only mildly with increasing number of nodes per subdomain. Computation of the initial nonhomogeneous term $\boldsymbol{\lambda}_*$ reduces to:

$$(C^T K C)\, \boldsymbol{\beta}_* = C^T\, P_0\, \mathbf{e},$$

so that $\boldsymbol{\lambda}_* = C(C^T K C)^{-1} C^T\, P_0\, \mathbf{e}$.

In this case, operator $Q = I - C(C^T K C)^{-1} C^T K$ and will be orthogonal only in the K induced inner product. A preconditioner for $P_0\, K$ can be sought within $\mathcal{G}_0$, so that the action of the inverse of the preconditioner has the form QNQ^T where N is symmetric (in the Euclidean inner product). In applications, however, only the action QN needs to be computed when the residuals from previous iterates lie in $\mathcal{G}_0$.

We may now summarize the FETI algorithm, employing the projection matrices P_0 and Q and a preconditioner, whose inverse has the form QNQ^T (though in practice, it will be sufficient to evaluate only QN) for a matrix N. The algorithm below includes the computation of $\boldsymbol{\lambda}_*$ and $\tilde{\boldsymbol{\lambda}}$.

Algorithm 4.3.1 *(FETI Algorithm to Solve (4.26))*

```
Let $\boldsymbol{\lambda}_0$ de a starting guess (for instance $\boldsymbol{\lambda}_0 = \mathbf{0}$)
1. Compute:
       $\begin{cases} \mathbf{e} \equiv M\, A_{\mathcal{E}\mathcal{E}}^{\dagger} \mathbf{f}_{\mathcal{E}} \\ \mathbf{g} \equiv Z^T \mathbf{f}_{\mathcal{E}} \end{cases}$
2. Solve the following system (using a pseudoinverse):
       $\begin{cases} G^T K \left( G\boldsymbol{\beta}_* + C\boldsymbol{\gamma}_* \right) + G^T G \boldsymbol{\delta}_* = G^T P_0 \left( \mathbf{e} - K\boldsymbol{\lambda}_0 \right) \\ C^T K \left( G\boldsymbol{\beta}_* + C\boldsymbol{\gamma}_* \right) + C^T G \boldsymbol{\delta}_* = C^T \left( P_0 \mathbf{e} - K\boldsymbol{\lambda}_0 \right) \\ G^T \left( G\boldsymbol{\alpha}_* + C\boldsymbol{\beta}_* \right) = G^T \mathbf{g}. \end{cases}$
3. Define:
       $\boldsymbol{\lambda}_* \leftarrow \boldsymbol{\lambda}_0 + G\boldsymbol{\beta}_* + C\boldsymbol{\gamma}_*.$
4. Compute the residual:
       $\mathbf{r}_0 \equiv P_0 (K\boldsymbol{\lambda}_* - \mathbf{e}).$
5. For $k = 1, 2, \cdots$ until convergence do:
       $\mathbf{z}_{k-1} = N\mathbf{r}_{k-1}$                          preconditioning
       $\mathbf{y}_{k-1} = Q\mathbf{z}_{k-1}$                          projection
       $\boldsymbol{\xi}_k = \mathbf{r}_{k-1}^T \mathbf{y}_{k-1}$
       $\mathbf{p}_k = \mathbf{y}_{k-1} + \frac{\xi_k}{\xi_{k-1}} \mathbf{p}_{k-1}$   $(\mathbf{p}_1 \equiv \mathbf{y}_0)$
       $\boldsymbol{\nu}_k = \frac{\xi_k}{\mathbf{p}_k^T P_0\, K \mathbf{p}_k}$
       $\boldsymbol{\lambda}_k = \boldsymbol{\lambda}_{k-1} + \nu_k \mathbf{p}_k$
       $\mathbf{r}_k = \mathbf{r}_{k-1} - \nu_k P_0\, K \mathbf{p}_k$
6. Endfor
7. Compute:
       $\begin{cases} \boldsymbol{\alpha} \equiv (G^T G)^{\dagger} G^T \left( K\boldsymbol{\lambda} - M A_{\mathcal{E}\mathcal{E}}^{\dagger} \mathbf{f}_{\mathcal{E}} \right) \\ \mathbf{u} = A_{\mathcal{E}\mathcal{E}}^{\dagger} \left( \mathbf{f}_{\mathcal{E}} - M^T \boldsymbol{\lambda} \right) + Z\boldsymbol{\alpha}. \end{cases}$
```

We next describe preconditioners of the form $Q\, N$ in the FETI algorithm.

4.3.2 Preconditioners for $P_0\, K$

We shall describe two preconditioners proposed in [FA15], for matrix $P_0\, K$. Since information will be *transferred globally* within the FETI algorithm in the projection step involving matrix Q, a coarse space term will be unnecessary in FETI preconditioners. Both the preconditioners considered below have a similar structure, and are motivated as follows.

Since matrices $A_{\mathcal{E}\mathcal{E}}$ and M have the following block structures:

$$\begin{cases} A_{\mathcal{E}\mathcal{E}} = \text{blockdiag}\big(A^{(1)}, \cdots, A^{(p)}\big), \\ \quad M = \big[\, M^{(1)} \cdots \; M^{(p)} \,\big], \end{cases}$$

matrix $K = M A^{\dagger}_{\mathcal{E}\mathcal{E}} M^T$ will formally satisfy:

$$K = \sum_{i=1}^{p} M^{(i)} A^{(i)^{\dagger}} M^{(i)^T}.$$

Each matrix $M^{(i)}$ will have the following block structures due to the ordering of interior and boundary nodes within each subdomain:

$$M^{(i)} = [M_I^{(i)} \;\; M_B^{(i)}] = [0 \;\; M_B^{(i)}]$$

where $M_I^{(i)} = 0$ since the continuity constraint involves only interface unknowns. Substituting this into the preceding expression for K yields:

$$\begin{cases} K = \sum_{i=1}^{p} M^{(i)} A^{(i)^{\dagger}} M^{(i)^T} \\ \quad = \sum_{i=1}^{p} \begin{bmatrix} 0 \\ M_B^{(i)^T} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix}^{\dagger} \begin{bmatrix} 0 \\ M_B^{(i)^T} \end{bmatrix} \\ \quad = \sum_{i=1}^{p} M_B^{(i)} S^{(i)^{\dagger}} M_B^{(i)^T}. \end{cases} \tag{4.41}$$

The last equation above follows easily when submatrix $A^{(i)}$ is nonsingular. When $A^{(i)}$ is singular, it can be verified by employing the block structure of $A^{(i)}$ and the algebraic definition of the pseudoinverse of a matrix.

The additive expression for K in (4.41) resembles a *subassembly identity*, heuristically, provided the boundary constraint matrices $M_B^{(i)}$ with entries from $\{-1, 0, 1\}$ are interpreted as boundary restriction matrices $\mathcal{R}_i$. This formal analogy suggests that preconditioners can be sought for K having a similar structure to Neumann-Neumann preconditioners [FA14, KL8]. For instance, given a two subdomain decomposition the constraints will be $M_B^{(1)} = I$ and $M_B^{(2)} = -I$, so that $K = \sum_{i=1}^{2} S^{(i)^{\dagger}}$. If matrix $A_{\mathcal{E}\mathcal{E}}$ is nonsingular, then the formal inverses of $S^{(1)^{\dagger}}$ and $S^{(2)^{\dagger}}$ will be spectrally equivalent to each other (independent of h). In this case, we may heuristically define the action of the inverse of a preconditioner for K by $QN = Q \sum_{i=1}^{2} S^{(i)}$.

Other preconditioners based on analogy with two subdomain Schur complement preconditioners are also possible. By construction, the resulting condition number will be independent of h. More generally, the heuristic similarity with Neumann-Neumann preconditioners suggests a preconditioner whose formal inverse has the structure:

$$\begin{cases} QN \equiv Q\left(\sum_{i=1}^{p} M_B^{(i)} S^{(i)} M_B^{(i)^T}\right) \\ \qquad = Q\left(\sum_{i=1}^{p} M_B^{(i)} \left(A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}\right) M_B^{(i)^T}\right). \end{cases} \tag{4.42}$$

In this case computing the action of QN will require the solution of a local Dirichlet problem on each subdomain, and the resulting preconditioner is referred to as a *Dirichlet* preconditioner, since computation of the action of $\left(A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}\right)$ requires the solution of a Dirichlet problem on each subdomain.

The action QL of the inverse of an alternative preconditioner, referred to as the *lumped* preconditioner, is obtained as follows:

$$QL \equiv Q\left(\sum_{i=1}^{p} M_B^{(i)} A_{BB}^{(i)} M_B^{(i)^T}\right).$$

This preconditioner does not require the solution of local Dirichlet problems and is obtained by approximating the local Schur complement $S^{(i)} \approx A_{BB}^{(i)}$. The following theoretical results will hold for the preconditioner $Q\,N$.

Theorem 4.20. *The following bounds hold for the Dirichlet preconditioner.*

1. *There exists $C > 0$ independent of h_0, h and jumps in the coefficients:*

$$\operatorname{cond}(P_0\, K, QN) \equiv \frac{\lambda_{max}(QNP_0\, K)}{\lambda_{min}(QNP_0\, K)} \leq C\,(1 + \log(h_0/h))^3.$$

Proof. See [MA25, KL8]. □

4.4 FETI-DP and BDDC Methods

In this section, we describe two popular variants of the FETI method to solve the saddle point problem (4.22) or its associated primal formulation. The FETI-DP (Dual-Primal) method solves a reduced version of (4.22) while BDDC (Balancing Domain Decomposition with Constraints) corresponds to a primal version of FETI-DP [FA11, FA10, ST4, DO, DO2, MA18, MA19]. Both methods work in a class of local solutions which are discontinuous across the subdomain boundaries, except for a family of chosen continuity *constraints*. For simplicity, we only consider simple continuity constraints across *cross points*, *edges* and *faces* of a subdomain boundary. More general constraints are considered in [TO10]. Both methods are CG based, and improve upon the scalability of the FETI algorithm in three dimensions, yielding robust convergence. The resulting preconditioned matrices have the same spectra, except for zeros or ones. In the following, we describe the reduction of system (4.22) to a smaller saddle point system and introduce notation, prior to formulating the FETI-DP and BDDC methods. To be consistent with preceding sections, our notation differs from that in [DO, DO2, MA18, MA19].

Reduced Saddle Point System. To reduce system (4.22) by elimination of the interior unknowns $\mathbf{u}_I^{(l)}$ for $1 \le l \le p$, we re-order the block vector $\mathbf{u}_{\mathcal{E}}$ as $(\mathbf{u}_I^T, \mathbf{u}_B^T)^T$ in the saddle point system (4.22), where:

$$\mathbf{u}_I = \left(\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T}\right)^T \quad \text{and} \quad \mathbf{u}_B = \left(\mathbf{u}_B^{(1)^T}, \ldots, \mathbf{u}_B^{(p)^T}\right)^T.$$

This will yield the following reordered system:

$$\begin{bmatrix} A_{II} & A_{IB} & 0 \\ A_{IB}^T & A_{BB} & M_B^T \\ 0 & M_B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ \mathbf{f}_B \\ \mathbf{0} \end{bmatrix}, \tag{4.43}$$

where the block submatrices A_{II}, A_{IB} and A_{BB} are defined as follows:

$$A_{II} = \begin{bmatrix} A_{II}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{II}^{(p)} \end{bmatrix}, \; A_{IB} = \begin{bmatrix} A_{IB}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{IB}^{(p)} \end{bmatrix}, \; A_{BB} = \begin{bmatrix} A_{BB}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{BB}^{(p)} \end{bmatrix},$$

with matrix $M_B = \begin{bmatrix} M_B^{(1)} & \cdots & M_B^{(p)} \end{bmatrix}$, while the load vectors satisfy:

$$\mathbf{f}_I = \left(\mathbf{f}_I^{(1)^T}, \ldots, \mathbf{f}_I^{(p)^T}\right)^T \text{ and } \mathbf{f}_B = \left(\mathbf{f}_B^{(1)^T}, \ldots, \mathbf{f}_B^{(p)^T}\right)^T.$$

Here, the matrices $A_{XY}^{(l)}$ and $M_B^{(l)}$, and vectors $\mathbf{u}_X^{(l)}$, $\mathbf{f}_X^{(l)}$ are as in (4.10) and (4.6) for $X, Y = I, B$. We solve for $\mathbf{u}_I = A_{II}^{-1}(\mathbf{f}_I - A_{IB}\mathbf{u}_B)$ using the first block row of (4.43). Substituting this expression into the second block row of (4.43) yields a reduced saddle point system for determining $\mathbf{u}_B$ and $\boldsymbol{\lambda}$.

The reduced saddle point system will be:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & M_B^T \\ M_B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_B \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \mathbf{0} \end{bmatrix}, \tag{4.44}$$

where $\tilde{\mathbf{f}}_B \equiv \left(\mathbf{f}_B - A_{IB}^T A_{II}^{-1} \mathbf{f}_I\right)$ and $S_{\mathcal{E}\mathcal{E}} = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$ satisfies:

$$S_{\mathcal{E}\mathcal{E}} = \begin{bmatrix} S^{(1)} & & 0 \\ & \ddots & \\ 0 & & S^{(p)} \end{bmatrix} \quad \text{where} \quad S^{(i)} \equiv (A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}).$$

The solution to (4.43) can be obtained by solving (4.44) for $\mathbf{u}_B$ and $\boldsymbol{\lambda}$, and subsequently $\mathbf{u}_I = A_{II}^{-1}(\mathbf{f}_I - A_{IB}\mathbf{u}_B)$. Here M_B and $M_B^{(l)}$ will be of size $m \times n_B$ and $m \times n_B^{(l)}$ respectively, and $S_{\mathcal{E}\mathcal{E}}$ of size $n_B = (n_B^{(1)} + \cdots + n_B^{(p)})$.

Primal and Dual Spaces, Restriction and Extension Maps. Given $\Omega_1, \ldots, \Omega_p$ let $B^{(l)} = \partial\Omega_l \cap \Omega$ denote the interior segments of subdomain boundaries, and let $B = \cup_{l=1}^{p} B^{(l)}$ be the interface. We employ the notation:

- Let $\mathcal{U} = \mathbb{R}^q$ be the space of nodal vectors associated with finite element functions on B in traditional substructuring. Here, q equals the number of nodes of the triangulation on B, and $\mathcal{U}$ parameterizes the global degrees of freedom on B. For $1 \leq l \leq p$ and $\mathbf{u} \in \mathcal{U}$, let $\mathcal{R}_l \mathbf{u}$ denote the *restriction* of the vector $\mathbf{u}$ of nodal values on B onto the indices of nodes on $B^{(l)}$. Thus $\mathcal{R}_l$ will be an $n_B^{(l)} \times q$ matrix with zero-one entries.
- Let $W_l \equiv \text{Range}(\mathcal{R}_l) = \mathbb{R}^{n_B^{(l)}}$ denote the space of local nodal vectors associated with displacements on $B^{(l)}$ and let $\mathcal{W} \equiv (W_1 \times \cdots \times W_p)$ be the space of extended local displacements with $\dim(\mathcal{W}) = (n_B^{(1)} + \cdots + n_B^{(p)})$. Let $\mathcal{R}_{\mathcal{E}} : \mathcal{U} \rightarrow \mathcal{W}$ denote the restriction matrix from $\mathcal{U}$ into $\mathcal{W}$:
$$\mathcal{R}_{\mathcal{E}}^T = \begin{bmatrix} \mathcal{R}_1^T & \cdots & \mathcal{R}_p^T \end{bmatrix} \quad \text{and} \quad \mathcal{R}_{\mathcal{E}} = \begin{bmatrix} \mathcal{R}_1^T & \cdots & \mathcal{R}_p^T \end{bmatrix}^T,$$
where $\mathcal{R}_{\mathcal{E}}$ is a matrix of size $n_B \times q$.
- Let $M_B : \mathcal{W} \rightarrow \Lambda$, where $M_B \mathbf{v}_B$ denotes the jump discontinuity in $\mathbf{v}_B$ across the subdomains, for $\mathbf{v}_B \in \mathcal{W}$. Here, M_B is of size $m \times n_B$ and $m \equiv \dim(\Lambda) \geq q$ also denotes the number of Lagrange multiplier variables.
- By construction Kernel$(M_B) = $ Range$(\mathcal{R}_{\mathcal{E}})$, thus $M_B \mathcal{R}_{\mathcal{E}} = \mathbf{0}$.
- Denote the *primal Schur complement* matrix S of size $q \times q$ as:
$$S \equiv \mathcal{R}_{\mathcal{E}}^T S_{\mathcal{E}\mathcal{E}} \mathcal{R}_{\mathcal{E}} = \sum_{l=1}^{p} \mathcal{R}_l^T S^{(l)} \mathcal{R}_l,$$
which is employed in *traditional iterative substructuring.*

We shall assume that $\Omega_1, \ldots, \Omega_p$ are *geometrically conforming*, so that B can be further partitioned into *globs*, such as *cross points* and *edges* for $\Omega \subset \mathbb{R}^2$, or *cross points*, *edges* and *faces* when $\Omega \subset \mathbb{R}^3$. We heuristically define globs such as cross points, edges and faces, in the following.

When $\Omega \subset \mathbb{R}^2$, we heuristically define an *edge* as any non-trivial segment $\text{int}(\partial\Omega_l \cap \partial\Omega_j)$ which can be mapped homeomorphically onto the open segment $(0, 1)$. We let n_E denote the number of distinct edges and enumerate them as $E_1, \ldots, E_{n_E}$. We define a *cross-point* as an endpoint within Ω of an edge. We let n_X denote the number of distinct cross points and enumerate them as $X_1, \ldots, X_{n_X}$. We assume that the cross points and edges partition B.

When $\Omega \subset \mathbb{R}^3$, we heuristically define a *face* as any non-trivial segment $\text{int}(\partial\Omega_l \cap \partial\Omega_j)$ which can be mapped homeomorphically onto the open square $(0, 1) \times (0, 1)$. We let n_F denote the number of distinct faces and enumerate them as $F_1, \ldots, F_{n_F}$. We define an *edge* as any non-trivial intersection in Ω of two faces $\text{int}(\overline{F}_l \cap \overline{F}_j)$ which can be homeomorphically mapped onto the open interval $(0, 1)$. We let n_E denote the number of distinct edges and enumerate them as $E_1, \ldots, E_{n_E}$. We define a cross point as any endpoint in Ω of an edge. We let n_X be the number of distinct cross points and enumerate them as $X_1, \ldots, X_{n_X}$. We assume the cross points, edges and faces partition B.

The disjoint cross points, edges and faces are referred to as *globs*, and we shall assume that the interface B can be partitioned into distinct globs. In the FETI-DP and BDDC methods, one *coarse degree of freedom* will be associated with each distinct glob in B, and one basis function with mean value *one* on each glob with zero nodal values outside the glob will employed in formulating the *primal space*. There will be as many coarse degrees of freedom or coarse basis functions as there are distinct globs in B, as described below.

Definition 4.21. *When $\Omega \subset \mathbb{R}^2$, let $q_0 = (n_E + n_X)$ denote the number of coarse degrees of freedom. We define Q_0 as an $q_0 \times q$ matrix which maps onto the coarse degrees of freedom. Each row of Q_0 will be associated with a distinct glob of B, in some chosen ordering of the globs.*

- *If the i'th row of Q_0 is associated with a cross point X_l then:*

$$(Q_0)_{ij} = \begin{cases} 1 & \text{if node } j \text{ in } B \text{ is the cross point } X_l \\ 0 & \text{otherwise} \end{cases}$$

- *If the i'th row of Q_0 is associated with the edge E_l then:*

$$(Q_0)_{ij} = \begin{cases} \frac{1}{|E_l|} & \text{if node } j \text{ in } B \text{ lies in } E_l \\ 0 & \text{if node } j \text{ in } B \text{ does not lie in } E_l \end{cases}$$

where $|E_l|$ denotes the number of nodes in E_l.

Thus, if $\mathbf{u} \in \mathcal{U} = \mathbb{R}^q$ is a nodal vector of global degrees of freedom on B, then $(Q_0\mathbf{u})_i$ will be the mean value of $\mathbf{u}$ on the glob associated with row i. The above weights are uniform within each glob, for simplicity. More generally, the entries of the local mass matrix on the glob must be divided by its row sum.

Definition 4.22. *When $\Omega \subset \mathbb{R}^3$, let $q_0 = (n_F + n_E + n_X)$ denote the number of coarse degrees of freedom on B. Define Q_0 as an $q_0 \times q$ matrix which maps onto the coarse degrees of freedom, as follows. Each row of Q_0 will be associated with a distinct glob of B, in some chosen ordering of the globs.*

- *If the i'th row of Q_0 is associated with cross point X_l then:*

$$(Q_0)_{ij} = \begin{cases} 1 & \text{if node } j \text{ in } B \text{ is the cross point } X_l \\ 0 & \text{otherwise} \end{cases}$$

- *If the i'th row of Q_0 is associated with the edge E_l then:*

$$(Q_0)_{ij} = \begin{cases} \frac{1}{|E_l|} & \text{if node } j \text{ in } B \text{ lies in } E_l \\ 0 & \text{if node } j \text{ in } B \text{ does not lie in } E_l \end{cases}$$

where $|E_l|$ denotes the number of nodes in E_l.

- *If the i'th row of Q_0 is associated with the face F_l then:*

$$(Q_0)_{ij} = \begin{cases} \frac{1}{|F_l|} & \text{if node } j \text{ in } B \text{ lies in } F_l \\ 0 & \text{if node } j \text{ in } B \text{ does not lie in } F_l \end{cases}$$

  where $|F_l|$ denotes the number of nodes in F_l.

Thus, if $\mathbf{u} \in \mathcal{U} = \mathbb{R}^q$ is a nodal vector of global degrees of freedom on B, then $(Q_0\mathbf{u})_i$ will be the mean value of $\mathbf{u}$ on the glob associated with row i. Here too, the weights are uniform within each glob, for simplicity. More generally, the entries of the local mass matrix on the glob must be divided by its row sum.

Since each coarse degree of freedom is associated with a distinct glob, and since by definition, each glob either lies entirely within a subdomain boundary segment $B^{(i)}$ or does not lie in $B^{(i)}$, only certain coarse degree of freedom will be non-zero on $B^{(i)}$. Let $q_0^{(i)}$ denote the number of globs in $B^{(i)}$. We then define a restriction matrix $\mathcal{R}_i^c$ of size $q_0^{(i)} \times q_0$ as a matrix with zero or one entries which picks the coarse degrees of freedom which are non-zero on $B^{(i)}$.

Definition 4.23. *Given a global ordering of the q_0 globs (and associated coarse degrees of freedom) on B and a local ordering of the $q_0^{(i)}$ globs on $B^{(i)}$, we define a restriction matrix $\mathcal{R}_i^c$ of size $q_0^{(i)} \times q_0$ as follows:*

$$(\mathcal{R}_i^c)_{lj} \equiv \begin{cases} 1 \text{ if glob } j \text{ in the global ordering is } l \text{ in the local ordering on } B^{(i)} \\ 0 \text{ otherwise.} \end{cases}$$

Remark 4.24. For instance, if $\overline{\Omega}_i \subset \Omega \subset \mathbb{R}^2$ is a rectangle, then there will be eight coarse degrees of freedom associated with $\partial\Omega_i$, with four cross points and four edges. If $\overline{\Omega}_i \subset \Omega \subset \mathbb{R}^3$ is a box, then there will be twenty six coarse degrees of freedom on $\partial\Omega_i$, with six faces, twelve edges and eight cross points.

Using the restriction matrices $\mathcal{R}_i^c$ and the coarse degrees of freedom matrix Q_0, we define a family of constraint matrices C_i of size $q_0^{(i)} \times n_B^{(i)}$ that will be employed to formulate the primal and dual spaces.

Definition 4.25. *We define a matrix $C_i \equiv \mathcal{R}_i^c Q_0 \mathcal{R}_i^T$ for $1 \le i \le p$. We also define $C \equiv \text{blockdiag}(C_1, \ldots, C_p)$ as the block diagonal matrix of size $(q_0^{(1)} + \cdots + q_0^{(p)}) \times n_B$ (where $n_B = (n_B^{(1)} + \cdots + n_B^{(p)})$):*

$$C \equiv \begin{bmatrix} C_1 & & 0 \\ & \ddots & \\ 0 & & C_p \end{bmatrix}.$$

Remark 4.26. Since $\mathcal{R}_i^T$ and $\mathcal{R}_i^c$ are matrices with zero or one entries, with at most one non-zero entry per row or column, if $\mathbf{w}_i \in W_i$, then $C_i\mathbf{w}_i$ will compute the average value of $\mathbf{w}_i$ on each of the $q_0^{(i)}$ distinct globs on $B^{(i)}$, in the local orderings. Thus, if $C_i\mathbf{w}_i = \mathbf{0}$, then $\mathbf{w}_i$ will be zero at all the cross points in $B^{(i)}$, with mean value zero on the edges and faces (if any) in $B^{(i)}$.

Definition 4.27. *We define a matrix $\mathcal{R}^c$ of size $(q_0^{(1)} + \cdots + q_0^{(p)}) \times q_0$ as:*

$$\mathcal{R}^c \equiv \begin{bmatrix} \mathcal{R}_1^c \\ \vdots \\ \mathcal{R}_p^c \end{bmatrix}$$

corresponding to a restriction of global coarse degrees of freedom on B onto the local coarse degrees of freedom onto each of the local boundaries $B^{(i)}$.

The FETI-DP and BDDC methods employ several subspaces $\mathcal{W}_0$, $\mathcal{W}_D$, $\mathcal{W}_P$ and $\mathcal{W}_*$ of $\mathcal{W}$. Recall that $\mathcal{W} = (W_1 \times \cdots \times W_p)$ denotes the space of nodal vectors on the boundaries, whose associated finite element functions are *discontinuous* across the subdomains. Below, we define $\mathcal{W}_* \subset \mathcal{W}$ as the space of local nodal vectors whose local coarse degrees of freedom are unique, i.e., continuous across the subdomain boundaries. The other degrees of freedom in $\mathcal{W}_*$ may be discontinuous across the subdomain boundaries.

Definition 4.28. *We define $\mathcal{W}_*$ as the following subspace of $\mathcal{W}$:*

$$\mathcal{W}_* \equiv \left\{ \mathbf{w}_B = (\mathbf{w}_B^{(1)^T}, \ldots, \mathbf{w}_B^{(p)^T})^T : C\,\mathbf{w}_B \in \text{Range}(\mathcal{R}^c) \right\},$$

i.e., for each $\mathbf{w}_B \in \mathcal{W}_*$ there must exist some $\mathbf{u} \in \mathbb{R}^{q_0}$ such that $C\,\mathbf{w}_B = \mathcal{R}^c\mathbf{u}$.

The space $\mathcal{W}_*$ can be further decomposed as a sum of two spaces:

$$\mathcal{W}_* = \mathcal{W}_D + \mathcal{W}_P,$$

where $\mathcal{W}_D$ is referred to as the *dual space* and involves local constraints, while the space $\mathcal{W}_P$, which is referred to as the *primal space*, involves global constraints. The primal space $\mathcal{W}_P$ will be employed as a coarse space.

Definition 4.29. *The dual space $\mathcal{W}_D \equiv \text{Kernel}(C) \subset \mathcal{W}_*$ will consist of local nodal vectors whose coarse degrees of freedom (mean value on each glob) are zero on each subdomain boundary:*

$$\mathcal{W}_D \equiv \text{Kernel}(C) = \left\{ \mathbf{w}_B = (\mathbf{w}_B^{(1)^T}, \ldots, \mathbf{w}_B^{(p)^T})^T : C_i\mathbf{w}_B^{(i)} = \mathbf{0} \ \text{ for } 1 \le i \le p \right\}.$$

The primal space $\mathcal{W}_P$ will be a subspace of $\mathcal{W}_*$ complementary to $\mathcal{W}_D$, and defined as the span of q_0 local basis functions whose coarse degrees of freedom are *continuous* across the subdomains.

Definition 4.30. *We define the primal space as* $\mathcal{W}_P \equiv \text{Range}(\Phi)$ *where:*

$$\Phi \equiv \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_p \end{bmatrix} \quad \text{where} \quad C_i \Phi_i = \mathcal{R}_i^c \quad \text{for} \ \ 1 \le i \le p,$$

where Φ_l *is of size* $n_B^{(l)} \times q_0$ *and* Φ *is of size* $n_B \times q_0$ *with* $\dim(\text{Range}(\Phi)) = q_0$.

Remark 4.31. By construction, if $\mathbf{v}_B \in \mathcal{W}_*$, then there exists $\mathbf{u} \in \mathbb{R}^{q_0}$ such that $C_i \mathbf{v}_B^{(i)} = \mathcal{R}_i^c \mathbf{u}$. Thus, $\mathbf{w}_i \equiv (\mathbf{v}_B^{(i)} - \Phi_i \mathbf{u})$ will satisfy $C_i \mathbf{w}_i = \mathbf{0}$ for $1 \le i \le p$, yielding that $(\mathbf{w}_1^T, \ldots, \mathbf{w}_p^T)^T \in \mathcal{W}_D$. Thus $\mathcal{W}_P$ and $\mathcal{W}_D$ are complementary:

$$\mathcal{W}_* = \mathcal{W}_P + \mathcal{W}_D.$$

As a result, each $\mathbf{u}_B \in \mathcal{W}_*$ may be decomposed and sought in the form:

$$\mathbf{u}_B = \mathbf{u}_D + \Phi\,\mathbf{u}_c \quad \text{where} \quad C\,\mathbf{u}_D = \mathbf{0} \quad \text{and} \ \ \Phi\,\mathbf{u}_c \in \mathcal{W}_P.$$

Remark 4.32. The subspace $\mathcal{W}_0 \equiv \text{Kernel}(M_B)$ satisfies:

$$\mathcal{W}_0 \subset \mathcal{W}_D \subset \mathcal{W}_* \subset \mathcal{W}.$$

Remark 4.33. The FETI-DP and BDDC methods will employ the following property. The minimization of $J_B(\mathbf{v}_B) = \frac{1}{2}\mathbf{v}_B^T S_{\mathcal{E}\mathcal{E}} \mathbf{v}_B - \mathbf{v}_B^T \tilde{\mathbf{f}}_B$ subject to the *constraint* that $C\,\mathbf{v}_B = \mathbf{0}$ can be reduced to p concurrent local problems, since $S_{\mathcal{E}\mathcal{E}} = \text{blockdiag}(S^{(1)}, \ldots, S^{(p)})$ and $C = \text{blockdiag}(C_1, \ldots, C_p)$. Indeed, let $\mathbf{v}_B = (\mathbf{v}_B^{(1)^T}, \ldots, \mathbf{v}_B^{(p)^T})^T$, $\tilde{\mathbf{f}}_B = (\tilde{\mathbf{f}}_B^{(1)^T}, \ldots, \tilde{\mathbf{f}}_B^{(p)^T})^T$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_p^T)^T$. Then, by reordering the system, the solution to:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_B \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \mathbf{0} \end{bmatrix} \tag{4.45}$$

reduces to the solution of:

$$\begin{bmatrix} S^{(i)} & C_i^T \\ C_i & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_B^{(i)} \\ \boldsymbol{\mu}_i \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B^{(i)} \\ \mathbf{0} \end{bmatrix} \quad \text{for} \quad 1 \le i \le p. \tag{4.46}$$

If $\tilde{\mathbf{f}}_B = S_{\mathcal{E}\mathcal{E}} \mathbf{w}_B$ and $S_{\mathcal{E}\mathcal{E}}$ is positive definite within $\mathcal{W}_D$, then it is easily verified that $\mathbf{u}_B = P_{\mathcal{W}_D} \mathbf{w}_B$ where $P_{\mathcal{W}_D}$ denotes the $S_{\mathcal{E}\mathcal{E}}$-orthogonal projection onto $\mathcal{W}_D$. Henceforth, we assume that matrix $S_{\mathcal{E}\mathcal{E}}$ is positive definite within $\mathcal{W}_*$.

FETI-DP Method. The FETI-DP method seeks the solution $(\mathbf{u}_B^T, \boldsymbol{\lambda}^T)^T$ to (4.44) by maximizing a dual function $\mathcal{F}(\boldsymbol{\lambda})$ associated with (4.44) using a PCG algorithm to determine $\boldsymbol{\lambda} \in \mathbb{R}^m$. It is based on the decomposition $\mathbf{u}_B = \mathbf{u}_D + \Phi\,\mathbf{u}_c$ where $\mathbf{w}_D \in \mathcal{W}_D$ with $C\,\mathbf{w}_D = \mathbf{0}$ and $\Phi\,\mathbf{u}_c \in \mathcal{W}_P$. We recall the saddle point problem (4.44) with $\tilde{\mathbf{f}}_B \equiv \left(\mathbf{f}_B - A_{IB}^T A_{II}^{-1} \mathbf{f}_I\right)$:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & M_B^T \\ M_B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_B \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \mathbf{0} \end{bmatrix}. \tag{4.47}$$

The Lagrangian function associated with the above saddle point problem is:

$$\mathcal{L}(\mathbf{u}_B, \boldsymbol{\lambda}) = \frac{1}{2}\,\mathbf{u}_B^T S_{\mathcal{E}\mathcal{E}} \mathbf{u}_B - \mathbf{u}_B^T \tilde{\mathbf{f}}_B + \boldsymbol{\lambda}^T M_B \mathbf{u}_B.$$

Since the constraint $M_B\,\mathbf{u}_B = \mathbf{0}$ yields $\mathbf{u}_B \in \mathcal{W}_0 \subset \mathcal{W}_*$, we may alternatively minimize the functional within $\mathcal{W}_*$ subject to the constraint $M_B\,\mathbf{u}_B = \mathbf{0}$. The FETI-DP method seeks $\mathbf{u}_B = \mathbf{u}_D + \mathbf{u}_P$ where $C\,\mathbf{u}_D = \mathbf{0}$ and $\mathbf{u}_P = \Phi\,\mathbf{u}_c$. The constraint $C\,\mathbf{u}_D = \mathbf{0}$ can be imposed by augmenting the Lagrangian with the term $\boldsymbol{\mu}^T C \mathbf{u}_D$ for $\boldsymbol{\mu} \in \mathbb{R}^{q_0^{(1)}+\cdots+q_0^{(p)}}$. This will alter $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, but not $\mathbf{u}_B$. Seeking the saddle point of the augmented Lagrangian:

$$\mathcal{L}_{aug}(\mathbf{u}_D, \mathbf{u}_c, \boldsymbol{\mu}, \boldsymbol{\lambda}) \equiv \mathcal{L}(\mathbf{u}_D + \Phi\,\mathbf{u}_c, \boldsymbol{\lambda}) + \boldsymbol{\mu}^T C \mathbf{u}_D$$

will yield the following saddle point system:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & S_{\mathcal{E}\mathcal{E}}\,\Phi & C^T & M_B^T \\ \Phi^T S_{\mathcal{E}\mathcal{E}} & \Phi^T S_{\mathcal{E}\mathcal{E}} \Phi & 0 & \Phi^T M_B^T \\ C & 0 & 0 & 0 \\ M_B & M_B \Phi & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_D \\ \mathbf{u}_c \\ \boldsymbol{\mu} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \Phi^T \tilde{\mathbf{f}}_B \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{4.48}$$

Rearranging the unknowns as $(\mathbf{u}_D^T, \boldsymbol{\mu}^T, \mathbf{u}_c^T, \boldsymbol{\lambda}^T)^T$ results in the system:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T & S_{\mathcal{E}\mathcal{E}} \Phi & M_B^T \\ C & 0 & 0 & 0 \\ \Phi^T S_{\mathcal{E}\mathcal{E}} & 0 & \Phi^T S_{\mathcal{E}\mathcal{E}} \Phi & \Phi^T M_B^T \\ M_B & 0 & M_B \Phi & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_D \\ \boldsymbol{\mu} \\ \mathbf{u}_c \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \mathbf{0} \\ \Phi^T \tilde{\mathbf{f}}_B \\ \mathbf{0} \end{bmatrix}. \tag{4.49}$$

The FETI-DP method solves the above system by solving a symmetric positive definite system for determining $\boldsymbol{\lambda} \in \Lambda = \mathbb{R}^m$ by a PCG method. In the following, we express system (4.49) more compactly as:

$$\begin{bmatrix} K & L^T \\ L & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{g} \\ \mathbf{0} \end{bmatrix} \tag{4.50}$$

where the matrices K and L and the vectors $\mathbf{x}$ and $\mathbf{g}$ are as described next.

$$K \equiv \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T & S_{\mathcal{E}\mathcal{E}} \Phi \\ C & 0 & 0 \\ \Phi^T S_{\mathcal{E}\mathcal{E}} & 0 & \Phi^T S_{\mathcal{E}\mathcal{E}} \Phi \end{bmatrix},\ L^T \equiv \begin{bmatrix} M_B^T \\ 0 \\ \Phi^T M_B^T \end{bmatrix},\ \mathbf{x} \equiv \begin{bmatrix} \mathbf{u}_D \\ \boldsymbol{\mu} \\ \mathbf{u}_c \end{bmatrix},\ \mathbf{g} \equiv \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \mathbf{0} \\ \Phi^T \tilde{\mathbf{f}}_B \end{bmatrix}. \tag{4.51}$$

The FETI-DP method seeks the solution to (4.44) by eliminating $\mathbf{x}$ and by solving the resulting reduced system $F\,\boldsymbol{\lambda} = \mathbf{d}$ for $\boldsymbol{\lambda}$ by a PCG method, where the inverse of the preconditioner is $M_D S_{\mathcal{E}\mathcal{E}} M_D^T$. Here, $F\,\boldsymbol{\lambda} = \mathbf{d}$ arises as the condition for maximizing the dual function $\mathcal{F}(\boldsymbol{\lambda}) = \inf_{\mathbf{x}} \mathcal{L}_{\text{aug}}(\mathbf{x}, \boldsymbol{\lambda})$:

$$F \equiv (LK^{-1}L^T) \quad \text{and} \quad \mathbf{d} \equiv (LK^{-1}\mathbf{g}).$$

Once $\boldsymbol{\lambda} \in \mathbb{R}^m$ is determined, we obtain $\mathbf{x} = K^{-1}(\mathbf{g} - L^T\,\boldsymbol{\lambda})$. By definition, matrix $F = F^T$, however, it will be positive definite, see [FA11, FA10, MA19] and Remark 4.36. In the following, we elaborate on the action of K^{-1}.

Remark 4.34. Matrix K is a saddle point matrix, and is indefinite. However, within the subspace $C\,\mathbf{v}_D = \mathbf{0}$, matrix K can be verified to be positive definite. A system of the form $K\,\mathbf{x} = \mathbf{g}$ can be solved by duality, as follows:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T & S_{\mathcal{E}\mathcal{E}}\,\Phi \\ C & 0 & 0 \\ \Phi^T S_{\mathcal{E}\mathcal{E}} & 0 & \Phi^T S_{\mathcal{E}\mathcal{E}}\,\Phi \end{bmatrix} \begin{bmatrix} \mathbf{u}_D \\ \boldsymbol{\mu} \\ \mathbf{u}_c \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{0} \\ \mathbf{g}_2 \end{bmatrix} \tag{4.52}$$

Given $\mathbf{u}_c$, we may solve the first two block rows above to obtain:

$$\begin{bmatrix} \mathbf{u}_D \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{g}_1 - S_{\mathcal{E}\mathcal{E}}\,\Phi\,\mathbf{u}_c \\ \mathbf{0} \end{bmatrix}. \tag{4.53}$$

Substituting this into the third block row yields the reduced system for $\mathbf{u}_c$:

$$\begin{cases} S_c\,\mathbf{u}_c = \mathbf{g}_c, \quad \text{where} \\ S_c = \left(\Phi^T S_{\mathcal{E}\mathcal{E}}\,\Phi - \begin{bmatrix} S_{\mathcal{E}\mathcal{E}}\,\Phi \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} S_{\mathcal{E}\mathcal{E}}\,\Phi \\ \mathbf{0} \end{bmatrix} \right) \\ \mathbf{g}_c = \mathbf{g}_2 - \begin{bmatrix} S_{\mathcal{E}\mathcal{E}}\,\Phi \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{0} \end{bmatrix}. \end{cases} \tag{4.54}$$

Once $S_c\,\mathbf{u}_c = \mathbf{g}_c$ is solved, we may determine $(\mathbf{u}_D^T, \boldsymbol{\mu}^T)^T$ by solving (4.53). Matrix S_c of size q_0 can be shown to be sparse and can be assembled in parallel, since $S_{\mathcal{E}\mathcal{E}}$ and C are both block diagonal, see Remark 4.38. By Remark 4.33, it will hold that $S_c = \Phi^T (I - P_{\mathcal{W}_D})^T S_{\mathcal{E}\mathcal{E}} (I - P_{\mathcal{W}_D})\Phi$, since $C\,\mathbf{u}_D = \mathbf{0}$. As a result, S_c will be positive definite within $\mathcal{W}_P$, see [DO, DO2, MA18, MA19].

Remark 4.35. The FETI-DP preconditioner F_0 for F is chosen so that both the FETI-DP and BDDC preconditioned matrices have the same spectra. Let $D = \text{blockdiag}(D^{(1)}, \ldots, D^{(p)}) : \mathcal{W} \rightarrow \mathcal{W}$ be a *discrete partition of unity*:

$$\mathcal{R}_{\mathcal{E}}^T\, D\, \mathcal{R}_{\mathcal{E}} = I,$$

where each $D^{(l)} : W_l \to W_l$ is a diagonal matrix with non-negative diagonal entries. Such weight matrices are employed in the BDDC method to average the solution on different subdomain boundaries. Diagonal *dual* weight matrices $\mathcal{D}_*^{(l)} : \Lambda \to \Lambda$, each of size m, are defined based on the entries of the matrices $D^{(j)}$ as follows. Recall that each row of M_B is associated with a matching requirement between nodal values on two distinct subdomains. Suppose that a node α on B lies on $B^{(l)} \cap B^{(j)}$, and that $\mathrm{ind}(\alpha, l, j)$ denotes the row index in M_B which enforces the matching between the local nodal values at α in $B^{(l)}$ and in $B^{(j)}$. Let $\mathrm{ind}(\alpha, j)$ denote the index of the node α in the local ordering in $B^{(j)}$. We define the diagonal dual matrix $\mathcal{D}_*^{(l)}$ for all $\alpha \in B$ as:

$$(\mathcal{D}_*^{(l)})_{\mathrm{ind}(\alpha,\mathrm{l},\mathrm{j})} \equiv \begin{cases} (D^{(j)})_{\mathrm{ind}(\alpha,\mathrm{j})} & \text{if } \alpha \in B^{(l)} \cap B^{(j)} \\ 0 & \text{if } \alpha \notin B^{(l)} \end{cases}$$

Let M_D be a matrix the same size as M_B defined by:

$$M_D \equiv \left[\mathcal{D}_*^{(1)} M^{(1)} \ \cdots \ \mathcal{D}_*^{(p)} M^{(p)} \right].$$

Then, it can be shown that $M_B M_D^T M_B = M_B$ and $M_D^T M_B + \mathcal{R}_\mathcal{E} \mathcal{R}_\mathcal{E}^T D = I$, see [RI5, KL10, FR]. The inverse F_0^{-1} of the FETI-DP preconditioner for F is:

$$F_0^{-1} \equiv M_D S_{\mathcal{E}\mathcal{E}} M_D^T \implies \mathrm{cond}(F_0, F) \le c\,(1 + \log^2(h_0/h)).$$

Remark 4.36. Matrix F can be verified to be positive definite as follows. We express $F = (LK^{-1})K(K^{-1}L^T)$ and for $\boldsymbol{\lambda} \in \mathbb{R}^m$ let:

$$\mathbf{x} = (\mathbf{w}_D^T, \tilde{\boldsymbol{\mu}}^T, \mathbf{w}_c^T)^T = K^{-1} L^T \boldsymbol{\lambda}.$$

Then, since $C\,\mathbf{w}_D = \mathbf{0}$, we will obtain that:

$$\mathbf{x}^T K \mathbf{x} = \begin{bmatrix} \mathbf{w}_D \\ \mathbf{w}_c \end{bmatrix}^T \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & S_{\mathcal{E}\mathcal{E}} \Phi \\ \Phi^T S_{\mathcal{E}\mathcal{E}} & \Phi^T S_{\mathcal{E}\mathcal{E}} \Phi \end{bmatrix} \begin{bmatrix} \mathbf{w}_D \\ \mathbf{w}_c \end{bmatrix}. \tag{4.55}$$

The latter will be positive provided $S_{\mathcal{E}\mathcal{E}}$ is positive definite within $\mathcal{W}_*$ (which we assume to hold) and provided $(\mathbf{w}_D^T, \mathbf{w}_c^T)^T \neq \mathbf{0}$ for $\boldsymbol{\lambda} \neq \mathbf{0}$.

BDDC Method. The BDDC method [DO, DO2, MA18, MA19] is a PCG method to solve the *primal* problem associated with system (4.44). Since $M_B \mathbf{u}_B = \mathbf{0}$, we may seek $\mathbf{u}_B = \mathcal{R}_\mathcal{E} \mathbf{u}$ for some $\mathbf{u} \in \mathcal{U} = \mathbb{R}^q$. Substituting this, the primal problem associated with (4.44) can easily be verified to be the Schur complement system arising in traditional substructuring:

$$S\,\mathbf{u} = \mathbf{f} \quad \text{where} \quad S \equiv (\mathcal{R}_\mathcal{E}^T S_{\mathcal{E}\mathcal{E}} \mathcal{R}_\mathcal{E}) \quad \text{and} \quad \mathbf{f} = (\mathcal{R}_\mathcal{E}^T \tilde{\mathbf{f}}_B). \tag{4.56}$$

The BDDC preconditioner S_0 is formulated using the same coarse space and local saddle point problems employed in the FETI-DP method. Matrix $S_0^{-1} S$

in the BDDC method has essentially the same spectrum as the preconditioned matrix $F_0^{-1}F$ in the FETI-DP method, except for zeros or ones.

The BDDC preconditioner employs a *discrete partition of unity* matrix on B, with $D = \text{blockdiag}(D^{(1)}, \ldots, D^{(p)})$, where each $D^{(l)} : W_l \rightarrow W_l$ is a diagonal matrix with non-negative entries, satisfying:

$$\mathcal{R}_{\mathcal{E}}^T \, D \, \mathcal{R}_{\mathcal{E}} = I.$$

In practice, the diagonal entries of $D^{(l)}$ are chosen as a weighted average of the diagonal entries of the stiffness matrices $A^{(j)}$. Let i denote the index of a node on $B^{(l)}$ and $j(i) = \text{ind}(B^{(j)}, i)$ the local index of node i in $B^{(j)}$. Then:

$$(D^{(l)})_{ii} = \frac{A_{ii}^{(l)}}{\sum_{\{j: B^{(j)} \cap B^{(l)} \neq \emptyset\}} A_{j(i)\, j(i)}^{(j)}}.$$

The BDDC preconditioner also employs a coarse basis Ψ of size $n_B \times q_0$ obtained by modifying the matrix Φ of size $n_B \times q_0$ which satisfies $C\,\Phi = \mathcal{R}^c$:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \Psi \\ G \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{R}^c \end{bmatrix}.$$

If we expand $\Psi = \Phi + \hat{\Phi}$, then since $C\,\Phi = \mathcal{R}^c$, matrix $\hat{\Phi}$ will satisfy:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \hat{\Phi} \\ G \end{bmatrix} = \begin{bmatrix} -S_{\mathcal{E}\mathcal{E}}\,\Phi \\ 0 \end{bmatrix}.$$

Solving for $\hat{\Phi}$, and computing $\Psi^T S_{\mathcal{E}\mathcal{E}} \Psi$ after algebraic simplification yields:

$$\Psi^T S_{\mathcal{E}\mathcal{E}} \Psi = \left(\Phi^T \, S_{\mathcal{E}\mathcal{E}} \, \Phi - \begin{bmatrix} S_{\mathcal{E}\mathcal{E}}\,\Phi \\ 0 \end{bmatrix}^T \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} S_{\mathcal{E}\mathcal{E}}\,\Phi \\ 0 \end{bmatrix} \right) = S_c.$$

Employing Remark 4.33, it follows that $\tilde{\Phi} = -P_{\mathcal{W}_D}\,\Phi$ and $\Psi = (I - P_{\mathcal{W}_D})\,\Phi$. The coarse matrix S_c, can be assembled using either expression above.

The BDDC preconditioner S_0 for S corresponds to an *additive Schwarz* preconditioner with inexact solvers, based on the following subspaces of $\mathcal{U}$:

$$\begin{cases} \mathcal{U}_0 = \text{Range}(\mathcal{R}_{\mathcal{E}}^T \, D^T \, \Psi) \\ \mathcal{U}_i = \left\{ \mathcal{R}_i^T D^{(i)} \mathbf{w}_i : C_i \, \mathbf{w}_i = \mathbf{0}, \ \ \mathbf{w}_i \in W_i \right\}, \quad \text{for } \ 1 \leq i \leq p. \end{cases}$$

The spaces Range(Ψ) and $\mathcal{W}$ consist of nodal vectors associated with finite element functions which are *discontinuous* across the subdomain boundaries. However, the weighted averaging using $\mathcal{R}_i^T D^{(i)}$ or $\mathcal{R}^T D^T$ yields nodal vectors in $\mathcal{U}$, associated with continuous finite element functions on B. The action S_0^{-1} of the inverse of the BDDC preconditioner for S is defined as:

$$S_0^{-1}\,\mathbf{r} \equiv \mathcal{R}_{\mathcal{E}}^T\,D\,\Psi\,S_c^{-1}\,\Psi^T\,D^T\,\mathcal{R}_{\mathcal{E}}\,\mathbf{r} + \begin{bmatrix} D^T\mathcal{R}_{\mathcal{E}} \\ 0 \end{bmatrix}^T \begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} D^T\mathcal{R}_{\mathcal{E}} \\ 0 \end{bmatrix} \mathbf{r}.$$

The block diagonal structure of $S_{\mathcal{E}\mathcal{E}}$ and C yields its parallel form:

$$S_0^{-1} = \mathcal{R}_{\mathcal{E}}^T\,D\,\Psi\,S_c^{-1}\,\Psi^T\,D^T\,\mathcal{R}_{\mathcal{E}} + \sum_{i=1}^{p} \begin{bmatrix} D^{(i)^T}\mathcal{R}_i \\ 0 \end{bmatrix}^T \begin{bmatrix} S^{(i)} & C_i^T \\ C_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} D^{(i)^T}\mathcal{R}_i \\ 0 \end{bmatrix}.$$

The following bounds will hold for the FETI-DP and BDDC methods.

Lemma 4.37. *The following convergence bounds will hold:*

$$\frac{\lambda_{\max}(S_0^{-1}S)}{\lambda_{\min}(S_0^{-1}S)} = \frac{\lambda_{\max}(F_0^{-1}F)}{\lambda_{\min}(F_0^{-1}F)} \le \kappa$$

and

$$\kappa \le \sup_{\mathbf{w}\in\mathcal{W}_*} \frac{\|M_D^T M_B \mathbf{w}\|^2_{\mathcal{S}_{\mathcal{E}\mathcal{E}}}}{\|\mathbf{w}\|^2_{\mathcal{S}_{\mathcal{E}\mathcal{E}}}} = \sup_{\mathbf{w}\in\mathcal{W}_*} \frac{\|\mathcal{R}_{\mathcal{E}}\mathcal{R}_{\mathcal{E}}^T D \mathbf{w}\|^2_{\mathcal{S}_{\mathcal{E}\mathcal{E}}}}{\|\mathbf{w}\|^2_{\mathcal{S}_{\mathcal{E}\mathcal{E}}}},$$

where $\kappa \le c(1+\log^2(h_0/h))$ and h_0 is the diameter of the subdomains.

Proof. See [DO, DO2, MA18, MA19].

Remark 4.38. The columns of matrix Ψ of size $n_B \times q_0$ can be constructed as follows. If $\mathbf{e}_j$ denotes the j'th column of the identity matrix I of size q_0, then the j'th column $\boldsymbol{\psi}_j$ of Ψ can be computed by solving:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\psi}_j \\ \boldsymbol{\mu}_j \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathcal{R}^c\,\mathbf{e}_j \end{bmatrix}.$$

The components of $\boldsymbol{\psi}_j$ will be non-zero only on the boundaries $B^{(l)}$ which intersect the glob associated with the j'th column of $\mathcal{R}^c$. Thus, using the block structure of $S_{\mathcal{E}\mathcal{E}}$ and C, only a few local problems need to be solved. The non-zero entries of the sparse matrix S_c of size q_0 can be computed as $(S_c)_{ij} = \boldsymbol{\psi}_i^T S_{\mathcal{E}\mathcal{E}} \boldsymbol{\psi}_j$ based on the support of $\boldsymbol{\psi}_i$ and $\boldsymbol{\psi}_j$.

Remark 4.39. In applications, each local saddle point problem:

$$\begin{bmatrix} S^{(i)} & C_i^T \\ C_i & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \boldsymbol{\mu}_i \end{bmatrix} = \begin{bmatrix} \mathbf{f}_i \\ \mathbf{g}_i \end{bmatrix}, \tag{4.57}$$

can be solved using the Schur complement method. On each $B^{(i)}$, the entries of $\mathbf{w}_i$ corresponding to the cross-points on $B^{(i)}$ can be eliminated. The entries of the Lagrange multiplier variables enforcing the constraints on the cross points can also be eliminated. Thus, the specified rows of $S^{(i)}$ and C_i^T and associated columns of $S^{(i)}$ and C_i must be eliminated. The resulting submatrix of $S^{(i)}$

will be non-singular (even if $S^{(i)}$ were singular). For notational convenience, we shall denote the resulting saddle point system as in the above. To solve for the remaining entries of $\mathbf{w}_i$ and $\boldsymbol{\mu}_i$, parameterize $\mathbf{w}_i$ in terms of $\boldsymbol{\mu}_i$ using the first block row. This formally yields:

$$\mathbf{w}_i = S^{(i)^{-1}}(\mathbf{f}_i - C_i^T \boldsymbol{\mu}_i).$$

Substituting this expression into the second block row yields:

$$T_i \, \boldsymbol{\mu}_i = \left(C_i S^{(i)^{-1}} \mathbf{f}_i - \mathbf{g}_i \right) \quad \text{where} \quad T_i \equiv (C_i S^{(i)^{-1}} C_i^T).$$

The Schur complement T_i of size $q_0^{(i)}$ can be assembled explicitly, $q_0^{(i)}$ will be at most eight for rectangular subdomains when $\Omega \subset \mathbb{R}^2$ or of size twenty-six when $\Omega \subset \mathbb{R}^3$. Once $\boldsymbol{\mu}_i$ is determined, $\mathbf{w}_i = S^{(i)^{-1}} (\mathbf{f}_i - C_i^T \boldsymbol{\mu}_i)$. Note that matrix $S^{(i)} = (A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)})$ need not be assembled. Instead, the solution of the system can be obtained by solving the sparse system:

$$\begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \\ A_{IB}^{(i)^T} & A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{y}_i \\ \mathbf{w}_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_i - C_i^T \boldsymbol{\mu}_i \end{bmatrix}.$$

Thus, the solution to (4.57) can be sought by solving two sparse symmetric positive definite systems and one dense symmetric positive definite system of a small size. See [DO, DO2, MA18, MA19] for alternative methods.

5

Computational Issues and Parallelization

In this chapter, we discuss several computational issues that arise with the implementation of domain decomposition algorithms. The first issue concerns the choice of a decomposition of a domain into non-overlapping or overlapping subdomains. When an algorithm is implemented using multiple processors, the number of interior unknowns per subdomain must be approximately the same, to ensure load balancing, while the number of boundary unknowns must be minimized to reduce inter-subdomain communication. We describe graph partitioning algorithms which partition a grid. The second issue concerns the expected parallel computation time and speed up when implementing a domain decomposition preconditioner on an idealized parallel computer architecture. We outline *heuristic* estimate for this using idealized models for the computational time and inter-processor data transfer times.

Chap. 5.1 presents background on grid generation and graph theory, and describes how the problem of partitioning a domain or an unstructured grid can be heuristically reduced to a graph partitioning algorithm. We then describe the Kernighan-Lin, recursive spectral bisection and multilevel graph partitioning algorithms for partitioning graphs. Following that, we brief discuss the implementation of Schwarz and Schur complement algorithms on unstructured grids. Some heuristic coarse spaces are also outlined for use on unstructured grids, with subdomains of irregular shapes.

Chap. 5.2 discusses background on the speed up and scalability of algorithms on parallel computers. Employing a heuristic model of an idealized parallel computer with distributed memory, we describe models for the computational time required for implementing various domain decomposition preconditioners. Under such idealized assumptions, it is shown that domain decomposition iterative algorithms have reasonable scalability.

5.1 Algorithms for Automated Partitioning of Domains

The term unstructured grid refers broadly to triangulations without any identifiable structure. Such grids arise in computational fluid dynamics [PE4] and aerodynamics computations [ST6, BA22, MA40, MA41], which require the triangulation and discretization of partial differential equations on regions having complex geometry. In such applications, the triangulation is generated by using grid generation software [GE6, HO2, MA40, MA41, HE9, TH3, OW]. The resulting grids are typically not quasiuniform, and lack the connectivity of uniform grids and the hierarchical structure of multigrids, see Fig. 5.1.[1] Additionally, the density of grid points and the number of elements incident to each node can vary significantly with location. As a result, algorithms are required to automate the partitioning of a domain into subdomains, so that the number of grid points per subdomain is approximately the same.

In this section, we discuss several practical techniques for implementing domain decomposition solvers on *unstructured* grids. We discuss the selection of subdomains so that *load balancing* constraints are satisfied, and so that the *communication* time between processors assigned to different subdomains is minimized. This issue is typically addressed by employing heuristic graph partitioning algorithms. We also discuss the formulation of heuristic *coarse spaces* for elliptic equations discretized on unstructured grids with subdomains having irregular boundaries, where traditional coarse spaces are not defined. Chap. 5.1.1 describes grid generation algorithms, followed by graph partitioning algorithms in Chap. 5.1.2. Chap. 5.1.3 describes the construction of subdomains, while a few coarse spaces are described for unstructured grids in Chap. 5.1.4. Comments on Schwarz, Schur complement and FETI algorithms are presented in Chap. 5.1.5 to Chap. 5.1.7.

5.1.1 Grid Generation Algorithms

Generating a triangulation $\mathcal{T}_h(\Omega)$ on a domain Ω in two or three dimensions with complex geometry, is generally a computationally intensive task. There is

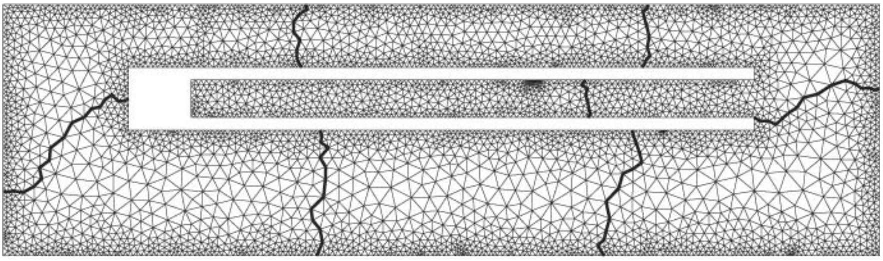


Fig. 5.1. An unstructured grid [BA23]

[1] The author thanks Dr. Timothy Barth for his kind permission to use Fig. 5.1.

an extensive literature on algorithms and software for automated generation of grids [GE6, HO2, MA40, HE9, MA41, TH3, OW]. Below, we list a few.

- *Grid based method.* In this method, a uniform or structured simplicial or box type grid $\mathcal{T}_h(\Omega^*)$ with a specified grid size h is overlaid on an extended domain $\Omega^* \supset \Omega$, and the triangulation $\mathcal{T}_h(\Omega^*)$ of Ω^* is modified to conform to the boundary $\partial\Omega$. The resulting triangulation of Ω will be of low cost, however, it can be of poor quality for numerical approximation.
- *Decomposition and mapping method.* One of the earliest methods, here the domain is decomposed into subregions, and each subregion is mapped onto one or more standard reference regions. A structured triangulation of each reference domain is then mapped back to triangulate the original subdomains. However, the subdomain triangulations may not match near their boundaries, so that the triangulations must be appropriately modified.
- *Advancing front method.* In this method, the boundary $\partial\Omega$ of the domain is first triangulated (for instance, by the decomposition and mapping method), yielding an initial *front* of the triangulation. The algorithm then advances (updates) these fronts by generating new nodes and elements of a desired size within the interior of the domain, and adjacent to the current front. The algorithm terminates when the entire domain is triangulated.
- *Delaunay triangulation method.* A Delaunay triangulation is a simplicial triangulation (triangles in $\mathbb{R}^2$ or tetrahedra in $\mathbb{R}^3$) such that any circumsphere (i.e., a sphere in $\mathbb{R}^3$ or a circle in $\mathbb{R}^2$ passing though the nodes of a tetrahedra or triangle) do not contain other nodes in the interior. Many Delaunay triangulation algorithms are available, some based on the computation of Voronoi cells (polyhedral cells consisting of all points in Euclidean space closest to a node). Typically, in the first phase nodes are placed on the boundary $\partial\Omega$ of the domain (for instance, by the decomposition and mapping method) and new nodes are introduced within the interior, using the advancing front method (or alternative methods). In the second phase, a Delaunay triangulation $\mathcal{T}_h(\Omega)$ of Ω is constructed using the given distribution of nodes.

Automatic mesh generation software may combine one or more of the above methods and include a phase of refinement or smoothing of the resulting grid, depending on the geometry and specifications for the grid size. As a result, the generated grid may not be quasiuniform or structured. Readers are referred to [GE6, HO2, MA40, HE9, MA41, TH3] for literature on unstructured meshes and to [OW] for a survey of software algorithms.

5.1.2 Graph Partitioning Algorithms

The problem of decomposing a domain Ω into subdomains, or partitioning an index set of nodes $\mathcal{I} = \{x_1, \ldots, x_n\}$ into subindex sets, can be formulated mathematically as a *graph partitioning* problem. Given a triangulation $\mathcal{T}_h(\Omega)$ of Ω, a graph [BO2] (or a weighted graph) can be constructed representing the connectivity of the triangulation (either connectivity of *elements* or of the *nodes* within the triangulation). A partition of the domain, or of the index set $\mathcal{I}$ of nodes in the triangulation, may then be obtained by partitioning this associated graph (or weighted graph) into subgraphs. Load balancing requirements can be incorporated by requiring that the subgraphs be approximately of equal size, while minimization of communication costs can be imposed by requiring that the number of edges cut between subgraphs in the partition is minimized [FA9, MA30]. Formally, this problem can be formulated as a combinatorial minimization of an objective functional incorporating the above requirements. Here, we introduce the graph partitioning problem, its associated combinatorial minimization problem, and describe three *heuristic* algorithms for its solution. The reader is referred to [PO3] for details.

Definition 5.1. *A graph $G = (V, E)$ consists of a collection V of n vertices*

$$V = \{v_1, \cdots, v_n\},$$

and a collection E of m edges

$$E = \{e_1, \cdots, e_m\},$$

where each edge represents adjacencies between pairs of vertices. Thus, if edge e_l is incident to vertices v_i and v_j we denote it as $e_l = (v_i, v_j) = (v_j, v_i) \in E$. The order of the graph, denoted by $|V|$, refers to the number n of vertices, while the size of the graph, denoted $|E|$, refers to the number m of edges.

Given a graph G of order n, the adjacencies in E may be represented using an $n \times n$ symmetric matrix M_G referred to as the *adjacency* matrix;

$$(M_G)_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{if } (v_i, v_j) \notin E. \end{cases}$$

If the edges in E are enumerated as $e_1, \ldots, e_m$, then the vertices incident to each edge may be summarized in an $n \times m$ *incidence* matrix N_G:

$$(N_G)_{lj} = \begin{cases} 1, & \text{if edge } e_j \text{ is incident with vertex } v_l \\ 0, & \text{otherwise.} \end{cases}$$

The number of edges incident to a vertex v_i is referred to as the *degree* of the vertex and will be denoted as $d(v_i)$. In various applications, it will be useful to assign *weights* to edges and vertices in a graph. Such graphs are referred to as weighted graphs.

Definition 5.2. *A weighted graph is a graph $G = (V, E)$ with weights w_{ij} assigned to each edge $(v_i, v_j) \in E$. Such weights can be summarized in an $n \times n$ symmetric weight matrix W. Weights may also be assigned to individual vertices $v_i \in V$ and denoted by $w(v_i)$. By default, weights can be assigned to any graph $G = (V, E)$ by defining $w_{ij} = w_{ji} = 1$ if $(M_G)_{ij} = 1$ and $w(v_i) = 1$.*

Associated with a graph $G = (V, E)$ with a nonnegative weight matrix W, we define an $n \times n$ graph *Laplacian* matrix L_G as follows:

$$(L_G)_{ij} \equiv \begin{cases} \sum_{l \neq i} w_{il}, & \text{if } j = i \\ -w_{ij}, & \text{if } (v_i, v_j) \in E \\ 0, & \text{if } (v_i, v_j) \notin E \text{ and } i \neq j. \end{cases} \tag{5.1}$$

If the graph is unweighted, then the default weights $w_{ij} = (M_G)_{ij}$ for $i \neq j$, given by the adjacency matrix, should be used, and in this case the diagonal entries $(L_G)_{ii} = d(v_i)$ will correspond to the degrees of the vertices. By definition, L_G will be symmetric and weakly diagonally dominant with zero row sums. Consequently L_G will be singular with eigenvector $\mathbf{x}_1 = (1, \cdots, 1)^T$ corresponding to eigenvalue $\lambda_1 = 0$. Due to symmetry and weak diagonal dominance of L_G, its eigenvalues $\{\lambda_i\}$ will be nonnegative. We assume that these eigenvalues are ordered as:

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Definition 5.3. *A graph $G = (V, E)$ is said to be connected if for any two vertices $v_i, v_j \in V$ there exists a "path" $v_i = x_1, x_2, \cdots, x_l = v_j$ such that all consecutive vertices are adjacent, i.e., $(x_r, x_{r+1}) \in E$ for $r = 1, \cdots, l-1$. In matrix terms, a graph G will be connected if and only if its adjacency matrix M_G is irreducible, i.e., any 2×2 block partitioning of matrix PM_GP^T must yield a nonzero off diagonal block for any permutation matrix P reordering the rows or columns.*

If a graph G is not connected, then the algebraic multiplicity of the zero eigenvalue of L_G will yield its number of connected components.

Definition 5.4. *For a connected graph G, the algebraic multiplicity of the zero eigenvalue of L_G will be one and $\lambda_2 > 0$. In this case, the eigenvector $\mathbf{x}_2$ of L_G corresponding to eigenvalue $\lambda_2 > 0$:*

$$L_G \, \mathbf{x}_2 = \lambda_2 \, \mathbf{x}_2,$$

is referred to as the Fiedler vector of the graph G.

The Fiedler vector of a connected graph can be employed to partition a graph into *two*, as shall be described later. This can be applied recursively.

In applications to the partitioning of a triangulation, two alternative graphs $G = (V, E)$ may be associated with a given triangulation Ω_h.

- In applications to Schwarz algorithms, let the vertices v_i in the graph correspond to nodes x_i of Ω_h. In this case vertices v_i and v_j can be defined to be adjacent if nodes x_i and x_j belong to the same element.
- In applications to Schur complement algorithms, it will be preferable to identify the vertices v_i of the graph with elements κ_i of triangulation Ω_h. Vertex v_i can be defined to be adjacent to vertex v_j if elements $\overline{\kappa}_i \cap \overline{\kappa}_j \neq \emptyset$.

More details of such associations will be described later.

Once a graph $G = (V, E)$ has been associated with a domain or the nodes in Ω_h, a partition of the domain or its nodes, can be obtained by *partitioning* the vertices of graph $G = (V, E)$ into p subsets $V_1, \cdots, V_p$ of order $n_1, \cdots, n_p$, respectively so that:

$$\begin{cases} V_1 \cup \cdots \cup V_p = V, \\ \qquad\quad V_i \cap V_j = \emptyset, \quad \text{if } i \neq j. \end{cases} \tag{5.2}$$

The *induced subgraph* on the vertices V_i (i.e., the adjacencies from E between vertices in V_i) will be required to be *connected.* The load balancing constraint can be heuristically approximated by requiring the number n_i of nodes within each subset V_i be approximately the same, as stated formally in the following.

Definition 5.5. *Given a graph $G = (V, E)$ and a parameter $\epsilon > 0$ chosen by the user, we define $\mathcal{K}_\epsilon$ as an admissible partition of V into p sets $V_1, \cdots, V_p$ of size $n_1, \cdots, n_p$, respectively, if the following hold:*

1. *If $n_i = |V_i|$ for $i = 1, \cdots, p$, then:*

$$(1-\epsilon)\frac{n}{p} \leq n_i \leq \frac{n}{p}(1+\epsilon), \qquad \textit{for } i = 1, \cdots, p.$$

2. *The induced subgraphs $G_i = (V_i, E_i)$ are connected, where each E_i denotes adjacencies from E between vertices in V_i.*

Remark 5.6. In some applications, it may be convenient to let each vertex in the graph represent more than one nodal unknown in the original triangulation Ω_h. In such cases, a weight $w(v_i)$ can be assigned to each vertex to denote the number of nodes that vertex v_i represents. Then, the number n_i of nodes which subset V_i represents should be computed as:

$$n_i = |V_i| = \sum_{v_l \in V_i} w(v_l). \tag{5.3}$$

This will reduce to the number of vertices in V_i if $w(v_l) = 1$.

If one processor is assigned to each subdomain defined by V_i, then the volume of communication between the different processors can be heuristically estimated in terms of the total number of edges between the vertices in different sets V_i in the partition. If weighted edges are used, this quantity may be replaced by the sum of the edge weights on edges between different sets V_i. The requirement that the communication between different subdomains be minimized may thus be approximated by minimizing the sum of such edge weights between distinct subsets V_i. Accordingly, we may define an objective functional $\delta(\cdot)$ which represents the sum of edge weights between distinct subsets V_i in the partition.

Definition 5.7. *Given a graph $G = (V, E)$ with weight matrix W and two disjoint vertex subsets V_i and V_j of V, we denote by $\delta(V_i, V_j)$ the total sum of edge weights between all pairs of vertices in V_i and V_j:*

$$\delta\left(V_i, V_j\right) \equiv \sum_{\{v_r \in V_i, v_s \in V_j\}} w_{rs}. \tag{5.4}$$

Given three or more disjoint vertex subsets of V, we define $\delta\left(V_1, \cdots, V_p\right)$ as the sum of edge weights between each distinct pair of subsets V_i and V_j:

$$\delta\left(V_1, \cdots, V_p\right) \equiv \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \delta\left(V_i, V_j\right). \tag{5.5}$$

The functional $\delta\left(V_1, \cdots, V_p\right)$ will represent the volume of communication between subsets in the partition.

If W is chosen by default with $w_{ij} = \left(M_G\right)_{ij}$, then $\delta\left(V_1, \cdots, V_p\right)$ will correspond to the total number of edges between all distinct pairs of subsets V_i and V_j of the partition of V. The problem of partitioning a graph G so that the load balancing constraint holds and so that the communication costs between subdomains is minimized, may formally be approximated by the following combinatorial minimization problem. Find a $\mathcal{K}_\epsilon$ partition $V_1, \ldots, V_p$ satisfying:

$$\delta\left(V_1, \cdots, V_p\right) = \min_{(\tilde{V}_1, \cdots, \tilde{V}_p) \in \mathcal{K}_\epsilon} \delta\left(\tilde{V}_1, \cdots, \tilde{V}_p\right). \tag{5.6}$$

Unfortunately, as with most combinatorial optimization problems, this is an NP hard discrete problem, see [PO3]. Consequently, no algorithm of polynomial complexity is known for determining the exact solution. We therefore restrict consideration to *heuristic* algorithms which approximate the solution to the above. The following three algorithms will be outlined in the following: the Kernighan-Lin algorithm, the recursive spectral bisection algorithm and the multilevel graph partitioning algorithm. The latter algorithm generally has the lowest complexity amongst the three.

Kernighan-Lin Algorithm. This algorithm [KE4], corresponds to a discrete descent method for the combinatorial minimization problem (5.6). Start with any initial partition $\tilde{V}_1, \cdots, \tilde{V}_p$ in $\mathcal{K}_\epsilon$. Repeatedly exchange pairs of vertices v_i and v_j for which the resulting partition is still within $\mathcal{K}_\epsilon$ and for which a reduction in the functional $\delta(\cdot)$ is obtained. If the vertex weights $w(v_i)$ are unitary then such an exchange will leave $n_1, \ldots, n_p$ unchanged, however, if nonunitary vertex weights are employed, this constraint must be checked. To avoid stagnation at a local minimum, the Kernighan-Lin algorithm permits a fixed number q_* of exchanges within $\mathcal{K}_\epsilon$ which increase the value of $\delta(\cdot)$. The algorithm must ideally be implemented for several selections of initial partitions, and the partition corresponding to the lowest value of $\delta(\cdot)$ must be stored. Once a prescribed number of iterations have been completed, this optimal stored partition can be chosen as an approximate solution of (5.6).

The complexity of the Kernighan-Lin algorithm is $O(n^2 \log(n))$ if a fixed number of iterations is implemented. However, the number of exchanges of vertices per iteration can be reduced significantly if only *boundary* vertices are exchanged, i.e., vertices which are adjacent to vertices in other sets of the partition. An $O(|E|)$ complexity algorithm is known for $p = 2, 4$, see [PO3].

Remark 5.8. To implement the Kernighan-Lin sweep, for any subset $\tilde{V}_i \subset V$ and vertex v_r define $d_{\tilde{V}_i}(v_r)$ as the sum of edge weights w_{rs} between vertex v_r and vertices v_s in $\tilde{V}_i$:

$$d_{\tilde{V}_i}(v_r) \equiv \sum_{\{(v_r,v_s)\in E : v_s \in \tilde{V}_i\}} w_{rs}.$$

Define the *gain* associated with exchanging $v_r \in \tilde{V}_i$ and $v_s \in \tilde{V}_j$ as follows:

$$\text{gain}(v_r, v_s) = \begin{cases} d_{\tilde{V}_i}(v_r) - d_{\tilde{V}_j}(v_r) + d_{\tilde{V}_j}(v_s) - d_{\tilde{V}_i}(v_s) & \text{if } (v_r, v_s) \notin E \\ d_{\tilde{V}_i}(v_r) - d_{\tilde{V}_j}(v_r) + d_{\tilde{V}_j}(v_s) - d_{\tilde{V}_i}(v_s) - 2w_{rs} & \text{if } (v_r, v_s) \in E. \end{cases}$$

If the gain is nonnegative, then the exchange should be accepted. At most q_* exchanges resulting in a negative gain should be accepted.

Recursive Spectral Bisection Algorithm. The recursive spectral bisection algorithm is a popular graph partitioning algorithm which repeatedly partitions a graph into *two* subgraphs [SI2, PO2, BA20, FI, FI2, BA21, BO3]. Each graph (or subgraph) is partitioned based on sorting the entries of the Fiedler vector of the graph (or subgraph). The partitions obtained by recursive spectral bisection are typically of very good quality as measured by $\delta(\cdot)$, however, the algorithm is ideally suited for $p \approx 2^J$ for integer $J \geq 1$, and is relatively expensive to implement due to computation of the Fiedler vector.

We motivate the spectral bisection algorithm by considering the partition of a graph $G = (V, E)$ with weight matrix W into *two* subgraphs so that (5.6) is minimized. For simplicity, we suppose that $|V|$ is an even integer and that all vertex weights $w(v_i)$ are unitary. In this case we seek $|V_1| = |V_2|$ and the partition is referred to as a *bisection*. Let L_G denote the weighted graph Laplacian matrix (5.1). Suppose V_1, V_2 is a solution to the graph bisection problem, then define a vector $\mathbf{q}$ as:

$$(\mathbf{q})_i \equiv \begin{cases} 1, & \text{if } v_i \in V_1 \\ -1, & \text{if } v_i \in V_2. \end{cases}$$

By construction, we obtain:

$$\begin{cases} \mathbf{q}^T L_G \mathbf{q} = \sum_{\{(v_i,v_j)\in E\}} w_{ij}(q_i - q_j)^2 \\ \qquad\quad = \sum_{\{v_i \in V_1, v_j \in V_2\}} w_{ij}\, 4 \\ \qquad\quad = 4\,\delta(V_1, V_2). \end{cases}$$

Additionally $\mathbf{q}^T\mathbf{1} = \sum_j q_j = 0$. Therefore, minimization of $\delta(V_1, V_2)$ over all admissible partitions, will be equivalent to the minimization of $\mathbf{q}^T L_G \mathbf{q}$ for $\mathbf{q} \in Q$ where:

$$Q \equiv \left\{ \tilde{\mathbf{q}} = (\tilde{q}_1, \ldots, \tilde{q}_n)^T : \tilde{q}_i = \pm 1, \;\; 1 \le i \le n, \; \tilde{\mathbf{q}}^T\mathbf{1} = 0 \right\}.$$

We may thus state the bisection problem as determining $\mathbf{q} \in Q$ such that:

$$\mathbf{q}^T L_G \mathbf{q} = \min_{\{\tilde{\mathbf{q}} \in Q\}} \tilde{\mathbf{q}}^T L_G \tilde{\mathbf{q}}. \tag{5.7}$$

This is called a quadratic assignment problem [PO3], and it is a discrete (combinatorial) optimization problem which may be heuristically *approximated* by a quadratic minimization problem over $\mathbb{R}^n$ (with appropriate constraints) as indicated next. Define $Q_* \equiv \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T\mathbf{x} = n, \; \mathbf{x}^T\mathbf{1} = 0\} \supset Q$. We obtain:

$$\begin{cases} \min_{\mathbf{q} \in Q} \mathbf{q}^T L_G \mathbf{q} \;\ge\; \min_{\mathbf{x}_2^T L_G \mathbf{x}_2} \mathbf{x}^T L_G \mathbf{x} & = \; \mathbf{x}_2^T L_G \mathbf{x}_2 \\ & = \; \lambda_2 \, \mathbf{x}_2^T \mathbf{x}_2 \\ & = \; \lambda_2 \, n, \end{cases}$$

where $\mathbf{x}_2$ is a *Fiedler* vector (i.e., an eigenvector of L_G corresponding to eigenvalue $\lambda_2 > 0$) scaled so that its Euclidean norm is $\sqrt{n}$.

We may approximate the discrete minimum of $\mathbf{q}^T L_G \mathbf{q}$ in Q as follows.

- Compute the Fiedler vector $\mathbf{x}_2$ (having norm $\sqrt{n}$) associated with L_G:

$$L_G \, \mathbf{x}_2 = \lambda_2 \, \mathbf{x}_2.$$

- Since the components $(\mathbf{x}_2)_i$ may not be in $\{+1, -1\}$, sort its entries in increasing order and let $\alpha_{1/2}$ denote a median value of the entries of $\mathbf{x}_2$.
- If $(\mathbf{x}_2)_i > \alpha_{1/2}$ define $\mathbf{q}_i = +1$ and if $(\mathbf{x}_2)_i < \alpha_{1/2}$ define $\mathbf{q}_i = -1$. If $(\mathbf{x}_2)_i = \alpha_{1/2}$ define $\mathbf{q}_i = \pm 1$, so that $(n/2)$ components have $+1$ entries.

The above algorithm is easily generalized when $|V|$ is not even and when the vertex weights are not unitary. Indeed, for any choice of nonnegative integers n_1 and n_2 satisfying $n_1 + n_2 = n$, we may extend the above partitioning by defining V_1 as the vertices corresponding to the first n_1 components of the Fiedler vector after sorting (taking into account nonunitary weights of vertices). The following theoretical result will hold.

Lemma 5.9. *Suppose the following assumptions hold.*

1. *Let G be a connected graph.*
2. *Let $\mathbf{x}_2$ denote the Fiedler vector of L_G.*
3. *For $\alpha \ge 0$ and $\beta \le 0$ define:*

$$\begin{aligned} \mathcal{I}_\alpha &\equiv \{i : (\mathbf{x}_2)_i \le \alpha\} \\ \mathcal{J}_\beta &\equiv \{i : (\mathbf{x}_2)_i \ge -\beta\}. \end{aligned}$$

Then the following results will hold:

1. *The induced graph associated with* $V_1 = \{v_i : i \in \mathcal{I}_\alpha\}$ *is connected.*
2. *The induced graph associated with* $V_2 = \{v_i : i \in \mathcal{J}_\beta\}$ *is connected.*
3. *For any* $\mathbf{p} \in Q$
$$\|\mathbf{x}_2 - \mathbf{q}\|_2 \leq \|\mathbf{x}_2 - \mathbf{p}\|_2.$$

Proof. For results 1 and 2 see [FI]. For result 3 see [PO3, CI6]. □

The *recursive* spectral bisection algorithm partitions a graph $G = (V, E)$ by repeatedly applying the spectral bisection algorithm to each of the subgraphs obtained from the previous applications of the spectral bisection algorithm. We summarize the algorithm below, and employ the notation $G_i^{(k)} = (V_i^{(k)}, E_i^{(k)})$ to denote the i'th subgraph at stage k. Weight matrices of subgraphs are defined as submatrices of the parent weight matrix corresponding to the indices in the subgraphs.

Algorithm 5.1.1 *(Recursive Spectral Bisection Algorithm)*
Let $p \approx 2^J$ *denote the number of sets in the partition*
Define $G_1^{(1)} = (V_1^{(1)}, E_1^{(1)}) \equiv G = (V, E)$

```
1. For $k = 1, \cdots, J-1$ do:
2.     Spectrally bisect each subgraph $G_i^{(k)}$ at level $k$ into two:
           $G_i^{(k)} \rightarrow \left\{ G_{I_1(i)}^{(k+1)},\ G_{I_2(i)}^{(k+1)} \right\} \qquad \text{for } 1 \leq i \leq 2^{k-1}.$
3.     Reindex $G_i^{(k+1)}$ so that indices $1 \leq i \leq 2^k$
4. Endfor
```

Here $I_1(i)$ and $I_2(i)$ denote temporary indices for the partitioned graphs, before reindexing. In practice, the Fiedler vector $\mathbf{x}_2$ (or an approximation of it) may be computed approximately by the Lanczos algorithm [GO4]. As mentioned earlier, the quality of spectral partitions are very good, though more expensive to compute. For additional details, see [PO3, CI8, CI7].

Multilevel Graph Partitioning Algorithm. The multilevel graph partitioning algorithm [SI2, BA20, VA3, HE7, KA3, KU] is motivated by graph compaction algorithms and multigrid methodology [BR22, HA2, MC2]. Given a graph $G = (V, E)$ with weight matrix W, this graph partitioning algorithm constructs a hierarchy of smaller order or "coarser" graphs $G^{(l)} = (V^{(l)}, E^{(l)})$ with weight matrices $W^{(l)}$, by repeated merging (agglomeration) of pairs of vertices within each parent graph. Each graph in the hierarchy is constructed to have approximately half the number of vertices as its parent graph. Once a weighted coarse graph of sufficiently small order has been constructed, a standard graph partitioning algorithm (such as recursive spectral bisection) is applied to partition the coarsest weighted graph by minimizing a suitably defined objective functional equivalent to (5.6). The partitioned subgraphs of the coarse graph are then "projected" onto the next finer level in the hierarchy by unmerging (deagglomeration) of the merged vertices. These projected partitions are improved at the finer level by applying a Kernighan-Lin type algorithm. This procedure is recursively applied till a partitioning of the original graph is obtained. Since the bulk of the computations are implemented on the coarsest graph, the computational cost is significantly reduced.

We describe additional details. Each graph in the multilevel hierarchy will be indexed as $l = 0, 1, \ldots, J$. In contrast with traditional multilevel notation, however, index $l = 0$ will denote the original and largest order graph in the hierarchy, while index $l = J$ will denote the coarsest and smallest order graph in the hierarchy. For $0 \leq l \leq J$, the graphs in the hierarchy will be denoted as $G^{(l)} = \left(V^{(l)}, E^{(l)}\right)$ with weight matrices $W^{(l)}$ of size n_l. The initial graph will be the original weighted graph $G^{(0)} \equiv G$ with $(V^{(0)}, E^{(0)}) \equiv (V, E)$, $W^{(0)} \equiv W$ and $n_0 \equiv n$. If the original graph $G = (V, E)$ is not weighted, then the default weight matrix W is employed with unitary weights $w(v_i) = 1$ assigned to the original vertices v_i in V.

Given a parent graph $G^{(l-1)} = \left(V^{(l-1)}, E^{(l-1)}\right)$, this algorithm defines a coarser (smaller order) graph $G^{(l)} = \left(V^{(l)}, E^{(l)}\right)$ by merging (agglomerating) pairs of vertices within $V^{(l-1)}$ by a procedure referred to as *maximal matching.*

Definition 5.10. *Given a graph $G^{(l)} = (V^{(l)}, E^{(l)})$ a matching is any subset of edges from $E^{(l)}$ such that no more than one edge is incident to each vertex. A maximal matching is a matching in which no additional edge can be added without violating the matching condition.*

A maximal matching can be constructed in graph $G^{(l-1)}$ as follows. Select one vertex randomly, say $v_r^{(l-1)}$, from the graph and determine an *unmatched* vertex adjacent to it (if it exists) with maximal edge weight, i.e., match $v_r^{(l-1)}$ with $v_s^{(l-1)}$ if $w_{rs}^{(l-1)}$ is largest amongst all the unmatched vertices $v_s^{(l-1)}$. If no adjacent unmatched vertex is found for $v_r^{(l-1)}$, then it is left as a singleton and matched with itself. To obtain a maximal matching, this procedure is repeated till there are no remaining unmatched vertices. We shall denote by $I_1(i, l)$ and $I_2(i, l)$ the indices of the two parent vertices at level $(l - 1)$ which are matched and merged to yield vertex $v_i^{(l)}$ at level l. If a vertex is matched with itself (i.e., is a singleton) then $I_1(i, l) = I_2(i, l)$ denotes the index at level $(l - 1)$ of vertex $v_i^{(l)}$. Since a vertex $v_i^{(l)}$ at level l is the agglomeration of vertices $v_{I_1(i,l)}^{(l-1)}$ and $v_{I_2(i,l)}^{(l-1)}$ from $V^{(l-1)}$, we express this as:

$$v_i^{(l)} = \{v_{I_1(i,l)} \cup v_{I_2(i,l)}\}.$$

Consequently, vertices in $V^{(l)}$ represent a subset of vertices from the original graph $V^{(0)} = V$. The new vertices $v_i^{(l)}$ in $V^{(l)}$ are assigned weights as follows:

$$w^{(l)}(v_i^{(l)}) = w^{(l-1)}(v_{I_1(i,l)}^{(l-1)}) + w^{(l-1)}(v_{I_2(i,l)}^{(l-1)}),$$

if $v_i^{(l)}$ is not a singleton. Otherwise $w^{(l)}(v_i^{(l)}) = w^{(l-1)}(v_{I_1(i,l)}^{(l-1)})$. The weight $w^{(l)}(v_i^{(l)})$ will denote the number of vertices of the original graph $V^{(0)}$ in $v_i^{(l)}$. Vertices $v_i^{(l)}$ and $v_j^{(l)}$ in $V^{(l)}$ will be defined to be *adjacent* in $E^{(l)}$ if any of the parent vertices of $v_i^{(l)}$ are adjacent to any parent vertices of $v_j^{(l)}$ at level $(l-1)$. A weight $w_{ij}^{(l)}$ will be assigned to adjacent vertices $v_i^{(l)}$ and $v_j^{(l)}$ by summing the weights on all edges between the parent nodes of $v_i^{(l)}$ and $v_j^{(l)}$ at level $(l-1)$. More specifically, if

$$v_i^{(l)} = v_{I_1(i,l)} \cup v_{I_2(i,l)} \quad \text{and} \quad v_j^{(l)} = v_{I_1(j,l)} \cup v_{I_2(j,l)},$$

we define:

$$w_{ij}^{(l)} = \sum_{r \in \{I_1(i,l), I_2(i,l)\}} \sum_{s \in \{I_1(j,l), I_2(j,l)\}} w_{rs}^{(l-1)}.$$

The first phase of the multilevel graph partitioning algorithm recursively applies maximal matching to compute coarser graphs till a coarse graph $G^{(J)}$ of sufficiently small order is constructed. Weights and edges are recursively defined by applying the preceding expressions. Before we describe the second phase in the multilevel graph partitioning algorithm, we discuss how a partition $V_1^{(l)}, \ldots, V_p^{(l)}$ of vertices in $V^{(l)}$ can be "projected" to yield a partition of $V^{(l-1)}$. We define a projection P_l^{l-1} as:

$$P_l^{l-1} V_i^{(l)} \equiv \cup_{w_j^{(l)} \in V_i^{(l)}} \left(w_{I_1(j,l)}^{(l-1)} \cup w_{I_2(j,l)}^{(l-1)} \right). \tag{5.8}$$

More generally, given indices $0 \le r < l$, we define a projection P_l^r recursively:

$$P_l^r V_i^{(l)} \equiv P_{r+1}^r \cdots P_l^{l-1} V_i^{(l)}. \tag{5.9}$$

Thus, a partition $V_1^{(l)}, \ldots, V_p^{(l)}$ of $V^{(l)}$ will yield a partition of $V^{(r)}$ by use of the projections $P_l^r V_1^{(l)}, \ldots, P_l^r V_p^{(l)}$, which will *deagglomerate* all the vertices $v_i^{(l)} \in V_k^{(l)}$. Formally, we obtain an expression similar to (5.8).

We next describe how to define an induced objective function $\delta^{(l)}(\cdot)$ which is equivalent to $\delta(\cdot)$ for a partition $V_1^{(l)}, \ldots, V_p^{(l)}$ at level l:

$$\begin{cases} \delta^{(l)}\left(V_i^{(l)}, V_j^{(l)}\right) \equiv \sum_{\{v_r^{(l)} \in V_i^{(l)},\, v_s^{(l)} \in V_j^{(l)}\}} w_{rs}^{(l)} \\ \delta^{(l)}\left(V_1^{(l)}, \ldots, V_p^{(l)}\right) \equiv \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \delta^{(l)}\left(V_i^{(l)}, V_j^{(l)}\right). \end{cases} \tag{5.10}$$

By construction, the preceding objective functionals will satisfy:

$$\delta^{(l)}\left(V_1^{(l)}, \ldots, V_p^{(l)}\right) = \delta^{(r)}\left(P_l^r V_1^{(l)}, \ldots, P_l^r V_p^{(l)}\right), \tag{5.11}$$

for $0 \le r < l$, where $\delta^{(0)}(\cdot) = \delta(\cdot)$. Hence, by construction, if a sequence of partitions are constructed on graph $G^{(l)}$ such that the value of the objective functional $\delta^{(l)}(\cdot)$ monotonically decreases, then the value of $\delta^{(0)}(\cdot)$ will also decrease monotonically for the projected partitions at level $l = 0$.

We now summarize the multilevel graph partitioning algorithm. Given $G^{(0)} \equiv G$, a hierarchy of graphs $G^{(1)}, \ldots, G^{(J)}$ are constructed by maximal matching. Then, the coarsest graph $G^{(J)}$ is partitioned using an effective graph partitioning algorithm, such as Kernighan-Lin or recursive spectral bisection, to minimize $\delta^{(J)}(\cdot)$. The resulting partition is then projected to the next finer level using P_J^{J-1} and refined by an application of several iterations of the Kernighan-Lin algorithm using $\delta^{(J-1)}(\cdot)$. This procedure is recursively applied till a partition is obtained on the finest graph. The algorithm is summarized next for an input graph $G^{(0)} = (V^{(0)}, E^{(0)})$ and J denoting the number of desired levels in the hierarchy.

Algorithm 5.1.2 *(Multilevel Graph Partitioning Algorithm)*

```
1. For $l = 1, \ldots, J$ do:
2.     Construct a coarser graph using maximal matching:
           $V^{(l)} \leftarrow V^{(l-1)}$
           $E^{(l)} \leftarrow E^{(l-1)}$.
3.     Define vertex and edge weights:
           $\begin{cases} w^{(l)}(v_i^{(l)}) = w^{(l-1)}(v_{I_1(i,l)}^{(l-1)}) + w^{(l-1)}(v_{I_2(i,l)}^{(l-1)}) \\ w_{ij}^{(l)} = \sum_{r \in \{I_1(i,l), I_2(i,l)\}} \sum_{s \in \{I_1(j,l), I_2(j,l)\}} w_{rs}^{(l-1)} \end{cases}$
4. Endfor
5. Partition: $V^{(J)} \rightarrow (V_1^{(J)}, \cdots, V_p^{(J)})$
6. For $l = J, \ldots, 1$ do:
7.     Project: $P_l^{l-1} V_i^{(l)} \rightarrow V_i^{(l-1)}$ for $i = 1, \ldots, p$
8.     Improve the partition employing Kernighan-Lin and $\delta^{(l)}(\cdot)$
9. Endfor
Output: $V_1^{(0)}, \ldots, V_p^{(0)}$
```

Numerical studies, see [SI2, BA20, HE7, KA3, PO2], indicate that the quality of multilevel partitions are comparable with that obtained by recursive spectral bisection as measured by $\delta(\cdot)$. Various software implementations of multilevel partitioning algorithms are available, see CHACO [HE8], METIS [KA3] and [KU]. For additional discussion, readers are referred to [PO3].

5.1.3 Construction of Subdomain Decomposition

Graph partitioning can be applied to either partition Ω into nonoverlapping subdomains $\Omega_1, \ldots, \Omega_p$, or to partition the index set $\mathcal{I}$ of nodes in $\mathcal{T}_h(\Omega)$ into subindex sets $\mathcal{I}_1, \ldots, \mathcal{I}_p$, so that load balancing and minimal communication constraints hold. To partition the index set $\mathcal{I} = \{x_1, \ldots, x_n\}$ of vertices in Ω_h. Define a graph $G = (V, E)$ with vertices $v_i \equiv x_i$ for $i = 1, \ldots, n$ where v_i is adjacent to v_j in E if vertex x_i and x_j belong to the same element $\kappa \in \mathcal{T}_h(\Omega)$. Assign unitary weights $w(v_i) = 1$ to the vertices and unitary weights $w_{ij} \equiv 1$ to the edges $(v_i, v_j) \in E$. Apply any of the partitioning algorithms to minimize $\delta(\cdot)$ within $\mathcal{K}_\epsilon$ (for a suitable $\epsilon > 0$) and partition V into $V_1, \ldots, V_p$. This yields a partition $\mathcal{I}_1, \ldots, \mathcal{I}_p$ of the index set $\mathcal{I}$. To obtain overlap amongst the index sets, for any $\beta > 0$ extend each index set $\mathcal{I}_i$ as:

$$\mathcal{I}_i^* \equiv \{l : \operatorname{dist}(x_l, x_j) \le \beta\, h_0, \quad \text{for } j \in \mathcal{I}_i\}, \tag{5.12}$$

where $\operatorname{dist}(x_l, x_j)$ denotes the Euclidean distance between x_l and x_j.

Remark 5.11. If n_j denotes the number of vertices in $\mathcal{I}_j$ and $n_j^* \ge n_j$ the number of vertices in $\mathcal{I}_j^*$, and if $\mathcal{T}_h(\Omega)$ is not quasiuniform, then n_i^* may vary significantly, violating load balancing requirements.

To partition Ω into nonoverlapping subdomains, let $\kappa_1, \ldots, \kappa_q$ denote an ordering of the elements in triangulation of Ω_h. Define a graph $G = (V, E)$ with vertices $v_i \equiv \kappa_i$ for $i = 1, \ldots, q$, where v_i is adjacent to v_j in E if $\overline{\kappa}_i \cap \overline{\kappa}_j \neq \emptyset$. We assign unitary vertex weights $w(v_i) = 1$ and unitary edge weights $w_{ij} = 1$ for $(v_i, v_j) \in E$. We may apply any of the partitioning algorithms to minimize $\delta(\cdot)$ within $\mathcal{K}_\epsilon$ (for $\epsilon > 0$) and partition V into $V_1, \ldots, V_p$. By construction, this will yield a partition of Ω into connected subdomains:

$$\Omega_i \equiv (\cup_{v_l \in V_i} \kappa_l), \qquad \text{for } 1 \le i \le p. \tag{5.13}$$

Overlap may be included amongst the subdomains by extending each subdomain Ω_i to Ω_i^* by including all elements adjacent within a distance $\beta\, h_0 > 0$, where $\operatorname{dist}(\kappa_r, \kappa_j)$ denotes the the distance between the centroids of elements κ_r and κ_j. The size of Ω_i^* and the associated number of nodes may vary significantly if $\mathcal{T}_h(\Omega)$ is not quasiuniform.

5.1.4 Coarse Spaces on Unstructured Grids

Traditional coarse spaces defined on a coarse grid $\mathcal{T}_{h_0}(\Omega)$ will not be applicable on unstructured grids, since $\mathcal{T}_h(\Omega)$ is not obtained by the refinement of $\mathcal{T}_{h_0}(\Omega)$. Instead, alternative coarse spaces may be employed to provide global transfer of information on such grids [WI6, CA4, CH17, CH3, SA11, SA12, SA13]. We shall outline the following coarse spaces:

- *Coarse space $V_{0,I}(\Omega)$ obtained by interpolation of an external space.*
- *Piecewise constant discrete harmonic finite element space $V_{0,P}(\Omega)$.*

We shall let $V_h(\Omega)$ denote the finite element space defined on the unstructured grid $\mathcal{T}_h(\Omega)$, and formulate coarse spaces either algebraically, corresponding to a subspace of nodal vectors in $\mathbb{R}^n$ associated with finite element functions.

Coarse Space Based on Interpolation. The finite element coarse space $V_{0,I}(\Omega) \subset V_h(\Omega)$ is defined by *interpolating* or *projecting* an external finite dimensional space $V_{h_0}(\Omega_*)$ of functions with desirable approximation properties onto the finite element space $V_h(\Omega)$, see [CA4, CH17, CH3, CA17]. Let $\Omega^* \supset \Omega$ and let $\{\phi_1^{(0)}(x), \cdots, \phi_{n_0}^{(0)}(x)\}$ denote n_0 basis functions in $V_{h_0}(\Omega_*) \subset H^1(\Omega^*)$ having desirable properties. The coarse space $V_{0,I}(\Omega)$ is defined as the subspace of $V_h(\Omega) \cap H_0^1(\Omega)$ spanned by interpolants (or projections) of these basis functions onto the finite element space:

$$V_0(\Omega) \equiv \text{span}\left\{I_h\phi_1^{(0)}(\cdot), \ldots, I_h\phi_{n_0}^{(0)}(\cdot)\right\} \subset V_h(\Omega), \tag{5.14}$$

where I_h denotes a finite element *interpolation* or *projection* map onto $V_h(\Omega)$.

A matrix representation of $V_{0,I}(\Omega)$ can be obtained using the standard interpolation map I_h as follows. Let $\mathcal{I} = \{x_1, \cdots, x_n\}$ denote an ordering of the interior nodes of $\mathcal{T}_h(\Omega)$. Then, an $n \times n_0$ extension matrix R_0^T is defined:

$$R_0^T \equiv \begin{bmatrix} \phi_1^{(0)}(x_1) & \cdots & \phi_{n_0}^{(0)}(x_1) \\ \vdots & & \vdots \\ \phi_1^{(0)}(x_n) & \cdots & \phi_{n_0}^{(0)}(x_n) \end{bmatrix}. \tag{5.15}$$

The functions $\{\phi_i^{(0)}(\cdot)\}_{i=1}^{n_0}$ should ideally be chosen so that the above matrix is of full rank. The restriction matrix R_0 will be the transpose of the extension matrix, and $A_0 \equiv R_0 A R_0^T$. We indicate two examples below.

Example 5.12. If $\Omega^* \supset \Omega$ is a polygonal or polyhedral domain covering Ω and triangulated by a *quasiuniform* grid $\mathcal{T}_{h_0}(\Omega^*)$, let $\{\phi_1^{(0)}(x), \cdots, \phi_{n_0}^{(0)}(x)\}$ denote a finite element nodal basis defined on triangulation $\mathcal{T}_{h_0}(\Omega^*)$. Such basis functions will be in $H^1(\Omega^*)$. To ensure that each coarse node in $\mathcal{T}_{h_0}(\Omega^*)$ corresponds to a true (nonredundant) degree of freedom, it will be assumed that the support of each nodal basis function defined on $\mathcal{T}_{h_0}(\Omega^*)$ intersects interior nodes of $\mathcal{T}_h(\Omega)$. A coarse space can be constructed as in (5.15), where R_0 will be sparse. Such a basis was tested in [CA4, CH17, CH3] and shown to yield a quasioptimal convergence rate under appropriate assumptions. It is more suited for Dirichlet boundary value problems.

Example 5.13. An alternative coarse space can be constructed by choosing a space $V_{n_0}(\Omega^*)$ of *polynomials* on $\Omega^* \supset \Omega$ and interpolating it onto the finite element space $V_h(\Omega)$. If $\{\phi_1^{(0)}(x), \cdots, \phi_{n_0}^{(0)}(x)\}$ denotes a monomial or Tchebycheff basis for polynomials of degree d or less on a rectangular domain $\Omega_* \supset \Omega$, then the matrices R_0, R_0^T and A_0 can be constructed as in (5.15). However, these matrices will not be sparse. In two dimensions, the monomials:

$$V_d(\Omega_*) \equiv \text{span}\,\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2, \ldots, x_1^d, \ldots, x_2^d\},$$

may be used, but a Tchebycheff basis would be preferable. Alternatively, a tensor product of one dimensional polynomials may be employed. Heuristics studies indicate reasonable convergence for Neumann boundary value problems [CA18], in which case nodes on $\Omega \cup \mathcal{B}_N$ must also be included in (5.15).

Coarse Space of Piecewise Discrete Harmonic Functions. We next describe a coarse space $V_{0,P}(\Omega) \subset V_h(\Omega)$ of piecewise discrete harmonic finite element functions. Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping subdomain decomposition of Ω, constructed by graph partitioning of the triangulation Ω_h of Ω. The coarse finite element space $V_{0,P}(\Omega)$ will consist of finite element functions which are discrete harmonic on each subdomain Ω_l with specially chosen boundary values on each $B^{(l)} = \partial\Omega_l \backslash \mathcal{B}_D$.

A matrix basis for V_0 can be constructed as follows [MA14, CO8, SA7]. Denote by $B = \cup_{i=1}^{p} B^{(i)}$ the common interface, We shall assume that the indices in $\mathcal{I}$ are grouped and ordered as $I \cup B$ corresponding to the nodes in the subdomains $\Omega_1, \cdots, \Omega_p$ and on interface B, with n_I and n_B denoting the number of nodes in I and B, respectively. Let y_i for $i = 1, \ldots, n_B$ denote the ordering of nodes on B. Then, for each node $y_i \in B$ define $N_G(y_i)$ as the number of subdomain boundaries $B^{(k)}$ with $y_i \in B^{(k)}$. Employ the block partitioning $\mathbf{w} = \left(\mathbf{w}_I^T, \mathbf{w}_B^T\right)^T$ as in Schur complement methods, resulting in the following block structure for A:

$$A \equiv \begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix},$$

where $A_{II} = \text{blockdiag}(A_{II}^{(1)}, \ldots, A_{II}^{(p)})$. The columns of R_0^T will be defined as piecewise discrete A-harmonic vectors corresponding to the following p specifically chosen interface data vectors $\mathbf{w}_B^{(k)}$ for $k = 1, \cdots, p$:

$$\left(\mathbf{w}_B^{(k)}\right)_i = \begin{cases} \frac{1}{N_G(y_i)}, & \text{if } y_i \in B^{(k)}, \qquad \text{for } i = 1, \ldots, n_B \\ 0, & \text{otherwise.} \end{cases}$$

Denote the discrete harmonic extension matrix as $E \equiv -A_{II}^{-1} A_{IB}$ and define matrix R_0^T as:

$$R_0^T \equiv \begin{bmatrix} E\mathbf{w}_B^{(1)} & \cdots & E\mathbf{w}_B^{(p)} \\ \mathbf{w}_B^{(1)} & \cdots & \mathbf{w}_B^{(p)} \end{bmatrix}.$$

The coarse finite element space $V_{0,P}(\Omega) \subset V_h(\Omega)$ will consist of finite element functions whose nodal vectors are in Range $\left(R_0^T\right)$. The restriction matrix R_0 will be the transpose of R_0^T and $A_0 \equiv R_0 A R_0^T$. Approximation properties of such spaces are described in [CO8, SA7, MA17].

Remark 5.14. The finite element functions in $V_{0,P}(\Omega)$ correspond to discrete harmonic extensions into the subdomains, of finite element functions in the piecewise constant coarse space $V_{0,P}(B)$ employed in the balancing domain decomposition preconditioner [MA17].

5.1.5 Schwarz Algorithms

We consider next the matrix implementation of Schwarz iterative algorithms on an unstructured grid $\mathcal{T}_h(\Omega)$. These algorithms can be formulated as before, based on suitable restriction and extension matrices. We shall assume that the index set $\mathcal{I}$ has been partitioned into subindex sets $\mathcal{I}_1, \ldots, \mathcal{I}_p$ using a graph partitioning algorithm which minimizes $\delta(\cdot)$, and that each index set $\mathcal{I}_l$ has been extended to $\mathcal{I}_l^*$ as described earlier in this section. Let the subindex sets $\mathcal{I}_1^*, \ldots, \mathcal{I}_p^*$ have $n_1^*, \ldots, n_p^*$ nodes in each set.

Define an index function with $\text{index}(i, \mathcal{I}_l^*)$ denoting the global index in $\mathcal{I}$ of the local index $1 \le i \le n_l^*$ in $\mathcal{I}_l^*$. Then, the entries of an $n_l^* \times n$ local restriction matrix R_l can be defined by:

$$(R_l)_{ij} = \begin{cases} 1, & \text{if index}(i, \mathcal{I}_l^*) = j \\ 0, & \text{if index}(i, \mathcal{I}_l^*) \neq j. \end{cases}$$

The extension matrices R_l^T will be transposes of the restriction matrices with $A_l \equiv R_l A R_l^T$. Once a coarse space has be chosen with restriction matrix R_0, the system $A\mathbf{u} = \mathbf{f}$ may be solved using matrix multiplicative, additive or hybrid Schwarz algorithms based on the restriction matrices $R_0, R_1, \ldots, R_p$. If the unstructured grid is quasiuniform, optimal convergence should be obtained, see [CH3, CH17, CA4, CA18] and [CO8, SA7]. Studies of the effects of partitioning algorithms, amount of overlap and other factors in unstructured grid applications are presented in [CI8].

5.1.6 Schur complement algorithms

When the grid is unstructured, the nonoverlapping subdomains $\Omega_1, \ldots, \Omega_p$ determined by a graph partitioning algorithm may have complex geometry, and traditional globs such as edges, faces and wirebaskets may be difficult to identify. However, the subdomain boundary segments $B^{(i)}$ will be well defined, so that Neumann-Neumann and balancing domain decomposition preconditioners can be applied based on $S^{(i)}$. Depending on whether $c(x) = 0$ or $c(x) \ge c_0 > 0$ in the elliptic equation, the local Schur complement $S^{(i)}$:

$$S^{(i)} = \left(A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)} \right),$$

may be singular when Ω_i is a floating subdomain. The singular vector will be $\mathbf{z}_i = (1, \cdots, 1)^T$ and the balancing domain decomposition algorithm can be applied. When $c(x) \ge c_0 > 0$, the subdomain stiffness matrices $A^{(i)}$ will not be singular. However, if no mechanism is employed for global transfer of information, then the convergence rate of the resulting Neumann-Neumann algorithm will deteriorate as $h_0^{-2}(1 + \log(h_0/h))^2$ if the grid $\mathcal{T}_h(\Omega)$ is quasiuniform and the subdomains are shape regular of diameter h_0. Instead, the balancing domain decomposition or traditional Neumann-Neumann algorithm can be employed with the coarse space $V_{0,P}(\Omega)$ described earlier for unstructured grids, see [SA7, MA15].

5.1.7 FETI Algorithms

As with Neumann-Neumann and balancing domain decomposition algorithms, FETI algorithms also require minimal geometric information about the subdomains on unstructured grids. If $c(x) = 0$ and Ω_i is floating, the subdomain stiffness matrices $A^{(i)}$ will be singular, while if $c(x) \geq c_0 > 0$, matrix $A^{(i)}$ will be non-singular. Appropriate versions of the FETI algorithm can be employed on unstructured grids [FA15].

5.2 Parallelizability of Domain Decomposition Solvers

In this section, we *heuristically* model the potential *parallel efficiency* of domain decomposition solvers [GR10, GR12, SK, CH15, FA9, SM4, GR16]. We do this by employing theoretical models, under highly idealized assumptions, for the execution times of representative domain decomposition solvers implemented on a parallel computer having p processors with distributed memory. We consider representative Schwarz or Schur complement preconditioners, with and without coarse space correction, and CG acceleration.

Our discussion will be organized as follows. In Chap. 5.2.1 we present background and notation on identities used for the parallel computation of matrix-vector products and inner products, and representative Schwarz and Schur complement preconditioners. In Chap. 5.2.2, we describe background on parallel computers and measures for assessing the speed up, efficiency and scalability of parallel algorithms. Chap. 5.2.3 describes a domain decomposition strategy for allocating memory and computations to individual processors, and derives heuristic estimates for the parallel execution times of representative solvers, with and without coarse space correction. In Chap. 5.2.4 we employ these bounds to obtain models for the parallel efficiency of various domain decomposition iterative solvers.

5.2.1 Background

Consider the following self adjoint and coercive elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)u = f(x), & \text{in } \Omega \subset \mathbb{R}^d \\ \qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{5.16}$$

with *smooth* coefficients $a(x) \geq a_0 > 0$ and $c(x) \geq 0$. Its discretization by a finite element method based on a quasiuniform triangulation $\tau_h(\Omega)$ of Ω, with grid size h, will yield the linear system:

$$A\,\mathbf{u} = \mathbf{f}, \tag{5.17}$$

where $A = A^T > 0$ is of size n.

We consider the solution of (5.17) by a preconditioned CG algorithm using an additive Schwarz or Neumann-Neumann preconditioner. Accordingly, we let $\Omega_1, \ldots, \Omega_{n_s}$ denote a nonoverlapping decomposition of of $\Omega \subset \mathbb{R}^d$ into n_s subdomains, each of diameter h_0 and volume (area) $|\Omega_i| = O(h_0^d)$. To obtain an overlapping decomposition, we extend each subdomain Ω_i by including all points of Ω within a distance of $\beta\, h_0$ from Ω_i, resulting in subdomain Ω_i^*. By construction, the volume (area) of the extended subdomains will satisfy $|\Omega_i^*| = O\left((1+\beta_*)\,|\Omega_i|\right)$ for $\beta_* \equiv (1+\beta)^d - 1$. Due to quasiuniformity of the underlying triangulation, if n denotes the number of interior nodes in Ω, then each nonoverlapping subdomain Ω_i will contain $O(n/n_s)$ unknowns while overlapping subdomains Ω_i^* will contain $O\left((1+\beta_*)n/n_s\right)$ unknowns.

Notation. We will employ the following notation. The pointwise nodal restriction map onto nodes in $\overline{\Omega}_i$ will be denoted $R^{(i)}$ and the local stiffness matrix on $\overline{\Omega}_i$ will be denoted $A^{(i)}$. Consequently, the subassembly identity can be expressed in the form:

$$A = \sum_{i=1}^{n_s} R^{(i)^T} A^{(i)} R^{(i)}.$$

Local load vectors will be denoted $\mathbf{f}^{(i)}$ for $1 \le i \le n_s$ so that the global load vector has the form $\mathbf{f} \equiv \sum_{i=1}^{n_s} R^{(i)^T} \mathbf{f}^{(i)}$. We shall assume there exists diagonal matrices $I^{(i)}$ which form a decomposition of the identity:

$$I = \sum_{i=1}^{n_s} R^{(i)^T} I^{(i)} R^{(i)},$$

where matrix $I^{(i)}$ has the same size as $A^{(i)}$ with nonnegative diagonal entries. Such matrices can be constructed by defining $(I^{(i)})_{kk} = 1$ if $x_k \in \Omega_i$ and $(I^{(i)})_{kk} = 1/N(x_k)$ if $x_k \in B^{(i)}$ where $N(x_k)$ denotes the number of subdomain boundaries to which node x_k belongs to.

If a Schur complement preconditioner is employed, then $\mathcal{R}_i$ will denote the pointwise restriction map from nodes on interface B onto the boundary segment $B^{(i)}$ of Ω_i. The local Schur complements will be denoted $S^{(i)}$, so that the subassembly identity has the form:

$$S = \sum_{i=1}^{n_s} \mathcal{R}_i^T S^{(i)} \mathcal{R}_i.$$

A decomposition of the identity on B of the form:

$$I = \sum_{i=1}^{n_s} \mathcal{R}_i^T I^{(i)} \mathcal{R}_i,$$

will also be assumed, where $I^{(i)}$ (with some abuse of notation) denotes a diagonal matrix of the same size as $S^{(i)}$ with nonnegative diagonal entries.

If a coarse space is employed, it will be spanned by the rows of $\mathcal{R}_0$ with $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T$ denoting the coarse Schur complement matrix.

Given overlapping subdomains Ω_i^*, we let R_i denote the pointwise restriction map onto nodes in Ω_i^*, so that $A_i = R_i A R_i^T$ will be a principal submatrix of A corresponding to nodes in Ω_i^*. If a coarse space is employed, then the row space of R_0 will span the coarse space, and the coarse space matrix will be denoted $A_0 = R_0 A R_0^T$.

5.2.2 Parallel Computation

We consider a parallel computer with an MIMD (multiple instruction, multiple data) architecture with distributed memory [HO, LE16, AL3, QU8, GR]. We will assume there are p identical processors, each with local memory and capable of executing programs independently, where τ_f denotes the time for a floating point operation. For simplicity, it will be assumed that data can be communicated directly between any pair of processors (though, in most domain decomposition applications it will be sufficient to pass data between neighboring processors, as specified by some adjacency matrix). We shall let $T_{\mathrm{comm}}(n) \equiv \tau_0 + n\,\tau_c$ denote the average time for transferring n units of data between two processors. Here τ_0 denotes the start up time, which we shall assume is *zero* for simplicity.

On a typical MIMD parallel computer, the speed of communication τ_c between processors will be significantly slower than the speed τ_f of floating point operations, i.e., $\tau_f \ll \tau_c$. This unfortunate fact places constraints on the types of parallel algorithms suitable for implementation on such hardware. In such cases, interprocessor communication must be kept to a minimum to obtain high speed up of algorithms. If several processors simultaneously send data to each other, then a suitable protocol such as message passing interface [GR15] may be employed. By design, large portions of domain decomposition algorithms involve computations which can be implemented independently without communication, provided each processor is assigned to implement the computations on one or more subdomains. The remaining portions typically require communication, either between adjacent subdomains or with a coarse space (if present). Algorithms having relatively large sections of independent computations with relatively small sections requiring communication are said to have *coarse granularity*, see [HO, LE16, AL3, QU8, GR], and are generally suited for implementation on MIMD architectures.

The performance of an algorithm on a parallel computer is typically assessed by a quantity referred to as the speed up, which measures the rate of reduction in its execution time as the number of processors is increased. Formally, if $T(p, n)$ denotes the execution time for implementing a parallel algorithm having problem size n using p processors, then its relative speed up is defined as the ratio of its execution time $T(1, n)$ on a serial computer to its execution time $T(p, n)$ on a parallel computer with p processors.

Definition 5.15. *The relative speed up of an algorithm implemented using* p *processors is defined as:*

$$S(p,n) \equiv \frac{T(1,n)}{T(p,n)}.$$

The speed up ratio has a theoretical maximum value of p *for a perfectly parallelizable algorithm with* $1 \leq S(p,n) \leq p$.

Remark 5.16. Even if the relative speed up of a parallel algorithm attains its maximum value, there may be other parallel implementations with shorter execution times. This is because the relative speed up ratio is not measured with reference to the *best* serial execution time. When the speed up is measured relative to the best serial execution time, the resulting speed up is referred to as total speed up. This is defined below, where $T_{\text{best}}(1,n)$ denotes the best serial execution time.

Definition 5.17. *The total speed up of an algorithm is defined as:*

$$\overline{S}(p,n) \equiv \frac{T_{\text{best}}(1,n)}{T(p,n)}.$$

When the best serial algorithm or execution time is not known, the relative speed up may be used as a measure of its parallel performance. In finite element applications, the lowest attainable complexity for the solution of a sparse linear system of size n arising from discretizations of elliptic equations will be denoted $\phi(n)$. In special cases, linear (or almost linear) order complexity may be attained for multigrid and fast Poisson solvers, depending on the elliptic equation, geometry and discretization. In such cases, $T_{\text{best}}(1,n) = C\, n\, \tau_f$, but we will assume $T_{\text{best}}(1,n) = \phi(n)\, \tau_f$ where $\phi(n) = c_0 n^{\alpha} + o(n^{\alpha})$ for $1 < \alpha \leq 3$.

Remark 5.18. The execution time $T(p,n)$ of domain decomposition algorithms may depend on other factors, such as the number n_s of subdomains, the amount β of overlap (if overlapping subdomains are employed), the stopping criterion ϵ, the complexity $\phi(\cdot)$ of the local solver, size n_0 of the coarse space, amongst other factors. If this dependence of the execution time on such additional factors needs to be emphasized, we shall denote the execution time as $T(p,n,n_s,\beta,\epsilon,n_0,\phi)$ and the relative speed up as $S(p,n,n_s,\beta,\epsilon,n_0,\phi)$ and the total speed up as $\overline{S}(p,n,n_s,\beta,\epsilon,n_0,\phi)$. In the following, we define the *parallel efficiency* of an algorithm as the percentage of the speed up relative to the maximum speed up of p.

Definition 5.19. *The relative parallel efficiency of an algorithm implemented using* p *processors is defined as:*

$$E(p,n) \equiv \frac{T(1,n)}{p\, T(p,n)} \times 100\%.$$

The total parallel efficiency of an algorithm is defined as:

$$\overline{E}(p,n) \equiv \frac{T_{\text{best}}(1,n)}{p\, T(p,n)} \times 100\%.$$

Amdahl's Law. In practice, there may be constraints on the maximal speed up attainable in an algorithm, regardless of the computer hardware, due to portions of the algorithm in which computations can only be executed sequentially. Such an upper bound on the speed up is given by *Amdahl's law*, which may be derived as follows. Let $0 < \alpha < 1$ denote the fraction of computations within an algorithm which are *serial* in nature. Then, assuming perfect parallelizability of the remaining portion of the algorithm, and ignoring overhead and communication costs, the following estimate can be obtained for the optimal execution times:

$$\begin{aligned} T(1,n) &= \alpha\, T(1,n) + (1-\alpha)\, T(1,n) \\ T(p,n) &= \alpha\, T(1,n) + (1-\alpha)\, T(1,n)/p. \end{aligned}$$

This yields the following upper bound for the speed up:

$$S(p,n) = \frac{T(1,n)}{T(p,n)} = \frac{1}{\alpha + (1-\alpha)/p} \le \frac{1}{\alpha}.$$

Thus, the parallel speed up of an algorithm cannot exceed the inverse of the fraction α of serial computations within the algorithm.

The fraction α of serial computations within an algorithm can be difficult to estimate and may vary with the problem size n. Amdahl's law yields a pessimistic bound in practice, due to the implicit assumption that the fraction α of serial computations remains fixed independent of n. Empirical evidence indicates that $\alpha(n)$ diminishes with increasing problem size n for most algorithms. A less pessimistic upper bound for the maximum speed up was derived by Gustafson-Barris as indicated below. The parallel execution time given p processors is decomposed as:

$$T(p,n) = A(n) + B(n),$$

where $A(n)$ denotes the execution time for the serial portion of the algorithm, while $B(n)$ denotes the parallel execution time for the parallelizable portion of the algorithm. This yields the following estimate for the serial execution time of the algorithm:

$$T(1,n) = A(n) + p\, B(n)$$

from which we estimate the speed up as:

$$S(p,n) = \frac{T(1,n)}{T(p,n)} = \frac{A(n) + p\, B(n)}{A(n) + B(n)} = \left(\frac{A(n)}{A(n)+B(n)}\right) + \left(\frac{B(n)}{A(n)+B(n)}\right) p.$$

Unlike the fixed bound given by Amdahl's law, the Gustafson-Baris bound for the speed up increases linearly with the number of processors.

In applications, it is often of interest to know whether parallel algorithms can be found which maintain their efficiency as the size n of the problem is scaled up. The *scalability* of a parallel algorithm, defined below, is a measure of how efficiently an algorithm makes use of additional processors.

Definition 5.20. *An algorithm is said to be scalable if it is possible to keep its efficiency constant by increasing the problem size as the number of processors increases. More specifically, an algorithm is scalable if given $m\,p$ processors where $m > 1$, the problem size can be increased to $n(m) > n$ such that:*

$$E(m\,p, n(m)) = E(p, n).$$

An algorithm is said to be perfectly scalable if its efficiency remains constant when the problem size n and the number of processors p are increased by the same factor m:

$$E(m\,p, m\,n) = E(p, n).$$

An algorithm is said to be *highly* scalable if its parallel efficiency depends only weakly on the number of processors as the problem size n and the number p of processors are increased by the same factor.

Remark 5.21. Using the definition of scalability, it is easily seen that the following will hold for an algorithm satisfying $E(m\,p, n(m)) = E(p, n)$:

$$T(m\,p, n(m)) = (\frac{T(1, n(m))}{m\,T(1, n)})\,T(p, n).$$

Here, the expression $T(1, n(m))/(m\,T(1, n))$ is the factor by which the computation time is increased or decreased, in relation to $T(p, n)$, as the number of processors is increased to $m\,p$ and the problem size is increased to $n(m)$.

5.2.3 Parallelization of PCG Algorithms

Each iteration in a PCG algorithm can be decomposed into two portions, a portion not involving the preconditioner (matrix-vector products, update of residuals, iterates and inner products), and a portion computing the action of the inverse of the preconditioner. When implementing a PCG algorithm on a parallel computer with distributed memory, it will be desirable to allocate memory to individual processors in a way compatible with both sections of the algorithm, thereby minimizing communication of additional data. Furthermore, if coarse space correction is employed within the preconditioner, care must exercised in the parallel implementation of the coarse problem. Typically, three alternative approaches may be employed for solving a coarse space problem in parallel in domain decomposition preconditioners:

- Parallelize the solution of the coarse problem (using all the processors) and store relevant data on each processor.
- Gather all the relevant coarse data on a specific processor and solve the coarse problem only on this processor, and broadcast the result to all other processors.
- Gather the coarse data on each processor, solve the coarse problem redundantly in parallel on each processor.

Generally, the latter two approaches are preferable on typical parallel architectures [GR10], though we shall consider only the second approach.

Motivated by the preceding, we shall heuristically consider the following strategy for allocating memory and computations to individual processors.

- Each of the p processors is assigned to handle all the computations corresponding to one or more subdomains or a coarse problem. Thus, if a coarse space is not employed, each processor will be assigned to handle (n_s/p) subdomains, and $(n_s/p)+1$ subproblems if a coarse space is employed.
- To ensure approximate load balancing, we shall require the number of unknowns $O(n/n_s)$ per nonoverlapping subdomain (or $O\left((1+\beta_*)n/n_s\right)$ per overlapping subdomain) to be approximately equal. If a coarse space is employed, we shall additionally require the number n_0 of coarse space unknowns not to exceed the number of unknowns per subdomain, yielding the constraint $n_0 \leq C(n/n_s)$.
- To reduce communication between the processors, we shall assume that the subdomain data are distributed amongst the different processors as follows. The processor which handles subdomain Ω_i should ideally store the current approximation of the local solution $\mathbf{u}^{(i)}$ on $\overline{\Omega}_i$, the local stiffness matrix $A^{(i)}$, local load vector $\mathbf{f}^{(i)}$ and matrix $I^{(i)}$. If overlapping subdomains Ω_i^* are used, then the local solution $\mathbf{u}_i$ on Ω_i^*, submatrix $A_i = R_i A R_i^T$, local load $R_i\mathbf{f}$, local residual $R_i\mathbf{r}$ and the components $R_i R_j^T$ for adjacent subdomains should also be stored locally. If a coarse space is employed, then the nonzero rows of $R_0 R^{(i)^T}$ and $R_0 R_i^T$ should also be stored locally.
- The processor which handles the coarse space should also store matrix $A_0 = R_0 A R_0^T$ and the nonzero entries of $R_j R_0^T$ for $1 \leq j \leq n_s$.

We shall let K denote the maximum number of adjacent subdomains.

When deriving theoretical estimates of execution times, we shall assume that an efficient sparse matrix solver having complexity $\phi(m) = c_0\, m^\alpha + o(m^\alpha)$ for some $1 < \alpha \leq 3$ is employed to solve all the subproblems of size m occurring within a domain decomposition preconditioner. Analysis in [CH15] suggests that if a *serial* computer is employed, then the *optimal* diameter h_0 of a traditional coarse grid must satisfy:

$$h_0 = O\left(h^{\alpha/(2\,\alpha-d)}\right) \quad \text{for} \quad \Omega \subset \mathbb{R}^d.$$

If a parallel computer is employed with p processors, then load balancing requires the number n_0 of coarse space unknowns to satisfy $n_0 \leq c(n/n_s)$. Since theoretical analysis indicates a coarse space must satisfy an approximation property of order h_0 for optimal or almost optimal convergence, this heuristically suggests $n_0 \approx n_s \approx n^{1/2}$ for traditional coarse spaces.

In the following, we outline parallel algorithms for evaluating matrix multiplication and inner products, and the action of additive Schwarz and

Neumann-Neumann preconditioners. We derive heuristic estimates for the parallel execution times of the resulting algorithms.

Parallelization of Matrix Vector Products. By assumption,we let a vector $\mathbf{w}$ be distributed amongst different processors with component $R^{(i)}\mathbf{w}$ (and $R_i\mathbf{w}$, if overlapping subdomains are employed) stored on the processor handling Ω_i. As a result, a matrix-vector product $A\,\mathbf{w}$ can be computed using the subassembly identity:

$$A\,\mathbf{w} = \sum_{i=1}^{n_s} R^{(i)^T} A^{(i)} R^{(i)} \mathbf{w},$$

and the result can be stored locally using the following steps.

1. In parallel, multiply each of the local vectors $R^{(i)}\mathbf{w}$ (assumed to be stored locally) using the local stiffness matrix $A^{(i)}$.
2. The processor handling Ω_i should send the data $R^{(j)} R^{(i)^T} \left(A^{(i)} R^{(i)} \mathbf{w}\right)$ to the processor handling Ω_j.
3. The processor handling Ω_j should sum the contributions it receives:

   $$R^{(j)} A\mathbf{w} = \sum_{i=1}^{n_s} R^{(j)} R^{(i)^T} A^{(i)} R^{(i)} \mathbf{w},$$

   from all (at most K) neighbors, and store the result locally.

If t_i denotes the parallel execution time for the i'th step above, it will satisfy:

$$\begin{cases} t_1 \le c_1\,(n_s/p)\,(n/n_s)\,\tau_f \\ t_2 \le c_2\,(n_s/p)\,K\,(n/n_s)^{(d-1)/d}\,\tau_c + \tau_0 \\ t_3 \le c_3\,(n_s/p)\,K\,(n/n_s)^{(d-1)/d}\,\tau_f. \end{cases}$$

Apart from τ_0, the other terms are inversely proportion to p.

Matrix-vector products involving the Schur complement matrix S can be computed similarly, based on an analogous subassembly identity:

$$S\mathbf{w}_B = \sum_{i=1}^{n_s} \mathcal{R}_i^T S^{(i)} \mathcal{R}_i \mathbf{w}_B.$$

Since $S^{(i)} = A_{II}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}$, such computations require the solution of local linear systems, with the solver of complexity $\phi(.)$. Thus, the parallel execution time for matrix multiplication by S will be bounded by a sum of $t_1 = c_1\,(n_s/p)\,\phi(n/n_s)\,\tau_f$, $t_2 = c_2\,(n_s/p)\,K\,(n/n_s)^{(d-1)/d}\,\tau_c + \tau_0$ and also $t_3 = c_3\,K\,(n_s/p)\,(n/n_s)^{(d-1)/d}\,\tau_f$. Again, apart from the start up time τ_0, the other terms are inversely proportion to p.

Parallelization of Inner Products. Inner products can be computed in parallel based on the distributed data stored on each processor. By assumption, given vectors $\mathbf{w}$ and $\mathbf{v}$, their components $R^{(i)}\mathbf{w}$ and $R^{(i)}\mathbf{v}$ will be stored on the processor handling Ω_i. Since matrix $I^{(i)}$ will also be stored locally, the inner product $\mathbf{w}^T\mathbf{v}$ can be computed using the identity:

$$\mathbf{w}^T\mathbf{v} = \sum_{i=1}^{n_s} \mathbf{w}^T R^{(i)^T} I^{(i)} R^{(i)} \mathbf{v}.$$

This computation may be distributed as follows.

1. In parallel, the processor handling Ω_i should compute the local inner products $\mathbf{w}^T R^{(i)^T} I^{(i)} R^{(i)} \mathbf{v}$.
2. Each processor should sum the (n_s/p) local inner products it handles and communicate the computed result to all the other processors.
3. Each processor should sum all the local inner products it receives and store the resulting answer locally.

If t_i denotes the execution time for the i'th step above, it will satisfy:

$$\begin{cases} t_1 \le c_1\,(n_s/p)\,(n/n_s)\,\tau_f \\ t_2 \le c_2\,(n_s/p)\,\tau_f + c_3\,p\,\tau_c + \tau_0 \\ t_3 \le c_4\,p\,\tau_f. \end{cases}$$

Except for $c_3\,p\,\tau_c + \tau_0$ and $c_4\,p\,\tau_f$, the other terms vary inversely with p.

Analogous estimates will hold for inner products in Schur complement algorithms, based on interface unknowns. The total execution time in this case will be bounded by the sum of $t_1 = c_1\,(n_s/p)\,(n/n_s)^{(d-1)/d}\,\tau_f$ along with $t_2 = c_2\,(n_s/p)\,\tau_f + c_3\,p\,\tau_c + \tau_0$ and $t_3 = c_4\,p\,\tau_f$. Except for $c_3\,p\,\tau_c + \tau_0$ and $c_4\,p\,\tau_f$, the other terms are inversely proportion to p.

Parallelization of an Additive Schwarz Preconditioner. If there is *no coarse space*, the inverse of such a preconditioner will have the form:

$$M^{-1} = \sum_{i=1}^{n_s} R_i^T A_i^{-1} R_i.$$

Computation of the action of M^{-1} on a residual vector $\mathbf{r}$ can be implemented in parallel as follows.

1. In parallel, solve $A_i\mathbf{w}_i = R_i\mathbf{r}$ using the locally stored residual vector $R_i\mathbf{r}$ and the locally stored submatrix A_i.
2. In parallel, the processor handling Ω_i^* should send $R_j R_i^T \mathbf{w}_i$ to each of the processors handling Ω_j^* for $\Omega_j^* \cap \Omega_i^* \neq \emptyset$.
3. In parallel, each processor should sum contributions of solutions from adjacent subdomains and store $R_j M^{-1}\mathbf{r} = \sum_{i=1}^{n_s} R_j R_i^T \mathbf{w}_i$ locally.

The computational time for each step can be estimated.

If t_i denotes the execution time for the i'th step, it will satisfy:

$$\begin{cases} t_1 \leq c_1\,(n_s/p)\,\phi\,((1+\beta_*)(n/n_s))\;\tau_f \\ t_2 \leq c_2\,K\,\beta_*(n/p)\,\tau_c + \tau_0 \\ t_3 \leq c_3\,K\,\beta_*\,(n/p)\tau_f. \end{cases}$$

Apart from τ_0, the terms are inversely proportional to p.

If a *coarse space* is included, the preconditioner will have the form:

$$M^{-1} = \sum_{i=1}^{n_s} R_i^T A_i^{-1} R_i + R_0^T A_0^{-1} R_0.$$

Care must be exercised when parallelizing the coarse grid correction term $R_0^T A_0^{-1} R_0$ since the computation of $R_0\mathbf{r}$ requires global communication between processors. We shall assume that the coarse space computations are performed on a processor assigned to the coarse space, however, they may alternatively be performed redundantly on each of the other processors in parallel. We shall not consider the parallelization of coarse space computations. By assumption, the nonzero rows of $R_0 R^{(i)^T}$, matrix $I^{(i)}$ and vector $R^{(i)}\mathbf{r}$ are stored locally on the processor handling Ω_i^*. Thus, the vector $R_0\mathbf{r}$ may be computed based on the following expression:

$$\begin{cases} R_0 = R_0\left(\sum_{i=1}^{n_s} R^{(i)^T} I^{(i)} R^{(i)}\right) \\ \quad\; = \sum_{i=1}^{n_s}\left(R_0 R^{(i)^T}\right)\left(I^{(i)} R^{(i)}\right). \end{cases}$$

Below, we summarize an algorithm for the parallel computation of $M^{-1}\mathbf{r}$.

1. The processor handling Ω_i^* should compute the nontrivial rows of the term $R_0 R^{(i)^T} I^{(i)} R^{(i)}\mathbf{r}$ using the locally stored vector $R^{(i)}\mathbf{r}$ and matrix $I^{(i)}$. Send these nontrivial rows to the processor handling coarse space correction. The processor handling the coarse space should sum the components:
$$R_0\,\mathbf{r} \equiv \sum_{i=1}^{n_s} R_0 R^{(i)^T} I^{(i)} R^{(i)}\,\mathbf{r}.$$
2. In parallel, solve $A_i \mathbf{w}_i = R_i\,\mathbf{r}$ for $0 \leq i \leq n_s$.
3. If $\Omega_i^* \cap \Omega_j^* \neq \emptyset$ then the processor handling Ω_i^* should send $R_j R_i^T \mathbf{w}_i$ to the processor handling Ω_j^*. The processor handling the coarse space should send relevant components of $R_0^T \mathbf{w}_0$ to the processor handling Ω_i^*.
4. In parallel, the processor handling Ω_i^* should sum the components:
$$R_i\,M^{-1}\,\mathbf{r} \equiv \sum_{j=0}^{n_s} R_i R_j^T \mathbf{w}_j.$$

The computational time for each step above can be estimated.

If t_i denotes the execution time for the i'th step above, it will satisfy:

$$\begin{cases} t_1 \le c_1\, K\,(1+\beta_*)(n/p)\,\tau_f + c_2\, K\, n_0\,\tau_c + \tau_0 + c_3\, K\, n_0 \tau_f \\ t_2 \le c_4\, \frac{(n_s+1)}{p}\, \phi\left((1+\beta_*)(n/n_s)\right)\, \tau_f \\ t_3 \le c_5\, K\,(1+\beta_*)\,(n/p)\,\tau_c + \tau_0 \\ t_4 \le c_6\, \frac{(n_s+1)}{p}\,(K+1)\,(1+\beta_*)\,(n/n_s)\,\tau_f, \end{cases}$$

provided that $n_0 \le (1+\beta_*)(n/n_s)$. Additionally, if n_s scales proportionally to p, then apart from τ_0, the other terms are inversely proportional to p.

Parallelization of the Neumann-Neumann Preconditioner. We next consider a Neumann-Neumann Schur complement preconditioner, in which the action of the inverse of the preconditioner has the form:

$$M^{-1} = \sum_{i=1}^{n_s} \mathcal{R}_i^T S^{(i)^\dagger} \mathcal{R}_i + \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0,$$

where $S_0 = \mathcal{R}_0 S \mathcal{R}_0^T \asymp A_0$. Care must be exercised when parallelizing the computation of $\mathcal{R}_0^T S_0^{-1} \mathcal{R}_0 \mathbf{r}_B$, since it requires global communication. It will be assumed that the nonzero rows of $\mathcal{R}_0 \mathcal{R}_i^T$ are stored on the processor handling Ω_i. The action of $\mathcal{R}_0$ on $\mathbf{r}_B$ can be computed using the identity:

$$\begin{cases} \mathcal{R}_0 = \mathcal{R}_0 \left(\sum_{i=1}^{n_s} \mathcal{R}_i^T I^{(i)} \mathcal{R}_i \right) \\ \quad\;\; = \sum_{i=1}^{n_s} \left(\mathcal{R}_0 \mathcal{R}_i^T \right) \left(I^{(i)} \mathcal{R}_i \right). \end{cases}$$

Below, we list the implementation of the Neumann-Neumann preconditioner.

1. In parallel, each processor handling Ω_i^* should compute the nontrivial rows of $\mathcal{R}_0 \mathcal{R}_i^T I^{(i)} \mathcal{R}_i \mathbf{r}_B$ using (the locally stored) $\mathcal{R}_i \mathbf{r}_B$ and matrix $I^{(i)}$. Send these nontrivial rows to the processor handling coarse space correction and then sum the components to obtain:
$$\mathcal{R}_0\, \mathbf{r}_B \equiv \sum_{i=1}^{n_s} \mathcal{R}_0 \mathcal{R}_i^T I^{(i)} \mathcal{R}_i\, \mathbf{r}_B.$$
2. In parallel, solve $S^{(i)} \mathbf{w}_i = \mathcal{R}_i \mathbf{r}_B$ for $0 \le i \le n_s$ where $S^{(0)} \equiv S_0$.
3. In parallel, if $\Omega_i^* \cap \Omega_j^* \neq \emptyset$, the processor handling Ω_i^* should send $\mathcal{R}_j \mathcal{R}_i^T \mathbf{w}_i$ to the processor handling Ω_j^*. The processor handling the coarse space should send $\mathcal{R}_i \mathcal{R}_0^T \mathbf{w}_0$ to the processor handling Ω_i^* for $1 \le i \le n_s$.
4. In parallel, the processor handling Ω_i^* should sum the components:
$$\mathcal{R}_i M^{-1}\, \mathbf{r}_B \equiv \sum_{j=0}^{n_s} \mathcal{R}_i \mathcal{R}_j^T\, \mathbf{w}_j.$$

The computation times for the above steps can be estimated.

If t_i denotes the execution time for the i'th step above, it will satisfy:

$$\begin{cases} t_1 \leq c_1 \frac{(n_s+1)}{p} K \, (n/n_s)^{(d-1)/d} \, \tau_f + c_2 \, K \, n_0 \, \tau_c + \tau_0 + c_3 \, K \, n_0 \tau_f \\ t_2 \leq c_4 \frac{(n_s+1)}{p} \, \phi\, (n/n_s) \; \tau_f \\ t_3 \leq c_5 \, K \, (n_s/p) \, (n/n_s)^{(d-1)/d} \, \tau_c + \tau_0 \\ t_4 \leq c_6 \frac{(n_s+1)}{p} \, (K+1) \, (n/n_s)^{(d-1)/d} \, \tau_f, \end{cases}$$

provided that $n_0 = O(n/n_s)$. If n_s is proportional to p, then apart from τ_0, the other terms vary inversely with p.

5.2.4 Estimation of the Total Execution Times

Using the preceding estimates, we may estimate the execution time $T(p, n, \epsilon)$ of CG algorithms for different choices of preconditioners. Here $T(p, n, \epsilon)$ is the total execution time for implementing a PCG algorithm to solve a problem of size n, on a p processor parallel computer, where the initial residual is reduced by a factor ϵ. The total execution time will be the product of the number $N(n, \epsilon)$ of iterations required to reduce the residual by the factor ϵ, and the parallel execution time $T_*(p, n)$ *per iteration*:

$$T(p, n, \epsilon) = N(n, \epsilon) \, T_*(p, n). \tag{5.18}$$

We shall suppress dependence on ϵ for convenience. The execution time $T_*(p, n)$ per iteration can be further decomposed as:

$$T_*(p, n) = G_*(p, n) + H_*(p, n), \tag{5.19}$$

where $G_*(p, n)$ denotes the execution time per iteration of the preconditioning step, while $H_*(p, n)$ denotes the execution time per iteration for the remaining computations (matrix-vector products, inner products, vector addition).

Estimates for $H_*(p, n)$ and $G_*(p, n)$ can be obtained by summing up the relevant execution time estimates t_i for appropriately chosen routines from the preceding pages. Employing the total execution times, we heuristically estimate the parallel efficiency of the additive Schwarz and Neumann-Neumann PCG algorithms, making several simplifying assumptions.

- We assume that the best serial execution time satisfies:

$$T_{\text{best}}(1, n) = \phi(n) \, \tau_f \leq c_0 \, n^{\alpha} \tau_f.$$

- We assume that:

$$\tau_0 = 0, \;\; p^2 \leq n, \;\; p \leq n_s, \;\; n_0 \leq (1 + \beta_*)(n/n_s).$$

- We omit lower order terms in expressions.

We shall express the efficiency in terms of n, p, d, α and $\gamma_c = (\tau_c/\tau_f) \gg 1$.

Additive Schwarz Preconditioner Without Coarse Space. Estimates of $H_*(p,n)$ to solve $A\mathbf{u} = \mathbf{f}$ using a CG algorithm can be obtained by summing the appropriately chosen quantities t_i from the preceding section for matrix-vector products and inner products routines. Estimates of $G_*(p,n)$ can be obtained similarly by summing the t_i from the preceding section for the additive Schwarz preconditioner without a coarse space. We assume that $\tau_0 = 0$, $p^2 \leq n$ and $p \leq n_s$, and omit all lower order terms.

$$\begin{cases} H_*(p,n) \leq d_1\,(n/p)\,\tau_f + d_2\,K\,(n/p)\,\tau_c \\ \qquad\qquad\quad +d_3\,(n_s/p)\,K\,(n/n_s)^{(d-1)/d}\,\tau_f \\ G_*(p,n) \leq c_0\,e_1\,(n_s/p)\,(1+\beta_*)^{\alpha}\,(n/n_s)^{\alpha}\,\tau_f \\ \qquad\qquad\quad +e_2\,K\,\beta_*\,(n/p)\,\tau_c + e_3\,K\,\beta_*\,(n/p)\,\tau_f. \end{cases} \tag{5.20}$$

Bounds from Chap. 2 for the condition number of the additive Schwarz PCG algorithm without coarse space correction yields:

$$\text{cond}(M,A) \leq C(\beta)\,h_0^{-2}.$$

Standard estimates for error reduction in PCG algorithms [GO4] yields:

$$N(n,h_0,\epsilon,\beta) \leq C(\epsilon,\beta)\,h_0^{-1},$$

for some $C(\epsilon,\beta)$ independent of n, n_s and p. Summing $H_*(p,n)$ and $G_*(p,n)$, and retaining only the highest order terms and substituting $h_0^{-1} = O(n_s^{1/d})$, (which holds since $n_s = O(|\Omega|\,h_0^{-d})$), yields:

$$T(p,n,n_s) \leq c_0\,n_s^{1/d}\left(C_1\,\gamma_c\,(n/p) + C_2\,(n_s/p)\,(n/n_s)^{\alpha} + C_3\,(n/p)^{(d-1)/d}\right)\tau_f$$

where $\gamma_c \equiv (\tau_c/\tau_f) \gg 1$. Here C_i may depend on all parameters excluding n, n_s, p and γ_c. Substituting that $T_{\text{best}}(1,n) = c_0\,n^{\alpha}\,\tau_f$ along with the preceding bound for $T(p,n,n_s)$ yields the following heuristic bound for the total efficiency when $p = n_s \leq n^{1/2}$, $\tau_0 = 0$ and $1 < \alpha \leq 3$:

$$\overline{E}(p,n) \geq \left(\frac{n^{\alpha}}{p^{(d+1)/d}\left(C_1\,\gamma_c\,(n/p) + C_2\,(n/p)^{\alpha} + C_3\,(n/p)^{(d-1)/d}\right)}\right).$$

By considering only the leading order terms as p increases, it may be noted that the value of n can be increased to maintain a constant efficiency, as p is varied. Thus the above algorithm is *scalable*. Heuristically, the value of p which minimizes the denominator will optimize the efficiency, for a fixed n.

Additive Schwarz Preconditioner with Coarse Space. Estimates of $G_*(p,n)$ and $H_*(p,n)$ can be obtained for the additive Schwarz preconditioner with coarse space correction by summing the appropriate t_i:

$$\begin{cases} H_*(p,n) \le d_1\,(n/p)\,\tau_f + d_2\,K\,(n/p)\,\tau_c \\ \qquad\qquad + d_3\,(n_s/p)\,K\,(n/n_s)^{(d-1)/d}\,\tau_f \\ G_*(p,n) \le c_0\,e_1\,(n_s/p)\,(1+\beta_*)^{\alpha}\,(n/n_s)^{\alpha}\,\tau_f \\ \qquad\qquad + e_2\,K\,\beta_*\,(n/p)\,\tau_c \end{cases} \tag{5.21}$$

where lower order terms and the start up time τ_0 have been omitted. Bounds from Chap. 2 yield the following estimate for the condition number of the additive Schwarz PCG algorithm with coarse space correction:

$$\text{cond}(M, A) \le C(\beta).$$

Standard estimates for error reduction in PCG algorithms [GO4] yields:

$$N(n, h_0, \epsilon, \beta) \le C(\epsilon, \beta),$$

where $C(\epsilon, \beta)$ is independent of n, n_s. Summing $H_*(p,n)$ and $G_*(p,n)$ and retaining only the highest order terms in $\phi(\cdot)$ yields the following bound:

$$T(p, n, n_s) \le c_0\,(C_1\,\gamma_c\,(n/p) + C_2\,(n_s/p)\,(n/n_s)^{\alpha})\;\tau_f$$

where $\gamma_c \equiv (\tau_c/\tau_f) \gg 1$, and C_i may depend on all parameters excluding n, p and n_s. Substituting $T_{\text{best}}(1,n) = c_0\,n^{\alpha}\,\tau_f$ and the preceding bounds for $T(p,n,n_s)$ yields a bound for $\overline{E}(p,n)$ when $p = n_s \le n^{1/2}$ and $\tau_0 = 0$:

$$\overline{E}(p,n) \ge \left(\frac{n^{\alpha}}{p\,(C_1\,\gamma_c\,(n/p) + C_2\,(n/p)^{\alpha})}\right).$$

The above bound is an improvement over the efficiency of the additive Schwarz algorithm without coarse space correction. By considering only the leading order terms, it is seen that as p is increased, the efficiency can be maintained. Thus, this algorithm is *scalable.* Heuristically, the value of p which minimizes the denominator optimizes the efficiency.

Neumann-Neumann Preconditioner for the Schur Complement. The terms $G_*(p,n)$ and $H_*(p,n)$ can be estimated for the Schur complement algorithm with Neumann-Neumann preconditioner by summing relevant estimates t_i for routines described in the preceding section:

$$\begin{cases} H_*(p,n) \le d_1\,(n_s/p)\,\phi(n/n_s)\,\tau_f + d_2\,K\,(n/p)\,\tau_c \\ G_*(p,n) \le e_1\,K\,(n_s/p)\,\phi(n/n_s)\,\tau_f \\ \qquad\qquad + e_2\,K\,(n_s/p)\,(n/n_s)^{(d-1)/d}\,\tau_c. \end{cases} \tag{5.22}$$

Here, lower order terms and the start up time τ_0 have been omitted. Bounds from Chap. 3 yield the following condition number estimate:

$$\text{cond}(M, A) \le C\,(1 + \log(h_0/h))^2\,,$$

for the Neumann-Neumann algorithm with coarse space correction. Bounds for the error reduction of PCG algorithms [GO4] yields:

$$N(n, h_0, \epsilon, \beta) \le C(\epsilon) \left(1 + \log(h_0/h)\right),$$

where $C(\epsilon)$ is independent of n, n_s. Since by assumption $h_0^{-1} = O(n_s^{1/d})$, it follows that $\log(h_0/h) = O(d^{-1}\log(n/n_s))$. Summing the terms $H_*(p,n)$ and $G_*(p,n)$, substituting $\log(h_0/h) = O(d^{-1}\log(n/n_s))$ and retaining only the highest order terms in $\phi(\cdot)$, yields the following bound for $T(p,n,n_s)$:

$$\begin{aligned} T(p,n,n_s) \le\, & c_0 \log(n/n_s) \left(C_1 (n_s/p) \, (n/n_s)^{\alpha} + C_2 (n/p)^{(d-1)/d} \gamma_c \right) \tau_f \\ & + c_0 \log(n/n_s) \left(C_3 (n/p) \, \gamma_c \right) \tau_f, \end{aligned}$$

where $\gamma_c = (\tau_c/\tau_f) \gg 1$. Substituting the estimate $T_{\text{best}}(1,n) = c_0 \, n^{\alpha} \, \tau_f$ and using the preceding bound for $T(p,n,n_s)$ yields the following lower bound for the total efficiency when $p = n_s \le n^{1/2}$ and $\tau_0 = 0$:

$$\overline{E}(p,n) \ge \left(\frac{n^{\alpha}}{p \log(n/n_s) \left(C_1 \, (n/p)^{\alpha} + C_2 \, (n/p)^{(d-1)/d} \gamma_c + C_3 \, (n/p) \, \gamma_c \right)} \right).$$

By considering only leading order terms, it is seen that as p is increased, a value of n can be determined so that the efficiency in maintained. Thus, this algorithm is *scalable*. An intermediate value of p will optimize the efficiency.

Remark 5.22. The preceding discussion shows that the representative domain decomposition solvers are *scalable*, though not perfectly scalable. Readers are referred to [GR10, GR12, SK, CH15, FA9, SM4, GR16] for additional discussion on the parallel implementation of domain decomposition algorithms.

6

Least Squares-Control Theory: Iterative Algorithms

In this chapter, we describe iterative algorithms formulated based on the least squares-control theory framework [LI2, GL, AT, GU2]. The methodology applies to non-self adjoint elliptic equations, however, for simplicity we shall describe a matrix formulation for the following self adjoint elliptic equation:

$$\begin{cases} L\,u \equiv -\nabla \cdot (a(x)\nabla u) + c(x)u = f, \text{ in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\quad u = 0, \text{ on } \partial\Omega, \end{cases} \tag{6.1}$$

where $c(x) \geq 0$. We denote a finite element discretization of (6.1) as:

$$A\,\mathbf{u} = \mathbf{f} \tag{6.2}$$

where $A = A^T > 0$ is the stiffness matrix of size n and $\mathbf{b} \in \mathbb{R}^n$.

Given a decomposition of Ω into two or more subdomains, a least squares-control formulation of (6.1) employs unknown functions on each subdomain. These unknowns solve the partial differential equation on each subdomain, with unknown boundary data that serve as *control* data. The control data must be chosen so that the subdomain solutions match with neighbors to yield a global solution to (6.1). This problem can be formulated mathematically as a constrained minimization problem, which seeks to minimize the difference between the local unknowns on the regions of overlap, subject to the constraint that the local unknowns solve the elliptic equation on each subdomain. In Chap. 6.1, we consider a decomposition of Ω into *two* overlapping subdomains, while Chap. 6.2 considers two non-overlapping subdomains. Although saddle point methodology may also be employed to solve this least squares-control problem, we reduce it to an *unconstrained* minimization problem, and solve it using a CG algorithm. Some extensions to multiple subdomains are discussed in Chap. 6.3. Our discussion is heuristic and described for its intrinsic interest, since the iterative algorithms based on the Schur complement, Schwarz and Lagrange multiplier formulations are more extensively studied. One of the algorithms elaborates an algorithm from Chap. 1.5.

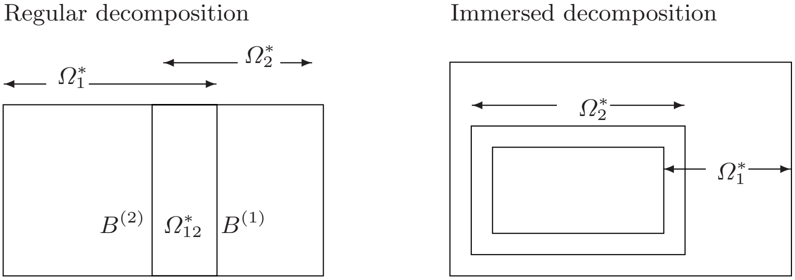


Fig. 6.1. Two overlapping subdomains

6.1 Two Overlapping Subdomains

In this section, we describe a least squares-control formulation of (6.1) based on a decomposition of Ω into two overlapping subdomains [AT]. Our focus will be on the iterative solution of system (6.2). Accordingly, we consider two subdomains Ω_1^* and Ω_2^* which form an overlapping decomposition of Ω with sufficient overlap, as in Fig. 6.1. We define $\Omega_{12}^* \equiv \Omega_1^* \cap \Omega_2^*$ as the region of overlap between the two subdomains. We define $B^{(i)} = \partial\Omega_i \cap \Omega$ as the internal boundary of each subdomain and $B_{[i]} = \partial\Omega_i \cap \partial\Omega$ as its external boundary. Let $\|\cdot\|_{\alpha,\Omega_{12}^*}$ be the fractional Sobolev norm $H^\alpha(\Omega_{12}^*)$ on Ω_{12}^*. for $0 \le \alpha \le 1$.

We shall employ the following functional in the overlapping case:

$$J(v_1, v_2) \equiv \|v_1 - v_2\|_{\alpha,\Omega_{12}^*}^2, \tag{6.3}$$

where $v_1(.)$ and $v_2(.)$ are defined on Ω_1^* and Ω_2^*, respectively. The least squares-control formulation of (6.1) seeks local functions $u_1(.)$ and $u_2(.)$ defined on the subdomains Ω_1^* and Ω_2^*, which minimizes the functional within V_*:

$$J(u_1, u_2) = \min_{(v_1,v_2)\in V_*} J(v_1, v_2) \tag{6.4}$$

where V_* consists of $v_1(.)$ and $v_2(.)$ solving:

$$\begin{cases} L\, v_i = f, & \text{in } \Omega_i^* \\ v_i = g_i, & \text{on } B^{(i)} \qquad \text{for } i = 1, 2 \\ v_i = 0, & \text{on } B_{[i]}. \end{cases} \tag{6.5}$$

Here g_i denotes the *unknown* local Dirichlet data. By construction, if the global solution u to (6.1) exists, then its restriction $u_i(.) \equiv u(.)$ on Ω_i^* for $i = 1, 2$ will minimize $\|u_1 - u_2\|_{\alpha,\Omega_{12}}^2$ with minimum value *zero.*

Remark 6.1. If the solution (u_1, u_2) to (6.4) and (6.5) satisfies $u_1(.) = u_2(.)$ on Ω_{12}^*, then it can easily be verified that $u_i(.)$ will match the true solution $u(.)$ on Ω_i^*. In this case, the Dirichlet boundary data $g_i(.)$ on $B^{(i)}$ can be regarded as *control* data which needs to be determined in order to minimize the square norm error term (6.4).

We shall formulate a matrix version of the above least squares-control formulation using the following notation. We shall order all the nodes in Ω and partition them based on the subregions $\Omega_1^* \setminus \overline{\Omega}_2^*$, $B^{(2)}$, Ω_{12}^*, $B^{(1)}$ and $\Omega_2^* \setminus \overline{\Omega}_1^*$ and define the associated set of indices as:

$$\begin{cases} \mathcal{I}_{11} &= \text{ indices of nodes in } \Omega_1^* \setminus \overline{\Omega}_2^* \\ B^{(2)} &= \text{ indices of nodes in } B^{(2)} \\ \mathcal{I}_{12} &= \text{ indices of nodes in } \Omega_{12}^* \\ B^{(1)} &= \text{ indices of nodes in } B^{(1)} \\ \mathcal{I}_{22} &= \text{ indices of nodes in } \Omega_2^* \setminus \overline{\Omega}_1^*. \end{cases} \tag{6.6}$$

Let n_{11}, $n_B^{(2)}$, n_{12}, $n_B^{(1)}$ and n_{22} denote the number of indices in $\mathcal{I}_{11}$, $B^{(2)}$, $\mathcal{I}_{12}$, $B^{(1)}$ and $\mathcal{I}_{22}$, respectively. Let $n_I^{(1)} = (n_{11} + n_{B^{(2)}} + n_{12})$ be the number of nodes in Ω_1^* and $n_I^{(2)} = (n_{12} + n_B^{(1)} + n_{22})$ the number of nodes in Ω_2^*. Define $n_i = (n_I^{(i)} + n_B^{(i)})$ and $n_{\mathcal{E}} = (n_1 + n_2)$. If v_i denotes a finite element function defined on subdomain $\overline{\Omega}_i^*$, we let $\mathbf{v}^{(i)} = (\mathbf{v}_I^{(i)^T}, \mathbf{v}_B^{(i)^T})^T \in \mathbb{R}^{n_i}$ denote the vector of its interior and boundary nodal values. The indices of nodes in Ω_1^* will be $I^{(1)} = \mathcal{I}_{11} \cup B^{(2)} \cup \mathcal{I}_{12}$ and $I^{(2)} = \mathcal{I}_{12} \cup B^{(1)} \cup \mathcal{I}_{22}$ in Ω_2^*.

We let $A_{II}^{(i)}$ denote a submatrix of A of size $n_I^{(i)}$ corresponding to the indices in $I^{(i)}$, representing coupling between interior nodes in Ω_i^*. Similarly, we let $A_{IB}^{(i)}$ denote an $n_I^{(i)} \times n_B^{(i)}$ submatrix of A representing coupling between nodes in $I^{(i)}$ and $B^{(i)}$, i.e., interior nodes in Ω_i^* with boundary nodes on $B^{(i)}$. A global extended vector consisting of the local subdomain nodal vectors will be denoted $\mathbf{v}_{\mathcal{E}} = (\mathbf{v}^{(1)^T}, \mathbf{v}^{(2)^T})^T \in \mathbb{R}^{n_{\mathcal{E}}}$. Given the original load vector $\mathbf{f} \in \mathbb{R}^n$, we define local interior load vectors $\mathbf{f}_I^{(i)} \in \mathbb{R}^{n_I^{(i)}}$ as the restriction of $\mathbf{f}$ onto the interior nodes in each subdomain, and $\mathbf{f}_B \in \mathbb{R}^{n_B}$ as the restriction of $\mathbf{f}$ onto the nodes on B. Given the ordering $B^{(2)} \cup \mathcal{I}_{12} \cup B^{(1)}$ of nodes in $\overline{\Omega}_{12}^*$, we let R_{12} denote an $n_{12} \times n_1$ restriction matrix which maps a nodal vector on $\overline{\Omega}_1^*$ into its subvector of nodal values on $\overline{\Omega}_{12}^*$. Similarly, we define a restriction matrix R_{21} as an $n_{12} \times n_2$ matrix mapping a vector of nodal values on $\overline{\Omega}_2^*$ into its subvector of nodal values on $\overline{\Omega}_{12}^*$. For $0 \leq \alpha \leq 1$ we let A_α denote a symmetric positive definite matrix of size n_{12} representing the finite element discretization of the $H^\alpha(\Omega_{12}^*)$ Sobolev inner product on $\overline{\Omega}_{12}^*$.

A *discrete* version of the least squares-control problem (6.4) and (6.5) can now be obtained by discretizing the square norm functional and constraints. Accordingly, if (v_1, v_2) are finite element functions defined on (Ω_1^*, Ω_2^*) with associated nodal vectors $\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right)$, we define $\mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right)$ as:

$$\begin{aligned} \mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right) &\equiv \tfrac{1}{2} \|v_1 - v_2\|_{\alpha,\Omega_{12}^*}^2 = \tfrac{1}{2} \|R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\|_{A_\alpha}^2 \\ &= \tfrac{1}{2} \left(R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\right)^T A_\alpha \left(R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\right). \end{aligned} \tag{6.7}$$

The constraints (6.5) can be discretized to yield the following linear system:

$$\begin{cases} A_{II}^{(i)}\mathbf{v}_I^{(i)} + A_{IB}^{(i)}\mathbf{v}_B^{(i)} = \mathbf{f}_I^{(i)} \\ \qquad\qquad\quad \mathbf{v}_B^{(i)} = \mathbf{g}^{(i)} \end{cases} \quad 1 \le i \le 2 \tag{6.8}$$

where $\mathbf{f}_I^{(i)}$ denotes the local internal load vector and $\mathbf{g}^{(i)}$ denotes the unknown discrete Dirichlet boundary data on $B^{(i)}$. Since the second block row above corresponds to a renaming of the Dirichlet boundary data, we eliminate $\mathbf{g}^{(i)}$ and shall henceforth employ $\mathbf{v}_B^{(i)}$.

Remark 6.2. By construction, if $\mathbf{J}\left(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right) = 0$, then the restrictions:

$$R_{12}\mathbf{u}^{(1)} = R_{21}\mathbf{u}^{(2)},$$

of the local nodal vectors will match on $\overline{\Omega}_{12}^*$, and hence their associated finite element functions u_1 and u_2 will also match on the region $\overline{\Omega}_{12}^*$ of overlap.

The objective functional $\mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right)$ and the linear constraints may be expressed compactly using matrix notation. We let K denote a *singular* matrix of size $n_{\mathcal{E}}$ having the following block structure:

$$K = \begin{bmatrix} R_{12}^T A_\alpha R_{12} & -R_{12}^T A_\alpha R_{21} \\ -R_{21}^T A_\alpha R_{12} & R_{21}^T A_\alpha R_{21} \end{bmatrix} \tag{6.9}$$

corresponding to the partitioning $\mathbf{v}_{\mathcal{E}} = (\mathbf{v}^{(1)^T}, \mathbf{v}^{(2)^T})^T$. Then functional $\mathbf{J}(\mathbf{v}_{\mathcal{E}})$ for $\mathbf{v}_{\mathcal{E}} = (\mathbf{v}^{(1)^T}, \mathbf{v}^{(2)^T})^T$ may be equivalently expressed as:

$$\mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right) = \frac{1}{2}\begin{bmatrix} \mathbf{v}^{(1)} \\ \mathbf{v}^{(2)} \end{bmatrix}^T \begin{bmatrix} R_{12}^T A_\alpha R_{12} & -R_{12}^T A_\alpha R_{21} \\ -R_{21}^T A_\alpha R_{12} & R_{21}^T A_\alpha R_{21} \end{bmatrix} \begin{bmatrix} \mathbf{v}^{(1)} \\ \mathbf{v}^{(2)} \end{bmatrix} = \frac{1}{2}\mathbf{v}_{\mathcal{E}}^T K \mathbf{v}_{\mathcal{E}}. \tag{6.10}$$

The constraints (6.8) may be expressed compactly as:

$$N\mathbf{v}_{\mathcal{E}} = \mathbf{f}_{\mathcal{E}}, \quad \text{where } N = \begin{bmatrix} N^{(1)} & 0 \\ 0 & N^{(2)} \end{bmatrix}, \; N^{(i)} \equiv \begin{bmatrix} A_{II}^{(i)} & A_{IB}^{(i)} \end{bmatrix}, \; \mathbf{f}_{\mathcal{E}} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \end{bmatrix} \tag{6.11}$$

where the local nodal vectors $\mathbf{v}^{(i)}$ satisfy $\mathbf{v}^{(i)} = (\mathbf{v}_I^{(i)^T}, \mathbf{v}_B^{(i)^T})^T$. Here N is an $(n_I^{(1)} + n_I^{(2)}) \times n$ rectangular matrix and $N^{(i)}$ is an $n_I^{(i)} \times n_i$ rectangular matrix, of full rank. The discrete least squares-control formulation seeks to minimize $\mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right)$ subject to constraint (6.11), as described in the following result.

Lemma 6.3. *Suppose the following assumptions hold.*

1. *Let* $\mathbf{u}$ *denote the solution of (6.2).*
2. *Let* $\mathbf{w}_{\mathcal{E}} = \left(\mathbf{w}^{(1)^T}, \mathbf{w}^{(2)^T}\right)^T$ *denote an extended nodal vector satisfying:*
$$\mathbf{J}\left(\mathbf{w}_{\mathcal{E}}\right) = \min_{\mathbf{v}_{\mathcal{E}} \in V_*} \mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right) \tag{6.12}$$
where
$$V_* = \left\{\mathbf{v}_{\mathcal{E}} : N\mathbf{v}_{\mathcal{E}} = \mathbf{f}_{\mathcal{E}}\right\}. \tag{6.13}$$
3. *For* $i = 1, 2$ *let* R_i *denote a restriction matrix mapping a nodal vector of the form* $\mathbf{v}$ *onto a vector of nodal values on* $\overline{\Omega}_i^*$.

Then, the following results will hold for $0 \leq \alpha \leq 1$

$$\mathbf{w}^{(i)} = R_i \mathbf{u} \quad \textit{for } i = 1, 2,$$

with $\mathbf{J}\left(\mathbf{w}_{\mathcal{E}}\right) = 0$.

Proof. Follows by construction. □

The constrained minimization problem (6.12) can be reformulated as a saddle point linear system. Indeed, define a Lagrangian function $\mathcal{L}\left(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\lambda}\right)$

$$\mathcal{L}\left(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\lambda}\right) \equiv \mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right) + \boldsymbol{\lambda}^T (N\mathbf{v}_{\mathcal{E}} - \mathbf{f}_{\mathcal{E}}), \tag{6.14}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{n_I^{(1)} + n_I^{(2)}}$ denotes a vector of Lagrange multiplier variables. Then, the saddle point linear system associated with (6.12) is easily derived by requiring the first variation of $\mathcal{L}\left(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\lambda}\right)$ to be zero, as described in Chap. 10:

$$\begin{bmatrix} K & N^T \\ N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_{\mathcal{E}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_{\mathcal{E}} \end{bmatrix}. \tag{6.15}$$

Here matrix K is a *singular* matrix of size $n_{\mathcal{E}}$ having low rank, while matrix N is an $(n_I^{(1)} + n_I^{(2)}) \times n_{\mathcal{E}}$ matrix of full rank. Traditional iterative algorithms based either on augmented Lagrangian formulations [GL7] or the projected gradient method (as in Chap. 4, see [FA14]) may be employed to solve (6.15). However, we shall describe an alternative approach.

Remark 6.4. We briefly outline why system (6.15) will be *nonsingular* even though matrix K is singular. General results in Chap. 10 show that a saddle point system is nonsingular when the following conditions hold.

- Matrix N should have full rank. This is equivalent to the inf sup condition which can easily be verified for (6.11) since $N^{(i)}$ are of full rank.
- Matrix K should be symmetric and *coercive* within the subspace V_0:
$$V_0 = \left\{\mathbf{v}_{\mathcal{E}} : N\mathbf{v}_{\mathcal{E}} = \mathbf{0}\right\}.$$

Suppose the coercivity of K within V_0 is violated, then due to the finite dimensionality of V_0, there must exist a *non-trivial* $\mathbf{v}_{\mathcal{E}} \in V_0$ satisfying $\mathbf{v}_{\mathcal{E}}^T K \mathbf{v}_{\mathcal{E}} = \|R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\|_{A_\alpha}^2 = 0$. By construction $N\mathbf{v}_{\mathcal{E}} = \mathbf{0}$ yields $N^{(i)}\mathbf{v}^{(i)} = \mathbf{0}$ for $i = 1, 2$, and so the restriction $R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}$ will be discrete harmonic on Ω_{12}^*. Since $\|R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\|_{A_\alpha}^2 = 0$, it will hold that $R_{12}\mathbf{v}^{(1)} = R_{21}\mathbf{v}^{(2)}$ and consequently, a global nodal vector $\mathbf{v}$ can be defined matching $\mathbf{v}^{(i)}$ on both subdomains and by construction $\mathbf{v}$ will satisfy $A\,\mathbf{v} = \mathbf{0}$, yielding that $\mathbf{v} = \mathbf{0}$ and $\mathbf{v}^{(i)} = \mathbf{0}$ for $i = 1, 2$. We arrive at a contradiction.

The minimum of $\mathbf{J}(\mathbf{v}_{\mathcal{E}})$ in V_* can alternatively be sought by *parameterizing* V_* and minimizing the resulting *unconstrained* functional. We describe this approach next. The general solution to the full rank system $N\mathbf{v}_{\mathcal{E}} = \mathbf{f}_{\mathcal{E}}$ can be parameterized in terms of the boundary data $\mathbf{v}_B^{(i)}$ by solving $N^{(i)}\mathbf{v}^{(i)} = \mathbf{f}_I^{(i)}$:

$$\mathbf{v}^{(i)} = \begin{bmatrix} -A_{II}^{(i)^{-1}} A_{IB}^{(i)} \\ I \end{bmatrix} \mathbf{v}_B^{(i)} + \begin{bmatrix} A_{II}^{(i)^{-1}} \mathbf{f}_I^{(i)} \\ \mathbf{0} \end{bmatrix} \quad \text{for} \quad i = 1, 2. \tag{6.16}$$

To simplify the expressions, denote the restrictions of such vectors to $\overline{\Omega}_{12}^*$ as:

$$\begin{cases} R_{12}\mathbf{v}^{(1)} = \mathcal{H}_1 \mathbf{v}_B^{(1)} + \mathbf{e}_1 \\ R_{21}\mathbf{v}^{(2)} = \mathcal{H}_2 \mathbf{v}_B^{(2)} + \mathbf{e}_2 \end{cases} \tag{6.17}$$

where

$$\begin{aligned} \mathcal{H}_1 &\equiv R_{12} \begin{bmatrix} -A_{II}^{(1)^{-1}} A_{IB}^{(1)} \\ I \end{bmatrix}, \quad \mathbf{e}_1 \equiv R_{12} \begin{bmatrix} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} \\ \mathbf{0} \end{bmatrix} \\ \mathcal{H}_2 &\equiv R_{21} \begin{bmatrix} -A_{II}^{(2)^{-1}} A_{IB}^{(2)} \\ I \end{bmatrix}, \quad \mathbf{e}_2 \equiv R_{21} \begin{bmatrix} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \\ \mathbf{0} \end{bmatrix}. \end{aligned} \tag{6.18}$$

Here, the subdomain Dirichlet data $\mathbf{v}_B^{(1)}$ and $\mathbf{v}_B^{(2)}$ represent control variables. Substituting this parameterization into the functional $\mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right)$ yields the following reduced functional $\mathbf{J}_B\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right) = \mathbf{J}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$:

$$\mathbf{J}_B\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right) \equiv \tfrac{1}{2} \left\| \left(\mathcal{H}_1 \mathbf{v}_B^{(1)} + \mathbf{e}_1\right) - \left(\mathcal{H}_2 \mathbf{v}_B^{(2)} + \mathbf{e}_2\right) \right\|_{A_\alpha}^2.$$

The new *unconstrained* minimization problem associated with (6.12) and (6.13) seeks boundary data $\mathbf{w}_B^{(1)}$ and $\mathbf{w}_B^{(2)}$ which minimizes:

$$\mathbf{J}_B\left(\mathbf{w}_B^{(1)}, \mathbf{w}_B^{(2)}\right) = \min_{\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right)} \mathbf{J}_B\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right). \tag{6.19}$$

Applying stationarity conditions to:

$$\mathbf{J}_B\left(\mathbf{w}_B^{(1)}, \mathbf{w}_B^{(2)}\right) = \frac{1}{2} \|\mathcal{H}_1 \mathbf{w}_B^{(1)} - \mathcal{H}_2 \mathbf{w}_B^{(2)} + \mathbf{e}_1 - \mathbf{e}_2\|_{A_\alpha}^2$$

yields the linear system:

$$\begin{cases} \frac{1}{2}\frac{\partial J_B}{\partial \mathbf{w}_B^{(1)}} = & \mathcal{H}_1^T A_\alpha \left(\mathcal{H}_1 \mathbf{w}_B^{(1)} - \mathcal{H}_2 \mathbf{w}_B^{(2)} + \mathbf{e}_1 - \mathbf{e}_2 \right) = 0 \\ \frac{1}{2}\frac{\partial J_B}{\partial \mathbf{w}_B^{(2)}} = & -\mathcal{H}_2^T A_\alpha \left(\mathcal{H}_1 \mathbf{w}_B^{(1)} - \mathcal{H}_2 \mathbf{w}_B^{(2)} + \mathbf{e}_1 - \mathbf{e}_2 \right) = 0. \end{cases}$$

Rewriting the above yields a block linear system:

$$\begin{bmatrix} \mathcal{H}_1^T A_\alpha \mathcal{H}_1 & -\mathcal{H}_1^T A_\alpha \mathcal{H}_2 \\ -\mathcal{H}_2^T A_\alpha \mathcal{H}_1 & \mathcal{H}_2^T A_\alpha \mathcal{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{w}_B^{(1)} \\ \mathbf{w}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \mathcal{H}_1^T A_\alpha (\mathbf{e}_2 - \mathbf{e}_1) \\ \mathcal{H}_2^T A_\alpha (\mathbf{e}_1 - \mathbf{e}_2) \end{bmatrix}. \tag{6.20}$$

Henceforth, we assume that the solution to (6.19) is obtained by solving system (6.20). We thus have the following equivalence between (6.12) and (6.19).

Lemma 6.5. *Suppose the following assumptions hold.*

1. *Let* $\left(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right)^T$ *denote the constrained minimum of* $\mathbf{J}(\cdot,\cdot)$*:*

$$\mathbf{J}\left(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right) = \min_{(\mathbf{v}^{(1)},\mathbf{v}^{(2)}) \in V_*} \mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right).$$

2. *Let* $\left(\mathbf{w}_B^{(1)}, \mathbf{w}_B^{(2)}\right)^T$ *denote the unconstrained minimum of* $J_B(\cdot,\cdot)$*:*

$$\mathbf{J}_B\left(\mathbf{w}_B^{(1)}, \mathbf{w}_B^{(2)}\right) = \min_{(\mathbf{v}_B^{(1)},\mathbf{v}_B^{(2)})} \mathbf{J}_B\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right).$$

Then, the following results will hold:

$$\mathbf{u}^{(i)} = \begin{bmatrix} A_{II}^{(i)^{-1}} \left(\mathbf{f}_I^{(i)} - A_{IB}^{(i)} \mathbf{w}_B^{(i)} \right) \\ \mathbf{w}_B^{(i)} \end{bmatrix} \qquad i = 1, 2.$$

Proof. Follows by direct substitution and algebraic simplification. □

Remark 6.6. The coefficient matrix in (6.20) is symmetric by construction. Importantly, it will also be positive definite, and an iterative method such as CG algorithm may be applied to solve (6.20). To verify that the coefficient matrix in (6.20) is positive definite, without loss of generality let $\mathbf{e}_i = \mathbf{0}$ for $i = 1, 2$. Since the coefficient matrix in (6.20) generates the quadratic form associated with a square norm, it will be positive *semidefinite*:

$$\mathbf{J}_B\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right) = \left\| \mathcal{H}_1 \mathbf{v}_B^{(1)} - \mathcal{H}_2 \mathbf{v}_B^{(2)} \right\|_{A_\alpha}^2 \geq 0.$$

To show definiteness, note that:

$$\mathbf{J}_B\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right) = \mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right) = \|R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\|_{A_\alpha}^2$$

where $\mathbf{v}^{(i)}$ is the discrete harmonic extension of the boundary data $\mathbf{v}_B^{(i)}$ when $\mathbf{e}_i = \mathbf{0}$. Suppose for contradiction that $\|R_{12}\mathbf{v}^{(1)} - R_{21}\mathbf{v}^{(2)}\|_{A_\alpha}^2 = 0$. Then $R_{12}\mathbf{v}^{(1)} = R_{21}\mathbf{v}^{(2)}$ and a global nodal vector $\mathbf{v}$ can be defined matching the two nodal vectors $\mathbf{v}^{(i)}$ on the local grids. By construction $\mathbf{v}$ will be discrete harmonic globally, and hence imply that $\mathbf{v}$ and $\mathbf{v}^{(i)}$ are identically zero. Definiteness follows by the preceding, due to finite dimensionality.

Next, using *heuristic* arguments we outline a preconditioner M for the coefficient matrix F in (6.20) when $\alpha = 1$. The preconditioner we describe can be motivated by considering a rectangular domain Ω with rectangular overlapping subdomains and a uniform triangulation. For a suitable ordering of the nodes, A will be block tridiagonal matrix of the form *blocktridiag* $(-I, T, -I)$, as described in Chap. 3.3. Matrix T will be diagonalized by the discrete sine transform. If the boundary data $\mathbf{v}_B^{(i)}$ on $B^{(i)}$ corresponds to an eigenvector $\mathbf{q}_l$ of matrix T, then $R_{12}\mathcal{H}_1\mathbf{v}_B^{(1)}$ will be scalar multiples of $\mathbf{q}_l$ along each vertical grid line in Ω_1^*. Similarly for $R_{21}\mathcal{H}_2\mathbf{v}_B^{(2)}$. It can then be verified that:

$$\begin{aligned} c\left(\|c_1\mathbf{q}_l\|^2_{1/2,B^{(1)}} + \|c_2\mathbf{q}_l\|^2_{1/2,B^{(2)}}\right) &\le \|R_{12}\mathcal{H}_1\, c_1\,\mathbf{q}_l - R_{21}\mathcal{H}_2\, c_2\,\mathbf{q}_l\|^2_{A_1} \\ \|R_{12}\mathcal{H}_1\, c_1\,\mathbf{q}_l - R_{21}\mathcal{H}_2\, c_2\,\mathbf{q}_l\|^2_{A_1} &\le C\left(\|c_1\mathbf{q}_l\|^2_{1/2,B^{(1)}} + \|c_2\mathbf{q}_l\|^2_{1/2,B^{(2)}}\right). \end{aligned}$$

Using superposition and orthogonality of the modes (in the Euclidean and discrete fractional Sobolev norms) we obtain for general $\left(\mathbf{v}_B^{(1)}, \mathbf{v}_B^{(2)}\right)$:

$$\begin{aligned} c\left(\|\mathbf{v}_B^{(1)}\|^2_{1/2,B^{(1)}} + \|\mathbf{v}_B^{(2)}\|^2_{1/2,B^{(2)}}\right) &\le \|R_{12}\mathcal{H}_1\,\mathbf{v}_B^{(1)} - R_{21}\mathcal{H}_2\,\mathbf{v}_B^{(2)}\|^2_{A_1} \\ \|R_{12}\mathcal{H}_1\,\mathbf{v}_B^{(1)} - R_{21}\mathcal{H}_2\,\mathbf{v}_B^{(2)}\|^2_{A_1} &\le C\left(\|\mathbf{v}_B^{(1)}\|^2_{1/2,B^{(1)}} + \|\mathbf{v}_B^{(2)}\|^2_{1/2,B^{(2)}}\right). \end{aligned}$$

Similar bounds are *heuristically* expected to hold for more general domains and operators, where $c < C$ are independent of h. Based on this observation, a suitable preconditioner M when $\alpha = 1$ can be obtained by *decoupling* the two boundary segments and defining a block diagonal preconditioner whose diagonal blocks correspond to discretizations of the fractional Sobolev norms:

$$M^{-1} = \begin{bmatrix} M_B^{(1)^{-1}} & 0 \\ 0 & M_B^{(2)^{-1}} \end{bmatrix},$$

where $M_B^{(i)}$ denotes any standard two subdomain interface preconditioner for the two subdomain Schur complement matrix $S_B^{(i)}$ on the interface $B^{(i)}$. An alternative preconditioner which *couples* the two boundary values may be constructed as follows. Suppose that the nodes in $\overline{\Omega}_{12}^*$ are partitioned and ordered according to $\mathcal{I}_{12}$, $B^{(1)}$and $B^{(2)}$. Define the action of the inverse of the preconditioner as:

$$M^{-1}\begin{bmatrix} \mathbf{r}_B^{(1)} \\ \mathbf{r}_B^{(2)} \end{bmatrix} \equiv \begin{bmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{bmatrix}^T A_\alpha^{-1} \begin{bmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{r}_B^{(1)} \\ \mathbf{r}_B^{(2)} \end{bmatrix}$$

where matrix A_α will have the following block structure for $\alpha = 1$:

$$A_1 = \begin{bmatrix} A_{II}^{(12)} & A_{IB^{(1)}}^{(12)} & A_{IB^{(2)}}^{(12)} \\ A_{IB^{(1)}}^{(12)^T} & A_{B^{(1)}B^{(1)}}^{(12)} & 0 \\ A_{IB^{(2)}}^{(12)^T} & 0 & A_{B^{(2)}B^{(2)}}^{(12)} \end{bmatrix}.$$

If F denotes the coefficient matrix in (6.20) when $\alpha = 1$, we heuristically expect $\text{cond}\,(M, F) \leq C$ where $C > 0$ is independent of h for both of the above preconditioners, when $\alpha = 1$. Below, we summarize the discrete least squares-control algorithm assuming that $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{f}_I^{(i)}$ and A_α have been assembled using (6.18). Subroutines for computing the action of R_{12}, R_{21}, $\mathcal{H}_1$ and $\mathcal{H}_2$ will also be required, based on the expressions in (6.18).

Algorithm 6.1.1 *(Least Squares-Control Overlapping Algorithm)*

1. *Solve using a preconditioned conjugate gradient method:*

$$\begin{bmatrix} \mathcal{H}_1^T A_\alpha \mathcal{H}_1 & -\mathcal{H}_1^T A_\alpha \mathcal{H}_2 \\ -\mathcal{H}_2^T A_\alpha \mathcal{H}_1 & \mathcal{H}_2^T A_\alpha \mathcal{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_B^{(1)} \\ \mathbf{u}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \mathcal{H}_1^T A_\alpha (\mathbf{e}_2 - \mathbf{e}_1) \\ \mathcal{H}_2^T A_\alpha (\mathbf{e}_1 - \mathbf{e}_2) \end{bmatrix}.$$

2. *Compute:*

$$\begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \end{bmatrix} = \begin{bmatrix} A_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - A_{IB}^{(1)} \mathbf{u}_B^{(1)} \right) \\ A_{II}^{(2)^{-1}} \left(\mathbf{f}_I^{(2)} - A_{IB}^{(2)} \mathbf{u}_B^{(2)} \right) \end{bmatrix}.$$

Define $\mathbf{u}^{(1)} = \left(\mathbf{u}_I^{(1)^T}, \mathbf{u}_B^{(1)^T} \right)^T$ *and* $\mathbf{u}^{(2)} = \left(\mathbf{u}_I^{(2)^T}, \mathbf{u}_B^{(2)^T} \right)^T$.

6.2 Two Non-Overlapping Subdomains

We next describe a least squares-control iterative algorithm for solving (6.1) based on a decomposition of Ω into *two* non-overlapping regions [GU2, GU3]. Let Ω_1 and Ω_2 denote a nonoverlapping decomposition of Ω with interface $B = \partial\Omega_1 \cap \partial\Omega_2$, and exterior boundary segments $B_{[i]} = \partial\Omega_i \cap \partial\Omega$ for $i = 1, 2$, as in Fig. 6.2. Our focus will be on a matrix implementation of the least square-control algorithm for solving (6.2) when $c(x) \geq c_0 > 0$ in (6.1). For index $0 \leq \alpha \leq \frac{1}{2}$, we shall let $\|\cdot\|_{\alpha,B}^2$ denote the square of the fractional Sobolev $H^\alpha(B)$ norm on the interface B.

We shall employ the following functional in the non-overlapping case:

$$J(v_1, v_2) \equiv \|v_1 - v_2\|_{\alpha,B}^2, \tag{6.21}$$

where $v_1(.)$ and $v_2(.)$ are defined on Ω_1 and Ω_2, respectively. The least squares-control formulation of (6.1) seeks local functions $u_1(.)$ and $u_2(.)$ defined on

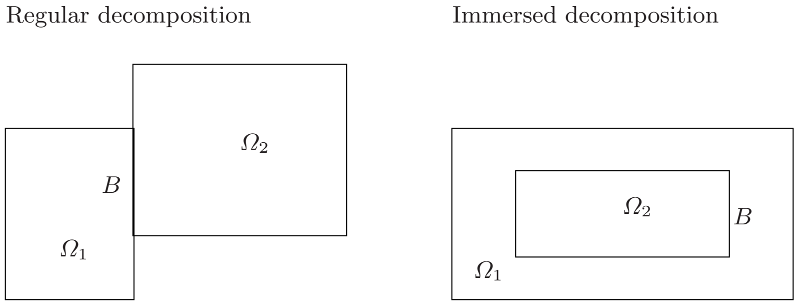


Fig. 6.2. Two subdomain non-overlapping decomposition

the subdomains Ω_1 and Ω_2, which minimizes the functional within V_*:

$$J(u_1, u_2) = \min_{(v_1,v_2)\in V_*} J(v_1, v_2) \tag{6.22}$$

where V_* consists of $v_1(.)$ and $v_2(.)$ satisfying:

$$\begin{cases} Lv_i = f, & \text{in } \Omega_i \\ v_i = 0, & \text{on } B_{[i]} \end{cases} \quad \text{and} \quad \mathbf{n}_1 \cdot (a(x)\nabla v_1) + \mathbf{n}_2 \cdot (a(x)\nabla v_2) = 0, \ \text{ on } B. \tag{6.23}$$

Here $\mathbf{n}_i$ denotes the unit exterior normal to $\partial\Omega_i$ on B. The last constraint $\mathbf{n}_1 \cdot (a(x)\nabla v_1) + \mathbf{n}_2 \cdot (a(x)\nabla v_2) = 0$ on B, enforces continuity of the local fluxes on B, and is a flux *transmission* condition. By minimizing $\|v_1 - v_2\|^2_{\alpha,B}$, the model seeks to enforce continuity of the local solutions across B.

By construction, if $u_i \equiv u$ on Ω_i denotes the restriction of the true solution u of (6.1) onto subdomain Ω_i, then the constraints in (6.23) will be satisfied and $\|u_1 - u_2\|^2_{\alpha,B}$ will attain a minimum value of zero. The constraints in (6.23) can be *parameterized* using the local fluxes on B. For $i = 1, 2$ denote by $g_i(x)$ the flux on B associated with the unknown v_i:

$$g_i(x) \equiv \mathbf{n}_i(x) \cdot (a(x)\nabla v_i). \quad \text{on } \ B, \tag{6.24}$$

Then, constraint (6.23) will be equivalent to:

$$\begin{cases} L\, v_i \equiv -\nabla \cdot (a(x)\nabla u) + c(x)u = f, & \text{in } \Omega_i \\ \mathbf{n}_i \cdot (a(x)\nabla v_i) = g_i, & \text{on } B \qquad \text{for } \ i = 1, 2 \\ v_i = 0, & \text{on } B_{[i]}, \end{cases} \tag{6.25}$$

along with the *transmission* condition requirement:

$$g_1(x) + g_2(x) = 0, \quad \text{on } \ B. \tag{6.26}$$

The Neumann data $g_i(x)$ parameterizing the local solutions can be regarded as *control* data which is to be chosen to minimize the square norm (6.22).

We shall employ the following notation to obtain a matrix formulation of the above least squares-control problem. The indices of the nodes on Ω will be partitioned as $I^{(1)}$, $I^{(2)}$ and B where:

$$\begin{cases} I^{(1)} = \text{ indices of nodes in } \Omega_1 \\ I^{(2)} = \text{ indices of nodes in } \Omega_2 \\ B \quad = \text{ indices of nodes in } B. \end{cases} \tag{6.27}$$

The number of indices in $I^{(1)}$, $I^{(2)}$, B will be denoted $n_I^{(1)}$, $n_I^{(2)}$ and n_B, respectively. We let $n_1 = (n_I^{(1)} + n_B)$ and $n_2 = (n_I^{(2)} + n_B)$ denote the number of nodes in $\overline{\Omega}_1$ and $\overline{\Omega}_2$, respectively, and $n = (n_I^{(1)} + n_I^{(2)} + n_B)$ the number of nodes in Ω. Given a finite element function v_i defined on subdomain $\overline{\Omega}_i$, we let $\mathbf{v}^{(i)} = (\mathbf{v}_I^{(i)^T}, \mathbf{v}_B^{(i)^T})^T \in \mathbb{R}^{n_i}$ denote a vector of its nodal values. We define $\mathbf{v}_{\mathcal{E}} = (\mathbf{v}^{(1)^T}, \mathbf{v}^{(2)^T})^T$ as an extended nodal vector consisting of the two subdomain nodal vectors (with nonmatching $\mathbf{v}_B^{(1)} \neq \mathbf{v}_B^{(2)}$). On each subdomain $\overline{\Omega}_i$ let $\mathbf{f}^{(1)} = (\mathbf{f}_I^{(i)^T}, \mathbf{f}_B^{(i)^T})^T$ denote the subdomain load vector, with global load vector given by $\mathbf{f} = (\mathbf{f}_I^{(1)^T}, \mathbf{f}_I^{(2)^T}, \mathbf{f}_B^T)^T$ for $\mathbf{f}_B = (\mathbf{f}_B^{(1)} + \mathbf{f}_B^{(2)})$. As in substructuring, we let $A_{II}^{(i)}$, $A_{IB}^{(i)}$ and $A_{BB}^{(i)}$ denote the submatrices of the subdomain stiffness matrix $A^{(i)}$. On the interface B, for $0 \leq \alpha \leq \frac{1}{2}$ we let A_α denote the finite element discretization of the fractional Sobolev inner product $H^\alpha(B)$.

A discrete version of the least squares-control problem (6.22) and (6.23) can be constructed by discretizing the square norm and constraints. Given finite element functions v_1 and v_2 on Ω_1 and Ω_2 with associated nodal vectors $(\mathbf{v}_I^{(1)^T}, \mathbf{v}_B^{(1)^T})^T$ and $(\mathbf{v}_I^{(2)^T}, \mathbf{v}_B^{(2)^T})^T$, we define an objective functional $\mathbf{J}(\mathbf{v}_{\mathcal{E}})$ for $\mathbf{v}_{\mathcal{E}} = (\mathbf{v}^{(1)^T}, \mathbf{v}^{(2)^T})^T$ as:

$$\begin{cases} \mathbf{J}\left(\mathbf{v}\right) \equiv \frac{1}{2} \left\| \mathbf{v}_B^{(1)} - \mathbf{v}_B^{(2)} \right\|_{A_\alpha}^2 \\ \qquad\;\; = \frac{1}{2} \left(\mathbf{v}_B^{(1)} - \mathbf{v}_B^{(2)} \right)^T A_\alpha \left(\mathbf{v}_B^{(1)} - \mathbf{v}_B^{(2)} \right). \end{cases} \tag{6.28}$$

The constraints in (6.23) may be discretized as:

$$\begin{cases} A_{II}^{(1)} \mathbf{v}_I^{(1)} + A_{IB}^{(1)} \mathbf{v}_B^{(1)} = \mathbf{f}_I^{(1)} \\ A_{II}^{(2)} \mathbf{v}_I^{(2)} + A_{IB}^{(2)} \mathbf{v}_B^{(2)} = \mathbf{f}_I^{(2)} \\ A_{IB}^{(1)^T} \mathbf{v}_I^{(1)} + A_{BB}^{(1)} \mathbf{v}_B^{(1)} + A_{IB}^{(2)^T} \mathbf{v}_I^{(2)} + A_{BB}^{(2)} \mathbf{v}_B^{(2)} = \mathbf{f}_B, \end{cases} \tag{6.29}$$

where $\mathbf{f}_B \neq \mathbf{0}$ to ensure that the global discretization is consistent with (6.2).

Remark 6.7. It will be computationally convenient to parameterize the local discrete fluxes based on (6.25) and (6.26). We thus express (6.29) as:

$$\begin{cases} A_{II}^{(i)} \mathbf{v}_I^{(i)} + A_{IB}^{(i)} \mathbf{v}_B^{(i)} \quad = \mathbf{f}_I^{(i)} \\ A_{IB}^{(i)^T} \mathbf{v}_I^{(i)} + A_{BB}^{(i)} \mathbf{v}_B^{(i)} = \mathbf{g}_i \end{cases} \quad \text{for } i = 1, 2 \tag{6.30}$$

where $\mathbf{g}_i \in \mathbb{R}^{n_B}$ denotes the unknown local fluxes which must satisfy:

$$\mathbf{g}_1 + \mathbf{g}_2 = \mathbf{f}_B. \tag{6.31}$$

We shall parameterize the local fluxes as:

$$\begin{cases} \mathbf{g}_1 = \mathbf{f}_B^{(1)} + \mathbf{g}_B, \\ \mathbf{g}_2 = \mathbf{f}_B^{(2)} - \mathbf{g}_B, \end{cases} \tag{6.32}$$

for some unknown flux vector $\mathbf{g}_B \in \mathbb{R}^{n_B}$. By construction, the sum of the second block rows in (6.30) yields the third block row in (6.29) using (6.32).

The objective functional and constraints may be expressed compactly as:

$$\mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right) = \frac{1}{2}\,\mathbf{v}_{\mathcal{E}}^T K \mathbf{v}_{\mathcal{E}} = \frac{1}{2} \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_B^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B^{(2)} \end{bmatrix}^T \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & A_\alpha & 0 & -A_\alpha \\ 0 & 0 & 0 & 0 \\ 0 & -A_\alpha & 0 & A_\alpha \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_B^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B^{(2)} \end{bmatrix}. \tag{6.33}$$

The constraint (6.29) may be expressed in matrix form as:

$$N\mathbf{v}_{\mathcal{E}} = \mathbf{f}, \text{ where } N = \begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} & 0 & 0 \\ 0 & 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ A_{IB}^{(1)^T} & A_{BB}^{(1)} & A_{IB}^{(2)^T} & A_{BB}^{(2)} \end{bmatrix}, \quad \mathbf{v}_{\mathcal{E}} = \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_B^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B^{(2)} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B. \end{bmatrix}. \tag{6.34}$$

The discrete least squares-control problem will seek to minimize $\mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right)$ subject to constraint (6.34), as described in the following result.

Lemma 6.8. *Suppose the following conditions hold.*

1. *Let* $\mathbf{u}$ *denote the solution of (6.2).*
2. *Let* $\mathbf{w}_{\mathcal{E}} = \left(\mathbf{w}^{(1)^T}, \mathbf{w}^{(2)^T}\right)^T$ *denote an extended nodal vector satisfying:*

$$\mathbf{J}\left(\mathbf{w}_{\mathcal{E}}\right) = \min_{\mathbf{v}_{\mathcal{E}} \in V_*} \mathbf{J}\left(\mathbf{v}_{\mathcal{E}}\right) \tag{6.35}$$

where

$$V_* = \left\{\mathbf{v}_{\mathcal{E}} \,:\, N\mathbf{v}_{\mathcal{E}} = \mathbf{f}\right\}. \tag{6.36}$$

3. *For* $i = 1, 2$ *let* R_i *denote a restriction matrix mapping a nodal vector of the form* $\mathbf{u}$ *onto a vector of nodal values on* $\overline{\Omega}_i$.

Then, the following results will hold for $0 \le \alpha \le \frac{1}{2}$

$$\mathbf{w}^{(i)} = R_i \mathbf{u} \qquad \text{for } i = 1, 2,$$

with $\mathbf{J}(\mathbf{w}_{\mathcal{E}}) = 0$.

Proof. Follows by construction. □

A saddle point formulation of (6.35) can be obtained by seeking the stationary point of the following Lagrangian functional:

$$\mathcal{L}\left(\mathbf{v}_{\mathcal{E}}, \boldsymbol{\lambda}\right) = \frac{1}{2}\, \mathbf{v}_{\mathcal{E}}^T K \mathbf{v}_{\mathcal{E}} + \boldsymbol{\lambda}^T (N\mathbf{v}_{\mathcal{E}} - \mathbf{f}),$$

with Lagrange multiplier variables $\boldsymbol{\lambda} \in \mathbb{R}^n$. The saddle point system will be:

$$\begin{bmatrix} K \; N^T \\ N \; 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_{\mathcal{E}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f} \end{bmatrix}, \tag{6.37}$$

where matrix K is a *singular* matrix of low rank. Iterative algorithms such as the augmented Lagrangian algorithm [GL7] or the projected gradient algorithm as in FETI [FA14], can be applied to solve (6.37). Instead, we describe an alternative algorithm based on parameterizing $V_* = \{\mathbf{v}_{\mathcal{E}} : N\mathbf{v}_{\mathcal{E}} = \mathbf{f}\}$.

The vectors in the constraint set V_* can be parameterized in terms of the vector $\mathbf{g}_B \in \mathbb{R}^{n_B}$ as described below. Eliminate the interior variables $\mathbf{v}_I^{(i)}$ using the first block equation in (6.30) to obtain:

$$\mathbf{v}_I^{(i)} = A_{II}^{(i)^{-1}} \left(\mathbf{f}_I^{(i)} - A_{IB}^{(i)} \mathbf{v}_B^{(i)}\right). \tag{6.38}$$

Substituting this into the second block row of (6.30) yields:

$$S^{(i)} \mathbf{v}_B^{(i)} = \mathbf{g}_i - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} \mathbf{f}_I^{(i)}, \tag{6.39}$$

where $S^{(i)} \equiv A_{BB}^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} A_{IB}^{(i)}$ denotes the local Schur complement matrix. When the local Schur complement matrix $S^{(i)}$ is nonsingular, the solution to the above Schur complement system can be expressed as:

$$\begin{cases} \mathbf{v}_B^{(1)} = S^{(1)^{-1}} \left(\mathbf{f}_B^{(1)} + \mathbf{g}_B - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)}\right) \\ \mathbf{v}_B^{(2)} = S^{(2)^{-1}} \left(\mathbf{f}_B^{(2)} - \mathbf{g}_B - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)}\right), \end{cases} \tag{6.40}$$

where we have substituted parameterization (6.32) of the local flux vectors $\mathbf{g}_1 = \mathbf{f}_B^{(1)} + \mathbf{g}_B$ and $\mathbf{g}_2 = \mathbf{f}_B^{(2)} - \mathbf{g}_B$. We express this more compactly as:

$$\begin{cases} \mathbf{v}_B^{(1)} = \mathcal{H}_1 \mathbf{g}_B + \mathbf{e}_1 \\ \mathbf{v}_B^{(2)} = \mathcal{H}_2 \mathbf{g}_B + \mathbf{e}_2, \end{cases} \tag{6.41}$$

where

$$\begin{cases} \mathcal{H}_1 \equiv \quad S^{(1)^{-1}} \\ \mathcal{H}_2 \equiv - S^{(2)^{-1}} \\ \mathbf{e}_i \;\equiv \quad S^{(i)^{-1}} \left(\mathbf{f}_B^{(i)} - A_{IB}^{(i)^T} A_{II}^{(i)^{-1}} \mathbf{f}_I^{(i)}\right), \text{ for } i = 1, 2. \end{cases} \tag{6.42}$$

An unconstrained minimization problem equivalent to the constrained minimization problem (6.35) can be obtained as follows. Define a functional

$\mathbf{J}_B(\mathbf{g}_B)$ equivalent to $\mathbf{J}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$ as follows:

$$\mathbf{J}_B(\mathbf{g}_B) \equiv \frac{1}{2} \, \|(\mathcal{H}_1 \mathbf{g}_B + \mathbf{e}_1) - (\mathcal{H}_2 \mathbf{g}_B + \mathbf{e}_2)\|_{A_\alpha}^2 . \tag{6.43}$$

Then, substituting the parameterization (6.41) and (6.38) of the constraint set V_* into the constrained minimization problem (6.35) yields the following minimization problem for the unknown flux vector $\mathbf{g}_B \in \mathbb{R}^{n_B}$:

$$\mathbf{J}_B(\mathbf{g}_B^*) = \min_{\mathbf{g}_B} \mathbf{J}_B(\mathbf{g}_B). \tag{6.44}$$

Once the desired control vector $\mathbf{g}_B^*$ has been determined, the components of the solution to (6.35) can be obtained as follows:

$$\begin{cases} \mathbf{u}_B^{(1)} &= S^{(1)^{-1}} \left(\mathbf{f}_B^{(1)} + \mathbf{g}_B^* - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} \right) \\ \mathbf{u}_I^{(1)} &= A_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - A_{IB}^{(1)} \mathbf{u}_B^{(1)} \right) \\ \mathbf{u}_B^{(2)} &= S^{(2)^{-1}} \left(\mathbf{f}_B^{(2)} - \mathbf{g}_B^* - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \right) \\ \mathbf{u}_I^{(2)} &= A_{II}^{(2)^{-1}} \left(\mathbf{f}_I^{(2)} - A_{IB}^{(2)} \mathbf{u}_B^{(2)} \right). \end{cases} \tag{6.45}$$

The following equivalence property will hold.

Lemma 6.9. *Suppose the following assumptions hold.*

1. *Let* $\mathbf{u}_{\mathcal{E}} = \left(\mathbf{u}^{(1)^T}, \mathbf{u}^{(2)^T} \right)^T$ *denote the constrained minimum of* $\mathbf{J}(\mathbf{v}_{\mathcal{E}})$*:*

$$\mathbf{J}\left(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right) = \min_{(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}) \in V_*} \mathbf{J}\left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right).$$

2. *Let* $\mathbf{g}_B^*$ *denote the unconstrained minimum of* $\mathbf{J}_B(\cdot)$*:*

$$\mathbf{J}_B\left(\mathbf{g}_B^*\right) = \min_{\mathbf{g}_B} \mathbf{J}_B\left(\mathbf{g}_B\right).$$

Then, the following result will hold:

$$\begin{cases} \mathbf{u}_B^{(1)} &= S^{(1)^{-1}} \left(\mathbf{f}_B^{(1)} + \mathbf{g}_B^* - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} \right) \\ \mathbf{u}_I^{(1)} &= A_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - A_{IB}^{(1)} \mathbf{u}_B^{(1)} \right) \\ \mathbf{u}_B^{(2)} &= S^{(2)^{-1}} \left(\mathbf{f}_B^{(2)} - \mathbf{g}_B^* - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \right) \\ \mathbf{u}_I^{(2)} &= A_{II}^{(2)^{-1}} \left(\mathbf{f}_I^{(2)} - A_{IB}^{(2)} \mathbf{u}_B^{(2)} \right). \end{cases}$$

Proof. Follows by direct substitution and algebraic simplification. □

A linear system for the control vector $\mathbf{g}_B$ can be obtained by applying stationarity conditions to the functional $J_B(\mathbf{g}_B)$:

$$\tfrac{\partial J_B}{\partial \mathbf{g}_B^*} = (\mathcal{H}_1 - \mathcal{H}_2)^T A_\alpha \left[(\mathcal{H}_1 - \mathcal{H}_2)\mathbf{g}_B + (\mathbf{e}_1 - \mathbf{e}_2)\right] = 0.$$

This may be rewritten as:

$$(\mathcal{H}_1 - \mathcal{H}_2)^T A_\alpha (\mathcal{H}_1 - \mathcal{H}_2)\, \mathbf{g}_B = (\mathcal{H}_1 - \mathcal{H}_2)^T A_\alpha (\mathbf{e}_2 - \mathbf{e}_1). \tag{6.46}$$

We next describe a choice of matrix A_α and a choice of preconditioner M for the coefficient matrix $F = (\mathcal{H}_1 - \mathcal{H}_2)^T A_\alpha (\mathcal{H}_1 - \mathcal{H}_2)$, based on the properties of Schur complements described in Chap. 3.

Since matrix $\mathcal{H}_1 - \mathcal{H}_2 = S^{(1)^{-1}} + S^{(2)^{-1}}$ is symmetric and positive definite when $c(x) \ge c_0 > 0$, if we define $A_\alpha = (\mathcal{H}_1 - \mathcal{H}_2)^{-1}$, then the following linear system will be obtained for the Neumann control vector $\mathbf{g}_B^*$:

$$\left(S^{(1)^{-1}} + S^{(2)^{-1}}\right) \mathbf{g}_B^* = (\mathbf{e}_2 - \mathbf{e}_1).$$

The choice $A_\alpha = (\mathcal{H}_1 - \mathcal{H}_2)^{-1}$ will be spectrally equivalent to the matrix arising from the discretization of the $H^{1/2}(B)$ inner product. In this case any choice of preconditioner M spectrally equivalent to S^{-1} (or $S^{(i)^{-1}}$) will yield a spectrally equivalent preconditioner for $(\mathcal{H}_1 - \mathcal{H}_2)$. The action M^{-1} of the preconditioner M would then correspond to multiplication by any preconditioner for the Schur complement matrix S, $S^{(1)}$ or $S^{(2)}$, requiring the solution of a Dirichlet problem (as in the FETI Dirichlet preconditioner [FA14]).

In the following, we summarize the preceding least squares-control algorithm assuming that the vectors $\mathbf{e}_1$, $\mathbf{e}_2$ have been assembled and that subroutines are available for computing the action of matrices $\mathcal{H}_1$ and $\mathcal{H}_2$. We shall summarize the algorithm below for the choice $A_\alpha = \left(S^{(1)^{-1}} + S^{(2)^{-1}}\right)^{-1}$.

Algorithm 6.2.1 *(Least Squares-Control Nonoverlapping Algorithm)*

1. *Solve using a preconditioned conjugate gradient method:*

$$(\mathcal{H}_1 - \mathcal{H}_2)\, \mathbf{g}_B^* = (\mathbf{e}_1 - \mathbf{e}_2).$$

2. *Compute:*

$$\begin{cases} \mathbf{u}_B^{(1)} = S^{(1)^{-1}} \left(\mathbf{f}_B^{(1)} + \mathbf{g}_B^* - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)}\right) \\ \mathbf{u}_I^{(1)} = A_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - A_{IB}^{(1)} \mathbf{u}_B^{(1)}\right) \\ \mathbf{u}_B^{(2)} = S^{(2)^{-1}} \left(\mathbf{f}_B^{(2)} - \mathbf{g}_B^* - A_{IB}^{(2)^T} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)}\right) \\ \mathbf{u}_I^{(2)} = A_{II}^{(2)^{-1}} \left(\mathbf{f}_I^{(2)} - A_{IB}^{(2)} \mathbf{u}_B^{(2)}\right). \end{cases}$$

By construction, the following convergence bounds will hold.

Lemma 6.10. *Suppose the following assumptions hold.*

1. *Let* $A_\alpha \equiv \left(S^{(1)^{-1}} + S^{(2)^{-1}}\right)^{-1}$.
2. *Let preconditioner* M *satisfy:*

$$c_0 \le \frac{\mathbf{w}_B^T S \mathbf{w}_B}{\mathbf{w}_B^T M^{-1} \mathbf{w}_B} \le c_1, \qquad \textit{for } \mathbf{w}_B \neq \mathbf{0}.$$

Then there will exist $C > 0$ *independent of* h *such that:*

$$\text{cond}(M, \mathcal{H}_1 - \mathcal{H}_2) \le C \left(\frac{c_1}{c_0}\right).$$

Proof. Follows trivially from the assumptions. We omit the details. □

Remark 6.11. The least squares-control formulation described here has been formulated for *self adjoint* coercive elliptic equations. However, most practical applications of the least squares-control formulation, are to nonlinear and heterogeneous problems [GL, GL13, AT, GU2]. The methodology described here, based on an explicit parameterization of the constraint set, may still be possible in such applications, however, the parametric map will be nonlinear when the underlying elliptic equation is nonlinear.

Remark 6.12. The *first order system least squares* method provides an alternative methodology for various classes of partial differential equations, see [AZ, BO, CA23]. The emphasis in such formulations is on the reduction of the original partial differential equation into a first order system of partial differential equations with a subsequent application of the least squares method to determine its solution.

Remark 6.13. If the local load vectors $\mathbf{f}_B^{(1)}$ and $\mathbf{f}_B^{(2)}$ are not available, we may employ $\mathbf{g}_1 = \mathbf{f}_B + \mathbf{g}_B$ and $\mathbf{g}_2 = -\mathbf{g}_B$ instead.

6.3 Extensions to Multiple Subdomains

We briefly consider *heuristic* extensions of the least squares-control method when there are more than two subdomains. We consider both overlapping and non-overlapping decompositions. Let $\Omega_1, \ldots, \Omega_p$ denote a non-overlapping decomposition of Ω with subdomains of width h_0. Also let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping decomposition of Ω, where:

$$\Omega_l^* \equiv \{x \in \Omega : \text{dist}(x, \Omega_l) < \beta\, h_0\} \qquad \text{for } 1 \le l \le p,$$

for some $0 < \beta < 1$. We let $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$. For each index l, we define:

$$\mathcal{O}(l) = \{j : B_{lj} \neq \emptyset\}.$$

By construction $j \in \mathcal{O}(l)$ iff $l \in \mathcal{O}(j)$. We let $\mathcal{I}(l)$ denote an index set:

$$\mathcal{I}(l) \subset \mathcal{O}(l)$$

such that either $j \in \mathcal{I}(l)$ or $l \in \mathcal{I}(j)$ but *not both.*

6.3.1 Multiple Overlapping Subdomains

Given an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ define $B^{(l)} \equiv \partial \Omega_l^* \cap \Omega$ as the interior boundary segment and $B_{[l]} \equiv \partial \Omega_l^* \cap \partial \Omega$ as the exterior boundary segment of $\partial \Omega_l^*$. To obtain a least squares-control hybrid formulation of (6.1) based on the overlapping decomposition, given $v_l(.)$ defined on Ω_l^* define:

$$J(v_1, \ldots, v_p) \equiv \frac{1}{2} \sum_{l=1}^{p} \sum_{j \in \mathcal{I}(l)} \|v_l - v_j\|_{\alpha, B_{lj}}^2$$

where $\|\cdot\|_{\alpha, B_{lj}}$ denotes the $H^{\alpha}(B_{lj})$ Sobolev norm on B_{lj}. One possible least squares-control hybrid formulation of (6.1) is to seek:

$$J(u_1, \ldots, u_p) = \min_{(v_1, \ldots, v_p) \in V_*} J(v_1, \ldots, v_p) \tag{6.47}$$

where V_* consists of $v_l(.)$ defined on Ω_l^* satisfying:

$$\begin{cases} L\, v_l = f, & \text{in } \Omega_l^* \\ \quad v_l = g_l, & \text{on } B^{(l)} \quad \text{for } 1 \le l \le p. \\ \quad v_l = 0, & \text{on } B_{[l]}. \end{cases} \tag{6.48}$$

Here g_l denotes *unknown* local Dirichlet data on $B^{(l)}$. By construction, if $u_l \equiv u$ restricted to Ω_l^*, then it will solve the minimization problem with $J(u_1, \ldots, u_p) = 0$. In principle, a multisubdomain iterative solver can be formulated to solve (6.2) based on the minimization of (6.47) subject to the constraints (6.48).

6.3.2 Multiple Non-Overlapping Subdomains

Given the non-overlapping decomposition $\Omega_1, \ldots, \Omega_p$ define $B^{(l)} \equiv \partial \Omega_l \cap \Omega$ as the interior boundary segment and $B_{[l]} \equiv \partial \Omega_l \cap \partial \Omega$ as the exterior boundary segment of $\partial \Omega_l$. To obtain a least squares-control hybrid formulation of (6.1) based on this decomposition, given $v_l(.)$ defined on Ω_l^*, define:

$$J(v_1, \ldots, v_p) \equiv \frac{1}{2} \sum_{l=1}^{p} \sum_{j \in \mathcal{I}(l)} \|v_l - v_j\|_{\alpha, B_{lj}}^2$$

where $\|\cdot\|_{\alpha, B_{lj}}$ denotes the $H^{\alpha}(B_{lj})$ Sobolev norm on B_{lj}. Let $g(.)$ denote a *flux* function defined on $\cup_{l=1}^{p} \partial \Omega_l$ and parameterizing the local solutions in a constraint set V_* as follows:

$$\begin{cases} L\, v_l \equiv -\nabla \cdot (a \nabla v_l) + c\, v_l = & f, \ \text{in } \Omega_l^* \\ \qquad\qquad \mathbf{n}_l \cdot (a \nabla v_l) = & g, \ \text{on } B_{lj} \ \text{for } \ j \in \mathcal{I}(l) \\ \qquad\qquad \mathbf{n}_l \cdot (a \nabla v_l) = & -g, \ \text{on } B_{lj} \ \text{for } \ l \in \mathcal{I}(j) \\ \qquad\qquad\qquad\qquad v_l = & 0, \ \text{on } B_{[l]}. \end{cases} \quad \text{for } 1 \le l \le p. \tag{6.49}$$

Then, the following least squares-control formulation will be heuristically equivalent to the original elliptic equation (6.1):

$$J(u_1, \ldots, u_p) = \min_{(v_1, \ldots, v_p) \in V_*} J(v_1, \ldots, v_p) \tag{6.50}$$

By construction, if $u_l \equiv u$ restricted to Ω_l, then it will solve the minimization problem with $J(u_1, \ldots, u_p) = 0$. In principle, a multisubdomain iterative solver can be formulated based on this hybrid formulation.

7

Multilevel and Local Grid Refinement Methods

In this chapter, we describe iterative methods for solving the discretization of a self adjoint and coercive elliptic equation on a grid with a *multilevel* structure. Such grids are obtained by the successive refinement of an initial coarse grid, either globally or locally. When the refinement is global, the resulting grid is *quasi-uniform*, while if the refinement is restricted to subregions, the resulting grid will *not* be quasi-uniform. We describe preconditioners formulated using multigrid methodology [BR22, HA4, HA2, BR36]. Multilevel preconditioners can yield optimal order performance, like multigrid methods, however, they are convergent only with Krylov space acceleration.

In Chap. 7.1, we describe multilevel preconditioners for a discretization of an elliptic equation on a globally quasi-uniform grid, obtained by J successive refinements of a coarse grid. The hierarchical basis preconditioner [YS2, BA16, ON, AX4], BPX preconditioner [XU, BR20] and a multilevel Schwarz preconditioner [ZH2] are described. From an algorithmic viewpoint, these preconditioners have the structure of an additive Schwarz preconditioner, where each subdomain restriction map is replaced by a restriction onto a subspace defined on the multilevel triangulation [XU3]. Additionally, the subspace projections are computed inexactly.

In Chap. 7.2, we describe iterative algorithms for solving a discretization of an elliptic equation on a non-quasi-uniform locally refined (composite) grid. Such grids are obtained by the repeated local refinement of a conforming grid within *selected subregions* [MC4, BR7, HA9, DR12]. More specifically, the grid refinement procedure refines elements only within selected subregions where the solution is irregular, leaving all other elements intact, see Fig. 7.2. As a result, locally refined grids violate standard element adjacency requirements along the boundaries of refined regions. Despite this, a conforming finite element discretization can be constructed on such grids, by introducing *slave variables*. We describe the BEPS, FAC and AFAC iterative algorithms for solving discretizations of elliptic equations on locally refined grids, see also [MC4, HA9, DR12, MA21, MA22, BR6, EW6, EW7, CH23].

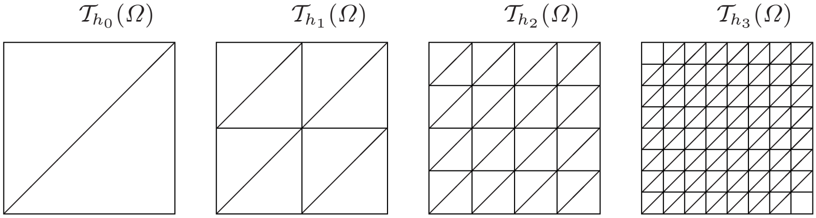

$\mathcal{T}_{h_0}(\Omega)$
$\mathcal{T}_{h_1}(\Omega)$
$\mathcal{T}_{h_2}(\Omega)$
$\mathcal{T}_{h_3}(\Omega)$

Fig. 7.1. Multilevel hierarchy of refined grids

7.1 Multilevel Iterative Algorithms

In this section, we shall describe the hierarchical basis preconditioner, the BPX preconditioner and the multilevel Schwarz preconditioner. for solving a discretization of the following self adjoint and coercive elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{7.1}$$

where $c(x) \geq 0$. From a matrix viewpoint, these preconditioners have the structure of matrix Schwarz preconditioners with approximate solvers on the hierarchy of grids. Let $\mathcal{T}_{h_0}(\Omega)$ denote an initial quasiuniform triangulation of Ω of mesh size h_0 containing n_0 interior nodes. For $1 \leq l \leq J$ let $\mathcal{T}_{h_l}(\Omega)$ denote a triangulation of Ω obtained by refinement of $\mathcal{T}_{h_{l-1}}(\Omega)$, see Fig. 7.1, with grid size $h_l = (h_{l-1}/2)$ and containing n_l interior nodes. For $0 \leq l \leq J$ we denote by $V_{h_l} \subset H_0^1(\Omega)$ the finite element space of dimension n_l defined on triangulation $\mathcal{T}_{h_l}(\Omega)$. By construction, the spaces V_{h_l} will be nested:

$$V_{h_0} \subset V_{h_1} \subset \cdots \subset V_{h_J}.$$

We enumerate the interior nodes of triangulation $\mathcal{T}_{h_l}(\Omega)$ as $x_1^{(l)}, \ldots, x_{n_l}^{(l)}$ and denote by $\phi_i^{h_l}(x)$ for $1 \leq i \leq n_l$ the standard finite element nodal basis for V_{h_l} satisfying $\phi_i^{h_l}(x_j^{(l)}) = \delta_{ij}$ where δ_{ij} is the Kronecker delta. On the finest grid level, we employ the notation $n = n_J$, $h = h_J$, $\mathcal{T}_h(\Omega) = \mathcal{T}_{h_J}(\Omega)$, $x_i = x_i^{(J)}$, $\phi_i(x) = \phi_i^{h_J}(x)$ and $V_h = V_{h_J}$. We discretize elliptic equation (7.1) using the finite element space V_h and denote the resulting symmetric and positive definite linear system of size n as:

$$A\mathbf{u} = \mathbf{f}. \tag{7.2}$$

For $0 \leq l \leq J$ we formally define an extension matrix $R_{h_l}^T$ as an $n \times n_l$ matrix:

$$R_{h_l}^T = \begin{bmatrix} \phi_1^{h_l}(x_1) & \cdots & \phi_{n_l}^{h_l}(x_1) \\ \vdots & & \vdots \\ \phi_1^{h_l}(x_n) & \cdots & \phi_{n_l}^{h_l}(x_n) \end{bmatrix}.$$

This matrix will be sparse and its action, or that of its transpose R_{h_l}, can be computed recursively in $O(n)$ complexity, see [YS2, ON]. For $0 \le l \le J$ we let G_{h_l} and A_{h_l} denote the mass and stiffness matrices, respectively, associated with the finite element space V_{h_l}, and denote the $L^2(\Omega)$-orthogonal projection onto V_{h_l} by Q_{h_l}. A matrix representation of Q_{h_l} can be obtained as:

$$Q_{h_l} = R_{h_l}^T G_{h_l}^{-1} R_{h_l} G_h,$$

where $G_{h_l} = R_{h_l} G_h R_{h_l}^T$. We shall omit the subindex when $l = J$.

7.1.1 Hierarchical Basis Preconditioner

The hierarchical basis preconditioner [YS2, BA16, ON] is motivated by a change of basis for V_h, from the standard nodal basis $\{\phi_i^h(x)\}$ to a basis referred to as the hierarchical basis, defined based on *interpolation* onto the hierarchy of grids and finite element spaces. The hierarchical basis preconditioner for matrix A will correspond to a diagonal (or block diagonal) matrix approximation of A, relative to this new basis. When $\mathcal{T}_h(\Omega) = \mathcal{T}_{h_J}(\Omega)$ has been obtained by J successive refinements of a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$ of Ω, we let I_{h_l} denote the nodal *interpolation map* onto the finite element space V_{h_l} for $0 \le l \le J$ as follows:

$$I_{h_l} w(x) = \sum_{i=1}^{n_l} w(x_i^{(l)})\, \phi_i^{h_l}(x). \tag{7.3}$$

By telescoping the interpolation maps, the following identity is obtained:

$$I_{h_J} = I_{h_0} + (I_{h_1} - I_{h_0}) + \cdots + \left(I_{h_J} - I_{h_{J-1}}\right), \tag{7.4}$$

since $I_{h_J} = I$ on V_h. The hierarchical basis preconditioner will formally correspond to an *approximate* abstract additive Schwarz preconditioner based on the subspaces $\text{Range}(I_{h_l} - I_{h_{l-1}}) \subset V_h$ for $0 \le l \le J$, where $I_{h_{-1}} \equiv 0$. A matrix representation of this preconditioner can be obtained as follows. Let $z_i^{(0)} = x_i^{(0)}$ for $1 \le i \le n_0$ denote the nodes on the coarsest grid. For $1 \le l \le J$ let $z_i^{(l)}$ for $1 \le i \le (n_l - n_{l-1})$ denote the nodes in $\mathcal{T}_{h_l}(\Omega)$ which are not in $\mathcal{T}_{h_{l-1}}(\Omega)$. Thus, by construction, all the nodes can be partitioned as follows:

$$\{x_1, \ldots, x_n\} = \cup_{l=0}^J \left\{ z_1^{(l)}, \ldots, z_{n_l - n_{l-1}}^{(l)} \right\}.$$

A hierarchical basis for V_h is then defined as follows.

- The first n_0 functions in the hierarchical basis will consist of the standard nodal basis for the coarsest space V_{h_0}. We shall denote them as $\{\psi_i^{(0)}(\cdot)\}$:

$$\psi_i^{(0)}(x) = \phi_i^{h_0}(x), \quad \text{for } 1 \le i \le n_0.$$

- The remaining hierarchical basis functions will be recursively defined for $1 \leq l \leq J$ as follows. If grid point $z_i^{(l)}$ in $\mathcal{T}_{h_l}(\Omega)$ (but not in $\mathcal{T}_{h_{l-1}}(\Omega)$) corresponds to node $x_j^{(l)}$, then define:

$$\psi_i^{(l)}(x) \equiv \phi_j^{(h_l)}(x) \in V_{h_l}, \text{ for } 1 \leq i \leq (n_l - n_{l-1}).$$

By construction $\left\{ \{\psi_i^{(l)}(x)\}_{i=1}^{n_l - n_{l-1}} \right\}_{l=0}^{J}$ will form a basis for V_h and it can be verified that the following will hold:

$$\begin{aligned} u_h(x) &= (I_{h_0} u_h)(x) + \textstyle\sum_{l=1}^{J} \left(I_{h_l} u_h - I_{h_{l-1}} u_h \right)(x) \\ = \textstyle\sum_{i=1}^{n_0} u_h(z_i^{(0)}) \psi_i^{(0)}(x) &+ \textstyle\sum_{l=1}^{J} \sum_{i=1}^{n_l - n_{l-1}} \left(u_h(z_i^{(l)}) - (I_{h_{l-1}} u_h)(z_i^{(l)}) \right) \psi_i^{(l)}(x). \end{aligned}$$

Note that $(I_{h_l} u_h - I_{h_{l-1}} u_{h_l})(x_j^{(l-1)}) = 0$ for $1 \leq j \leq n_{l-1}$ so that we may expand $(I_{h_l} u_h - I_{h_{l-1}} u_{h_l})$ solely in terms of the basis $\{\psi_k^{(l)}\}_{k=1}^{n_l - n_{l-1}}$. Theoretical estimates show that the off diagonal entries of the stiffness matrix, relative this hierarchical basis, decay in magnitude [YS2, BA16, ON]. Motivated by this, the hierarchical basis preconditioner for A is chosen to be a diagonal (or a block diagonal matrix) relative to this new basis. However, relative to the standard nodal basis, the action of the inverse of the hierarchical basis preconditioner M will have the following form:

$$M^{-1} = R^T D^{-1} R, \tag{7.5}$$

where D denotes a diagonal or block diagonal matrix of size n and R denotes a matrix of size n representing the transformation from the standard nodal basis to hierarchical basis. The structure of the matrices R and D are described below, corresponding to a matrix additive Schwarz preconditioner.

Using the hierarchical basis functions $\{\psi_i^{(l)}(x)\}_{i=1}^{n_l - n_{l-1}}$, for $0 \leq l \leq J$ define an extension matrix R_l^T of size $n \times (n_l - n_{l-1})$ as follows:

$$R_l^T = \begin{bmatrix} \psi_1^{(l)}(x_1) & \cdots & \psi_{n_l - n_{l-1}}^{(l)}(x_1) \\ \vdots & & \vdots \\ \psi_1^{(l)}(x_n) & \cdots & \psi_{n_l - n_{l-1}}^{(l)}(x_n) \end{bmatrix}.$$

Define the subspaces $V_l = \text{span}(R_l^T) \subset \mathbb{R}^n$ for $0 \leq l \leq J$. Then $\dim(V_0) = n_0$ and $\dim(V_l) = n_l - n_{l-1}$ with $n = n_0 + (n_1 - n_0) + \cdots + (n_J - n_{J-1})$. The hierarchical basis preconditioner corresponds to a matrix additive Schwarz preconditioner based on the above extension matrices.

The action of the inverse of the hierarchical basis preconditioner is:

$$M^{-1} = \sum_{l=0}^{J} R_l^T D_l^{-1} R_l, \tag{7.6}$$

where ideally $D_l = R_l A R_l^T$. In practice D_0 is replaced by a coarse grid approximation of A_{h_0} while D_l is chosen as some diagonal approximation of $R_l A R_l^T$ for $1 \le l \le J$. Furthermore, the action of R_l can be implemented algorithmically using tree data structures in $O(n \log(n))$ flops. For details on the implementation of R_l and the construction of D_l, see [YS2, ON, BA16].

Lemma 7.1. *The condition number of the hierarchical basis preconditioned system satisfies:*

$$\operatorname{cond}(M, A) \le \begin{cases} C\,\left(1 + \log^2(h_J)\right), & \text{if } \Omega \subset \mathbb{R}^2 \\ C\,h_J^{-1}, & \text{if } \Omega \subset \mathbb{R}^3. \end{cases}$$

where $C > 0$ is independent of $h_i$and J.

Proof. The theoretical bound depends on estimates of strengthened Cauchy-Schwarz inequalities between the subspaces $V_{h_l} = \text{Range}(I_{h_l} - I_{h_{l-1}})$, see [YS2, BA16, ON]. □

Remark 7.2. The convergence rate of the hierarchical basis preconditioned system deteriorates in three dimensions. However, modified preconditioners with improved convergence properties have been constructed, see [AX4]. Additionally, hierarchical basis preconditioners may also be formulated analogously for Schur complement matrices, provided triangulation Ω_h has a multilevel structure, see [SM6, TO].

7.1.2 BPX Preconditioner

The BPX preconditioner of [XU, BR20] is a parallel multilevel preconditioner for A with an *optimal order* convergence rate independent of the mesh parameters $h_0, \ldots, h_J$ and levels J. It is motivated by an important Sobolev norm equivalence property which holds when the underlying finite element space has a multilevel structure. For $u \in V_{h_J}$ the following norms are equivalent:

$$\|u\|^2_{H^1(\Omega)} \asymp h_0^{-2} \|\mathcal{Q}_{h_0} u\|^2_{L^2(\Omega)} + \sum_{l=1}^{J} h_l^{-2} \, \|(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}})u\|^2_{L^2(\Omega)} \tag{7.7}$$

with equivalence parameters independent of the number of levels J and the mesh parameters $h_0, \ldots, h_J$, see [BR20, OS2, BO6, OS3, LO]. Here $\mathcal{Q}_{h_l}$ denotes the $L^2(\Omega)$-orthogonal projection onto V_{h_l}. Let $\mathcal{A}_h$ denote the linear map generating the form $a(\cdot, \cdot)$ on V_h endowed with the $L^2(\Omega)$-inner product:

$$a(u, v) = \int_\Omega (a(x)\nabla u \cdot \nabla u + c(x)uv)\, dx \equiv (\mathcal{A}_h u, v)_{L^2(\Omega)}\,, \quad u, v \in V^h. \tag{7.8}$$

Then, since the form $(\mathcal{A}_h u, u)_{L^2(\Omega)}$ is spectrally equivalent to $\|u\|^2_{H^1(\Omega)}$, the norm equivalence (7.7) suggests the following approximation $\mathcal{M}_h$ of $\mathcal{A}_h$ on V_h

equipped with the $L^2(\Omega)$ inner product:

$$\mathcal{M}_h = h_0^{-2}\, \mathcal{Q}_{h_0} + \sum_{l=1}^{J} h_l^{-2} \left(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}} \right) \tag{7.9}$$

where $\mathcal{Q}_{h_0}$, $(\mathcal{Q}_{h_1} - \mathcal{Q}_{h_0}), \ldots, (\mathcal{Q}_{h_J} - \mathcal{Q}_{h_{J-1}})$ are $L^2(\Omega)$-orthogonal projections which are mutually orthogonal. Formally, the map $\mathcal{M}_h$ defines the BPX preconditioner. Its eigenspaces in V_{h_J} consist of Range$(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}}) \subset V_{h_l}$ of dimension $(n_l - n_{l-1})$ associated with eigenvalue h_l^{-2} for $0 \le l \le J$. Its formal inverse $\mathcal{M}_h^{-1}$ in V_h equipped with $(\cdot,\cdot)_{L^2(\Omega)}$ will be:

$$\mathcal{M}_h^{-1} = h_0^{2}\, \mathcal{Q}_{h_0} + \sum_{l=1}^{J} h_l^{2} \left(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}} \right) \tag{7.10}$$

as can be verified directly by multiplying $\mathcal{M}_h$ and $\mathcal{M}_h^{-1}$.

To obtain a matrix representation M_h^{-1} of the inverse of $\mathcal{M}_h^{-1}$ relative to the Euclidean inner product, given $u \in V_h$, we evaluate the energy associated with the right hand side of (7.10) in the $L^2(\Omega)$ inner product and simplify:

$$\begin{cases} \left(\mathcal{M}_h^{-1} u, u\right)_{L^2(\Omega)} = h_0^2 \, \|\mathcal{Q}_{h_0} u\|^2_{L^2(\Omega)} + \sum_{l=1}^{J} h_l^2 \, \|(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}}) u\|^2_{L^2(\Omega)} \\ \quad = h_0^2 \, \|\mathcal{Q}_{h_0} u\|^2_{L^2(\Omega)} + \sum_{l=1}^{J} h_l^2 \, \left(\|\mathcal{Q}_{h_l} u\|^2_{L^2(\Omega)} - \|\mathcal{Q}_{h_{l-1}} u\|^2_{L^2(\Omega)} \right) \\ \quad = \dfrac{3}{4} \left(\sum_{l=0}^{J-1} h_l^2 \, \|\mathcal{Q}_{h_l} u\|^2_{L^2(\Omega)} \right) + h_J^2 \, \|\mathcal{Q}_{h_J} u\|^2_{L^2(\Omega)} \\ \quad = \dfrac{3}{4} \left(\sum_{l=0}^{J-1} h_l^2 \, \left(\mathcal{Q}_{h_l} u, u\right)_{L^2(\Omega)} \right) + h_J^2 \, \left(\mathcal{Q}_{h_J} u, u\right)_{L^2(\Omega)}. \end{cases} \tag{7.11}$$

If $\mathbf{u}$ denotes the nodal vector associated with a finite element function u and G_{h_l} denotes the Gram (mass) matrix associated with V_{h_l}, then substituting $\|\mathcal{Q}_{h_l} u\|^2_{L^2(\Omega)} = \mathbf{u}^T G_h R_{h_l}^T G_{h_l}^{-1} R_{h_l} G_h \mathbf{u}$ into the above yields the following representation of M_h^{-1} relative the Euclidean inner product:

$$M_h^{-1} = \frac{3}{4} \left(\sum_{l=0}^{J-1} h_l^2 \, G_h R_{h_l}^T G_{h_l}^{-1} R_{h_l} G_h \right) + h_J^2 \, G_{h_J} R_{h_J}^T G_{h_J}^{-1} R_{h_J} G_{h_J}. \tag{7.12}$$

We may further approximate the action of the inverse of the BPX preconditioner by replacing the factors $\frac{3}{4}$ by 1. Additionally, when $\Omega \subset \mathbb{R}^d$, the Gram matrices G_{h_l} will be diagonally dominant and can be approximated by $h_l^d \, I$ where I denotes an identity matrix of size n_l.

The preceding approximations yield:

$$M_h^{-1} \asymp h_J^{2d} \, R_{h_0}^T h_0^{2-d} R_{h_0} + \sum_{l=1}^{J} h_J^{2d} \, R_{h_l}^T h_l^{2-d} R_{h_l}. \tag{7.13}$$

As a scaling of M_h will not alter cond(M, A_h), the term h_J^{2d} may be omitted. The following theoretical bound will hold for cond$(\mathcal{M}_h, \mathcal{A}_h)$.

Lemma 7.3. *The BPX preconditioner* $\mathcal{M}_h$ *defined formally by:*

$$\mathcal{M}_h = h_0^{-2} \mathcal{Q}_{h_0} + \sum_{i=1}^{J} h_i^{-2} \left(\mathcal{Q}_{h_i} - \mathcal{Q}_{h_{i-1}} \right),$$

will satisfy $\mathrm{cond}(\mathcal{M}_h, \mathcal{A}_h) \leq C$ *independent of* $h_0, \ldots, h_J$ *and* J.

Proof. The original proof in [BR20] yielded a nonoptimal bound of $C\,J^2$ (or a bound $C\,J$ assuming full elliptic regularity of (7.1)). An improved bound of $C\,J$ was obtained in [ZH2, ZH] without elliptic regularity assumptions. An optimal order bound C was obtained in [OS2, OS3]. Alternative proofs of the optimal order bounds are given in [GR3, BO6]. □

Remark 7.4. The BPX preconditioner may be regarded as an abstract additive Schwarz preconditioner on the finite element space V_h based on the subspaces $V_0 = V_{h_0}$ and $V_l = \mathrm{Range}(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}})$ for $1 \leq l \leq J$, where a scalar multiple of the identity operator approximates the subspace problem. The action of the BPX preconditioner can be computed recursively in linear order complexity. The BPX preconditioner can also be generalized to precondition discretizations of other fractional Sobolev norms [BR17]. For instance, the BPX norm equivalence (7.7) also holds for $0 \leq \alpha \leq 1$ and $u \in V_h$:

$$\|u\|_{H^\alpha(\Omega)}^2 \asymp h_0^{-2\alpha} \, \|\mathcal{Q}_{h_0} u\|_{L^2(\Omega)}^2 + \sum_{l=1}^{J} h_l^{-2\alpha} \, \|(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}})u\|_{L^2(\Omega)}^2. \quad (7.14)$$

This suggests the following spectral equivalences: in the inner product $(\cdot, \cdot)_{L^2(\Omega)}$:

$$\begin{cases} \mathcal{M}_\alpha \;\; = h_0^{-2\alpha} \, \mathcal{Q}_{h_0} + \sum_{l=1}^{J} h_l^{-2\alpha} \, (\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}}) \\ \mathcal{M}_\alpha^{-1} = h_0^{2\alpha} \, \mathcal{Q}_{h_0} + \sum_{l=1}^{J} h_l^{2\alpha} \big(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}}\big). \end{cases}$$

A heuristic Euclidean approximation of $\mathcal{M}_\alpha^{-1}$ for $0 < \alpha \leq 1$ is:

$$M_\alpha^{-1} \approx h_J^{2d} \sum_{l=0}^{J} R_{h_l}^T h_l^{2\alpha - d} R_{h_l}.$$

Efficient ways to compute the action of M_α^{-1} are described in [BR17]. In domain decomposition, if a two subdomain interface $B = \partial\Omega_1 \cap \partial\Omega_2$ has a multilevel grid, the Schur complement S will satisfy $M_{1/2} \asymp S$, see [BR17].

Remark 7.5. The BPX preconditioning methodology can also be applied to construct preconditioners for implicit discretizations of parabolic equations, without dependence on mesh parameters, number of levels or the time step. If A_h denotes the positive definite stiffness matrix arising from a discretization of a self adjoint elliptic operator, then an implicit discretization of its associated parabolic equation will yield a singularly perturbed elliptic equation with coefficient matrix $G_h + \tau A_h$ for $0 < \tau \ll 1$. An optimal order BPX type preconditioner can be formulated for $\mathcal{M}_\tau = \mathcal{G}_h + \tau \mathcal{A}_h$ as:

$$\mathcal{M}_\tau^{-1} = \left(1 + \tau h_0^{-2}\right)^{-1} \mathcal{Q}_{h_0} + \sum_{l=1}^{J} \left(1 + \tau h_l^{-2}\right)^{-1} \left(\mathcal{Q}_{h_l} - \mathcal{Q}_{h_{l-1}}\right),$$

with spectral bounds independent of τ, h_i and J. Its matrix form will be:

$$M_\tau^{-1} \approx (G_h + \tau A_h)^{-1} \approx h_J^{2d} \sum_{l=0}^{J} R_{h_l}^T \left(\frac{3\,\tau h_l^{-2-d}}{1+5\,\tau\, h_l^2 + 4\tau^2\, h_l^{-4}}\right) R_{h_l}.$$

The scaling factor h_J^{2d} can be omitted in the above preconditioner. The reader is referred to [BR17] for more efficient approximations of M_τ^{-1}.

Remark 7.6. The BPX preconditioner does not directly take into the account coefficient variations in $a(x)$. See [XU] for a variant of BPX which takes coefficients into account. The multilevel Schwarz algorithm, which is described next, generalizes the BPX preconditioner and two level Schwarz algorithms, and incorporates coefficient variation.

7.1.3 Multilevel Schwarz Preconditioner

Multilevel Schwarz preconditioners [ZH, ZH2, WA2] are abstract Schwarz subspace preconditioners which employ subdomain problems on different grid levels. Suppose that triangulation $\mathcal{T}_h(\Omega) = \mathcal{T}_{h_J}(\Omega)$ is obtained by J successive refinements of a coarse triangulation $\mathcal{T}_{h_0}(\Omega)$. Then, given an overlapping decomposition $\{\Omega_l^*\}_{l=1}^p$, a *two level* Schwarz algorithm will employ the fine grid spaces $V_{h_J} \cap H_0^1(\Omega_l^*)$ for $1 \leq l \leq p$ and a coarse space $V_{h_m} \cap H_0^1(\Omega)$ defined on $\mathcal{T}_{h_m}(\Omega)$ where $m < J$, for global transfer of information. If the dimension of the coarse space V_{h_m} is large, it may be advantageous to *recursively* decompose this coarse problem on level m using subspaces of the form $V_{h_m} \cap H_0^1(\Omega_j^*)$ and a coarser space $V_{h_{m_2}}$ where $m_2 < m_1 \equiv m < J$. The multilevel Schwarz algorithm formalizes such a recursive procedure, by incorporating domain decomposition and multigrid methodology, to involve subproblems on various subdomains of various grid sizes in the hierarchy.

Let $\mathcal{T}_h(\Omega) = \mathcal{T}_{h_J}(\Omega)$ denote the finest triangulation with associated finite element space V_{h_J}. Let $0 = l_1 < l_2 < \cdots < l_k = J$ denote integer values representing grid levels to be employed in a k-level Schwarz algorithm.

- On each level l_j except $l_1 = 0$, decompose the domain into m_j overlapping subdomains $\Omega_1^{(l_j)}, \ldots, \Omega_{m_j}^{(l_j)}$ of Ω whose boundaries align with $\mathcal{T}_{h_{l_j}}(\Omega)$:

$$\Omega \subset \left(\Omega_1^{(l_j)} \cup \cdots \cup \Omega_{m_j}^{(l_j)}\right).$$

- For $l_j \neq 0$ define the finite element spaces $V_{h_{l_j}}^{(i)} = V_{h_{l_j}} \cap H_0^1(\Omega_i^{(l_j)})$ for $1 \leq i \leq m_j$. When $l_j = 0$, let $m_j = 1$ and define $V_{h_0}^{(1)} = V_{h_0}$.

- Let $R^T_{h_{l_j},i}$ denote the extension matrix which maps nodal values at nodes of $\mathcal{T}_{h_{l_j}}(\Omega)$ in $\Omega_i^{(l_j)}$ onto fine grid nodal values on $\mathcal{T}_h(\Omega)$. Its transpose $R_{h_{l_j},i}$ will denote an appropriate restriction matrix.
- Let $A^i_{h_{l_j}} = R^T_{h_{l_j},i} A_{h_J} R_{h_{l_j},i}$ denote submatrices of $A_{h_{l_j}} = R_{h_{l_j}} A_h R^T_{h_{l_j}}$, the stiffness matrix corresponding to $V_{h_{l_j}}$.

Multilevel Schwarz algorithms are matrix Schwarz algorithms based on the finite element spaces $V^i_{h_{l_j}}$ for $1 \le i \le m_j$ and $1 \le j \le k$. Thus, for instance the action of the inverse of a multilevel additive Schwarz preconditioner is:

$$M^{-1} = \sum_{j=1}^{k} \sum_{i=1}^{m_j} R^T_{h_{l_j},i} \left(A^i_{h_{l_j}} \right)^{-1} R_{h_{l_j},i}.$$

Similarly for multilevel multiplicative Schwarz algorithms [WA2].

Remark 7.7. A hybrid multilevel Schwarz preconditioner can be formulated using a symmetrized sequential Schwarz on each level, additive on different levels. The BPX and multilevel additive Schwarz preconditioners have some similarities. For instance, the BPX preconditioner can be obtained as a special case of the multilevel additive Schwarz preconditioner, provided $l_j = j$ for $0 \le j \le J$ and each subdomain $\Omega_i^{(l_j)}$ is defined to consist of elements of $\mathcal{T}_{h_{l_j}}(\Omega)$ adjacent to some interior node in $\mathcal{T}_{h_{l_j}}(\Omega)$, so that $A^i_{h_{l_j}}$ are scalar matrices with entries of magnitude $O(h_{l_j}^{-2})$. Conversely, if each scaling factor h_l^2 employed in (7.10) is formally replaced by the inverse of the submatrix of A_h corresponding to Range$(Q_{h_l} - Q_{h_{l-1}})$ in the BPX algorithm, see [XU], it will be a special case of the multilevel additive Schwarz preconditioner.

Multilevel Schwarz algorithms can have optimal convergence bounds.

Lemma 7.8. *Suppose that* $(h_{l_j}/h_{l_{j-1}}) \le r$ *for* $r < 1$ *and that the subdomain diameter satisfies* $\mathrm{diam}\big(\Omega_i^{(l_j)}\big) \approx h_{l_{j-1}}$. *Then* $\mathrm{cond}(M, A_{h_J}) \le C(r,a)$, *where* $C(r,a)$ *is independent of* h_i *and* J *but dependent on* r *and* $a(x)$.

Proof. See [ZH2]. □

7.2 Iterative Algorithms for Locally Refined Grids

In this section, we describe multilevel iterative methods for solving the linear system that arises from the discretization of an elliptic equation on a *locally refined composite grid*. Such grids are obtained by the repeated partial refinement of a conforming grid within specified nested subregions [MC4, BR7, HA9, DR12, MA21, MA22, BR6, EW6, EW7, CH23]. Such grids result in *non-conforming* triangulations, since the elements are refined

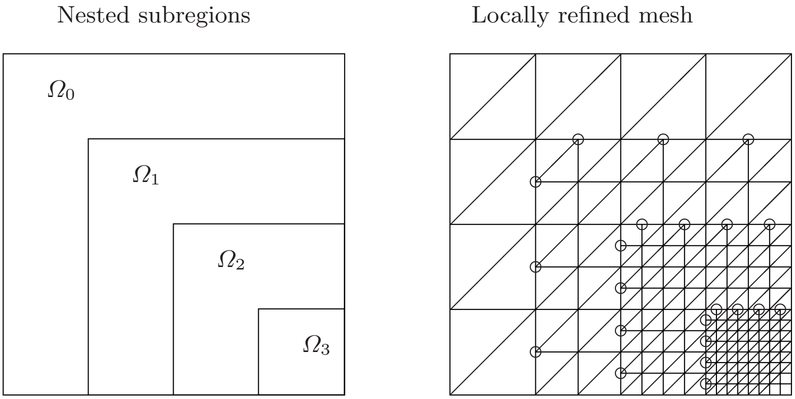


Fig. 7.2. Locally refined composite grid

only in specified subregions where the solution is irregular, leaving all other elements intact. The refined elements thus violate element adjacency requirements along the boundaries of refined regions, see Fig.7.2. Despite the non-conforming nature of locally refined composite grids, a *conforming* finite element discretization can be constructed on such grids, by introducing *slave variables*. In special cases, when the refinement regions are rectangular, the local grids may be chosen to be Cartesian, resulting in significant computational advantages. Related composite grids are described in [BA5, BA8, BE15, KE9, HE9].

Our discussion will be organized as follows. We first describe background on locally refined composite grids and the conforming discretization of an elliptic equation on such grids. We then describe the BEPS iterative algorithm for a two level composite grid. We conclude our discussion with a description of the FAC (Fast Adaptive Composite grid) and AFAC (Asynchronous Fast Adaptive Composite grid) iterative algorithms for multilevel composite grids.

7.2.1 Local Grid Refinement

We consider the following self adjoint and coercive elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{7.15}$$

where $c(x) \geq 0$. We assume that a family of nested subregions is specified:

$$\Omega \equiv \Omega_0 \supset \Omega_1 \supset \cdots \supset \Omega_p,$$

such that the solution $u(\cdot)$ of (7.15) is increasingly irregular (nonsmooth) within the subregions. If such subregions are not specified *a priori*, they may be *adaptively estimated*, as outlined later. Initially, a quasiuniform grid is constructed on Ω_0. At the l'th stage of refinement, elements within subregion Ω_l

are refined, leaving all other elements intact. This procedure is then repeatedly applied on each of the remaining nested subregions, yielding a *non-conforming* triangulation as in Fig. 7.2. We shall denote by:

$$\Phi_l \equiv \begin{cases} \Omega_l \backslash \overline{\Omega}_{l+1}, & \text{for } 0 \le l \le (p-1) \\ \Omega_p, & \text{for } l = p, \end{cases} \tag{7.16}$$

so that subregions $\Phi_0, \ldots, \Phi_p$ form a nonoverlapping decomposition of Ω, where the triangulation on each region Φ_l is locally quasiuniform. Once the refined grids have been constructed, a *conforming* finite element space can be defined on the global grid, as described in the next section. To ensure that the discretization is stable, however, it will be assumed that the area (volume) ratios $|\Omega_l| / |\Omega_{l+1}|$ are bounded uniformly.

We now elaborate on the procedure. Additional details may be found in [BA5, MC4, HA9, MC2]. Let $\epsilon > 0$ denote the user specified tolerance for the desired global discretization error of the finite element solution u_h on Ω:

$$|u - u_h|_{H^1(\Omega)} \le \epsilon. \tag{7.17}$$

To construct a locally refined composite grid and a discretization so that the finite element solution satisfies the above error bound, we shall start with a suitably chosen *quasiuniform* triangulation $\mathcal{T}_{h_0}(\Omega_0)$ of domain $\Omega_0 \equiv \Omega$ with grid size h_0. If the finite element solution u_{h_0} to the discretization of (7.15) on $\mathcal{T}_{h_0}(\Omega_0)$ satisfies $|u - u_{h_0}|_{H^1(\Omega)} > \epsilon$, then further grid refinement will be necessary. Since the true solution $u(\cdot)$ is not known, heuristic methods must be employed to estimate the discretization error $|u - u_{h_0}|_{H^1(\Omega)}$, see [BA8, BA5, JO2]. We outline one such finite element strategy below [JO2]:

$$\begin{cases} |u - u_{h_0}|_{H^1(\Omega)} \le C\, |u - \pi_{h_0} u|_{H^1(\Omega)} \\ \qquad\qquad\qquad\quad = C\Big(\sum_{\kappa \in \mathcal{T}_{h_0}} |u - \pi_{h_0} u|^2_{H^1(\kappa)} \Big)^{1/2} \\ \qquad\qquad\qquad\quad \le C\Big(\sum_{\kappa \in \mathcal{T}_{h_0}} h_\kappa^2 |u|^2_{H^2(\kappa)} \Big)^{1/2}, \end{cases} \tag{7.18}$$

where $\pi_{h_0} u$ denotes the standard nodal interpolation of the true solution $u(\cdot)$ onto the grid $\mathcal{T}_{h_0}(\Omega_0)$ and h_κ denotes the diameter of element $\kappa \in \mathcal{T}_{h_0}(\Omega)$.

Since the true solution $u(\cdot)$ is not known, $|u|_{H^2(\kappa)} \approx |u_{h_0}|_{H^2(\kappa)}$ may be employed heuristically as an *approximation*, using the solution u_{h_0}. If piecewise linear finite elements are used then $|u_{h_0}|_{H^2(\kappa)}$ will be zero within each element and such an estimate for $|u|_{H^2(\kappa)}$ will be inaccurate. In such a case, a local estimate for $|u|_{H^2(\kappa)}$ can be sought as a suitable difference quotient of ∇u_h evaluated at the centroids of adjacent elements, see [BA8, BA5, JO2], and Remark 7.10. Once an estimate for $|u|_{H^2(\kappa)}$ has been obtained, each element $\kappa \in \mathcal{T}_{h_0}(\Omega_0)$ for which the following condition holds, should be *flagged*:

$$h_\kappa^2\, |u_{h_0}|^2_{H^2(\kappa)} > \frac{\epsilon^2}{N_0\, C^2} \tag{7.19}$$

where N_0 denotes the total number of elements of $\mathcal{T}_{h_0}(\Omega_0)$. The union of all such flagged elements can be used to define the subregion Ω_1 requiring further refinement (or a union of rectangular regions enclosing such elements, provided Ω_1 aligns with the elements of $\mathcal{T}_{h_0}(\Omega_0)$). We denote by $\mathcal{T}_{h_0}(\Omega_1)$ the restriction of triangulation $\mathcal{T}_{h_0}(\Omega_0)$ to subregion Ω_1.

Once Ω_1 has been determined, all elements $\kappa \in \mathcal{T}_{h_0}(\Omega_1)$ should be refined uniformly, with local grid size $h_1 = (h_0/2)$, leaving all elements in Φ_0 intact. This procedure can then be applied recursively. Thus, in the lth stage of the local grid refinement procedure, for $1 \le l \le p$, let $u_{h_{l-1}}$ denote the finite element solution corresponding to a discretization of (7.15) on the current locally refined grid on Ω. Then, the discretization error $|u - u_{h_{l-1}}|_{H^1(\Omega_{l-1})}$ can be estimated on each element $\kappa \in \mathcal{T}_{h_{l-1}}(\Omega_{l-1})$ using difference quotients of $u_{h_{l-1}}$ as mentioned earlier based on the approximation $|u|_{H^2(\kappa)} \approx |u_{h_{l-1}}|_{H^2(\kappa)}$. All elements $\kappa \in \mathcal{T}_{h_{l-1}}(\Omega_{l-1})$ requiring refinement should be flagged if:

$$h_\kappa^2 \, |u_{h_{l-1}}|^2_{H^2(\kappa)} > \frac{\epsilon^2 \, |\Omega|}{N_0 \, |\Omega_{l-1}| \, 2^{d\,(l-1)} \, C^2} \tag{7.20}$$

for $h_0 = 2^l h_l$ and $\Omega_{l-1} \subset \mathbb{R}^d$. Subregion Ω_l can be defined as the union of all flagged elements (or a suitable enlargement of it, in which case it will be assumed that Ω_l aligns with the elements of $\mathcal{T}_{h_{l-1}}(\Omega_{l-1})$). Next, all elements κ of $\mathcal{T}_{h_{l-1}}(\Omega_{l-1})$ in Ω_l should be *uniformly refined* resulting in elements of size $h_l = (h_{l-1}/2)$. We denote the resulting quasiuniform triangulation of Ω_l by $\mathcal{T}_{h_l}(\Omega_l)$. This local refinement strategy can be terminated when there are no new flagged elements in the refined region.

Remark 7.9. Since only the elements within the nested subregions are refined, by construction the global "triangulation" will violate standard element adjacency requirements near nodes marked "$\circ$" on the subdomain boundaries $B^{(l)} \equiv \partial\Omega_l \backslash \partial\Omega$. Such nodes will be referred to as *slave nodes.* They will not represent true degrees of freedom, and care must be exercised to ensure that the finite element functions defined on either side of slave nodes, match, so that a conforming finite element space is obtained.

Remark 7.10. More accurate estimates of the finite element discretization error $|u - u_{h_l}|_{H^1(\Omega_l)}$ may be obtained [BA8, BA5]. For instance, all elements $\kappa \in \mathcal{T}_{h_l}(\Omega_l)$ of size h_l in Ω_l can be uniformly refined yielding elements of size $(h_l/2)$, and the discrete solution $u^*_{h_l}$ to (7.15) can be computed using continuous piecewise linear finite elements on the nonconforming grid on Ω with elements of size $(h_l/2)$ in Ω_l. On each element $\kappa \in \mathcal{T}_{h_l}(\Omega_l)$ let $I_Q u^*_{h_l}$ denote the quadratic interpolant of $u^*_{h_l}$ onto κ using the nodes of the refined elements. Then, the norm $|u|_{H^2(\kappa)}$ of the unknown solution u can be estimated as $|I_Q u^*_{h_l}|_{H^2(\kappa)}$. As before, all elements $\kappa \in \mathcal{T}_{h_l}(\Omega_l)$ for which $|I_Q u^*_{h_l}|_{H^2(\kappa)}$ is sufficiently large can be flagged for refinement:

$$h_\kappa^2 \, |I_Q u^*_{h_l}|^2_{H^2(\kappa)} > \frac{\epsilon^2 \, |\Omega|}{N_0 \, |\Omega_l| \, 2^{d\,l} \, C^2}. \tag{7.21}$$

Such an estimate for the discretization error will typically be more accurate than the estimate obtained using a difference quotient of ∇u_{h_l} on adjacent elements. However, this procedure will also be computationally more expensive as it requires determining $u^*_{h_l}$, see [MC2]

Notation. We shall employ the following notation for the nodes in the locally refined grid. For $0 \le l \le p$ let n_l denote the number of nodes of $\mathcal{T}_{h_l}(\Omega_l)$ in the *interior* of Ω_l, and q_l the number of nodes of $\mathcal{T}_{h_l}(\Omega_l)$ in Ω_{l+1}. For convenience, define $q_p = 0$. We enumerate the nodes of $\mathcal{T}_{h_l}(\Omega_l)$ in the interior of Ω_l as $\{x_i^{(l)}\}_{i=1}^{n_l}$ and assume they are ordered so that for $0 \le l \le p-1$:

$$x_i^{(l)} \in \begin{cases} \overline{\Phi}_l, & \text{for } 1 \le i \le n_l - q_l \\ \Omega_{l+1}, & \text{for } n_l - q_l + 1 \le i \le n_l. \end{cases}$$

For $1 \le l \le p$, we let m_l denote the number of nodes of $\mathcal{T}_{h_l}(\Omega_l)$ on the interface $B^{(l)} = \partial\Omega_l \backslash \partial\Omega$, and enumerate these nodes for $1 \le l \le p$ as:

$$x_i^{(l)} \in B^{(l)}, \quad \text{for } n_l + 1 \le i \le n_l + m_l.$$

For $1 \le l \le p$, we refer to the nodes $x_i^{(l)} \in \mathcal{T}_{h_l}(\Omega_l)$ which lie on $B^{(l)}$ but which do not belong to $\mathcal{T}_{h_{l-1}}(\Omega_l)$ as *slave* nodes. These are marked "∘" in Fig. 7.2. All remaining nodes will be referred to as *master* nodes.

Since locally refined grids violate element adjacency requirements along $B^{(l)}$, care must be exercised when constructing a $H^1(\Omega)$ conforming finite element space globally. More specifically, since the locally refined grid is nonconforming only at the slave nodes, it will be sufficient to require that the functions defined on elements adjacent to slave nodes, also match at the slave nodes. When piecewise linear finite elements are employed, this will be equivalent to requiring that the nodal value of a finite element function at a *slave node* be a linear combination of its nodal value on adjacent *master nodes*.

Thus, the slave nodes will not represent true degrees of freedom in a locally refined grid. Instead, only the master nodes will represent the true degrees of freedom. Using our chosen ordering of nodes, the total number n of master nodes is easily seen to satisfy:

$$n \equiv (n_0 - q_0) + (n_1 - q_1) + \cdots + (n_{p-1} - q_{p-1}) + n_p.$$

Henceforth, let $\{y_j\}_{j=0}^n$ denote a reordering of all the master nodes in Ω:

$$\{y_j\}_{j=0}^n = \left\{ \left\{ x_i^{(l)} \right\}_{i=1}^{n_l - q_l} \right\}_{l=0}^p,$$

based on the local orderings.

Remark 7.11. The software package DAGH (distributed adaptive grid hierarchy) implements adaptive grid refinement based on rectangular (Cartesian) subregions, see [MI2]. It is based on an adaptive mesh refinement algorithm [BE15] which was formulated originally for hyperbolic equations.

7.2.2 Discretization on Locally Refined Grids

On each subregion Ω_l for $1 \le l \le p$, let $V_{h_l}(\Omega_l) \subset H_0^1(\Omega_l)$ denote a *local* finite element space consisting of continuous, piecewise linear finite element functions on triangulation $\mathcal{T}_{h_l}(\Omega_l)$, vanishing outside Ω_l. We shall let $\{\phi_i^{(l)}(x)\}_{i=1}^{n_l}$ denote the standard continuous piecewise linear nodal basis function for $V_{h_l}(\Omega_l)$ satisfying:

$$\phi_i^{(l)}(x_j^{(l)}) = \delta_{ij},$$

where δ_{ij} denotes the Kronecker delta. The *global* composite finite element space $V_{h_0,\ldots,h_p}(\Omega)$ which will be employed to discretize elliptic equation (7.15) will be defined as the sum of the local finite element spaces:

$$V_{h_0,\ldots,h_p}(\Omega) = V_{h_0}(\Omega_0) + \cdots + V_{h_p}(\Omega_p). \tag{7.22}$$

By construction $V_{h_0,\ldots,h_p}(\Omega) \subset H_0^1(\Omega)$ so the space will be *conforming.* A nodal basis $\{\psi_i(x)\}_{i=1}^n$ of continuous piecewise linear functions can be constructed for $V_{h_0,\ldots,h_p}(\Omega)$ satisfying $\psi_i(y_j) = \delta_{ij}$, at all *master* nodes $\{y_j\}$. A global finite element discretization of (7.15) can be constructed on the composite grid [MC4, MC3, DR12, HA9], by seeking $u_h \in V_{h_0,\ldots,h_p}(\Omega)$ satisfying:

$$a(u_h, v_h) = F(v_h), \qquad \forall v_h \in V_{h_0,\ldots,h_p}(\Omega) \tag{7.23}$$

where the bilinear form $a(u,v)$ and functional $F(v)$ are defined as:

$$\begin{cases} a(u,v) \equiv \int_\Omega (a(x)\nabla u \cdot \nabla v + c(x)uv)\, dx \\ \;\; F(v) \equiv \int_\Omega f(x)v(x)dx. \end{cases} \tag{7.24}$$

Formally representing the finite element function $u_h(x) \in V_{h_0,\ldots,h_p}(\Omega)$ as:

$$\begin{cases} u_h(x) = \sum_{j=1}^n u_h(y_j)\, \psi_j(x) \\ \qquad\quad\; = \sum_{j=1}^n (\mathbf{u})_j \;\psi_j(x) \end{cases}$$

and choosing $v_h = \psi_j$ for $j = 1, \ldots, n$ yields the linear system:

$$A\mathbf{u} = \mathbf{f}, \tag{7.25}$$

where the stiffness matrix A, solution $\mathbf{u}$, and load vector $\mathbf{f}$ are defined by:

$$\begin{cases} (A)_{ij} \equiv a(\psi_i, \psi_j), & \forall i, j \\ \;(\mathbf{u})_i \equiv u_h(y_i), & \forall i \\ \;(\mathbf{f})_i \equiv F(\psi_i), & \forall i. \end{cases} \tag{7.26}$$

The stiffness matrix A and load vector $\mathbf{f}$ need not be assembled explicitly based on the basis $\{\psi_l(\cdot)\}$. Instead, they may be computed using a *subassembly* procedure involving the subdomain stiffness matrices, as described next.

Since each of the nonoverlapping subregions Φ_l are triangulated by the quasiuniform local grid $\mathcal{T}_{h_l}(\Phi_l)$ of grid size h_l for $0 \le l \le p$, it may be computationally advantageous to evaluate matrix-vector products $A\,\mathbf{v}$ based on the *subassembly identity* involving the nonoverlapping decomposition:

$$\overline{\Omega} = \overline{\Phi}_0 \cup \overline{\Phi}_1 \cup \cdots \overline{\Phi}_{p-1} \cup \overline{\Phi}_p.$$

Accordingly, we define local bilinear forms and functionals:

$$\begin{cases} a_{\Phi_l}(u,v) \equiv \int_{\Phi_l} \left(a(x)\nabla u(x) \cdot \nabla v(x) + c(x)u(x)v(x)\right) dx, & \text{for } 0 \le l \le p \\ F_{\Phi_l}(v) \equiv \int_{\Phi_l} f(x)v(x)dx, & \text{for } 0 \le l \le p, \end{cases}$$

and obtain the following decomposition:

$$\begin{cases} a\left(\sum_{i=1}^n \mathbf{u}_i\psi_i, \psi_j\right) = \sum_{l=0}^p a_{\Phi_l}\left(\sum_{i=1}^n \mathbf{u}_i\psi_i, \psi_j\right) \\ F(\psi_j) = \sum_{l=0}^p F_{\Phi_l}(\psi_j). \end{cases}$$

In order to express the preceding in matrix form, we shall introduce the following notation for the nodes within each of the subregions Φ_l. For each $0 \le l < p$ let r_l denote the number of nodes in $\overline{\Phi_l}$ from $\mathcal{T}_{h_l}(\Omega_l)$ with $r_p = n_p$. Enumerate and order these nodes locally as:

$$z_i^{(l)} \in \overline{\Phi}_l, \text{ for } 1 \le i \le r_l,$$

and denote the associated nodal basis functions on triangulation $\mathcal{T}_{h_l}(\Phi_l)$ as:

$$\chi_i^{(l)}(z_j^{(l)}) = \delta_{ij}, \text{ for } 1 \le i,j \le r_l.$$

Assemble the subdomain stiffness matrices $A^{(\Phi_l)}$ of size r_l on each of the subdomains Φ_l, including those nodes on its internal boundary segments $\partial\Phi_l$:

$$\left(A^{(\Phi_l)}\right)_{ij} \equiv a_{\Phi_l}\left(\chi_i^{(l)}, \chi_j^{(l)}\right), \quad \text{for } 1 \le i,j \le r_l.$$

Similarly, for $0 \le l \le p$ define subdomain load vectors $\mathbf{f}^{(\Phi_l)}$ by:

$$\left(\mathbf{f}^{(\Phi_l)}\right)_i \equiv F_{\Phi_l}\left(\chi_i^{(l)}\right), \quad \text{for } 1 \le i \le r_l.$$

Next, for each $0 \le l \le p$ define an extension matrix $R_{\Phi_l}^T$ between nodal vectors on $\overline{\Phi}_l$ and master nodal vectors as the following matrix of size $n \times r_l$:

$$R_{\Phi_l}^T \equiv \begin{bmatrix} \chi_1^{(l)}(y_1) & \cdots & \chi_{r_l}^{(l)}(y_1) \\ \vdots & & \vdots \\ \chi_1^{(l)}(y_n) & \cdots & \chi_{r_l}^{(l)}(y_n) \end{bmatrix}.$$

The sparsity of these extension matrices will depend on the number of master nodes within the support of the l'th level nodal basis functions. Associated restriction maps R_{Φ_l} are obtained by taking the transpose of the extension maps. The *subassembly identity* for the evaluation of matrix-vector products and computation of the load vector $\mathbf{f}$ can now be stated as:

$$\begin{cases} A\,\mathbf{v} = \sum_{l=0}^{p} R_{\Phi_l}^T A^{(\Phi_l)} R_{\Phi_l} \mathbf{v} \\ \quad \mathbf{f} = \sum_{l=0}^{p} R_{\Phi_l}^T \mathbf{f}^{(\Phi_l)}. \end{cases} \tag{7.27}$$

Before we describe iterative algorithms for solving the linear system (7.25) resulting from the global discretization, we introduce additional notation.

Definition 7.12. *For $0 \le l \le p$, we define the $n \times n_l$ extension matrix R_{Ω_l,h_l}^T:*

$$R_{\Omega_l,h_l}^T = \begin{bmatrix} \phi_1^{(l)}(y_1) & \cdots & \phi_{n_l}^{(l)}(y_1) \\ \vdots & & \vdots \\ \phi_1^{(l)}(y_n) & \cdots & \phi_{n_l}^{(l)}(y_n) \end{bmatrix}.$$

The sparsity of these matrices will depend on the number of master nodes within the support of each l'th level nodal basis function. Restriction matrices R_{Ω_l,h_l} will be transposes of the above matrices. We define a local stiffness matrix of size n_l corresponding to the Dirichlet problem on Ω_l discretized using $V_{h_l}(\Omega_l) \cap H_0^1(\Omega_l)$ as:

$$\left(A_{II}^{(\Omega_l,h_l)}\right)_{ij} \equiv a_{\Omega_l}\left(\phi_i^{(l)}, \phi_j^{(l)}\right), \qquad \text{for } 1 \le i, j \le n_l.$$

Similarly, we define additional extension and local stiffness matrices.

Definition 7.13. *For $1 \le l \le p$ define the $n \times q_{l-1}$ extension matrix $R_{\Omega_l,h_{l-1}}^T$:*

$$R_{\Omega_l,h_{l-1}}^T = \begin{bmatrix} \phi_{n_{l-1}-q_{l-1}+1}^{(l-1)}(y_1) & \cdots & \phi_{n_{l-1}}^{(l-1)}(y_1) \\ \vdots & & \vdots \\ \phi_{n_{l-1}-q_{l-1}+1}^{(l-1)}(y_n) & \cdots & \phi_{n_{l-1}}^{(l-1)}(y_n) \end{bmatrix}.$$

Again, the sparsity of these matrices will depend on the number of master nodes within the support of the $(l-1)$'th level nodal basis functions. Restriction matrices $R_{\Omega_l,h_{l-1}}$ will be transposes of the extension matrices. For $1 \le l \le p$, we denote the stiffness matrix of size q_l associated with the Dirichlet problem on Ω_l discretized on $V_{h_{l-1}}(\Omega_l) \cap H_0^1(\Omega_l)$ as:

$$\left(A_{II}^{(\Omega_l,h_{l-1})}\right)_{ij} \equiv a_{\Omega_l}\left(\phi_{n_l-q_l+i}^{(l-1)},\ \phi_{n_l-q_l+j}^{(l-1)}\right), \qquad \text{for } 1 \le i, j \le q_l.$$

We shall now describe iterative algorithms for (7.25).

7.2.3 BEPS Algorithm for Two Level Composite Grids

We shall first consider a preconditioner of [BR7] for the stiffness matrix A in (7.25) when $p = 2$. In this case, there will be only one level of grid refinement with $\Omega_1 \subset \Omega_0$. The refined grid $\mathcal{T}_{h_1}(\Omega_1)$ will be obtained by refinement of elements of $\mathcal{T}_{h_0}(\Omega_0)$ in Ω_1. The BEPS preconditioner will correspond to a *symmetrized multiplicative Schwarz preconditioner* based on the finite element subspaces $V_{h_1}(\Omega_1)$ and $V_{h_0}(\Omega_0)$. More specifically, the action of the inverse M^{-1} of the BEPS preconditioner M will be the output obtained after one symmetrized multiplicative Schwarz iteration with zero initial iterate, based on the column spaces $R^T_{\Omega_1,h_1}$ and $R^T_{\Omega_0,h_0}$. We summarize the preconditioner.

Algorithm 7.2.1 *(BEPS Preconditioner)*
The action $\mathbf{z} = M^{-1}\mathbf{r}$ *is given below.*

1. *Solve:*
$$A_{II}^{(\Omega_1,h_1)}\mathbf{v}_1 = R_{\Omega_1,h_1}\mathbf{r}.$$
2. *Solve:*
$$A_{II}^{(\Omega_0,h_0)}\mathbf{v}_2 = R_{\Omega_0,h_0}\left(\mathbf{r} - A\,R^T_{\Omega_1,h_1}\mathbf{v}_1\right).$$
3. *Solve:*
$$A_{II}^{(\Omega_1,h_1)}\mathbf{v}_3 = R_{\Omega_1,h_1}\left(\mathbf{r} - A\,R^T_{\Omega_1,h_1}\mathbf{v}_1 - A\,R^T_{\Omega_0,h_0}\mathbf{v}_2\right).$$

Output: $\mathbf{z} \equiv R^T_{\Omega_1,h_1}\mathbf{v}_3 + R^T_{\Omega_0,h_0}\mathbf{v}_2 + R^T_{\Omega_1,h_1}\mathbf{v}_1$.

Remark 7.14. Step 1 in this preconditioner may be omitted if the residual vector $\mathbf{r}$ satisfies $R_{\Omega_1,h_1}\mathbf{r} = \mathbf{0}$. This can be ensured if the initial iterate $\mathbf{v}^{(0)}$ in the conjugate gradient algorithm is chosen so that $\mathbf{r} = \mathbf{f} - A\,\mathbf{v}^{(0)}$ satisfies the constraint $R_{\Omega_1,h_1}\mathbf{r} = \mathbf{0}$. Then, provided the preconditioning involves steps 2 through 3, all subsequent residuals will satisfy this constraint. The initial iterate $\mathbf{v}^{(0)}$ may be chosen in the form $R^T_{\Omega_1,h_1}\boldsymbol{\alpha}$ for $\boldsymbol{\alpha} \in \mathbb{R}^{n_1}$ chosen so that $\left(R_{\Omega_1,h_1}AR^T_{\Omega_1,h_1}\right)\boldsymbol{\alpha} = R_{\Omega_1,h_1}\mathbf{f}$.

The BEPS preconditioner requires the solution of a linear system with coefficient matrix $A_{II}^{(\Omega_0,h_0)}$ once and coefficient matrix $A_{II}^{(\Omega_1,h_1)}$ twice (once if $R_{\Omega_1,h_1}\mathbf{r} = \mathbf{0}$). The following convergence bound will hold.

Theorem 7.15. *There exists* $C > 0$ *independent of* h_0 *and* h_1*, such that* *cond*$(M, A) \le C$.

Proof. See [BR7]. □

The BEPS preconditioner is sequential in nature. A more parallelizable variant of this preconditioner is described in [BR6].

7.2.4 FAC and AFAC Algorithms for Multilevel Composite Grids

We next consider multilevel locally refined grids. We shall describe the *Fast Adaptive Composite* (FAC) grid and *Asynchronous Fast Adaptive Composite* (AFAC) grid algorithms, see [MC4, MA21, MA22, DR12, MC3], for solving system (7.25). These algorithms are multilevel generalizations of the two level BEPS algorithm and formally correspond to matrix Schwarz algorithms. We first describe the FAC algorithm [HA9, MA21, DR12, MA22, MC3], which corresponds to an *unaccelerated* multiplicative Schwarz algorithm based on the subspaces $V_{h_l}(\Omega_l) \cap H_0^1(\Omega_l)$ for $0 \le l \le p$. In matrix form, it corresponds to a matrix Schwarz algorithm based on the column spaces $R^T_{\Omega_l,h_l}$ for $0 \le l \le p$.

Algorithm 7.2.2 *(Sequential FAC Algorithm)*
Let $\mathbf{v}^{(0)}$ denote a starting iterate

```
1. Define: $\mathbf{v}^* \leftarrow \mathbf{v}^{(0)}$.
2. For $k = 1, 2, \cdots$ until convergence do:
3.     For $l = p, p-1, \cdots, 0$ do:
           $$\mathbf{v}^* \leftarrow \mathbf{v}^* + R^T_{\Omega_l,h_l} \left( A_{II}^{(\Omega_l,h_l)} \right)^{-1} R_{\Omega_l,h_l} \left( \mathbf{f} - A\,\mathbf{v}* \right).$$
4.     Endfor
5. Endfor
```

Output: $\mathbf{v}^*$.

The FAC algorithm is *sequential.* The following convergence bound holds.

Theorem 7.16. *The convergence factor ρ of the sequential FAC iteration is independent of the mesh sizes h_l and the number p of levels. It depends on the ratio* $\max\{(h_l/h_{l-1})\}$ *and the ratio of areas (or volumes)* $\max\{(|\Omega_{l-1}|/|\Omega_l|)\}$.

Proof. See [HA9, MA21, DR12, MA22, MC3]. □

We next describe a *parallel* version of the FAC algorithm, which we express as an additive Schwarz preconditioner.

Algorithm 7.2.3 *(Parallel FAC Preconditioner)*
The action M^{-1} of the inverse of the parallel FAC preconditioner M is:

$$M^{-1}\mathbf{r} \equiv \sum_{l=0}^{p} R^T_{\Omega_l,h_l} \left(A_{II}^{(\Omega_l,h_l)} \right)^{-1} R_{\Omega_l,h_l}\mathbf{r}.$$

Unfortunately, the convergence rate for the above preconditioner deteriorates as the number p of grid levels increases, as the following result indicates.

Theorem 7.17. *There exists $C > 0$ independent of the mesh sizes h_l and the number p of levels, such that:*

$$\mathrm{cond}(M, A) \le C\,p,$$

Proof. See [DR12, MA21, MA22, MC3]. □

Theoretical analysis indicates that this deterioration in the condition number is due to the *redundant* projections of the solution onto the subspaces $V_{h_l}(\Omega_l)$ and $V_{h_{-1}}(\Omega_l)$ for $1 \leq l \leq p$. The following preconditioner removes such redundancy, and restores optimal order convergence, at the cost of additional computations. The resulting preconditioner is referred to as the AFAC preconditioner, and corresponds formally to an additive Schwarz preconditioner based on the subspaces $V_{h_0}(\Omega_0)$ and $V_{h_l}(\Omega_l) \cap V_{h_{l-1}}(\Omega_l)^{\perp}$ for $1 \leq l \leq\ p$. Here $V_{h_{l-1}}(\Omega_l)^{\perp}$ denotes the orthogonal complement in the $a(.,.)$ inner product.

Algorithm 7.2.4 *(AFAC preconditioner))*
The action M_{AFAC}^{-1} of the inverse of the AFAC preconditioner M_{AFAC} is:

$$
\begin{aligned}
M_{\mathrm{AFAC}}^{-1}\mathbf{r} \equiv\ & R_{\Omega_0,h_0}^T \left(A_{II}^{(\Omega_0,h_0)}\right)^{-1} R_{\Omega_0,h_0}\mathbf{r} \\
+ \textstyle\sum_{l=1}^{p} & \left(R_{\Omega_l,h_l}^T \left(A_{II}^{(\Omega_l,h_l)}\right)^{-1} R_{\Omega_l,h_l} - R_{\Omega_l,h_{l-1}}^T \left(A_{II}^{(\Omega_l,h_{l-1})}\right)^{-1} R_{\Omega_l,h_{l-1}} \right)\mathbf{r}.
\end{aligned}
$$

On each subdomain Ω_l for $1 \leq l \leq p$, the AFAC preconditioner requires the solution of two subproblems, involving matrices $A_{II}^{(\Omega_l,h_l)}$ and $A_{II}^{(\Omega_l,h_{l-1})}$.

We have the following convergence bound.

Theorem 7.18. *There exists $C > 0$ independent of the mesh sizes h_l and the number p of levels, but dependent on the ratios of the mesh sizes (h_{l-1}/h_l) and the ratios of the areas (or volumes) of the refined regions, such that:*

$$
\mathrm{cond}(M_{\mathrm{AFAC}}, A) \leq C.
$$

Proof. See [DR12, MA21, MA22, MC3]. □

Remark 7.19. We have tacitly assumed that each grid $\mathcal{T}_{h_{l-1}}(\Omega_l)$ was refined so that $h_l = (h_{l-1}/2)$. In practice, elements of size h_{l-1} in Ω_l can be refined to yield $h_l = (h_{l-1}/2^{\alpha})$ for $\alpha \geq 1$. In this case, there will be additional slave nodes. Local grid refinement methodology can also be applied to time varying (hyperbolic or parabolic type) problems, see [BE15, EW5].

8

Non-Self Adjoint Elliptic Equations: Iterative Methods

In this chapter, we describe domain decomposition methods for preconditioning the *nonsymmetric* linear systems arising from the discretization of non-self adjoint advection-diffusion elliptic equations. Under appropriate assumptions, such discretizations have the following form:

$$A\,\mathbf{u} = \mathbf{f}, \quad \text{with} \quad A = H + N, \;\; H^T = H \geq 0 \quad \text{and} \quad N^T = -N,$$

where H is Hermitian and *positive semi-definite* and N is skew-Hermitian. The eigenvalues of matrix A will typically be *complex* and occur in conjugate pairs, since A is real. They will have non-negative real parts when $H \geq 0$. Since A is non-symmetric, the CG algorithm cannot be employed to solve $A\,\mathbf{u} = \mathbf{f}$. Instead, the GMRES (or QMR, CGNR or CGNE) method can be employed with preconditioner M, see [AS3, FR5, SA2, AX]. These Krylov space algorithms typically require more storage than the CG method.

Domain decomposition methodology is less developed for non-self adjoint elliptic equations. Typically, the effectiveness of a preconditioner M for the non-symmetric matrix $A = H + N$ depends on the relative magnitude of the diffusion and advection terms, i.e., H and N respectively. A bound for its rate of convergence typically depends on the minimal eigenvalue of the Hermitian part of $M^{-1}H$ and the maximal singular value of $M^{-1}A$, see [SA2].

Our discussion is organized as follows. Section 8.1 presents background on non-self adjoint elliptic equations and their discretizations. Section 8.2 describes Schwarz and Schur complement preconditioners for diffusion dominated non-self adjoint elliptic equations. Section 8.3 considers the advection dominated case, in which the underlying elliptic equation is of *singular perturbation* type (exhibiting boundary layers). Two subdomain preconditioners are described motivated by heterogenous domain decomposition methods. Section 8.4 considers the implicit time stepping of non-self adjoint parabolic equations. Section 8.5 presents a selection of energy norm based theoretical results for non-self adjoint elliptic equations (see Chap. 15 for theoretical results in the maximum norm).

8.1 Background

We shall consider the following advection-diffusion elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f(x), & \text{in} \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad u = 0, & \text{on} \partial\Omega, \end{cases} \tag{8.1}$$

where $a(x) \geq a_0 > 0$ is a scalar diffusion (viscosity) coefficient, $\mathbf{b}(x)$ is a vector (advection or convection) field, and $c(x)$ is a reaction coefficient (we shall assume that $c(x) \geq 0$ except when specified otherwise). Equation (8.1) is typically referred to as being *diffusion dominated*, when $\|\mathbf{b}\|_\infty = O(a_0)$, and as being *advection dominated* when $\|a\|_\infty \ll \|\mathbf{b}\|_\infty$.

In the diffusion dominated case, equation (8.1) will have elliptic character throughout the domain, and its solution will be smooth when the forcing, boundary data, and geometry are sufficiently smooth. In the advection dominated case, however, equation (8.1) may be of singular perturbation type [KE5, LA5], exhibiting a *layer* region $\Omega_2 \subset \Omega$ in which the solution has “steep gradients” (derivatives of large magnitude), where the equation is of *elliptic* character, and a complementary region (“nonlayer”) $\Omega_1 \subset \Omega$, in which the diffusion term may be neglected (to an approximation), yielding *hyperbolic* character locally. In the advection dominated case, care must be exercised in discretizing (8.1) and in formulating iterative algorithms.

Weak Formulation. The standard weak formulation of (8.1) is obtained by multiplying the equation by a test function $v \in H_0^1(\Omega)$, and integrating the diffusion term by parts on Ω. It seeks $u \in H_0^1(\Omega)$ satisfying:

$$\begin{cases} \mathcal{A}(u, v) = F(v), \qquad \forall v \in H_0^1(\Omega), \quad \text{where} \\ \mathcal{A}(u, v) \equiv \int_\Omega (a(x)\, \nabla u \cdot \nabla v + (\mathbf{b}(x) \cdot \nabla u)\, v + c(x)\, u\, v)\, dx \\ \quad F(v) \equiv \int_\Omega f\, v\, dx. \end{cases} \tag{8.2}$$

Using integration by parts, $\int_\Omega (\mathbf{b}(x) \cdot \nabla u)\; v\, dx$ can be expressed as:

$$\begin{aligned} \int_\Omega (\mathbf{b}(x) \cdot \nabla u)\; v\, dx &= -\int_\Omega u\, \nabla \cdot (\mathbf{b}(x)\, v)\; dx \\ &= -\int_\Omega u\; (\mathbf{b}(x) \cdot \nabla v)\; dx - \int_\Omega (\nabla \cdot \mathbf{b}(x))\, u\, v\, dx, \end{aligned}$$

for $u,\, v \in H_0^1(\Omega)$. Taking a weighted arithmetic average of both expressions, with weight parameter $0 \leq \theta \leq 1$, yields the following expression:

$$\begin{aligned} \int_\Omega (\mathbf{b}(x) \cdot \nabla u)\; v\, dx = (1-\theta) \int_\Omega (\mathbf{b}(x) \cdot \nabla u)\; v\, dx - \theta \int_\Omega u\; (\mathbf{b}(x) \cdot \nabla v)\; dx \\ - \theta \int_\Omega (\nabla \cdot \mathbf{b}(x))\, u\, v\, dx. \end{aligned}$$

Substituting this into (8.2) yields several equivalent expressions for $\mathcal{A}(.,.)$:

$$\begin{aligned} \mathcal{A}^\theta(u, v) \equiv\; & \int_\Omega (a(x)\, \nabla u \cdot \nabla v + (c(x) - \theta\, \nabla \cdot \mathbf{b}(x))\, u\, v)\, dx \\ & + (1-\theta) \int_\Omega (\mathbf{b}(x) \cdot \nabla u)\; v\, dx - \theta \int_\Omega u\; (\mathbf{b}(x) \cdot \nabla v)\; dx, \end{aligned} \tag{8.3}$$

where $\mathcal{A}^{\theta}(u,v) = \mathcal{A}(u,v)$ for $u,\, v \in H_0^1(\Omega)$ and $0 \le \theta \le 1$. Commonly $\theta = 0$ is used for Galerkin approximation of (8.1). Other choices include $\theta = \frac{1}{2}$ or 1.

Choosing $\theta = \frac{1}{2}$ yields the following expression for $\mathcal{A}(.,.) = \mathcal{A}^{\frac{1}{2}}(.,.)$:

$$\begin{aligned}\mathcal{A}^{\frac{1}{2}}(u,v) \equiv\; &\textstyle\int_{\Omega} \big(a(x)\,\nabla u \cdot \nabla v + (c(x) - \frac{1}{2}\,\nabla \cdot \mathbf{b}(x))\,u\,v\big)\,dx \\ &+ \textstyle\frac{1}{2}\int_{\Omega} \left((\mathbf{b}(x)\cdot\nabla u)\,v - u(\mathbf{b}(x)\cdot\nabla v)\right)\,dx.\end{aligned} \tag{8.4}$$

From this, a *splitting* of the bilinear form $\mathcal{A}(.,.) = \mathcal{A}^{\frac{1}{2}}(.,.)$ into its *Hermitian* and *skew-Hermitian* parts can be obtained as follows:

$$\begin{cases} \mathcal{A}(u,v) = \mathcal{H}(u,v) + \mathcal{N}(u,v), \quad \text{where} \\ \mathcal{H}(u,v) = \int_{\Omega} \big(a(x)\,\nabla u \cdot \nabla v + (c(x) - \frac{1}{2}\nabla \cdot \mathbf{b})\,u\,v\big)\,dx \\ \mathcal{N}(u,v) = \frac{1}{2}\int_{\Omega} \left((\mathbf{b}(x)\cdot\nabla u)\,v - u\,(\mathbf{b}(x)\cdot\nabla v)\right)\,dx, \end{cases} \tag{8.5}$$

for $u,\, v \in H_0^1(\Omega)$. By construction, $\mathcal{H}(.,.)$ is *symmetric*. It will also satisfy:

$$\mathcal{H}(u,u) = \int_{\Omega} a(x)\,|\nabla u|^2\,dx + \int_{\Omega} \left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) u^2 dx \ge 0, \tag{8.6}$$

when $\big(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\big) \ge 0$. The form $\mathcal{N}(u,v)$ will be *skew-symmetric*, i.e.:

$$\mathcal{N}(u,v) = -\mathcal{N}(v,u), \quad \text{for} \quad u,\, v \in H_0^1(\Omega),$$

yielding that $\mathcal{N}(u,u) = 0$ for $u \in H_0^1(\Omega)$ and that $\mathcal{A}(u,u) = \mathcal{H}(u,u) \ge 0$. When $a_0 > 0$ and $\big(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\big) \ge 0$, the existence of solutions to (8.2) is guaranteed by the nonsymmetric Lax-Milgram lemma [CI2].

In domain decomposition applications, it is sometimes necessary to pose (8.1) on a subdomain $\Omega_i \subset \Omega$ with *Robin boundary conditions* on its interior boundary segment $\partial\Omega_i \cap \Omega$. In such applications, let $\mathcal{A}^{\theta}_{\Omega_i}(.,.)$ denote:

$$\begin{aligned}\mathcal{A}^{\theta}_{\Omega_i}(u,v) \equiv\; &\textstyle\int_{\Omega_i} \left(a(x)\,\nabla u \cdot \nabla v + (c(x) - \theta\,\nabla\cdot\mathbf{b}(x))\,u\,v\right)dx \\ &+ (1-\theta)\textstyle\int_{\Omega_i} (\mathbf{b}(x)\cdot\nabla u)\;v\,dx - \theta\,\textstyle\int_{\Omega_i} u\;(\mathbf{b}(x)\cdot\nabla v)\;dx,\end{aligned} \tag{8.7}$$

the local contribution to $\mathcal{A}^{\theta}(.,.)$. Formally integrating $\mathcal{A}^{\theta}_{\Omega_i}(.,.)$ by parts yields:

$$\begin{aligned}\mathcal{A}^{\theta}_{\Omega_i}(u,v) = &\textstyle\int_{\Omega_i} \left(-\nabla\cdot(a\,\nabla u) + \mathbf{b}(x)\cdot\nabla u + c(x)\,u\right)\;v\,dx \\ &+ \textstyle\int_{\partial\Omega_i} \mathbf{n}_i(x)\cdot\left(a(x)\,\nabla u - \theta\,\mathbf{b}(x)\,u\right)\;v\,ds_x,\end{aligned} \tag{8.8}$$

where $\mathbf{n}_i(x)$ denotes the unit exterior normal to Ω_i at $x \in \partial\Omega_i$. Thus, it will hold that $\mathcal{A}^{\theta}_{\Omega_i}(u,v) = \mathcal{A}_{\Omega_i}(u,v) - \theta\,\int_{\partial\Omega_i} \mathbf{n}_i(x)\cdot\mathbf{b}(x)\,u\,v\,ds_x$. Consider the following local functional, given $f(\cdot) \in L^2(\Omega)$ and $g(\cdot) \in H^{-1/2}(\partial\Omega_i)$:

$$F_{\Omega_i}(v) = \int_{\Omega_i} f\,v\,dx + \int_{\partial\Omega_i} g\,v\,ds_x, \quad \forall v \in V_i,$$

where $V_i = \left\{v \in H^1(\Omega_i) : v = 0 \text{ on } \partial\Omega_i \cap \partial\Omega\right\}$, and seek $w \in V_i$ such that:

$$\mathcal{A}^{\theta}_{\Omega_i}(w, v) = F_{\Omega_i}(v), \quad \forall v \in V_i. \tag{8.9}$$

This will correspond to a weak formulation of the boundary value problem:

$$\begin{cases} -\nabla \cdot (a(x)\nabla w) + \mathbf{b}(x) \cdot \nabla w + c(x)w = f(x), & \text{in } \Omega_i \\ \mathbf{n}_i(x) \cdot (a(x)\, \nabla w - \theta\, \mathbf{b}(x)\, w) = g(x), & \text{on } \partial\Omega_i \cap \Omega \\ w = 0, & \text{on } \partial\Omega_i. \cap \partial\Omega \end{cases} \tag{8.10}$$

This problem enforces Dirichlet boundary conditions on $\partial\Omega_i \cap \partial\Omega$, while on $\partial\Omega_i \cap \Omega$ it enforces a Neumann condition for $\theta = 0$ and a Robin condition for $0 < \theta \leq 1$. To replace the Robin condition in (8.10) by a β-Robin condition:

$$\mathbf{n}_i(x) \cdot (a(x)\, \nabla w) + \beta\, w = g, \quad \text{on} \quad \partial\Omega_i \cap \Omega,$$

replace $\mathcal{A}^{\theta}_{\Omega_i}(w, v)$ by $\mathcal{A}^{\theta}_{\Omega_i}(w, v) + \int_{\partial\Omega_i \cap \Omega} (\beta(x) + \theta\, \mathbf{n}_i(x) \cdot \mathbf{b}(x))\, w\, v\, ds_x$ in (8.9). When $\theta = \frac{1}{2}$, the resulting bilinear form is easily seen to be *coercive*, i.e.,

$$\mathcal{A}^{\frac{1}{2}}_{\Omega_i}(w, w) + \int_{\partial\Omega_i \cap \Omega} (\beta(x) + \frac{1}{2}\, \mathbf{n}_i(x) \cdot \mathbf{b}(x))\, w\, w\, ds_x \geq 0,$$

provided $\left(\beta(x) + \frac{1}{2}\, \mathbf{n}_i(x) \cdot \mathbf{b}(x)\right) \geq 0$.

Stable Discretizations. Traditional Galerkin finite element and centered finite difference discretizations of (8.1) are *unstable* for the *hyperbolic* equation obtained when $a(x) = 0$, see [JO2]. This instability also manifests itself as $\|a\|_\infty \to 0^+$, more specifically, when $\|a\|_\infty \ll h\, \|\mathbf{b}\|_\infty$, where h denotes the grid size. However, if h satisfies a cell *Peclet* restriction:

$$h\, \|\mathbf{b}\|_\infty \leq C\, \|a\|_\infty, \tag{8.11}$$

for some $C > 0$, independent of h, then the traditional Galerkin and centered finite difference discretizations of (8.1) will be *stable*. If $\|a\|_\infty \ll \|\mathbf{b}\|_\infty$, this may require an extremely small mesh size h, making the Galerkin and centered schemes computationally expensive in the advection dominated case. Instead, for such problems a *streamline-diffusion*, *stabilized* finite element or *upwind* finite difference discretization will yield a stable discretization without constraints on the mesh size [JO2, FR3, FR2].

Stable discretizations are typically constructed by adding a small diffusion term, depending on a parameter $0 < \delta \ll 1$, along the streamline direction, so that the discretization is stable even when $a(x) = 0$. For instance, the streamline-diffusion discretization [JO2] employs test functions of the form $v + \delta\, \mathbf{b}(x) \cdot \nabla v$ (for a small $\delta > 0$), and replaces the forms $\mathcal{A}(u, v)$ and $F_(v)$ by the modified forms $\tilde{\mathcal{A}}_\delta(u, v)$ and $\tilde{F}_\delta(v)$:

$$
\begin{cases}
\tilde{\mathcal{A}}_\delta(u,v) = & \sum_{\kappa\in\Omega_h}\int_\kappa \left(a(x)\,\nabla u\cdot\nabla v + (\mathbf{b}\cdot\nabla u)\,v + c(x)\,u\,v\right)\,dx \\
& -\delta\, \sum_{\kappa\in\Omega_h}\int_\kappa \nabla\cdot(a(x)\nabla u)\ (\mathbf{b}\cdot\nabla v)\,dx \\
& +\delta \sum_{\kappa\in\Omega_h}\int_\kappa (\mathbf{b}\cdot\nabla u)\,(\mathbf{b}\cdot\nabla v)\,dx \\
& +\delta \sum_{\kappa\in\Omega_h}\int_\kappa c(x)\,u\,(\mathbf{b}\cdot\nabla v)\,dx \\
\tilde{F}_\delta(v) \equiv & \sum_{\kappa\in\Omega_h}\int_\kappa f(x)\,(v+\delta\,\mathbf{b}\cdot\nabla v)\,dx.
\end{cases}
\tag{8.12}
$$

For alternative stabilizations, see [FR3, FR2].

We shall assume that a stable finite element discretization of (8.1) is employed. The resulting system will be denoted:

$$A\,\mathbf{u} = \mathbf{f}. \tag{8.13}$$

For instance, if a Galerkin discretization of (8.5) is employed satisfying a cell Peclet restriction, and $\{\phi_1,\ldots,\phi_n\}$ denotes a nodal basis for the finite element space $V_h \subset H_0^1(\Omega)$, then we can express $A = H + N$ where:

$$(A)_{ij} = \mathcal{A}(\phi_i,\phi_j),\ \ (H)_{ij} = \mathcal{H}(\phi_i,\phi_j),\ \ (N)_{ij} = \mathcal{N}(\phi_i,\phi_j),\ \ (\mathbf{f})_i = F(\phi_i),$$

with discrete solution $u_h(x) = \sum_i (\mathbf{u})_i\,\phi_i(x)$. Matrix H will be symmetric due to the symmetry of $\mathcal{H}(.,.)$. Furthermore, if $a_0 > 0$ and $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) \geq 0$, or if $a(x) = 0$ and $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) \geq \beta > 0$, then matrix H will also be *positive definite*. Similarly, matrix N will be real and skew-symmetric with $N^T = -N$ (due to skew-symmetry of $\mathcal{N}(.,.)$).

Remark 8.1. We may also decompose $\mathcal{A}(u,v) = \tilde{\mathcal{H}}(u,v) + \tilde{\mathcal{N}}(u,v)$, where $\tilde{\mathcal{H}}(u,v) \equiv \int_\Omega (a(x)\,\nabla u\cdot\nabla v + c(x)\,u\,v)\,dx$ and $\tilde{\mathcal{N}}(u,v) \equiv \int_\Omega (\mathbf{b}(x)\cdot\nabla u)\ v\,dx$. If $\nabla\cdot\mathbf{b} = 0$, then $\tilde{\mathcal{H}}(.,.) = \mathcal{H}(.,.)$ and $\tilde{\mathcal{N}}(.,.) = \mathcal{N}(.,.)$. Furthermore, if $c(x) \geq 0$, with $\left(\tilde{H}\right)_{ij} = \tilde{H}(\phi_i,\phi_j)$ and $\left(\tilde{N}\right)_{ij} = \tilde{N}(\phi_i,\phi_j)$, it can easily be verified that $\tilde{H}^T = \tilde{H} \geq 0$ and $\tilde{N}^T = -\tilde{N}$.

Remark 8.2. If a streamline-diffusion discretization is employed [JO2] based on the forms $\tilde{\mathcal{A}}_\delta(u,v)$ and $\tilde{F}_\delta(v)$ defined in (8.12), and piecewise linear finite elements are employed, with $a(x) \equiv a$, then the term $\int_\kappa \nabla\cdot(a(x)\nabla u)\ (\mathbf{b}\cdot\nabla v)\,dx$ will be zero. Additionally, if $c(x) = 0$ and $\nabla\cdot\mathbf{b}(x) = 0$, then we may decompose $\tilde{\mathcal{A}}_\delta(u,v) = \tilde{\mathcal{H}}_\delta(u,v) + \tilde{\mathcal{N}}_\delta(u,v)$, where:

$$
\begin{cases}
\tilde{\mathcal{H}}_\delta(u,v) = & \sum_{\kappa\in\Omega_h}\int_\kappa \left(a(x)\,\nabla u\cdot\nabla v\right)\,dx \\
& +\delta \sum_{\kappa\in\Omega_h}\int_\kappa (\mathbf{b}\cdot\nabla u)\,(\mathbf{b}\cdot\nabla v)\,dx \\
\tilde{\mathcal{N}}_\delta(u,v) = & \sum_{\kappa\in\Omega_h}\int_\kappa (\mathbf{b}\cdot\nabla u)\,v\,dx,
\end{cases}
$$

where $\tilde{\mathcal{H}}_\delta(\cdot,\cdot)$ is self-adjoint and coercive, and $\tilde{\mathcal{N}}_\delta(\cdot,\cdot)$ is skew-symmetric. For stabilized finite element discretizations of (8.1), see [FR3, FR2].

Spectrum of A. Eigenvalues of the nonsymmetric matrix A in (8.13) will generally be complex. The next result, referred to as Bendixson's lemma, describes a rectangular region in the complex plane $\mathbb{C}$ which encloses the eigenvalues of a matrix $C^{-1}A$, when $A = H + N$ is any complex matrix of size m and C is a Hermitian positive definite matrix of size m. Such methods may be employed to estimate the rate of convergence of the GMRES algorithm to solve $C^{-1}A\mathbf{u} = C^{-1}\mathbf{f}$, see [SA3, GO4, SA2, AX], where the commonly used bound (8.16) depends on the smallest eigenvalue of $(C^{-1}A + A^*C^{-*})/2$ and the largest singular value of $C^{-1}A$. We shall let $i = \sqrt{-1}$ denote the imaginary unit and $\mathbb{C}$ the complex field. Given a complex matrix F, let $F^* \equiv \overline{F}^T$ denote its conjugate transpose, where conjugation $\overline{\gamma + i\,\delta} = \gamma - i\,\delta$ for $\gamma, \delta \in \mathbb{R}$.

Lemma 8.3 (Bendixson). *Suppose the following conditions hold.*

1. *Let A be a complex matrix of size m with Hermitian part $H = \frac{1}{2}(A + A^*)$ and skew-Hermitian part $N = \frac{1}{2}(A - A^*)$.*
2. *Let C be a Hermitian positive definite matrix of size m such that the following bounds hold for $\mathbf{z} \in \mathbb{C}^m \setminus \{0\}$:*
$$\gamma_1 \leq \frac{\mathbf{z}^* H \mathbf{z}}{\mathbf{z}^* C \mathbf{z}} \leq \gamma_2 \quad \text{and} \quad \delta_1 \leq \frac{1}{i}\left(\frac{\mathbf{z}^* N \mathbf{z}}{\mathbf{z}^* C \mathbf{z}}\right) \leq \delta_2, \tag{8.14}$$
where $\gamma_1 \leq \gamma_2$ and $\delta_1 \leq \delta_2$ are real.
3. *Let λ be an eigenvalue of $C^{-1}A$.*

Then, the following bounds will hold:
$$\gamma_1 \leq \mathrm{Re}(\lambda) \leq \gamma_2 \quad \text{and} \quad \delta_1 \leq \mathrm{Im}(\lambda) \leq \delta_2,$$
where Re(λ) and $\mathrm{Im}(\lambda)$ *denote the real and imaginary parts of λ.*

Proof. The eigenvalues of $C^{-1}A$ will be contained in the *field of values* of the generalized Rayleigh quotient $R(\mathbf{z}) = \mathbf{z}^*A\mathbf{z}/\mathbf{z}^*C\mathbf{z}$ for $\mathbf{z} \neq \mathbf{0}$. Decompose $\mathbf{z}^*A\mathbf{z} = \mathbf{z}^*H\mathbf{z} + \mathbf{z}^*N\mathbf{z}$, and substitute this into the generalized Rayleigh quotient, use that $\mathbf{z}^*H\mathbf{z}$ is real and that $\mathbf{z}^*N\mathbf{z}$ is imaginary, and apply (8.14) to obtain the desired result. See [SA2]. □

As a Corollary of the preceding Lemma, we obtain the following result.

Corollary 8.4. *Suppose the following assumptions hold.*

1. *Let $0 < a_0 \leq a(x) \leq \|a\|_\infty$ and $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \geq 0$.*
2. *Let A denote the Galerkin matrix with $H = \frac{1}{2}(A+A^*)$ and $N = \frac{1}{2}(A-A^*)$.*
3. *Let $C = G$ denote the mass (Gram) matrix.*

Then, the following bounds will hold for $\mathbf{z} \in \mathbb{C}^m \setminus 0$:
$$\gamma_1 \leq \frac{\mathbf{z}^* H \mathbf{z}}{\mathbf{z}^* G \mathbf{z}} \leq \gamma_2 \quad \text{and} \quad \delta_1 \leq \frac{1}{i}\left(\frac{\mathbf{z}^* N \mathbf{z}}{\mathbf{z}^* G \mathbf{z}}\right) \leq \delta_2, \tag{8.15}$$
for $\gamma_1 = O(a_0)$, $\gamma_2 = O(\|a\|_\infty\, h^{-2})$, and $-\delta_1 = \delta_2 = O(\|\mathbf{b}\|_\infty\, a_0^{-1}\, h^{-1})$.

Proof. Apply the preceding Lemma using $C = G$. The bounds $\gamma_1 = O(a_0)$ and $\gamma_2 = O(\|a\|_\infty h^{-2})$ follow from standard finite element theory, since H corresponds to a stiffness matrix for a self adjoint problem.

To obtain bounds for δ_1 and δ_2, choose *complex* $u_h, v_h \in V_h$ with associated complex nodal vectors $\mathbf{u}$ and $\mathbf{v}$. Apply the Schwartz inequality:

$$|\mathcal{N}(u_h, v_h)| \leq (c_1 \|\mathbf{b}\|_\infty) \|u_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)},$$

for some $c_1 > 0$. Expressing the norms in matrix terms yields:

$$\left|\mathbf{v}^T N \mathbf{u}\right| \leq \left(c_1 \|\mathbf{b}\|_\infty a_0^{-1}\right) \|\mathbf{u}\|_H \|\mathbf{v}\|_G,$$

where $\|\mathbf{u}\|_H^2 = \mathbf{u}^* H \mathbf{u}$ and $\|\mathbf{v}\|_G^2 = \mathbf{v}^* G \mathbf{v}$ denote the matrix norms generated by H and G, respectively. Choosing $u_h = v_h$, and employing the inverse inequality $\|\mathbf{u}\|_H \leq C h^{-1} \|\mathbf{u}\|_G$ yields the desired result. □

Remark 8.5. The preceding result immediately yields bounds for the real and imaginary parts of the spectrum λ of A, since the Galerkin mass matrix G is well conditioned with $G \asymp h^d I$ for $\Omega \subset \mathbb{R}^d$:

$$\gamma_1 \leq \mathrm{Re}(\lambda) \leq \gamma_2, \qquad \delta_1 \leq \mathrm{Im}(\lambda) \leq \delta_2,$$

where $\gamma_1 = O(a_0 h^d)$, $\gamma_2 = O(\|a\|_\infty h^{-2+d})$, $-\delta_1 = \delta_2 = O(\|\mathbf{b}\|_\infty a_0^{-1} h^{-1+d})$. For *finite difference* discretizations, it can be shown that $\gamma_1 = O(a_0)$, while $\gamma_2 = \|a\|_\infty h^{-2}$ and $-\delta_1 = \delta_2 = O(\|\mathbf{b}\|_\infty a_0^{-1} h^{-1})$.

8.1.1 GMRES Bounds

The following bound [SA2] will be satisfied when the GMRES algorithm is used to solve $A\mathbf{u} = \mathbf{f}$, provided that A is diagonalizable with $A = U \Lambda U^{-1}$:

$$\|A\,\mathbf{u}^{(k)} - \mathbf{f}\| \leq \mathrm{cond}(U) \left(\min_{\{p(\cdot) \in \mathcal{P}_k \,:\, p(0)=1\}} \max_{\lambda \in Sp\,(A)} |p_k(\lambda)| \right) \|A\,\mathbf{u}^{(0)} - \mathbf{f}\|,$$

where $\mathbf{u}^{(k)}$ is the k'th GMRES iterate, $\mathcal{P}_k$ denotes polynomials of degree k and $Sp\,(A)$ denotes the spectrum of A (diagonal entries of Λ). For alternative estimates based on the "ε-pseudospectrum" of A, see [TR], or the angle between invariant subspaces of A, see [SI3]. When matrix A is *positive definite* (i.e., $H^T = H > 0$), the following minimum residual bound will hold:

$$\|A\,\mathbf{u}^{(k)} - \mathbf{f}\| \leq \left(1 - \frac{\lambda_{\min}(H)^2}{\sigma_{\max}(A)^2}\right)^{\frac{k}{2}} \|A\,\mathbf{u}^{(0)} - \mathbf{f}\|, \tag{8.16}$$

which estimates the residual norm of the k'th GMRES iterate $\mathbf{u}^{(k)}$. Here $\|\cdot\|$ denotes the Euclidean norm, $H = \frac{A^*+A}{2}$ denotes the symmetric part of A, and $\lambda_{\min}(H)$ denotes the minimal eigenvalue of H, while $\sigma_{\max}(A) = \|A\|$ denotes the maximal singular value of A, see [SA2].

8.2 Diffusion Dominated Case

Most domain decomposition preconditioners generalize to non-self adjoint problems. We first describe preconditioners for A in the diffusion dominated and *coercive* case, i.e. when $\|\mathbf{b}\|_\infty = O(a_0)$ and $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \ge 0$ in (8.1). In this case, it can be shown that the positive definite matrix H effectively preconditions A independent of h (however, the bounds deteriorate when $a_0 \ll \|\mathbf{b}\|_\infty$). Below, we describe several preconditioners based on H.

8.2.1 A Positive Definite Preconditioner

In the diffusion dominated and coercive case, matrix H will dominate the matrix N. As shown in Lemma 8.6, this motivates using the symmetric positive definite matrix H (or H_0 spectrally equivalent to H) as a preconditioner for the nonsymmetric matrix $A = H + N$, see [YS, VA10].

Lemma 8.6. *Suppose the following assumptions hold.*

1. *Let A be the stiffness matrix with $H = \frac{1}{2}(A + A^*)$, $N = \frac{1}{2}(A - A^*)$, and let H_0 be a symmetric positive definite preconditioner for H satisfying:*
$$\gamma_1 \le \frac{\mathbf{z}^* H \mathbf{z}}{\mathbf{z}^* H_0 \mathbf{z}} \le \gamma_2, \tag{8.17}$$
where $0 < \gamma_1 < \gamma_2$ are independent of h.
2. *Let $|\mathbf{v}^* N \mathbf{u}| \le C\, \|\mathbf{u}\|_H\, \|\mathbf{v}\|_G$ and $\|\mathbf{v}\|_G \le \nu\, \|\mathbf{v}\|_H$.*

Then, the following bounds will hold:
$$\gamma_1 \le \frac{\left(\mathbf{z}, H_0^{-1} A \mathbf{z}\right)_{H_0}}{\left(\mathbf{z}, \mathbf{z}\right)_{H_0}} \le \gamma_2 \quad \text{and} \quad \|H_0^{-1} A\|_{H_0} \le \delta, \tag{8.18}$$
for $\delta = \gamma_2(1 + C\,\gamma_2\,\nu)$, where $(\mathbf{v}, \mathbf{w})_{H_0} \equiv \mathbf{v}^* H_0 \mathbf{w}$ and $\|\cdot\|_{H_0}^2 = (\cdot, \cdot)_{H_0}$.

Proof. See [YS, VA10]. The first bound follows immediately upon substitution of $(\cdot, \cdot)_{H_0}$ and by employing the assumptions on H_0.

To obtain a bound for $\|H_0^{-1} A\|_{H_0}$, we separately estimate $\|H_0^{-1} H\|_{H_0}$ and $\|H_0^{-1} N\|_{H_0}$ in $H_0^{-1} A = H_0^{-1} H + H_0^{-1} N$. Accordingly, consider the term:
$$\|H_0^{-1} H\|_{H_0}^2 = \sup_{\mathbf{v} \ne \mathbf{0}} \frac{\mathbf{v}^* H H_0^{-1} H \mathbf{v}}{\mathbf{v}^* H_0 \mathbf{v}} = \sup_{\mathbf{v} \ne \mathbf{0}} \frac{\|H_0^{-1/2} H \mathbf{v}\|^2}{\|H_0^{1/2} \mathbf{v}\|^2},$$
where $\|\cdot\|$ denotes the Euclidean norm. The numerator satisfies:
$$\begin{aligned} \|H_0^{-1/2} H \mathbf{v}\| = \sup_{\mathbf{w} \ne \mathbf{0}} \frac{\left|\mathbf{w}^* H_0^{-1/2} H \mathbf{v}\right|}{\|\mathbf{w}\|} &= \sup_{\mathbf{u} \ne \mathbf{0}} \frac{|\mathbf{u}^* H \mathbf{v}|}{\|H_0^{1/2} \mathbf{u}\|} \\ &\le \frac{\|\mathbf{u}\|_H\, \|\mathbf{v}\|_H}{\|\mathbf{u}\|_{H_0}} \\ &\le \gamma_2^{1/2}\, \|\mathbf{v}\|_H. \end{aligned}$$

Substituting this bound yields the estimate:

$$\|H_0^{-1}H\|_{H_0}^2 \le \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\gamma_2\,\|\mathbf{v}\|_H^2}{\|H_0^{1/2}\mathbf{v}\|^2} = \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\gamma_2\,\|\mathbf{v}\|_H^2}{\|\mathbf{v}\|_{H_0}^2} \le \gamma_2^2.$$

Thus, $\|H_0^{-1}H\|_{H_0} \le \gamma_2$. Next, the second term can be estimated as:

$$\|H_0^{-1}N\|_{H_0}^2 = \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^* N^* H_0^{-1} N\mathbf{v}}{\mathbf{v}^* H_0\mathbf{v}} = \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\|H_0^{-1/2}N\mathbf{v}\|^2}{\|H_0^{1/2}\mathbf{v}\|^2}.$$

As before, the numerator may be estimated as:

$$\|H_0^{-1/2}N\mathbf{v}\| = \sup_{\mathbf{w}\neq\mathbf{0}} \frac{\left|\mathbf{w}^* H_0^{-1/2} N\mathbf{v}\right|}{\|\mathbf{w}\|} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{|\mathbf{u}^* N\mathbf{v}|}{\|H_0^{1/2}\mathbf{u}\|} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{|\mathbf{u}^* N\mathbf{v}|}{\|\mathbf{u}\|_{H_0}}.$$

Substituting the bound $|\mathbf{v}^* N\mathbf{u}| \le C\,\|\mathbf{u}\|_H\,\|\mathbf{v}\|_G$ yields:

$$\|H_0^{-1}N\mathbf{v}\|_{H_0}^2 = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{|\mathbf{u}^* N\mathbf{v}|^2}{\|\mathbf{u}\|_{H_0}^2} \le \sup_{\mathbf{u}\neq\mathbf{0}} \frac{C^2\|\mathbf{u}\|_H^2\,\|\mathbf{v}\|_G^2}{\|\mathbf{u}\|_{H_0}^2} \le C^2\,\gamma_2^2\,\|\mathbf{v}\|_G^2.$$

Employing this bound, and substituting $\|\mathbf{v}\|_G \le \nu\,\|\mathbf{v}\|_H$, we obtain:

$$\|H_0^{-1}N\|_{H_0}^2 = \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\|H_0^{-1/2}N\mathbf{v}\|^2}{\|\mathbf{v}\|_{H_0}^2} \le \sup_{\mathbf{v}\neq\mathbf{0}} \frac{C^2\,\gamma_2^2\,\|\mathbf{v}\|_G^2}{\|\mathbf{v}\|_{H_0}^2} \le C^2\gamma_2^4\,\nu^2,$$

where we applied a Poincare-Freidrichs inequality $\|\mathbf{v}\|_G^2 \le \nu^2\|\mathbf{v}\|_H^2$. Each term is independent of h, so summing both terms yields the desired bound. □

Remark 8.7. The preceding result yields bounds for the rate of convergence of the GMRES algorithm to solve $H_0^{-1}A\mathbf{u} = H_0^{-1}\mathbf{f}$ in the H_0-inner product. The expressions $\lambda_{\min}(H_0^{-1}A) \le \gamma_1$ and $\sigma_{\max}(H_0^{-1}A) \le \gamma_2(1 + C\,\gamma_2\,\nu)$, can be substituted into (8.16), provided $\|\cdot\|$ is replaced by $\|\cdot\|_{H_0}$. As a result, the convergence will depend only on the parameters γ_1, γ_2, C and ν. In applications to (8.1), we may estimate C and ν as follows. Applying integration by parts and Schwartz's inequality yields:

$$|\mathbf{u}^* N\mathbf{v}| = |\mathcal{N}(u_h, v_h)| \le \|\mathbf{b}\|_\infty\,\|u_h\|_{H^1(\Omega)}\,\|v_h\|_{L^2(\Omega)} \le C\,\|\mathbf{u}\|_H\,\|\mathbf{v}\|_G,$$

for some $C > 0$ independent of h, but dependent on $a_0 \le a(x)$. The bound $\|\mathbf{v}\|_G \le \nu\|\mathbf{v}\|_H$ will hold by the Poincare-Freidrichs inequality for some ν independent of h, but dependent on a_0. Provided H_0 is spectrally equivalent to H, the parameters γ_1 and γ_2 will be independent of h. When equation (8.1) is *advection dominated*, the parameters C and ν can become large. For instance, if $a(x) \equiv \varepsilon \ll \|\mathbf{b}\|_\infty$ and $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) = 0$, and if a Galerkin discretization satisfying a Peclet condition is employed, we shall obtain $C = O(\varepsilon^{-1})$ and $\nu = O(\varepsilon^{-1})$. If H_0 is chosen to be spectrally equivalent to H, independent of h and ϵ, then $\lambda_{\min}\left(H_0^{-1}A\right)$ will be independent of h and ϵ. However, $\sigma_{\max}(H_0^{-1}A)$ will deteriorate as $O(\varepsilon^{-2})$.

8.2.2 An Additive Preconditioner of [XU8]

When $\big(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\big)$ is *negative*, matrix $H = \frac{1}{2}(A + A^*)$ can be *indefinite.* In this case, Lemma 8.6 will not be applicable. Despite this, we may modify a symmetric positive definite preconditioner H_0 for A by including an additive but *indefinite* coarse space correction term [XU8]. Let $V_0 \subset V_h$ denote a coarse subspace based on a coarse triangulation with mesh size $h_0 > h$, and basis $\{\psi_1^{(0)}, \ldots, \psi_m^{(0)}\}$. Let R_0 and R_0^T denote the nodal restriction and extension maps associated with the coarse space, and $A_0 = R_0 A R_0^T$. If H_0 is any symmetric positive definite preconditioner for $L_* u = -\nabla \cdot (a(x)\nabla u)$, then the action of the inverse M^{-1} of the additive preconditioner of [XU8] is:

$$M^{-1} = H_0^{-1} + R_0^T A_0^{-1} R_0.$$

For sufficiently small h_0, this preconditioner will yield a rate of convergence independent of h in GMRES, see [XU8] and Chap. 8.5.

8.2.3 Schwarz Preconditioners

Given the non-symmetric stiffness matrix A obtained by discretization of $\mathcal{A}^0(.,.)$, Schwarz preconditioners can be constructed by formal analogy with the symmetric case. In the non-symmetric case, however, additional preconditioners can be formulated which employ a *discrete partition of unity.* The latter preconditioners will be non-symmetric even when A is symmetric.

Let $\Omega_1^*, \ldots, \Omega_p^*$ form an overlapping decomposition of Ω, such that each Ω_i^* aligns with a triangulation Ω_h of Ω. We shall assume that A is of dimension n. For $1 \le i \le p$ let n_i denote the number of interior nodes in Ω_i^*. Corresponding to each subdomain Ω_i^*, let R_i denote a matrix of size $n_i \times n$ with zero-one entries, which restricts a vector of nodal values on Ω to its subvector corresponding to nodes in Ω_i^*, in the local ordering of nodes. As in the self adjoint case, let R_0^T denote a matrix of size $n \times n_0$ whose columns span a coarse space. For $1 \le i \le p$ we define $A_i \equiv R_i A R_i^T$ as the submatrix of A of size n_i, corresponding to nodes in Ω_i^*, and $A_0 = R_0 A R_0^T$ as the coarse space matrix. By construction, each matrix A_i will be nonsymmetric.

The action of the inverse M^{-1} of an *additive Schwarz* preconditioner for the non-symmetric matrix A is analogous to the symmetric case:

$$M^{-1} = R_0^T A_0^{-1} R_0 + \sum_{i=1}^{p} R_i^T A_i^{-1} R_i. \tag{8.19}$$

Alternate additive preconditioners can be obtained using a discrete partition of unity. Let $\Omega_1, \ldots, \Omega_p$ form a non-overlapping decomposition of Ω such that:

$$\Omega_i^* = \{x \in \Omega \,:\, \operatorname{dist}(x, \Omega_i) < \beta_i\},$$

i.e., each Ω_i^* contains all points in Ω within a distance of some $\beta_i > 0$ of Ω_i. Let $D^{(i)}$ denote a diagonal matrix of dimension n_i with *nonnegative* diagonal entries that form a discrete partition of unity satisfying $I = \sum_{i=1}^p R_i^T D^{(i)} R_i$. For instance, if $x_l^{(i)}$ denotes the l'th interior grid point in Ω_i^* in the chosen local ordering of nodes, and $N(x_l^{(i)})$ denotes the number of non-overlapping subdomains $\overline{\Omega}_i$ containing $x_l^{(i)}$, then we can define:

$$\left(D^{(i)}\right)_{ll} \equiv \begin{cases} \frac{1}{N(x_l^{(i)})}, & \text{if } x_l^{(i)} \in \overline{\Omega}_i \\ 0, & \text{if } x_l^{(i)} \notin \overline{\Omega}_i. \end{cases}$$

Given $D^{(i)}$, the *restricted Schwarz* preconditioner will have the form:

$$\begin{cases} M^{-1}\mathbf{r} = R_0^T A_0^{-1} R_0\,\mathbf{r} + \sum_{i=1}^p R_i^T D^{(i)} A_i^{-1} R_i\,\mathbf{r}, & \text{or} \\ M^{-1}\mathbf{r} = R_0^T A_0^{-1} R_0\,\mathbf{r} + \sum_{i=1}^p R_i^T A_i^{-1} D^{(i)} R_i\,\mathbf{r}, \end{cases} \tag{8.20}$$

see [CA19, FR8] and Chap. 15. The motivation for using the discrete partition of unity matrices $D^{(i)}$ is that it helps to reduce the redundancy in the sum of the local solutions $(R_i^T A_i^{-1} R_i)\,\mathbf{r}$ on the regions of overlap.

Under suitable assumptions, such as $c(x) \geq c_0 > 0$, sufficient overlap between the subdomains, and a discrete maximum principle holding for the discretization, the following *unaccelerated* version of the restricted Schwarz algorithm to solve $A\mathbf{u} = \mathbf{f}$ without coarse space correction will converge in the maximum norm at a rate independent of h, but dependent on c_0 and the amount of overlap, see [MA33] and Chap. 15. Without a coarse space correction term, however, the rate of convergence will deteriorate as the number of subdomains increases, as $c_0 \to 0$, or as the overlap decreases.

Algorithm 8.2.1 *(Unaccelerated Restricted Schwarz Algorithm)*
Input: $\mathbf{w}^{(0)}$ *and* $\mathbf{f}$

```
1. For $k = 0, \cdots$ until convergence do:
       $\mathbf{w}^{(k+1)} = \mathbf{w}^{(l)} + \sum_{i=1}^{p} R_i^T D^{(i)} A_i^{-1} R_i \left(\mathbf{f} - A\mathbf{w}^{(k)}\right).$
2. Endfor
Output: $\mathbf{w}^{(k)}$
```

The local solves in step 1 can be implemented in parallel. Furthermore, the associated preconditioner will not be symmetric even if A were symmetric. Below, we describe the nonsymmetric (and possibly indefinite) multiplicative Schwarz preconditioner for use with Krylov space acceleration. Its associated unaccelerated fixed point iteration may not be convergent. Below, we list the *multiplicative Schwarz* preconditioner M.

Algorithm 8.2.2 *(Multiplicative Schwarz Preconditioner)*

```
Input: $\mathbf{w}^{(0)} \equiv \mathbf{0}$ and $\mathbf{r}$
1. For $i = 0, \cdots, p$ do:
       $\mathbf{w}^{(\frac{i+1}{p+1})} = \mathbf{w}^{(\frac{i}{p+1})} + R_i^T A_i^{-1} R_i \left(\mathbf{r} - A\mathbf{w}^{(\frac{i}{p+1})}\right).$
2. Endfor
Output: $M^{-1}\mathbf{r} \equiv \mathbf{w}^{(1)}$
```

The following convergence bounds will hold in the Euclidean norm $\|\cdot\|$, even when $(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x))$ is possibly negative.

Lemma 8.8. *Let M denote either the additive or sequential Schwarz preconditioner and let $\mathbf{u}^{(k)}$ denote the k'th GMRES iterate to solve the preconditioned system $M^{-1}A\mathbf{u} = M^{-1}\mathbf{f}$. If the grid size h_0 associated with the coarse space $\text{Range}(R_0^T)$ is sufficiently small (depending on $(\|a\|_\infty / \|\mathbf{b}\|_\infty)$ and h), then, the iterates $\mathbf{u}^{(k)}$ in the preconditioned GMRES algorithm will satisfy:*

$$\|A\mathbf{u}^{(k)} - \mathbf{f}\| \leq (1 - \delta_0)^k \|A\mathbf{u}^{(0)} - \mathbf{f}\|,$$

for some $0 \leq \delta_0 < 1$ independent of h.

Proof. See [CA20, XU8, CA21, WA3]. □

Below, we list the *unaccelerated* multiplicative Schwarz algorithm to solve $A\mathbf{u} = \mathbf{f}$, omitting the coarse space correction step $i = 0$ in step 2.

Algorithm 8.2.3 *(Unaccelerated Multiplicative Schwarz Algorithm)*

```
Input: $\mathbf{w}^{(0)}$ and $\mathbf{f}$
1. For $k = 0, \cdots$ until convergence do:
2.     For $i = 1, \cdots, p$ do:
           $\mathbf{w}^{(k+\frac{i}{p})} = \mathbf{w}^{(k+\frac{i-1}{p})} + R_i^T A_i^{-1} R_i \left(\mathbf{f} - A\mathbf{w}^{(k+\frac{i}{p})}\right).$
3.     Endfor
4. Endfor
Output: $\mathbf{u}^{(k)}$
```

Remark 8.9. If $c(x) \geq c_0 > 0$ and the overlap between Ω_i^* is sufficiently large, if A is an M-matrix and the initial iterate satisfies $\mathbf{w}^{(0)} \geq \mathbf{u}$ or $\mathbf{w}^{(0)} \leq \mathbf{u}$ componentwise, then the iterates $\mathbf{w}^{(k)}$ of the *unaccelerated* restricted Schwarz algorithm and the *unaccelerated* multiplicative Schwarz algorithm will converge *monotonically* to $\mathbf{u}$ in the maximum norm, i.e., each iterate will also satisfy $\mathbf{w}^{(k)} \geq \mathbf{u}$ or $\mathbf{w}^{(k)} \leq \mathbf{u}$ componentwise. The rate of convergence will be independent of h, but dependent on the number of subdomain, the amount of overlap and c_0, see [MA33, FR7, FR8] and Chap. 15.

Remark 8.10. When $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \geq 0$, then $A = A^T > 0$. The step:

$$\mathbf{w}^{(k+\frac{i}{p})} = \mathbf{w}^{(k+\frac{i-1}{p})} + R_i^T A_i^{-1} R_i \left(\mathbf{f} - A\mathbf{w}^{(k+\frac{i-1}{p})} \right),$$

will minimize the energy $J(\mathbf{w}) \equiv \frac{1}{2}\mathbf{w}^T A\mathbf{w} - \mathbf{w}^T\mathbf{f}$ within $\mathrm{Range}(R_i^T)$:

$$J\left(\mathbf{w}^{(k+\frac{i}{p})}\right) = \min_{\mathbf{v}_i \in \mathrm{Range}(R_i^T)} J\left(\mathbf{w}^{(k+\frac{i-1}{p})} + \mathbf{v}_i\right),$$

i.e., the energy $J(.)$ will be *non-increasing* each iteration. When A is *nonsymmetric*, the sequential Schwarz algorithm can be *modified* so that the square norm of the residual is minimized within the subspace $\mathrm{Range}(R_i^T)$ during the i'th fractional step. Indeed, define a functional $J_R(\mathbf{w})$ as:

$$J_R(\mathbf{w}) \equiv \frac{1}{2}\|A\mathbf{w} - \mathbf{f}\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. Then, if we require the update at the i'th fractional step to minimize $J_R(\cdot)$ within subspace $\mathrm{Range}(R_i^T)$:

$$J_R\left(\mathbf{w}^{(k+\frac{i}{p})}\right) = \min_{\mathbf{v}_i \in \mathrm{Range}(R_i^T)} J_R\left(\mathbf{w}^{(k+\frac{i-1}{p})} + \mathbf{v}_i\right),$$

the update will satisfy:

$$\mathbf{w}^{(k+\frac{i}{p})} = \mathbf{w}^{(k+\frac{i-1}{p})} + R_i^T \left(R_i A^T A R_i^T\right)^{-1} R_i A^T \left(\mathbf{f} - A\mathbf{w}^{(k+\frac{i-1}{p})}\right).$$

The resulting algorithm will be a sequential Schwarz algorithm to solve the normal equations $A^T A\mathbf{u} = A^T\mathbf{f}$, and $J_R(\cdot)$ will be *nonincreasing.*

8.2.4 Schur Complement Preconditioners

Schur complement preconditioners can be formulated for *nonsymmetric* diffusion dominated problems by analogy with the symmetric case. We shall let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of Ω into p subregions, with subdomain boundary $B^{(l)} = \partial\Omega_l \cap \Omega$ and $B_{[l]} = \partial\Omega_l \cap \partial\Omega$, and interface $B = \cup_{l=1}^p B^{(l)}$. We define the following subdomain forms:

$$\begin{cases} \mathcal{A}^0_{\Omega_l}(u, v) = \int_{\Omega_l} \left(a(x)\nabla u \cdot \nabla v + (\mathbf{b}(x) \cdot \nabla u)\, v + c(x)\, u\, v\right) dx \\ F_{\Omega_l}(v) \quad\;\; = \int_{\Omega_l} f\, v\, dx. \end{cases}$$

Given a nonoverlapping decomposition, let $\mathbf{u}_I^{(l)}$ and $\mathbf{u}_B^{(l)}$ denote nodal vectors corresponding to the discrete solution at nodes in the interior of Ω_l and on $B^{(l)}$, respectively. On each subdomain Ω_l, we define a local stiffness matrix:

$$A_{ij}^{(l)} = \mathcal{A}^0_{\Omega_l}(\phi_i, \phi_j).$$

Galerkin discretization of the following weak problem on subdomain Ω_l:

$$\mathcal{A}^0_{\Omega_l}(u,v) = F_{\Omega_l}(v) + \langle g, v\rangle\,, \quad \text{where } \langle g, v\rangle = \int_{\partial\Omega_l} g\, v\, ds_x, \quad \forall u, v \in H^1_0(\Omega),$$

will yield a linear system of the following form:

$$\begin{bmatrix} A^{(l)}_{II} & A^{(l)}_{IB} \\ A^{(l)}_{BI} & A^{(l)}_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(l)}_I \\ \mathbf{u}^{(l)}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(l)}_I \\ \mathbf{g}^{(l)}_B \end{bmatrix}, \tag{8.21}$$

where $\mathbf{f}^{(l)}_I$ and $\mathbf{g}^{(l)}_B$ denote local discretizations of $F_{\Omega_l}(\cdot)$ and $F_{\Omega_l}(\cdot) + \langle g, \cdot\rangle$, respectively. For sufficiently smooth u, v, integration by parts in Ω_l yields:

$$\begin{aligned} \mathcal{A}^0_{\Omega_l}(u,v) = \quad & \textstyle\int_{\Omega_l} \left(-\nabla\cdot(a(x)\nabla u) + (\mathbf{b}(x)\cdot\nabla u)\, v + c(x)\, u\, v\right) dx \\ & + \textstyle\int_{\partial\Omega_l} \mathbf{n}_l(x)\cdot(a(x)\nabla u)\; v\, ds_x. \end{aligned}$$

This shows that each *subdomain* problem above discretizes:

$$\begin{cases} Lu \equiv -\nabla\cdot(a(x)\nabla u) + \mathbf{b}(x)\cdot\nabla u + c(x)\, u = f, & \text{in}\Omega_l \\ u = 0, & \text{on}B_{[l]} \\ \mathbf{n}_l\cdot(a\nabla u) = g, & \text{on}B^{(l)}, \end{cases} \tag{8.22}$$

thereby enforcing *Neumann* boundary conditions on $B^{(l)}$. The symmetric and skew-symmetric components of $\mathcal{A}^0_{\Omega_l}(u,v) = \mathcal{H}^0_{\Omega_l}(u,v) + \mathcal{N}^0_{\Omega_l}(u,v)$ will be:

$$\begin{cases} \mathcal{H}^0_{\Omega_l}(u,v) = & \int_{\Omega_l} a(x)\nabla u\cdot\nabla v\, dx + \int_{\Omega_l}\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) u\, v\, dx \\ & + \frac{1}{2}\int_{\partial\Omega_l} \mathbf{n}_l(x)\cdot\mathbf{b}(x)\, u\, v\, ds_x \\ \mathcal{N}^0_{\Omega_l}(u,v) = & \frac{1}{2}\int_{\Omega_l}\left((\mathbf{b}(x)\cdot\nabla u)\, v - u\,(\mathbf{b}(x)\cdot\nabla v)\right) dx, \end{cases} \tag{8.23}$$

which can be verified easily using integration by parts. Thus, $\mathcal{A}^0_{\Omega_l}(u,u)$ will be *coercive* if $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) \geq 0$ in Ω_l and $\mathbf{n}_l(x)\cdot\mathbf{b}(x) \geq 0$ on $\partial\Omega_l$.

Let $\mathbf{u}_I = \left(\mathbf{u}^{(1)^T}_I, \ldots, \mathbf{u}^{(p)^T}_I\right)^T$ and $\mathbf{u}_B$ denote nodal vectors associated with the discrete solution in $\cup_{l=1}^p \Omega_l$ and B, respectively. Using this ordering of nodal variables, we may block partition the system $A\mathbf{u} = \mathbf{f}$ as follows:

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ \mathbf{f}_B \end{bmatrix}. \tag{8.24}$$

When $(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}) \geq 0$, it will hold that A is non-singular. In this case, we may express $\mathbf{u}_I = A^{-1}_{II}\left(\mathbf{f}_I - A_{IB}\mathbf{u}_B\right)$ and substitute this into the second block row to obtain a non-symmetric Schur complement system:

$$S\mathbf{u}_B = \tilde{\mathbf{f}}_B, \tag{8.25}$$

where $S = \left(A_{BB} - A_{BI}A^{-1}_{II}A_{IB}\right)$, and $\tilde{\mathbf{f}}_B = \left(\mathbf{f}_B - A_{BI}A^{-1}_{II}\mathbf{f}_I\right)$. The solution to (8.24) can be obtained by solving system (8.25) for $\mathbf{u}_B$ by a preconditioned GMRES algorithm, and then determining $\mathbf{u}_I = A^{-1}_{II}\left(\mathbf{f}_I - A_{IB}\mathbf{u}_B\right)$. The following algebraic properties will hold in the *nonsymmetric* case.

Lemma 8.11. *Let A_{II} be non-singular and let $E\,\mathbf{u}_B \equiv -A_{II}^{-1} A_{IB}\,\mathbf{u}_B$ denote a "discrete harmonic" extension. Then, the following results will hold.*

1. *The Schur complement $S = \left(A_{BB} - A_{BI} A_{II}^{-1} A_{IB}\right)$ will satisfy:*

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{bmatrix} \begin{bmatrix} E\,\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ S\,\mathbf{u}_B \end{bmatrix}.$$

2. *The energy associated with the nonsymmetric matrix S will satisfy:*

$$\mathbf{u}_B^T S \mathbf{u}_B \;=\; \begin{bmatrix} E\,\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}^T \begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{bmatrix} \begin{bmatrix} E\,\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix}.$$

Proof. Follows by direct substitution. □

Remark 8.12. As a corollary of the preceding Lemma, it will hold that S is positive definite when A is positive definite. More specifically, when $A > 0$ it will hold that $\mathbf{u}_B^T S \mathbf{u}_B \ge \mathbf{u}_B^T S_0 \mathbf{u}_B$ where $S_0 = (H_{BB} - H_{IB}^T H_{II}^{-1} H_{IB})$ is the Schur complement of matrix $H = \frac{1}{2}(A + A^T)$. To show this, note that:

$$\mathbf{u}_B^T S \mathbf{u}_B \;=\; (E\mathbf{u}_B, \mathbf{u}_B)^T\, A\, (E\mathbf{u}_B, \mathbf{u}_B) \;=\; (E\mathbf{u}_B, \mathbf{u}_B)^T\, H\, (E\mathbf{u}_B, \mathbf{u}_B)\,.$$

Since $E\mathbf{u}_B \;=\; -A_{II}^{-1} A_{IB} \mathbf{u}_B$ is not the "discrete harmonic" extension relative to H, the minimization property of the discrete harmonic extension relative to H will yield that $(E\mathbf{u}_B, \mathbf{u}_B)^T\, H\, (E\mathbf{u}_B, \mathbf{u}_B) \;\ge\; \mathbf{u}_B^T (H_{BB} - H_{IB}^T H_{II}^{-1} H_{IB}) \mathbf{u}_B$. Substituting this into preceding yields that $\mathbf{u}_B^T S \mathbf{u}_B \;\ge\; \mathbf{u}_B^T S_0 \mathbf{u}_B$.

The following M-matrix property will also hold in the nonsymmetric case.

Lemma 8.13. *If A is an M-matrix, then S will also be an M-matrix.*

Proof. See Chap. 3. □

We next outline Neumann-Neumann and Robin-Robin preconditioners for S in the diffusion dominated case.

A *nonsymmetric Neumann-Neumann* preconditioner can be formulated for S as follows. On each subdomain Ω_l, let $S^{(l)} = A_{BB}^{(l)} - A_{BI}^{(l)} A_{II}^{(l)} A_{IB}^{(l)}$ denote the subdomain Schur complement. As in the symmetric case, let the columns of $\mathcal{R}_0^T$ form a basis for a coarse space on B, and define $S_0 \equiv \mathcal{R}_0 S \mathcal{R}_0^T$. Then, the inverse of the non-symmetric Neumann-Neumann preconditioner for S is:

$$\begin{aligned} M^{-1} &= \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0 + \textstyle\sum_{l=1}^p \mathcal{R}_l^T D^{(l)} {S^{(l)}}^{-1} \mathcal{R}_l, \quad \text{or} \\ M^{-1} &= \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0 + \textstyle\sum_{l=1}^p \mathcal{R}_l^T {S^{(l)}}^{-1} D^{(l)} \mathcal{R}_l, \end{aligned} \tag{8.26}$$

where $\mathcal{R}_l$ denotes the nodal restriction map which restricts a nodal vector on B onto nodes on $B^{(l)}$, its transpose $\mathcal{R}_l^T$ extends a nodal vector on $B^{(l)}$ to a nodal vector on B (extension by zero), and $D^{(l)}$ forms a discrete partition of unity on B with nonnegative diagonal entries satisfying $I = \sum_{l=1}^{p} \mathcal{R}_l^T D^{(l)} \mathcal{R}_l$. Note that each discrete partition of unity matrix $D^{(l)}$ is employed only once. The action of the inverse of $S^{(l)}$ can be computed using the expression:

$$S^{(l)^{-1}} = \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{BI}^{(l)} & A_{BB}^{(l)} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix},$$

and will require solving a local problem of the form (8.22). Each local problem will be *coercive* if $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \geq 0$ in Ω_l and $\mathbf{n}_l(x) \cdot \mathbf{b}(x) \geq 0$ on $\partial\Omega_l$. See [CO4] for a non-symmetric version of the balancing preconditioner.

Remark 8.14. Analysis of a *two-subdomain* Neumann-Neumann algorithm shows that the rate of convergence can be sensitive to the magnitude of $a(x)$ and the direction of $\mathbf{b}(x)$, see [AC7, AL4, RA3]. Adding the mass matrix term $M_{\gamma,BB}^{(l)} \geq 0$ to $S^{(l)}$ can increase the coercivity of each subdomain Schur complement, and improve the convergence of the Neumann-Neumann algorithm [GA15, GA14, AC7, AL4, QU6, RA3]. Given $\gamma(x) \geq 0$ on $\partial\Omega_l$ and the finite element basis $\{\psi_i\}$ on $B^{(l)}$, let $M_{\gamma,BB}^{(l)}$ denote the mass matrix with entries $\left(M_{\gamma,BB}^{(l)}\right)_{ij} = \int_{B^{(l)}} \gamma(x)\, \psi_i(x)\, \psi_j(x)\, ds_x$. Then, the *Robin-Robin* preconditioner without a discrete partition of unity has the form [AC7, RA3]:

$$M^{-1} = \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0 + \sum_{l=1}^{p} \mathcal{R}_l^T \left(S^{(l)} + M_{\gamma,BB}^{(l)}\right)^{-1} \mathcal{R}_l, \tag{8.27}$$

where the action of $\left(S^{(l)} + M_{\gamma,BB}^{(l)}\right)^{-1}$ can be computed using:

$$\left(S^{(l)} + M_{\gamma,BB}^{(l)}\right)^{-1} = \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{BI}^{(l)} & A_{BB}^{(l)} + M_{\gamma,BB}^{(l)} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

Each local problem will enforce the discretization of a *Robin* boundary condition of the form $\mathbf{n}_l(x) \cdot (a\nabla u_l) + \gamma(x)u = g$ on $B^{(l)}$ for some $g(.)$.

8.3 Advection Dominated Case

The advection dominated case poses many computational challenges. We shall illustrate the issues by considering the advection dominated elliptic equation:

$$\begin{cases} -\varepsilon\Delta u_\varepsilon + \mathbf{b}(x) \cdot \nabla u_\varepsilon + c(x)\, u_\varepsilon = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\quad u_\varepsilon = 0, & \text{on } \partial\Omega, \end{cases} \tag{8.28}$$

where $\varepsilon \ll 1$ denotes a viscosity parameter and $\|\mathbf{b}\| = O(1)$. For notational convenience, we shall use $u(x)$ instead of $u_\varepsilon(x)$. To ensure well posedness of (8.28) as $\varepsilon \to 0^+$, we shall require $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \geq \beta > 0$. As $\varepsilon \to 0^+$, singular perturbation theory suggests that the solution $u_\varepsilon(x)$ can develop "steep gradients" (i.e., derivatives of large magnitude) within some *layer subregion*, see [KE5, LA5]. If the solution is to be resolved in the layer, such regions must be identified and the mesh size must be chosen appropriately small locally. Furthermore, a stable finite element [JO2, FR3, FR2] or upwind discretization of (8.28) must be employed as $\varepsilon \to 0^+$. Subsequently, an iterative algorithm can be formulated to solve the resulting system.

Multisubdomain Schwarz and Schur complement preconditioners can be applied in the advection dominated case with coarse space correction, however theoretical convergence bounds deteriorate as $\varepsilon \to 0^+$ unless the coarse grid size $h_0 \to 0^+$, see [CA20, XU8, CA21, WA3]. Instead, we shall focus on *two subdomain* Schwarz and Schur complement algorithms without coarse space correction. They will have the disadvantage of involving large subproblems, but yield convergence less sensitive to ε and h. Motivated by singular perturbation methodology, let $\Omega_1 \subset \Omega$ be a subdomain such that:

$$\varepsilon\,|\Delta u(x)| \ll |\mathbf{b}(x) \cdot \nabla u(x) + c(x)\,u(x)|\,, \quad \text{for } x \in \Omega_1. \tag{8.29}$$

We define Ω_2 as its complementary region $\left(\Omega \setminus \overline{\Omega}_1\right)$, and shall refer to it as a *layer* region, see [KE5, LA5]. By assumption, the term $-\varepsilon\Delta u$ may be omitted within Ω_1 to an approximation, however, omitting such a term within Ω_2 may introduce significant errors due to large gradients locally. Typically, the layer region Ω_2 will have smaller area (or volume) relative to Ω_1. For instance, if the layer is a "boundary layer", then Ω_2 will typically be a region of width $O(\varepsilon)$ surrounding $\partial\Omega$ when $\mathbf{b}(x) \neq \mathbf{0}$, while Ω_2 will be a region of width $O(\sqrt{\varepsilon})$ surrounding $\partial\Omega$ when $\mathbf{b}(x) = \mathbf{0}$, see [KE5, LA5] and Chap. 15. If the solution is to be resolved in the layer region Ω_2, then it may be necessary to choose a grid size $h_2 \ll h_1$ in Ω_2. For instance, if the layer region is of width $O(\varepsilon)$, then $h_2 = O(\varepsilon)$, while if the layer region is of width $O(\sqrt{\varepsilon})$, then $h_2 = O(\sqrt{\varepsilon})$.

Denote the linear system resulting from the discretization of (8.28) as:

$$A\,\mathbf{u} = \mathbf{f}, \quad \text{with} \quad A = \varepsilon\,H_0 + H_1 + N \tag{8.30}$$

where $H_0 \geq 0$ and $H_1 \geq 0$ are symmetric matrices and N is skew-symmetric. When $(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)) = 0$, it will hold that $H_1 = 0$ and $A = \varepsilon\,H_0 + N$. If H_0 is employed as a preconditioner for $A = \varepsilon\,H_0 + N$, an application of Lemma 8.6 will yield $\lambda_{\min}(H_0^{-1}A) = O(\varepsilon)$ and $\sigma_{\max}(H_0^{-1}A) = O(1)$, and so the convergence factor $\left(1 - O(\varepsilon^2)\right)$ in (8.16), to solve $H_0^{-1}A\mathbf{u} = H_0^{-1}\mathbf{f}$ deteriorates. If $(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)) \geq \beta > 0$, matrix H_1 will be spectrally equivalent to the mass matrix G with $(\varepsilon\,H_0 + H_1) \geq \beta\,G$. If $K_0 = K_0^T > 0$ is spectrally equivalent to $(\varepsilon\,H_0 + H_1)$ and used as a preconditioner for A, an application of Lemma 8.6 together with the bound $\left|\mathbf{v}^T N \mathbf{u}\right| \leq \tilde{C}\,h^{-1}\|\mathbf{v}\|_G \|\mathbf{u}\|_G$ (which will hold for $\tilde{C} = O(\|\mathbf{b}\|_\infty)$ by the inverse inequality), will yield $\lambda_{\min}(K_0^{-1}A) = O(1)$ and

$\sigma_{\max}(K_0^{-1}A) = O(h^{-1})$. The convergence factor in (8.16) will be $\left(1 - O(h^2)\right)$. However, if $\mathbf{b}(x)$ satisfies $\|\mathbf{b}\|_\infty = O(\tau) = O(h)$ (as in an implicit discretization of a parabolic equation with time step $\tau = O(h)$), an optimal order bound $\sigma_{\max}(M_0^{-1}A) = O(\tau\, h^{-1}) = O(1)$ will hold, with convergence factor $\left(1 - O(h/\tau)^2\right)$ in (8.16). In this case, the additive Schwarz preconditioner without coarse space correction will also satisfy similar bounds [WU2].

8.3.1 Hermitian-Skew-Hermitian Splittings

If efficient solvers are available for $(\alpha I + H)$ and $(\alpha I + N)$, then an algebraic two step unaccelerated splitting algorithm of [BA12] can be employed to solve $A\,\mathbf{u} = \mathbf{f}$. It is based on the Hermitian part $H = \frac{1}{2}(A + A^*)$ and the skew-Hermitian part $N = \frac{1}{2}(A - A^*)$ of A. Given an iterate $\mathbf{u}^{(k)}$, an updated iterate $\mathbf{u}^{(k+1)}$ is computed in two fractional steps as follows:

$$\begin{cases} (\alpha I + H)\,\mathbf{u}^{(k+\frac{1}{2})} = \mathbf{f} + (\alpha I - N)\,\mathbf{u}^{(k)} \\ (\alpha I + N)\,\mathbf{u}^{(k+1)} = \mathbf{f} + (\alpha I - H)\,\mathbf{u}^{(k+\frac{1}{2})}, \end{cases}$$

where $\alpha > 0$ is a parameter that must be chosen appropriately to ensure convergence. The error is contracted as follows:

$$(\mathbf{u} - \mathbf{u}^{(k+1)}) = \left((\alpha I + N)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - N)\right)(\mathbf{u} - \mathbf{u}^{(k)}).$$

The *spectral radius* $\rho\left((\alpha I + N)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - N)\right)$ will equal $\rho\left((\alpha I - H)(\alpha I + H)^{-1}(\alpha I - N)(\alpha I + N)^{-1}\right)$ by *similarity*, yielding:

$$\begin{aligned} &\rho\left((\alpha I + N)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - N)\right) \\ &\leq \|\,(\alpha I - H)\,(\alpha I + H)^{-1}\,\|\|\,(\alpha I - N)\,(\alpha I + N)^{-1}\,\|. \end{aligned}$$

Since N is skew-symmetric, it holds that $\|\,(\alpha I - N)\,(\alpha I + N)^{-1}\,\| \leq 1$ for any $\alpha \in \mathbb{R}$. Parameter α may be chosen to minimize $\|\,(\alpha I - H)\,(\alpha I + H)^{-1}\,\|$ and will depend on $\kappa(H)$, see [BA12], and the algorithm will converge robustly. GMRES acceleration will not be necessary, however, efficient solvers will be required for $(\alpha\, I + H)$ and $(\alpha\, I + N)$.

8.3.2 Schwarz Preconditioners

In the *advection dominated* case, multisubdomain Schwarz algorithms can be formulated as in the diffusion dominated case. However, unless the coarse grid size is sufficiently small, the energy norm convergence factor may not be independent of ε, h and the subdomain size h_0, [CA, CA20, XU8, CA21, WA3]. Thus, when h_0 is small, coarse space correction can result in a subproblem of a large size. Importantly, if $c(x) \geq c_0 > 0$, coarse space correction may sometimes be omitted. For instance, if the discretization satisfies a discrete maximum principle, the maximum norm convergence factor of the sequential and restricted Schwarz algorithms can be shown to be independent of h and ϵ, but dependent on h_0 (the size of the subdomains) and the amount of overlap, see [LI6, LI7, MA33] and Chap. 15. In this case, a uniform convergence factor can be obtained if we either choose a small number of subdomains, or increase the overlap between subdomains, resulting in large subproblems.

Our discussion will focus on *two subdomain* algorithms motivated by singular perturbation methodology [GL13, GA8, GA9, AS2, GA10, MA33, GA12]. Such methods will have the disadvantage of involving large subproblems, but can yield a uniform convergence factor in h and ε. We shall consider regions Ω_1 and Ω_2 that form a nonoverlapping decomposition of Ω so that (8.29) is satisfied. Define $\Omega_1^* \equiv \Omega_1$ and $\Omega_2^* \supset \Omega_2$ as extended subdomains:

$$\varepsilon \, |\Delta u(x)| \ll |\mathbf{b}(x) \cdot \nabla u(x) + c(x)\, u(x)| \,, \qquad \text{for } x \in \Omega_1^*. \tag{8.31}$$

If $u_l(x) \equiv u(x)$ in Ω_l^* denotes the local solution on Ω_l^* for $l = 1, 2$, then $u_1(x)$ and $u_2(x)$ will solve the following hybrid formulation:

$$\begin{cases} Lu_1 = f(x), & \text{in } \Omega_1^* \\ u_1 = u_2, & \text{on } B^{(1)} \\ u_1 = 0, & \text{on } B_{[1]}, \end{cases} \quad \text{and} \quad \begin{cases} Lu_2 = f(x), & \text{in } \Omega_2^* \\ u_2 = u_1, & \text{on } B^{(2)} \\ u_2 = 0, & \text{on } B_{[2]}, \end{cases} \tag{8.32}$$

where $B^{(l)} \equiv \partial\Omega_l^* \cap \Omega$ and $B_{[l]} \equiv \partial\Omega_l^* \cap \partial\Omega$ and $Lu \equiv -\varepsilon\, \Delta u + \mathbf{b}(x) \cdot \nabla u + c(x) u$. If $u_l(x) \to w_l(x)$ as $\epsilon \to 0^+$, heuristically, $w_1(x)$ and $w_2(x)$ will satisfy:

$$\begin{cases} L_0 w_1 = f(x), & \text{in } \Omega_1^* \\ w_1 = w_2, & \text{on } B_{\text{in}}^{(1)} \\ w_1 = 0, & \text{on } B_{[1],in}, \end{cases} \quad \text{and} \quad \begin{cases} L_0 u_2 = f(x), & \text{in } \Omega_2^* \\ w_2 = w_1, & \text{on } B_{\text{in}}^{(2)} \\ w_2 = 0, & \text{on } B_{[2],in}, \end{cases} \tag{8.33}$$

provided (8.31) holds on $B^{(l)}$ for $l = 1, 2$, where $L_0 w \equiv \mathbf{b}(x) \cdot \nabla w + c(x) w$. Here, the *inflow* boundary segments are defined as:

$$\begin{aligned} B_{in}^{(l)} &\equiv \{ x \in \partial\Omega_1^* \cap \Omega \, : \, \mathbf{n}_l(x) \cdot \mathbf{b}(x) < 0 \} \,, \\ B_{[l],in} &\equiv \{ x \in \partial\Omega_1^* \cap \partial\Omega \, : \, \mathbf{n}_l(x) \cdot \mathbf{b}(x) < 0 \} \end{aligned}$$

where $\mathbf{n}_l(x)$ denotes the unit exterior normal to $\partial\Omega_l^*$.

When $c(x) \geq c_0 > 0$ and the discretization of (8.1) satisfies a discrete maximum principle, yielding an M-matrix A, the unaccelerated multiplicative Schwarz algorithm 8.2.3 and the unaccelerated restricted Schwarz algorithm 8.2.1 can be employed to solve $A\mathbf{u} = \mathbf{f}$ without *acceleration*, provided the overlap between the subdomains is sufficiently large. For instance, the two-subdomain unaccelerated restricted Schwarz algorithm will be:

Algorithm 8.3.1 *(Unaccelerated Restricted Schwarz Algorithm)*
Input: $\mathbf{w}^{(0)}$ *and* $\mathbf{f}$

```
1. For $k = 0, \cdots$ until convergence do:
       $$\mathbf{w}^{(k+1)} = \mathbf{w}^{(l)} + \sum_{i=1}^{2} R_i^T D^{(i)} A_i^{-1} R_i \left( \mathbf{f} - A\mathbf{w}^{(k)} \right)$$
2. Endfor
```

Output: $\mathbf{w}^{(k)}$

where $R_1^T D^{(1)} R_1 + R_2^T D^{(2)} R_2 = I$ forms a discrete of unity for $\overline{\Omega}_1$ and $\overline{\Omega}_2$. If GMRES acceleration is employed, then a preconditioner can be formulated corresponding to one sweep of the unaccelerated algorithm with a trivial starting iterate. The following convergence result will hold for two subdomain decompositions. Importantly, assumption (8.31) does not need to hold.

Lemma 8.15. *Suppose the following assumptions hold.*

1. *Let Ω_1^* and Ω_2^* form an overlapping decomposition of Ω with overlap β.*
2. *Let the discretization of (8.1) satisfy a discrete maximum principle, so that matrix A is an M-matrix.*

Then, the following results will hold.

1. *If $c(x) \geq c_0 > 0$ and h is sufficiently small, the iterates will satisfy:*
$$\|\mathbf{u} - \mathbf{u}^{(k)}\|_\infty \leq \rho^k \|\mathbf{u} - \mathbf{u}^{(0)}\|_\infty,$$
for algorithms 8.2.3 and 8.2.1 with $\rho \leq \rho_0 < 1$ independent of ε and h.
2. *If $c(x) \geq 0$ and h is sufficiently small, and if $\mathbf{b}(x) \cdot \mathbf{n}_1(x) \leq -b_0 < 0$ on $B_{in}^{(1)}$ for some $b_0 > 0$, then the iterates of algorithm 8.2.3 will satisfy:*
$$\|\mathbf{u} - \mathbf{u}^{(k)}\|_\infty \leq \rho^k \|\mathbf{u} - \mathbf{u}^{(0)}\|_\infty,$$
for $\rho \leq \rho_0 = e^{-\frac{C}{\varepsilon}}$ with $C > 0$ independent of h and ε.

Proof. See [LI6, LI7, GA9, GA12, MA33] and Chap. 15.

Remark 8.16. A disadvantage of the *two subdomain* Schwarz algorithm is that it can be computationally expensive to implement due to the large size of the submatrices A_1 and A_2. However, if assumption 8.31 holds, then Ω_2^* will be of width $O(\epsilon)$. If h_l denotes the local grid size in Ω_l^*, then we may choose $h_2 \ll h_1$. So, the size of matrix A_1 will be reduced. Additionally, since the layer region Ω_2^* will be of width $O(\varepsilon)$, it may be possible to reorder the unknowns so that an efficient *band solver* can be employed for matrix A_2.

Remark 8.17. When assumption (8.31) holds, a *heterogeneous* Schwarz preconditioner can be constructed as follows. If the amount of overlap between subdomains Ω_1^* and Ω_2^* is *minimal*, let I_1 denote the indices of interior nodes in $\Omega_1^* \equiv \Omega_1$ and I_2 the indices of all remaining nodes in Ω, so that $I_1 \cup I_2$ forms a partition of all interior nodes. Then, matrix A can be block partitioned as:

$$\begin{bmatrix} A_{I_1I_1} & A_{I_1I_2} \\ A_{I_2I_1} & A_{I_2I_2} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{I_1} \\ \mathbf{u}_{I_2} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{I_1} \\ \mathbf{f}_{I_2} \end{bmatrix}, \tag{8.34}$$

where $\mathbf{u}_{I_l}$ and $\mathbf{f}_{I_l}$ denote nodal vectors corresponding to indices in I_l. Use the splitting $A = \varepsilon\, H_0 + H_1 + N$ and define $\tilde{H} = H_0$ and $\tilde{N} = H_1 + N$, so that:

$$\begin{bmatrix} \varepsilon\, \tilde{H}_{I_1I_1} + \tilde{N}_{I_1I_1} & \varepsilon\, \tilde{H}_{I_1I_2} + \tilde{N}_{I_1I_2} \\ \varepsilon\, \tilde{H}_{I_2I_1} + \tilde{N}_{I_2I_1} & \varepsilon\, \tilde{H}_{I_2I_2} + \tilde{N}_{I_2I_2} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{I_1} \\ \mathbf{u}_{I_2} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{I_1} \\ \mathbf{f}_{I_2} \end{bmatrix}. \tag{8.35}$$

If $\mathbf{u}^*$ denotes the nodal vector obtained by restricting the exact solution of (8.28) to the nodes in Ω_h, then by the choice of Ω_1, it will hold that $\varepsilon|\tilde{H}_{I_1I_1}\mathbf{u}^*_{I_1}| \ll |\tilde{N}_{I_1I_1}\mathbf{u}^*_{I_1}|$ component wise. We shall thus obtain:

$$\begin{bmatrix} \tilde{N}_{I_1I_1} & A_{I_1I_2} \\ A_{I_2I_1} & A_{I_2I_2} \end{bmatrix} \begin{bmatrix} \mathbf{u}^*_{I_1} \\ \mathbf{u}^*_{I_2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{f}_{I_1} - \varepsilon\, \tilde{H}_{I_1I_1}\mathbf{u}^*_{I_1} \\ \mathbf{f}_{I_2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{f}_{I_1} \\ \mathbf{f}_{I_2} \end{bmatrix}. \tag{8.36}$$

If $h_1 \gg h_2$, we may define a heterogeneous Schwarz preconditioner M for A corresponding to a modified block Gauss-Seidel preconditioner:

$$M = \begin{bmatrix} \tilde{N}_{I_1I_1} & 0 \\ A_{I_2I_1} & A_{I_2I_2} \end{bmatrix}. \tag{8.37}$$

This preconditioner is a special case of the *asymptotically motivated* domain decomposition preconditioner [AS2]. If the heterogeneous discretization (8.36) is stable when $\varepsilon \to 0^+$, the error arising from the omission of the term $\varepsilon\, H_{I_1I_1}\mathbf{u}^*_{I_1}$ will be bounded by $O(\varepsilon)$, and the discretization error will be bounded by the sum of the original truncation error and the magnitude of the omitted term (which will be $O(\varepsilon)$), see Chap. 12. Ideally, preconditioner M in (8.37) will be better suited to precondition the *heterogeneous* linear system (8.36), since the spectral properties of $\tilde{N}_{I_1I_1}$ and $A_{I_1I_1} = \varepsilon\, \tilde{H}_{I_1I_1} + \tilde{N}_{I_1I_1}$ may differ significantly, when the local mesh size h_1 in Ω_1^* is sufficiently small, even though $A_{I_1I_1}\mathbf{u}^*_{I_1} \approx \tilde{N}_{I_1I_1}\mathbf{u}^*_{I_1}$.

8.3.3 Schur Complement Preconditioners

In the *advection dominated* case also, multisubdomain Schur complement algorithms can be formulated. However, due to the hyperbolic character of the limiting advection equation, each subproblem will be sensitive to the local *inflow* direction of $\mathbf{b}(x)$ and the convergence factor will deteriorate as the subdomain size h_0 decreases, unless the coarse grid size is sufficiently small (yielding a large subproblem). We will focus on *two subdomain* algorithms without coarse space correction, motivated by singular perturbation methodology [GA15, GA14, NA5, AC7, AL4, QU6, RA3, AU]. Such algorithms will involve large subproblems, but can yield convergence uniform in h and ε.

The two subdomain Schur complement algorithms that we describe will be generalizations of the Dirichlet-Neumann algorithm 1.3.1, motivated by *inflow* transmission conditions of the limiting hyperbolic equation. Let Ω_1 and Ω_2 form a non-overlapping decomposition of Ω, such that $B = \partial\Omega_1 \cap \partial\Omega_2$ does not lie in the layer region. Let $u_l(x) \equiv u(x)$ on Ω_l for $l = 1, 2$. Then, $u_1(x)$ and $u_2(x)$ will solve the following hybrid formulation for $\varepsilon > 0$:

$$\begin{cases} L\, u_1 = f(x), & \text{in } \Omega_1 \\ u_1 = 0, & \text{on } B_{[1]} \\ \mathbf{n}_1 \cdot (\varepsilon\, \nabla u_1) = \mathbf{n}_1 \cdot (\varepsilon \nabla u_2), & \text{on } B \end{cases} \quad \text{and} \quad \begin{cases} L\, u_2 = f(x), & \text{in } \Omega_2 \\ u_2 = 0, & \text{on } B_{[2]} \\ u_2 = u_1, & \text{on } B \end{cases} \tag{8.38}$$

where $B = \partial\Omega_1 \cap \partial\Omega_2$ and $B_{[l]} = \partial\Omega_l \cap \partial\Omega$, and $\mathbf{n}_1(x)$ is the unit exterior normal to $\partial\Omega_1$. Alternative transmission conditions will be described.

We employ a stable discretization of (8.28) based on the bilinear form $\mathcal{A}^0(.,.)$, see [JO2, FR3, FR2], and let $\mathbf{u} = (\mathbf{u}_I^{(1)^T}, \mathbf{u}_I^{(2)^T}, \mathbf{u}_B^T)^T$ denote a block partition of the vector $\mathbf{u}$ of nodal unknowns corresponding to nodes on Ω_1, Ω_2 and B, respectively. The resulting linear system will be denoted as:

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)} \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ A_{BI}^{(1)} & A_{BI}^{(2)} & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B \end{bmatrix}. \tag{8.39}$$

Then, the Schur complement system will have the form:

$$\begin{cases} S\mathbf{u}_B = \tilde{\mathbf{f}}_B, \qquad \text{where} \\ S = A_{BB} - A_{BI}^{(1)} A_{II}^{(1)^{-1}} A_{IB}^{(1)} - A_{BI}^{(2)} A_{II}^{(2)^{-1}} A_{IB}^{(2)} \\ \tilde{\mathbf{f}}_B = \mathbf{f}_B - A_{BI}^{(1)} A_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} - A_{BI}^{(2)} A_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)}. \end{cases}$$

We may further decompose $S = S^{(1)} + S^{(2)}$, where $S^{(l)} = A_{BB}^{(l)} - A_{BI}^{(l)} A_{II}^{(l)^{-1}} A_{IB}^{(l)}$, for $l = 1, 2$, using $A_{BB} = A_{BB}^{(1)} + A_{BB}^{(2)}$. In the following, we describe different unaccelerated nonsymmetric generalizations of the Dirichlet-Neumann algorithm, based on modifications of the matrix $S^{(1)}$ or $S^{(2)}$.

Adaptive Robin-Neumann and Related Algorithms. The effectiveness of the Dirichlet-Neumann algorithm 1.3.1 can deteriorate as $\varepsilon \to 0^+$, see [CA31, CI9, TR2, GA14]. To *heuristically* understand this, consider the following *inflow*, *null flow* and *outflow* segments on $B = \partial\Omega_1 \cap \partial\Omega_2$, see Fig. 8.1:

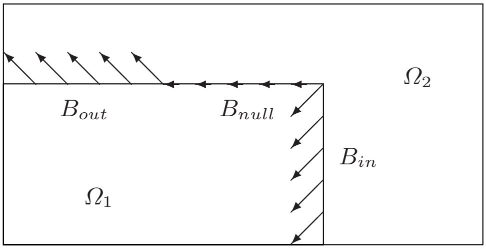


Fig. 8.1. Inflow, outflow and nullflow segments of B

$$\begin{aligned} B_{in} &\equiv \{x \in B \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0\} \\ B_{null} &\equiv \{x \in B \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) = 0\} \\ B_{out} &\equiv \{x \in B \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) > 0\}\,, \end{aligned}$$

where $\mathbf{n}_1(x)$ denotes the unit exterior normal to Ω_1 for $x \in B$. When $\varepsilon \to 0^+$, hybrid formulation (8.38) formally reduces to the following coupled system of hyperbolic equations, see [GA15, GA14, AC7, AL4, QU6, RA3]:

$$\begin{cases} L_0\, w_1 = f, & \text{in } \Omega_1 \\ \quad w_1 = w_2, & \text{on } B_{in} \\ \quad w_1 = 0, & \text{on } B_{[1],in} \end{cases} \quad \text{and} \quad \begin{cases} L_0\, w_2 = f, & \text{in } \Omega_2 \\ \quad w_2 = w_1, & \text{on } B_{out} \\ \quad w_2 = 0, & \text{on } B_{[2],in}, \end{cases}$$

with local inflow boundary conditions. Here $L_0\, w \equiv \mathbf{b} \cdot \nabla w + c\, w$ and:

$$\begin{aligned} B_{[1],in} &\equiv \left\{x \in B_{[1]} \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0\right\} \\ B_{[2],in} &= \left\{x \in B_{[2]} \,:\, \mathbf{n}_2(x) \cdot \mathbf{b}(x) < 0\right\}. \end{aligned}$$

As $\varepsilon \to 0^+$, a Neumann condition $\varepsilon\, \mathbf{n}_1 \cdot \nabla u_1 = \varepsilon\, \mathbf{n}_1 \cdot \nabla u_2$ on B_{in} does not reduce to an inflow condition $u_1 = u_2$ on B_{in}. However, a Robin condition $\varepsilon\, \mathbf{n}_1(x) \cdot \nabla u_1 + \gamma(x)\, u_1(x) = \varepsilon\, \mathbf{n}_1(x) \cdot \nabla u_2 + \gamma(x)\, u_2(x)$ formally reduces to $u_1(x) = u_2(x)$ on B_{in} as $\varepsilon \to 0^+$, provided $\gamma(x) \neq 0$ on B_{in}. A Dirichlet condition on B_{in} will also reduce to an inflow condition. These observation motivate different versions of *adaptive* algorithms [CA31, CI9, TR2, GA14]. Such adaptive algorithms employ an *equivalent* pair of transmission conditions on $B_{in} \cup B_{null}$ and B_{out}, from the list below, requiring that one of them formally reduces to an inflow condition on B_{in}:

$$\begin{cases} \text{Dirichlet:} & u_1 = u_2, & \text{on } B \\ \text{Neumann:} & \mathbf{n}_1 \cdot (\varepsilon \nabla u_1) = \mathbf{n}_1 \cdot (\varepsilon \nabla u_2), & \text{on } B \\ \text{Robin:} & \mathbf{n}_1 \cdot (\varepsilon \nabla u_1 - \mathbf{b}\, u_1) = \mathbf{n}_1 \cdot (\varepsilon \nabla u_2 - \mathbf{b}\, u_2), & \text{on } B \\ \gamma\text{-Robin:} & \mathbf{n}_1 \cdot (\varepsilon \nabla u_1) + \gamma(x)\, u_1 = \mathbf{n}_1 \cdot (\varepsilon \nabla u_2) + \gamma(x)\, u_2, & \text{on } B. \end{cases}$$

Each *adaptive* algorithm has a similar structure. The interface B is partitioned into $B_{in} \cap B_{null}$ and B_{out}, based on the *inflow* and *outflow* segments of B, relative to Ω_1, for the advection field $\mathbf{b}(x)$. A pair of equivalent (complementary) transmission conditions are chosen on B, such as Dirichlet and Neumann, or Neumann and Robin. As $\varepsilon \to 0^+$, one of the transmission conditions is required to reduce to an inflow condition on B_{in} when solving a subproblem on Ω_1. The complementary transmission condition is required to reduce to an inflow condition on B_{out} when solving on Ω_2. For instance, in the *adaptive Dirichlet-Neumann* (ADN) algorithm, when solving on Ω_1, Dirichlet conditions are imposed on $B_{in} \cap B_{null}$ and Neumann conditions are imposed on B_{out}. This ensures that inflow conditions are imposed in the limiting advection equations on Ω_1 as $\varepsilon \to 0^+$. When solving on Ω_2, Neumann conditions are imposed on $B_{in} \cap B_{null}$ and Dirichlet conditions are imposed on B_{out}, which ensures that inflow conditions are imposed on B_{out} (the inflow segment for Ω_2) as $\varepsilon \to 0^+$. Since Dirichlet and Neumann conditions are complementary, together they are equivalent to the original transmission conditions. Below, we list the continuous version of the ADN algorithm to update $w_1^{(k)}$ and $w_2^{(k)}$ as follows. Let $\psi_D(w) \equiv w$ denotes a Dirichlet boundary operator on B and $\psi_N^{(l)}(w) \equiv \varepsilon\, (\mathbf{n}_l \cdot \nabla w)$ a Neumann boundary operator on B, where $\mathbf{n}_l$ denotes the unit exterior normal to Ω_l. Let $0 < \theta < 1$ and $0 < \delta < 1$.

Algorithm 8.3.2 *(Adaptive Dirichlet-Neumann Algorithm)*
Input: $w_1^{(0)}$ *and* $w_2^{(0)}$

```
1. For $k = 0, \cdots$ until convergence do:
2.     Define $\lambda^{(k)} \equiv \theta\, \psi_D(w_1^{(k)}) + (1-\theta)\, \psi_D(w_2^{(k)})$ on $B_{in} \cup B_{null}$ and solve:
       $$\begin{cases} L w_1^{(k+1)} = f, & \text{in } \Omega_1 \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \\ \psi_D(w_1^{(k+1)}) = \lambda^{(k)}, & \text{on } B_{in} \cup B_{null} \\ \psi_N^{(1)}(w_1^{(k+1)}) = \psi_N^{(1)}(w_2^{(k)}), & \text{on } B_{out}, \end{cases}$$
3.     Define $\mu^{(k+1)} \equiv \delta\, \psi_D(w_1^{(k+1)}) + (1-\delta)\, \psi_D(w_2^{(k)})$ on $B_{out}$ and solve:
       $$\begin{cases} L w_2^{(k+1)} = f, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \psi_D(w_2^{(k+1)}) = \mu^{(k+1)}, & \text{on } B_{out} \\ \psi_N^{(2)}(w_2^{(k+1)}) = \psi_N^{(2)}(w_1^{(k+1)}), & \text{on } B_{in} \cup B_{null}, \end{cases}$$
4. Endfor
5. Output: $w_1^{(k)}$ and $w_2^{(k)}$
```

By construction, if $(w_1^{(k)}, w_2^{(k)})$ converges, its limit (w_1, w_2) will satisfy the original transmission conditions on B and solve hybrid formulation (8.38).

As mentioned earlier, the Dirichlet and Neumann conditions in the *ADN* algorithm can be replaced by other *complementary* transmission conditions, provided the operator replacing $\psi_D(.)$ reduces to a (weighted) Dirichlet condition on $B_{in} \cup B_{null}$, as $\varepsilon \to 0^+$. In the *adaptive* Robin-Neumann (ARN) algorithm, $\psi_D(.)$ is replaced by a Robin operator $\psi_R^{(l)}(w) \equiv \varepsilon\, \mathbf{n}_l \cdot (\nabla w - \mathbf{b}\, w)$ on Ω_l, while $\psi_N^{(l)}(.)$ is used as before. Robin and Neumann conditions will be complementary on B, provided $\mathbf{n}_l(x) \cdot \mathbf{b}(x) \neq 0$ for $x \in B$. This requirement also ensures that $\psi_R(w_1) = \psi_R(w_2)$ formally reduces to $w_1 = w_2$, as $\varepsilon \to 0^+$. In the *adaptive* γ-Robin-Neumann ($AR_\gamma N$) algorithm, $\psi_D(.)$ is replaced by $\psi_{R_\gamma}^{(l)}(w) \equiv \varepsilon\, (\mathbf{n}_l \cdot \nabla w) + \gamma\, w$ when solving on Ω_l. Again, $\psi_N(.)$ and $\psi_{R_\gamma}(.)$ will be complementary on B, if $\gamma(x) \neq 0$ on B. The γ-Robin map $\psi_{R_\gamma}^{(l)}(.)$ reduces to the Robin map $\psi_R(.)$ for the choice $\gamma(x) = -\mathbf{n}_l(x) \cdot \mathbf{b}(x)$.

Below, we indicate a matrix version of the $AR_\gamma N$ algorithm. Let m denote the number of nodal unknowns on B, with $m = (m_1 + m_2)$, where m_1 denotes the number of nodal unknowns on $B_{in} \cup B_{null}$ and m_2 denotes the number of nodal unknowns on B_{out}. We shall assume that the nodes in $B_{in} \cup B_{out}$ are ordered prior to the nodes in B_{out}. Let $\{\phi_1(.), \ldots, \phi_m(.)\}$ denote a finite element nodal basis on B. Then, we define the following mass matrices on B:

$$\begin{aligned} \left(M_{\gamma,BB}^{(1)}\right)_{ij} &= \int_{B_{in} \cup B_{null}} \gamma(x)\, \phi_i(x)\, \phi_j(x)\, ds_x \\ \left(M_{\gamma,BB}^{(2)}\right)_{ij} &= \int_{B_{out}} \gamma(x)\, \phi_i(x)\, \phi_j(x)\, ds_x. \end{aligned} \tag{8.40}$$

Let $\mathcal{R}_1$ denote an $m_1 \times m$ matrix *restriction* map which maps a nodal vector on B into nodal values on $B_{in} \cup B_{null}$. Its transpose $\mathcal{R}_1^T$ will extend nodal values on $B_{in} \cup B_{null}$ by zero to B. Similarly, we let $\mathcal{R}_2$ denote the $m_2 \times m$ *restriction* map of nodal values on B into nodal values on B_{out}. Its transpose $\mathcal{R}_2^T$ will extend nodal values on B_{out} by zero to B. We shall let $\mathbf{v}^{(k)}$ and $\mathbf{w}^{(k)}$ denote the discrete nodal vectors corresponding to $w_1^{(k)}$ and $w_2^{(k)}$, with $\mathbf{v}^{(k)} = \left(\mathbf{v}_I^{(k)^T}, \mathbf{v}_B^{(k)^T}\right)^T$ and $\mathbf{w}^{(k)} = \left(\mathbf{w}_I^{(k)^T}, \mathbf{w}_B^{(k)^T}\right)^T$. The matrix implementation of step 2 in the $AR_\gamma N$ algorithm to solve (8.39) will have the form:

$$\begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} \\ A_{BI}^{(1)} & M_{\gamma,BB}^{(1)} + A_{BB}^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(k+1)} \\ \mathbf{v}_B^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B + \mathbf{r}_B^{(k+1)} \end{bmatrix},$$

where the components $\mathcal{R}_1 \mathbf{r}_B^{(k+1)}$ and $\mathcal{R}_2 \mathbf{r}_B^{(k+1)}$ are chosen as:

$$\begin{cases} \mathcal{R}_1 \mathbf{r}_B^{(k+1)} = & \theta\, \mathcal{R}_1 \left(A_{BI}^{(1)} \mathbf{v}_I^{(k)} + (M_{\gamma,BB}^{(1)} + A_{BB}^{(1)}) \mathbf{v}_B^{(k)} - \mathbf{f}_B \right) \\ & + (1-\theta)\, \mathcal{R}_1 \left(-A_{BI}^{(2)} \mathbf{w}_I^{(k)} + (M_{\gamma,BB}^{(1)} - A_{BB}^{(2)}) \mathbf{w}_B^{(k)} \right) \\ \mathcal{R}_2 \mathbf{r}_B^{(k+1)} = & \mathcal{R}_2 \left(-A_{BI}^{(2)} \mathbf{w}_I^{(k)} + (M_{\gamma,BB}^{(1)} - A_{BB}^{(2)}) \mathbf{w}_B^{(k)} \right). \end{cases}$$

Our choice of the forcing terms $\mathbf{r}_B^{(k+1)}$ and $\mathbf{z}_B^{(k+1)}$ yield equations consistent with the original system. Step 3 can be implemented analogously.

Below, we list the matrix implementation of the $AR_\gamma N$ algorithm.

Algorithm 8.3.3 *(Adaptive γ-Robin-Neumann Algorithm)*
Input: $\mathbf{v}^{(0)}$, $\mathbf{w}^{(0)}$, $0 < \theta < 1$ *and* $0 < \delta < 1$

```
1. For $k = 0, \cdots$ until convergence do:
2.     Define the components $\mathcal{R}_1 \mathbf{r}_B^{(k+1)}$ and $\mathcal{R}_2 \mathbf{r}_B^{(k+1)}$ of $\mathbf{r}_B^{(k+1)}$ as follows:
       $\begin{cases} \mathcal{R}_1 \mathbf{r}_B^{(k+1)} = & \theta \, \mathcal{R}_1 \left( A_{BI}^{(1)} \mathbf{v}_I^{(k)} + (M_{\gamma,BB}^{(1)} + A_{BB}^{(1)}) \mathbf{v}_B^{(k)} - \mathbf{f}_B \right) \\ & + (1-\theta) \, \mathcal{R}_1 \left( -A_{BI}^{(2)} \mathbf{w}_I^{(k)} + (M_{\gamma,BB}^{(1)} - A_{BB}^{(2)}) \mathbf{w}_B^{(k)} \right) \\ \mathcal{R}_2 \mathbf{r}_B^{(k+1)} = & \mathcal{R}_2 \left( -A_{BI}^{(2)} \mathbf{w}_I^{(k)} + (M_{\gamma,BB}^{(1)} - A_{BB}^{(2)}) \mathbf{w}_B^{(k)} \right). \end{cases}$
3.     Solve:
       $\begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} \\ A_{BI}^{(1)} & M_{\gamma,BB}^{(1)} + A_{BB}^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(k+1)} \\ \mathbf{v}_B^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B + \mathbf{r}_B^{(k+1)} \end{bmatrix}$
4.     Define the components $\mathcal{R}_1 \mathbf{z}_B^{(k+1)}$ and $\mathcal{R}_2 \mathbf{z}_B^{(k+1)}$ of $\mathbf{z}_B^{(k+1)}$ as follows:
       $\begin{cases} \mathcal{R}_1 \mathbf{z}_B^{(k+1)} = & \mathcal{R}_1 \left( -A_{BI}^{(1)} \mathbf{v}_I^{(k+1)} + (M_{\gamma,BB}^{(2)} - A_{BB}^{(1)}) \mathbf{v}_B^{(k+1)} \right) \\ \mathcal{R}_2 \mathbf{z}_B^{(k+1)} = & \delta \, \mathcal{R}_2 \left( A_{BI}^{(2)} \mathbf{w}_I^{(k)} + (M_{\gamma,BB}^{(2)} + A_{BB}^{(2)}) \mathbf{w}_B^{(k)} - \mathbf{f}_B \right) \\ & + (1-\delta) \, \mathcal{R}_2 \left( -A_{BI}^{(1)} \mathbf{v}_I^{(k+1)} + (M_{\gamma,BB}^{(2)} - A_{BB}^{(1)}) \mathbf{v}_B^{(k+1)} \right). \end{cases}$
5.     Solve:
       $\begin{bmatrix} A_{II}^{(2)} & A_{IB}^{(2)} \\ A_{BI}^{(2)} & M_{\gamma,BB}^{(2)} + A_{BB}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(k+1)} \\ \mathbf{w}_B^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(2)} \\ \mathbf{f}_B + \mathbf{z}_B^{(k+1)} \end{bmatrix}$
6. Endfor
7. Output: $\mathbf{v}^{(k+1)}$ and $\mathbf{w}^{(k+1)}$
```

Remark 8.18. A *parallel* version of the above *sequential* algorithm can be obtained by replacing $\mathbf{v}_I^{(k+1)}$ and $\mathbf{v}_B^{(k+1)}$ by $\mathbf{v}_I^{(k)}$ and $\mathbf{v}_B^{(k)}$, respectively, in step 4. As noted in remark 8.19, the steps involved in updating $\mathbf{v}_B^{(k+1)}$ and $\mathbf{w}_B^{(k+1)}$ can be viewed as an unaccelerated iterative method to solve the Schur complement system $(S^{(1)} + S^{(2)})\mathbf{u}_B = \tilde{\mathbf{f}}_B$. When matrix $\left(S^{(l)} + M_{\gamma,BB}^{(l)} \right)$ is positive definite for $l = 1, 2$, the preceding algorithm can be shown to be convergent for a suitable choice of parameters $0 < \theta < 1$ and $0 < \delta < 1$, see [GA14, QU6]. Under additional assumptions, the convergence will be uniform in h and ε. If GMRES acceleration is employed, a preconditioner can be formulated for S by applying one iteration of the $AR_\gamma N$ algorithm with *zero* starting guess.

Remark 8.19. When $\mathbf{f}_I^{(1)} = \mathbf{0}$, $\mathbf{f}_I^{(2)} = \mathbf{0}$, the $AR_\gamma N$ algorithm will correspond to a "modified" *block Gauss-Seidel* algorithm to solve the extended system:

$$\begin{cases} \left(S^{(1)} + M_{\gamma,BB}^{(1)}\right) \mathbf{v}_B + \left(S^{(2)} - M_{\gamma,BB}^{(1)}\right) \mathbf{w}_B = \tilde{\mathbf{f}}_B \\ \left(S^{(1)} - M_{\gamma,BB}^{(2)}\right) \mathbf{v}_B + \left(S^{(2)} + M_{\gamma,BB}^{(2)}\right) \mathbf{w}_B = \tilde{\mathbf{f}}_B, \end{cases}$$

with *partial relaxation* (applied only to the components $\mathcal{R}_1 \mathbf{v}_B$ and $\mathcal{R}_2 \mathbf{w}_B$). When $\gamma(x) \neq 0$ on B, the above system will yield $\mathbf{v}_B = \mathbf{w}_B$, where $\mathbf{v}_B$ solves the original Schur complement system $(S^{(1)} + S^{(2)})\mathbf{u}_B = \tilde{\mathbf{f}}_B$. Furthermore, when $\mathbf{f}_I^{(1)} = \mathbf{0}$ and $\mathbf{f}_I^{(2)} = \mathbf{0}$, the different steps in the $AR_\gamma N$ algorithm can be expressed in terms of the Schur complement matrices as follows:

$$\begin{array}{ll} (S^{(1)} + M_{\gamma,BB}^{(1)})\mathbf{v}_B^{(k+1)} = (\mathbf{f}_B + \mathbf{r}_B^{(k+1)}) & \text{in step 3} \\ (S^{(2)} + M_{\gamma,BB}^{(2)})\mathbf{w}_B^{(k+1)} = (\mathbf{f}_B + \mathbf{z}_B^{(k+1)}) & \text{in step 5.} \end{array}$$

The vector $\mathbf{r}_B^{(k+1)}$ in step 2 and $\mathbf{z}_B^{(k+1)}$ in step 4 satisfy:

$$\begin{aligned} \mathcal{R}_1 \mathbf{r}_B^{(k+1)} &= \mathcal{R}_1 \left(\theta\, (S^{(1)} + M_{\gamma,BB}^{(1)})\mathbf{v}_B^{(k)} + (1-\theta)\, (-S^{(2)} + M_{\gamma,BB}^{(1)})\, \mathbf{w}_B^{(k)} - \theta\, \mathbf{f}_B\right) \\ \mathcal{R}_2 \mathbf{r}_B^{(k+1)} &= \mathcal{R}_2 (M_{\gamma,BB}^{(1)} - S^{(2)})\, \mathbf{w}_B^{(k)} \\ \mathcal{R}_1 \mathbf{z}_B^{(k+1)} &= \mathcal{R}_1 (M_{\gamma,BB}^{(2)} - S^{(1)})\mathbf{v}_B^{(k+1)} \\ \mathcal{R}_2 \mathbf{z}_B^{(k+1)} &= \mathcal{R}_2 \left(\delta\, (M_{\gamma,BB}^{(2)} + S^{(2)})\mathbf{w}_B^{(k)} + (1-\delta)\, (M_{\gamma,BB}^{(2)} - S^{(1)})\mathbf{v}_B^{(k+1)} - \delta\, \mathbf{f}_B\right). \end{aligned}$$

The original system (8.39) can be reduced to the case $\mathbf{f}_I^{(1)} = \mathbf{0}$ and $\mathbf{f}_I^{(2)} = \mathbf{0}$ as follows. Split $\mathbf{u}_I^{(l)} = \mathbf{y}_I^{(l)} + \tilde{\mathbf{u}}_I^{(l)}$ for $l = 1, 2$, where $\mathbf{y}_I^{(l)} = A_{II}^{(l)^{-1}} \mathbf{f}_I^{(l)}$ for $l = 1, 2$. Then, $\left(\tilde{\mathbf{u}}_I^{(1)^T}, \tilde{\mathbf{u}}_I^{(2)^T}, \mathbf{u}_B^T\right)^T$ will solve:

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)} \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ A_{BI}^{(1)} & A_{BI}^{(2)} & A_{BB} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}_I^{(1)} \\ \tilde{\mathbf{u}}_I^{(2)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \tilde{\mathbf{f}}_B \end{bmatrix}, \tag{8.41}$$

where $\tilde{\mathbf{f}}_B = \left(\mathbf{f}_B - A_{BI}^{(1)}\mathbf{y}_I^{(1)} - A_{BI}^{(2)}\mathbf{y}_I^{(2)}\right)$. By construction $S\,\mathbf{u}_B = \tilde{\mathbf{f}}_B$, and once $\mathbf{u}_B$ is known, we can determine $\tilde{\mathbf{u}}_I^{(l)} = -A_{II}^{(l)^{-1}} A_{IB}^{(l)}\mathbf{u}_B$ for $l = 1, 2$.

Remark 8.20. In practice, the Robin coefficient $\gamma(x)$ must be chosen to ensure that matrix $\left(S^{(l)} + M_{\gamma,BB}^{(l)}\right)$ is positive definite for $l = 1, 2$. When matrix A is positive definite, the Schur complement matrix $S = \left(S^{(1)} + S^{(2)}\right)$ is positive definite by remark 8.12, however, the subdomain Schur complement matrices $S^{(l)} = \left(A_{BB}^{(l)} - A_{BI}^{(l)} A_{II}^{(l)^{-1}} A_{IB}^{(l)}\right)$ need not be positive definite. To derive conditions on $\gamma(x)$ for coercivity of $\left(S^{(l)} + M^{(l)}_{\gamma,BB}\right)$, note that matrix $A^{(l)}$ is obtained by Galerkin discretization of $\mathcal{A}^{0}_{\Omega_l}(u,v)$, which differs from the *coercive* bilinear form $\mathcal{A}^{\frac{1}{2}}_{\Omega_l}(u,v)$ only by a boundary term:

$$\mathcal{A}^{\frac{1}{2}}_{\Omega_l}(u,v) = \mathcal{A}^{0}_{\Omega_l}(u,v) - \tfrac{1}{2}\int_B (\mathbf{n}_l(x)\cdot\mathbf{b}(x)\,u)\; v\,ds_x, \quad \forall u,\, v \in H^1_0(\Omega), \tag{8.42}$$

where $\mathcal{A}^{0}_{\Omega_l}(u,v)$, and $\mathcal{A}^{\frac{1}{2}}_{\Omega_l}(u,v)$ are defined by:

$$\begin{cases} \mathcal{A}^{0}(u,v) = \int_{\Omega_l} (\varepsilon\,\nabla\cdot\nabla v + \mathbf{b}(x)\cdot\nabla u\, v + c(x)\,u\,v)\,dx \\ \mathcal{A}^{\frac{1}{2}}_{\Omega_l}(u,v) = \int_{\Omega_l} \left(\varepsilon\,\nabla u\cdot\nabla v + (c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x))\,u\,v\right)dx \\ \qquad\qquad\quad +\frac{1}{2}\int_{\Omega_l} ((\mathbf{b}(x)\cdot\nabla u)\,v - u\,(\mathbf{b}(x)\cdot\nabla v))\;dx. \end{cases}$$

If $\left(\gamma(x) + \frac{1}{2}\,\mathbf{n}_l(x)\cdot\mathbf{b}(x)\right) \geq 0$ on B, employing expression (8.42) yields:

$$\begin{aligned} \left(\mathcal{A}^{0}_{\Omega_l}(u,u) + \int_B \gamma(x)\,u^2\,ds_x\right) &= \mathcal{A}^{\frac{1}{2}}_{\Omega_l}(u,u) + \int_B \left(\gamma(x) + \tfrac{1}{2}\,\mathbf{n}_l(x)\cdot\mathbf{b}(x)\right)\,u^2\,ds_x \\ &\geq \mathcal{A}^{\frac{1}{2}}_{\Omega_l}(u,u) \\ &= \int_{\Omega_l} \left(\varepsilon\,\nabla u\cdot\nabla u + (c(x) - \tfrac{1}{2}\,\nabla\cdot\mathbf{b}(x)\right)\,dx, \end{aligned}$$

which will be *coercive* when $\left(c(x) - \frac{1}{2}\,\nabla\cdot\mathbf{b}(x)\right) \geq c_0 > 0$. Importantly, coercivity of $\mathcal{A}^{0}_{\Omega_l}(u,u) + \int_B \gamma(x)\,u^2\,ds_x$ yields coercivity of $(S^{(l)} + M^{(l)}_{\gamma,BB})$, due to the relation between the Schur complement and the stiffness matrix. The choice $\gamma(x) = \frac{1}{2}\,|\mathbf{n}_l(x)\cdot\mathbf{b}(x)|$ is special, since it ensures that $\left(S^{(l)} + M^{(l)}_{\gamma,BB}\right)$ is coercive for both $l = 1, 2$.

Robin-Robin and Related Algorithms. The non-symmetric Robin-Robin algorithm and the related Dirichlet-Robin algorithm [AL4, AC7, RA3, QU6] are generalizations of the two subdomain Dirichlet-Neumann algorithm. They share similarities with adaptive algorithms and employ two complementary transmission conditions on the interface B, of which one is a *Robin* condition (as the name suggests), however, the interface B is *not* partitioned into inflow and outflow regions. The Robin boundary condition employed is a γ-Robin condition based on the boundary operator:

$$\psi^{(l)}_R(w) = \mathbf{n}_l\cdot\left(\varepsilon\nabla w - \frac{1}{2}\mathbf{b}(x)\right) + \tilde{\gamma}(x)\,w, \quad \text{on } B, \tag{8.43}$$

for some coefficient $\tilde{\gamma}(x) \geq 0$ chosen by the user, where $\mathbf{n}_l(x)$ is the exterior normal to $\partial\Omega_l$. The Robin-Robin algorithm employs two complementary γ-Robin transmission conditions on B, while the Dirichlet-Robin algorithm employs Dirichlet and γ-Robin transmission conditions on B.

Analysis of adaptive algorithms [GA14, QU6, AL4, AC7, RA3], shows that each subdomain problem must be *coercive* to ensure convergence. When the coefficients satisfy $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \geq c_0 > 0$ and $\varepsilon > 0$, the subdomain bilinear form $\mathcal{A}^{\frac{1}{2}}_{\Omega_l}(.,.)$ will be coercive for $l = 1, 2$ (see remark 8.20). Since $\psi_R(.)$ with $\tilde{\gamma}(x) = 0$ is the boundary operator associated with $\mathcal{A}^{\frac{1}{2}}(.,.)$, the coercivity of $\mathcal{A}^{\frac{1}{2}}(.,.)$ will ensure the coercivity of a subdomain Robin problem, provided $\tilde{\gamma}(x) \geq 0$. In applications, it will be assumed that $\tilde{\gamma}(x) \geq \gamma_0 > 0$.

The Dirichlet-Robin algorithm modifies the Dirichlet-Neumann by replacing the Neumann condition on B by a γ-Robin condition, for some $\tilde{\gamma}(x) \geq \gamma_0 > 0$. Given $w_1^{(k)}$ and $w_2^{(k)}$ an update $w_1^{(k+1)}$ is computed on Ω_1 by solving the Dirichlet problem:

$$\begin{cases} Lw_1^{(k+1)} = f, & \text{in } \Omega_1 \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \\ w_1^{(k+1)} = \lambda^{(k)}, & \text{on } B. \end{cases} \tag{8.44}$$

This is followed by the solution of a γ-Robin problem on Ω_2:

$$\begin{cases} Lw_2^{(k+1)} = f, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \psi_R^{(2)}(w_2^{(k+1)}) = \psi_R^{(2)}(w_1^{(k+1)}), & \text{on } B. \end{cases} \tag{8.45}$$

The Dirichlet data $\lambda^{(k)}$ is defined as follows:

$$\lambda^{(k)} \equiv \theta\, w_1^{(k)} + (1-\theta)\, w_2^{(k)}, \text{ on } B. \tag{8.46}$$

using a relaxation parameter $0 < \theta < 1$.

The Robin-Robin algorithm employs the following iteration:

$$\begin{cases} Lw_1^{(k+1)} = f, & \text{in } \Omega_1 \\ w_1^{(k+1)} = 0, & \text{on } B_{[1]} \\ \psi_R^{(1)}(w_1^{(k+1)}) = \lambda^{(k)}, & \text{on } B, \end{cases} \tag{8.47}$$

followed by the solution of:

$$\begin{cases} Lw_2^{(k+1)} = f, & \text{in } \Omega_2 \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]} \\ \psi_R^{(2)}(w_2^{(k+1)}) = \psi_R^{(2)}(w_1^{(k+1)}), & \text{on } B. \end{cases} \tag{8.48}$$

The Robin data is defined by:

$$\lambda^{(k)} \equiv \theta\, \psi_R^{(1)}(w_1^{(k)}) + (1-\theta)\psi_R^{(1)}(w_2^{(k)}), \text{ on } B, \tag{8.49}$$

for some relaxation parameter $0 < \theta < 1$. Each subproblem requires the solution of a problem with γ-Robin conditions on B. Different choices of the coefficient function $\tilde{\gamma}(x)$ yields different algorithms. For instance:

$$\begin{cases} \tilde{\gamma}(x) = \frac{1}{2}\left|\mathbf{b}(x)\cdot\mathbf{n}(x)\right|, & \textit{ARN}\text{ Alg.} \\ \tilde{\gamma}(x) = \frac{1}{2}\sqrt{\left|\mathbf{b}(x)\cdot\mathbf{n}(x)\right|^2 + 4\left(c(x) - \nabla\cdot\mathbf{b}(x)\right)\epsilon}, & \text{Alg. of [NA5]} \\ \tilde{\gamma}(x) = \frac{1}{2}\sqrt{\left|\mathbf{b}(x)\cdot\mathbf{n}(x)\right|^2 + 4\,\kappa\,\epsilon}, & \text{Alg. of [AU],} \end{cases}$$

where $\kappa > 0$ and $\varepsilon > 0$, see [QU6].

Remark 8.21. To obtain a matrix representation of the Robin-Robin algorithm with matrices A and $A^{(l)}$ obtained by discretization of $\mathcal{A}^0(.,.)$ and $\mathcal{A}^0_{\Omega_l}(.,.)$, respectively, given $\tilde{\gamma}(x)$, we define a subdomain mass matrix $G^{(l)}_{\tilde{\gamma},BB}$ as follows:

$$\left(G^{(l)}_{\tilde{\gamma},BB}\right)_{ij} = \int_B \left(\tilde{\gamma}(x) - \frac{1}{2}\mathbf{n}_l(x)\cdot\mathbf{b}(x)\right)\,\phi_i(x)\,\phi_j(x)\,ds_x, \tag{8.50}$$

where $\{\phi_1(.),\ldots,\phi_j(.)\}$ denotes the finite element nodal basis restricted to B (in an appropriate order).

Then, a matrix version of the γ-Robin-Robin algorithm is:

Algorithm 8.3.4 *(γ-Robin-Robin Algorithm)*
Input: $\mathbf{v}^{(0)}$, $\mathbf{w}^{(0)}$, $0 < \theta < 1$

```
1. For $k = 0, \cdots$ until convergence do:
2.     Define:
$$\begin{aligned} \mathbf{r}_B^{(k+1)} = \quad & \theta\left(A^{(1)}_{BI}\mathbf{v}_I^{(k)} + (G^{(1)}_{\tilde{\gamma},BB} + A^{(1)}_{BB})\mathbf{v}_B^{(k)} - \mathbf{f}_B\right) \\ & + (1-\theta)\left(-A^{(2)}_{BI}\mathbf{w}_I^{(k)} + (G^{(1)}_{\tilde{\gamma},BB} - A^{(2)}_{BB})\mathbf{w}_B^{(k)}\right) \end{aligned}$$
3.     Solve:
$$\begin{bmatrix} A^{(1)}_{II} & A^{(1)}_{IB} \\ A^{(1)}_{BI} & G^{(1)}_{\tilde{\gamma},BB} + A^{(1)}_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(k+1)} \\ \mathbf{v}_B^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B + \mathbf{r}_B^{(k+1)} \end{bmatrix}$$
4.     Define:
$$\mathbf{z}_B^{(k+1)} = -A^{(1)}_{BI}\mathbf{v}_I^{(k+1)} + (G^{(2)}_{\tilde{\gamma},BB} - A^{(1)}_{BB})\mathbf{v}_B^{(k+1)}.$$
5.     Solve:
$$\begin{bmatrix} A^{(2)}_{II} & A^{(2)}_{IB} \\ A^{(2)}_{BI} & G^{(2)}_{\tilde{\gamma},BB} + A^{(2)}_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(k+1)} \\ \mathbf{w}_B^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(2)} \\ \mathbf{f}_B + \mathbf{z}_B^{(k+1)} \end{bmatrix}$$
6. Endfor
7. Output: $\mathbf{v}^{(k+1)}$ and $\mathbf{w}^{(k+1)}$
```

In the preceding, the Robin-Robin subproblem on Ω_1 was:

$$\begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} \\ A_{BI}^{(1)} & G_{\tilde{\gamma},BB}^{(1)} + A_{BB}^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(k+1)} \\ \mathbf{v}_B^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B + \mathbf{r}_B^{(k+1)} \end{bmatrix},$$

where:

$$\begin{aligned} \mathbf{r}_B^{(k+1)} = \quad & \theta \left(A_{BI}^{(1)} \mathbf{v}_I^{(k)} + (G_{\tilde{\gamma},BB}^{(1)} + A_{BB}^{(1)}) \mathbf{v}_B^{(k)} - \mathbf{f}_B \right) \\ & + (1-\theta) \left(-A_{BI}^{(2)} \mathbf{w}_I^{(k)} + (G_{\tilde{\gamma},BB}^{(1)} - A_{BB}^{(2)}) \mathbf{w}_B^{(k)} \right). \end{aligned}$$

Analogous expressions hold for the Robin subproblem on Ω_2.

As in remark 8.19, system (8.39) can be reduced to the case $\mathbf{f}_I^{(1)} = \mathbf{0}$, $\mathbf{f}_I^{(2)} = \mathbf{0}$. In this case, the preceding γ-Robin-Robin algorithm will correspond to a *block Gauss-Seidel* algorithm to solve the extended system:

$$\begin{cases} \left(S^{(1)} + G_{\tilde{\gamma},BB}^{(1)} \right) \mathbf{v}_B + \left(S^{(2)} - G_{\tilde{\gamma},BB}^{(1)} \right) \mathbf{w}_B = \tilde{\mathbf{f}}_B \\ \left(S^{(1)} - G_{\tilde{\gamma},BB}^{(2)} \right) \mathbf{v}_B + \left(S^{(2)} + G_{\tilde{\gamma},BB}^{(2)} \right) \mathbf{w}_B = \tilde{\mathbf{f}}_B, \end{cases}$$

with *relaxation* of $\mathbf{v}_B$ using a parameter $0 < \theta < 1$. When $\tilde{\gamma}(x) \geq \gamma_0 > 0$ on B, the above system will yield $\mathbf{v}_B = \mathbf{w}_B$, where $\mathbf{v}_B$ solves the original Schur complement system $(S^{(1)} + S^{(2)})\mathbf{u}_B = \tilde{\mathbf{f}}_B$. The Dirichlet-Robin will correspond to a *block Gauss-Seidel* algorithm to solve the extended system:

$$\begin{cases} \mathbf{v}_B - \mathbf{w}_B = \mathbf{0} \\ \left(S^{(1)} - G_{\tilde{\gamma},BB}^{(2)} \right) \mathbf{v}_B + \left(S^{(2)} + G_{\tilde{\gamma},BB}^{(2)} \right) \mathbf{w}_B = \tilde{\mathbf{f}}_B, \end{cases}$$

using *relaxation* of $\mathbf{v}_B$ with a parameter $0 < \theta < 1$.

When $\mathbf{f}_I^{(1)} = \mathbf{0}$ and $\mathbf{f}_I^{(2)} = \mathbf{0}$, the steps in the γ-Robin-Robin algorithm can be expressed in terms of the Schur complement matrices as follows:

$$\begin{aligned} (S^{(1)} + G_{\tilde{\gamma},BB}^{(1)}) \mathbf{v}_B^{(k+1)} &= (\mathbf{f}_B + \mathbf{r}_B^{(k+1)}) \text{ in step 3} \\ (S^{(2)} + G_{\tilde{\gamma},BB}^{(2)}) \mathbf{w}_B^{(k+1)} &= (\mathbf{f}_B + \mathbf{z}_B^{(k+1)}) \text{ in step 5.} \end{aligned}$$

The vector $\mathbf{r}_B^{(k+1)}$ in step 2 and $\mathbf{z}_B^{(k+1)}$ in step 4 satisfy:

$$\begin{aligned} \mathbf{r}_B^{(k+1)} &= \left(\theta\, (S^{(1)} + G_{\tilde{\gamma},BB}^{(1)}) \mathbf{v}_B^{(k)} - \theta\, \mathbf{f}_B + (1-\theta)\, (-S^{(2)} + \tilde{M}_{\tilde{\gamma},BB}^{(1)})\, \mathbf{w}_B^{(k)} \right) \\ \mathbf{z}_B^{(k+1)} &= (\tilde{M}_{\tilde{\gamma},BB}^{(2)} - S^{(1)}) \mathbf{v}_B^{(k+1)}. \end{aligned}$$

Once $\mathbf{u}_B$ is determined, we can compute $\tilde{\mathbf{u}}_I^{(l)} = -A_{II}^{(l)^{-1}} A_{IB}^{(l)} \mathbf{u}_B$ for $l = 1, 2$.

Remark 8.22. In practice, the Schur complement system can be solved using GMRES acceleration, with a preconditioner obtained by applying one iteration of the above algorithm with zero initial iterate, yielding robust convergence [AC7, RA3]. A *parallel* Robin-Robin preconditioner S_0 for S will be:

$$S_0^{-1} \equiv \begin{bmatrix} 0 \\ I \end{bmatrix}^T \left(\begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} \\ A_{BI}^{(1)} & G_{\tilde{\gamma},BB}^{(1)} + A_{BB}^{(1)} \end{bmatrix}^{-1} + \begin{bmatrix} A_{II}^{(2)} & A_{IB}^{(2)} \\ A_{BI}^{(2)} & G_{\tilde{\gamma},BB}^{(2)} + A_{BB}^{(2)} \end{bmatrix}^{-1} \right) \begin{bmatrix} 0 \\ I \end{bmatrix},$$

corresponding to an analog of the Neumann-Neumann preconditioner.

Remark 8.23. A disadvantage of the two subdomain Robin-Robin preconditioner is that the local problems can be computationally expensive. However, if subdomain Ω_2 corresponds to a layer region, and $h_1 \gg h_2$, then matrix $A^{(1)}$ in Ω_1 may be of smaller size and solved using a direct solver. If the layer region Ω_2 is of width $O(\varepsilon)$, a band solver may be obtained for $A^{(2)}$ by reordering the unknowns along the thin layer region. For other advection dominated algorithms, see [CH26, HE2, SC7, NA5, NA4, AC4] and Chap. 12.

8.4 Time Stepping Applications

The preconditioners described in the preceding sections can be simplified when applied to solve the non-symmetric linear system arising from the implicit discretization of a non-selfadjoint parabolic equation. In this section, we remark on some simplifications, see Chap. 9 for a more detailed discussion. Implicit schemes result in linear systems having the form:

$$(G + \alpha\, \tau\, A)\, \mathbf{u} = \mathbf{f}, \tag{8.51}$$

where G is a mass matrix for finite element discretizations or an identity matrix for finite difference discretizations, $A = H + N$ denotes a discretization of the elliptic operator (with H symmetric positive semi-definite and N skew-symmetric), $\tau > 0$ denotes a time step, with $\alpha > 0$ a parameter (which we shall henceforth absorb into τ, for convenience). For appropriately small τ, the convergence factor of iterative algorithms to solve system (8.51) typically improve, and coarse space correction may not be necessary. This arises due to the spectral properties of the limiting matrix G, and when the skew-symmetric part $\tau\, N$ is dominated by the symmetric part $(G + \tau\, H)$.

A Symmetric Positive Definite Preconditioner. Given $A = H + N$, matrix $(G + \tau\, A)$ can be split as follows:

$$(G + \tau\, A) = (G + \tau\, H) + \tau\, N,$$

where $(G + \tau\, H)$ is symmetric and positive definite and N is skew-symmetric. This *splitting* immediately yields the following unaccelerated iteration:

$$(G + \tau\, H)\mathbf{u}^{(k+1)} = \mathbf{f} - \tau\, N\mathbf{u}^{(k)},$$

which will be convergent in a norm $\|\cdot\|$, provided:

$$\|\tau(G + \tau H)^{-1} N\| < 1.$$

If the Euclidean norm is employed, it will be sufficient to require that:

$$\tau \left(\frac{\sigma_{\max}(N)}{\lambda_{\min}(G + \tau H)} \right) < 1, \tag{8.52}$$

where $\sigma_{\max}(N)$ denotes the maximum singular value of N and $\lambda_{\min}(G + \tau H)$ the minimum eigenvalue of $(G + \tau H)$.

Remark 8.24. For a finite difference discretization $\sigma_{\max}(N) = O(h^{-1})$ and $\lambda_{\min}(G+\tau H) = O(1)$, while $\sigma_{\max}(N) = O(h^{-1+d})$ and $\lambda_{\min}(G+\tau H) = O(h^d)$ for a finite element discretization (when $\Omega \subset \mathbb{R}^d$). In either case, if $\tau = C\,h$, condition (8.52) will hold if $C > 0$ is sufficiently small. However, this can be too restrictive on τ. If $\tau \leq C\,h$ but condition (8.52) is violated, and K_0 is a symmetric positive definite preconditioner for $(G + \tau\, H)$, we may employ K_0 as a preconditioner for $(G + \tau\, A)$. Lemma 8.6 will then yield GMRES bounds for $K_0^{-1}(G + \tau\, A)$ independent of h and τ, (but dependent on C). If K_0 is a domain decomposition preconditioner, coarse space correction can be omitted if $\tau \leq \tilde{C}\,h_0^2$ for some $\tilde{C} > 0$.

Schwarz Preconditioners. Typically, Schwarz algorithms yield rapid convergence when applied to solve a well conditioned system $G\,\mathbf{u} = \mathbf{f}$. As a result, when applied to solve $(G + \tau\, A)\,\mathbf{u} = \mathbf{f}$, we expect the convergence to improve as $\tau \to 0^+$. Convergence analysis of Schwarz algorithms indicates that if the time step τ satisfies a constraint of the form $\tau \leq C\,h_0^2$ (where h_0 denotes the size of the subdomains), then the *coarse space* correction term can be *omitted* in Schwarz preconditioners for $(G + \tau\, A)$ without adverse deterioration in its convergence rate (yielding optimal or poly-logarithmic bounds), see [CA, CA3, KU3, KU6]. The additive Schwarz preconditioner without coarse space correction, has the form:

$$M^{-1} = \sum_{l=1}^{p} R_l^T \left(R_l (G + \tau A) R_l^T \right)^{-1} R_l,$$

where R_l denotes the nodal restriction map onto Ω_l^*. The coarse space correction step may also be omitted in the multiplication Schwarz algorithm provided $\tau \leq C\,h_0^2$. In Chap. 9, we describe variants of Schwarz algorithms which require only *one* iteration each time step, without compromising the stability or accuracy of the original implicit scheme, provided the overlap between subdomains is sufficiently large.

Schur Complement Preconditioners. The Schur complement associated with non-symmetric system (8.51) can be preconditioned by a Neumann-Neumann or Robin-Robin preconditioner, given a non-overlapping decomposition $\Omega_1, \ldots, \Omega_p$ of diameter h_0 and interface $B = \cup_{l=1}^{p} B^{(l)}$. As $\tau \to 0^+$, we expect the convergence rate to improve, see [DR5]. Matrix $S(\tau)$ will have the form:

$$S(\tau) \equiv (G_{BB} + \tau A_{BB}) - (G_{BI} + \tau A_{BI})(G_{II} + \tau A_{II})^{-1}(G_{IB} + \tau A_{IB}).$$

If matrix A is obtained by discretization of $\mathcal{A}^0(.,.)$, the Neumann-Neumann preconditioner for $S(\tau)$ will have the following matrix form:

$$M_0^{-1} = \mathcal{R}_0^T S_0^{-1} \mathcal{R}_0 + \sum_{l=1}^{p} \mathcal{R}_l^T \left(S^{(l)}(\tau) \right)^{-1} \mathcal{R}_l,$$

where $\mathcal{R}_l$ is the nodal restriction from B onto $B^{(l)}$ and $\mathcal{R}_l^T$ is its transpose:

$$S^{(l)}(\tau) = \left((G_{BB}^{(l)} + \tau\, A_{BB}^{(l)}) - (G_{BI}^{(l)} + \tau\, A_{BI}^{(l)})(G_{II}^{(l)} + \tau\, A_{II}^{(l)})^{-1}(G_{IB}^{(l)} + \tau\, A_{IB}^{(l)}) \right).$$

Heuristically, if a constraint of the form $\tau \le C\, h_0^2$ is satisfied for subdomains of size h_0, then the coarse grid correction term $\mathcal{R}_0^T S_0^{-1} \mathcal{R}_0$ can be omitted, retaining poly-logarithmic bounds. In Chap. 9, we describe *heuristic* Schur complement algorithms which require only *one iteration* each time step, without affecting the stability of the original implicit scheme. For a comparison of Schwarz and Schur complement algorithms, see [CA10, KE8].

8.5 Theoretical Results

In this section, we present selected theoretical results on non-self adjoint elliptic equations [SC4, YS, YS3, CA, VA10, CA20, XU8, CA21, WA3], and apply them to analyze the Sobolev norm convergence of an algorithm of [XU8].

8.5.1 Background

We consider the non-self adjoint, possibly *indefinite*, elliptic equation:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x) u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{8.53}$$

where $0 < a_0 \le a(x) \le a_1$. Here $c(x)$ is permitted to be *negative.* The weak formulation of the above problem will seek $u \in H_0^1(\Omega)$ such that:

$$\begin{cases} \mathcal{A}(u, \phi) = F(\phi), \qquad \forall \phi \in H_0^1(\Omega), \text{ where} \\ \mathcal{A}(u, \phi) \equiv \int_\Omega (a(x)\nabla u \cdot \nabla \phi + (\mathbf{b}(x) \cdot \nabla u)\, \phi + c(x)\, u\, \phi)\, dx \\ \quad F(\phi) \equiv \int_\Omega f\, \phi\, dx. \end{cases} \tag{8.54}$$

Its Galerkin discretization will seek $u_h \in V_h \subset H_0^1(\Omega)$ satisfying:

$$\mathcal{A}(u_h, \phi_h) = F(\phi_h), \quad \forall \phi_h \in V_h, \tag{8.55}$$

where $V_h \subset H_0^1(\Omega)$ is a finite element space defined on a quasiuniform triangulation Ω_h of Ω. Expanding $u_h = \sum_{i=1}^n (\mathbf{u})_i \phi_i$ in a basis for V_h yields a linear system: $A\mathbf{u} = \mathbf{f}$ where $(A)_{ij} = \mathcal{A}(\phi_i, \phi_j)$ and $(\mathbf{f})_i = F(\phi_i)$.

The following is Garding's inequality for the bilinear form $\mathcal{A}(u, v)$.

Lemma 8.25. *There exists $\alpha_* > 0$ and $\gamma > 0$ such that the bilinear form $\mathcal{A}(u,v)$ satisfies the Garding inequality:*

$$a_0\,\alpha_*\,\|u\|_1^2 - \gamma\|u\|_0^2 \le \mathcal{A}(u,u), \quad \forall u \in H_0^1(\Omega).$$

Proof. Assuming that $\mathbf{b}(x)$ is smooth, integrate by parts to obtain:

$$\begin{aligned}\mathcal{A}(u,u) &= \int_\Omega a(x)\,|\nabla u|^2 dx - \int_\Omega \left(\tfrac{1}{2}\nabla\cdot\mathbf{b}(x) - c(x)\right) u^2 dx \\ &\ge a_0 \int_\Omega |\nabla u|^2 dx - \gamma \int_\Omega u^2 dx,\end{aligned}$$

where $\gamma \equiv \max_\Omega \left|\frac{1}{2}\nabla\cdot\mathbf{b}(x) - c(x)\right|$. Next, apply the Poincaré-Freidrichs inequality which guarantees the existence of the positive constant α_* such that:

$$\alpha_*\|u\|_{1,\Omega}^2 \le |u|_{1,\Omega}^2 = \int_\Omega |\nabla u|^2 dx, \quad \forall u \in H_0^1(\Omega),$$

to obtain:

$$\mathcal{A}(u,u) \ge a_0\,\alpha_*\,\|u\|_1^2 - \gamma\|u^2\|_{0,\Omega}^2.$$

Rearranging terms yields the desired result. □

We shall split the nonsymmetric bilinear form $\mathcal{A}(u,v)$ as:

$$\mathcal{A}(u,v) = \mathcal{H}_0(u,v) + \mathcal{N}_0(u,v), \tag{8.56}$$

where:

$$\begin{cases}\mathcal{H}_0(u,v) \equiv \int_\Omega a(x)\nabla u\cdot\nabla v\,dx, & \forall u,v \in H_0^1(\Omega)\\ \mathcal{N}_0(u,v) \equiv \int_\Omega \left(\mathbf{b}(x)\cdot\nabla u + c(x)\,u\right)\,v\,dx, & \forall u,v \in H_0^1(\Omega).\end{cases} \tag{8.57}$$

Here $\mathcal{H}_0(.,.)$ is a symmetric bilinear form corresponding to the *principal* part of the elliptic operator, while $\mathcal{N}_0(.,.)$ is a nonsymmetric bilinear form corresponding to the *lower order* terms. The following bounds will hold.

Lemma 8.26. *There exists a positive constant C such that:*

$$\begin{aligned}\mathcal{A}(u,\phi) &\le C\|u\|_1\|\phi\|_1, & \forall u,v \in H_0^1(\Omega)\\ \mathcal{H}_0(u,\phi) &\le C\|u\|_1\|\phi\|_1, & \forall u,\phi \in H_0^1(\Omega)\\ \mathcal{N}_0(u,\phi) &\le C\|u\|_1\|\phi\|_0, & \forall u,\phi \in H_0^1(\Omega)\\ \mathcal{N}_0(u,\phi) &\le C\|u\|_0\|\phi\|_1, & \forall u,\phi \in H_0^1(\Omega).\end{aligned}$$

Proof. All the results, except the last inequality, follow trivially by the Schwartz inequality. To obtain the last inequality, integrate by parts and shift the derivatives before applying Schwartz's inequality. □

In the *indefinite* case, elliptic equation (8.53) may not always be solvable. For instance, if $\mathbf{b}(x) \equiv 0$ and $c(x) = -\lambda$ is the negative of an eigenvalue of the principal part of the elliptic operator, then the nonhomogeneous problem will be solvable only if $f(x)$ is orthogonal to the corresponding eigenfunction. However, whenever $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) \ge 0$, the quadratic form $\mathcal{A}(u,u)$ will be *coercive*, and the Lax-Milgram lemma [CI2] will guarantee solvability of (8.53).

Lemma 8.27. *Let* $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \geq 0$*, and let* $0 < c_0 \leq c_1$ *be such that:*

$$c_0\, \mathcal{H}_0(u,u) \leq \mathcal{A}(u,u) \;\; \textit{and} \;\; \mathcal{A}(u,v) \leq c_1 \mathcal{H}_0^{1/2}(u,u)\, \mathcal{H}_0^{1/2}(v,v), \; \forall u, v \in H_0^1(\Omega).$$

Then, equation (8.53) will be uniquely solvable.

Proof. We outline a proof based on Riesz isometry [CI2, EV]. Let $X = H_0^1(\Omega)$ and X' denote its dual space, with duality pairing $\langle \cdot, \cdot \rangle$ between X and X'. Let $A : X \rightarrow X'$ and $H_0 : X \rightarrow X'$ denote the induced maps:

$$\langle Au, v\rangle = \mathcal{A}(u,v) \quad \langle H_0 u, v\rangle = \mathcal{H}_0(u,v), \quad \forall u,\, v \in X.$$

Given $f \in X'$ the solvability of (8.53) reduces to that of $A\,u = f$ in X. When $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\right) \geq 0$, the principal part H_0 will be invertible by the Lax-Milgram lemma. Define $T : X \rightarrow X$ by $T\,u \equiv u + \theta\, H_0^{-1}\,(f - A\,u)$. For appropriately chosen θ (depending on c_0 and c_1), it can easily be verified that T is a *contraction*, and by construction, its fixed point will solve (8.53). □

Henceforth, we shall assume that elliptic equation (8.53) and its adjoint are uniquely solvable, where, given $f(x) \in L^2(\Omega)$, the formal adjoint to (8.54) seeks $v \in H_0^1(\Omega)$ such that:

$$\mathcal{A}(\phi, v) = (f, \phi), \quad \forall \phi \in H_0^1(\Omega). \tag{8.58}$$

Below, we state a result [SC4], referred to as Schatz's lemma, which can be employed to show the *coercivity* of the *indefinite* bilinear form $\mathcal{A}(\cdot,\cdot)$ within a certain subspace of $H_0^1(\Omega)$, particularly when $c(x)$ is negative.

Lemma 8.28 (Schatz's Lemma). *Suppose the following conditions hold:*

1. *For each* $f(x) \in L^2(\Omega)$*, let the adjoint problem (8.58) be uniquely solvable with* $v \in H^{1+\alpha}(\Omega)$ *for some* $0 < \alpha \leq 1$ *satisfying:*

$$\|v\|_{1+\alpha} \leq C\|f\|_0. \tag{8.59}$$

2. *Let* $w \in H_0^1(\Omega)$ *satisfy:*

$$\mathcal{A}(w, \phi) = 0, \quad \forall \;\; \phi \in V_{h_0}, \tag{8.60}$$

for a finite element space $V_{h_0} \subset H_0^1(\Omega)$.

Then, there exists $K_* > 0$ *independent of* h_0 *such that:*

$$\|w\|_0 \leq K_* h_0^{\alpha} \|w\|_1.$$

Proof. The proof employs Nietzsche's trick, see [CI2, JO2]. Given w satisfying (8.60), solve the adjoint problem using $f(x) = w(x)$:

$$\mathcal{A}(\phi, v) = (w, \phi), \quad \forall \phi \in H_0^1(\Omega).$$

Choosing $\phi = w$ as a test function yields:

$$\mathcal{A}(w, v) = (w, w).$$

We can replace $\mathcal{A}(w, v)$ by $\mathcal{A}(w, v - \phi_{h_0})$ for $\phi_{h_0} \in V_{h_0}$, since $\mathcal{A}(w, \phi_{h_0}) = 0$. Consequently, by applying boundedness of the bilinear form, we obtain:

$$\begin{aligned}(w, w) \;=\; \mathcal{A}(w, v - \phi_{h_0}) \;\leq\; C_1 \|w\|_1 \|v - \phi_{h_0}\|_1 &\leq C_2 \|w\|_1 h_0^{\alpha} \|v\|_{1+\alpha} \\ &\leq C_3 \|w\|_1 h_0^{\alpha} \|w\|_0.\end{aligned}$$

Here $\phi_{h_0} = I_{h_0} v \subset V_{h_0}$ denotes the finite element nodal interpolant, and we have used standard finite element interpolation error estimates and the bound $\|v\|_{1+\alpha} \leq C \|w\|_0$ (which holds by assumption on the regularity of the adjoint problem since $f = w$). It now follows that:

$$\|w\|_0 \leq K_* h_0^{\alpha} \|w\|_1,$$

for some $K_* > 0$, which is the desired result. □

Remark 8.29. By regularity theory for elliptic partial differential equations, bound (8.59) will hold true when Ω is a convex polyhedron, see [GR8].

Next, employing Schatz's lemma, we describe conditions which guarantee the *non-singularity* of the linear system $A\mathbf{u} = \mathbf{f}$ arising from a Galerkin discretization of (8.53). We also describe how $u - u_h$ can be estimated.

Lemma 8.30. *Suppose that (8.53) and its adjoint are uniquely solvable with solutions which are $H^{1+\alpha}(\Omega)$ regular when $f(x) \in L^2(\Omega)$. Then, there exists an $h_0 > 0$ such that for $h < h_0$ the matrix A, arising in the discretization of (8.53) based on the finite element space $V_h \subset H_0^1(\Omega)$, will be non-singular.*

Proof. Let $u(x) = 0$ denote the unique solution to the homogeneous problem (8.53) when $f(x) = 0$, and let $w_h \in V_h$ be any solution of its Galerkin discretization. We shall show that $w_h = 0$ is the only discrete homogeneous solution, provided h is sufficiently small.

Accordingly, consider the equation satisfied by w_h:

$$\mathcal{A}(w_h, \phi) = 0, \quad \forall \phi \in V_h.$$

By construction, the error $u - w_h$ will satisfy:

$$\mathcal{A}(u - w_h, \phi) = 0, \quad \forall \phi \in V_h.$$

Applying Garding's inequality to $u - w_h$ yields:

$$\begin{aligned}\alpha_*\, a_0 \|u - w_h\|_1^2 - \gamma\, \|u - w_h\|_0^2 &\leq \mathcal{A}(u - w_h, u - w_h) \\ &= \mathcal{A}(u - w_h, u),\end{aligned}$$

since $\mathcal{A}(u-w_h,w_h)=0$. The boundedness of $\mathcal{A}(\cdot,\cdot)$ yields:

$$\alpha_*\, a_0\, \|u-w_h\|_1^2 - \gamma\, \|u-w_h\|_0^2 \le C\|u-w_h\|_1\|u\|_1.$$

Applying Schatz's inequality to $u-w_h$ yields:

$$\|u-w_h\|_0 \le K_*\, h^\alpha \|u-w_h\|_1,$$

and substituting this into the inequality preceding it yields:

$$\alpha_*\, a_0\, \|u-w_h\|_1^2 - \gamma K_*^2 h^{2\alpha} \|u-w_h\|_1^2 \le C\|u-w_h\|_1\|u\|_1.$$

Dividing throughout by $\|u-w_h\|_1$ yields:

$$\left(\alpha_*\, a_0 - \gamma\, K_*^2\, h^{2\alpha}\right) \|u-w_h\|_1 \le C\|u\|_1.$$

When $h < h_0 = \left(\alpha_*\, a_0/\gamma K_*^2\right)^{1/2\alpha}$ this shows that $w_h = 0$ since $u = 0$. Since the homogeneous problem has a unique solution, it follows that A is non-singular, and that the discretization is uniquely solvable. □

Another consequence of Schatz's lemma is the *coercivity* of the bilinear from $\mathcal{A}(\cdot,\cdot)$ within any subspace of $H_0^1(\Omega)$ oblique $\mathcal{A}(.,.)$-orthogonal to V_{h_0} for sufficiently small h_0. Importantly, this result will hold even when $c(x) < 0$.

Lemma 8.31. *Let $V_{h_0} \subset H_0^1(\Omega)$ be a finite element space for which Schatz's lemma holds with:*

$$\|u\|_0 \le K_* h_0^\alpha |u|_1,$$

for u satisfying

$$\mathcal{A}(u,\phi_{h_0}) = 0, \qquad \forall \phi_{h_0} \in V_{h_0}. \tag{8.61}$$

Then for $h < h_0 = \left(a_0/2\,\gamma\, K_*^2\right)^{1/2\alpha}$, the following will hold:

$$\mathcal{A}(u,u) \ge \frac{1}{2}\mathcal{H}_0(u,u),$$

where $\mathcal{H}_0(u,u) = \int_\Omega a(x)|\nabla u|^2 dx$ and $\gamma = \max_{x\in\Omega} |c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)|$.

Proof. Let u satisfy (8.61). Employ the bound $\|u\|_0^2 \le K_*^2 h_0^{2\alpha}\, |u|_1^2$ from Schatz's lemma to obtain:

$$\begin{aligned}
\mathcal{A}(u,u) &= \mathcal{H}_0(u,u) + \textstyle\int_\Omega \left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) u^2\, dx \\
&\ge \mathcal{H}_0(u,u) - \max_{x\in\Omega} |c(x) - \tfrac{1}{2}\nabla\cdot\mathbf{b}(x)|\, \textstyle\int_\Omega u^2 dx \\
&\ge \mathcal{H}_0(u,u) - K_*^2 h_0^{2\alpha} \max_\Omega |c(x) - \tfrac{1}{2}\nabla\cdot\mathbf{b}(x)|\, |u|_1^2 \\
&\ge \mathcal{H}_0(u,u) - K_*^2 h_0^{2\alpha}\, \gamma\, |u|_1^2 \\
&\ge \left(1 - (\gamma\, K_*^2 h_0^{2\alpha})/a_0\right) \mathcal{H}_0(u,u).
\end{aligned}$$

The last term will be coercive provided h_0 is sufficiently small so that:

$$1 - (\gamma\, K_*^2 h_0^{2\alpha} / a_0) \geq \frac{1}{2}.$$

This establishes the coercivity of the nonsymmetric bilinear form $\mathcal{A}(.,.)$ within the subspace oblique $\mathcal{A}(.,.)$-orthogonal to V_{h_0} for h_0 sufficiently small. □

The next result establishes an upper bound for $\mathcal{A}(.,.)$ in terms of the principal part $\mathcal{H}_0(.,.)$ within a subspace.

Lemma 8.32. *Let $u \in H_0^1(\Omega)$ satisfy: $\mathcal{A}(u, \phi) = 0$ for all $\phi \in V_{h_0}$ and let Schatz's lemma hold with:*

$$\|u\|_0 \leq K_* h_0^{\alpha} |u|_1.$$

Then there exists $c_* > 0$ independent of h_0 such that:

$$|\mathcal{A}(u, w)| \leq (1 + c_* h_0^{\alpha})\, \mathcal{H}_0(u, u)^{1/2} \mathcal{H}_0(w, w)^{1/2},$$

for any $w \in H_0^1(\Omega)$.

Proof. Integrate by parts, apply Schatz's lemma and bounds for $\mathcal{N}_0(.,.)$:

$$\begin{aligned}
\mathcal{A}(u, w) &= \mathcal{H}_0(u, w) + \mathcal{N}_0(u, w) \\
&\leq \mathcal{H}_0(u, w) + C \|u\|_0 |w|_1 \\
&\leq \mathcal{H}_0(u, w) + C\, K_* h_0^{\alpha} |u|_1 |w|_1 \\
&\leq \left(1 + \tfrac{C\, K_* h_0^{\alpha}}{a_0}\right) \mathcal{H}_0(u, u)^{1/2} \mathcal{H}_0(w, w)^{1/2},
\end{aligned}$$

where the last inequality follows by the Schwartz inequality. The desired result follows for $c_* = (C\, K_* / a_0)$. □

The preceding upper and lower bounds indicate that on any subspace of $H_0^1(\Omega)$ oblique $\mathcal{A}(.,.)$-orthogonal to V_{h_0} for sufficiently small h_0, the nonsymmetric and indefinite quadratic form $\mathcal{A}(w, w)$ will be *coercive* and equivalent to the quadratic form associated with its principal part $\mathcal{H}_0(w, w)$:

$$c\, \mathcal{H}_0(w, w) \leq \mathcal{A}(w, w) \leq C\, \mathcal{H}_0(w, w).$$

This heuristically motivates the additive preconditioner [XU8] for A based on the discretization of $\mathcal{A}(.,.)$ in subspace V_{h_0} and any symmetric positive definite preconditioner spectrally equivalent to $\mathcal{H}_0(.,.)$. We shall describe convergence bounds for such a preconditioner in the next section.

We conclude this subsection by indicating how traditional error estimates can be obtained for a Galerkin discretization of (8.53) even when $\mathcal{A}(.,.)$ is nonsymmetric and possibly indefinite. Let u denote the weak solution and u_h its Galerkin approximation. Then, by construction:

$$\mathcal{A}(u - u_h, \phi_h) = 0, \quad \forall \phi_h \in V_h.$$

If h is sufficiently small, Schatz's lemma will hold, and $\mathcal{A}(.,.)$ will be coercive for $u - u_h$, so that:

$$\begin{aligned} c\,\mathcal{H}_0(u - u_h, u - u_h) \,\le\, \mathcal{A}(u - u_h, u - u_h) &= \mathcal{A}(u - u_h, u) \\ &= \mathcal{A}(u - u_h, u - I_h u) \\ &\le C \|u - u_h\|_1 \|u - I_h u\|_1. \end{aligned}$$

Standard error bounds now follow immediately.

8.5.2 Bounds for the Additive Preconditioner [XU8]

We shall now analyze the convergence of a preconditioned algorithm of [XU8], for diffusion dominated problems. Given a finite element space $V_h \subset H_0^1(\Omega)$, we shall employ the splitting $A = H_0 + N_0$ where H_0 and N_0 are discretizations of $\mathcal{H}_0(.,.) = (H_0 \cdot, \cdot)$ and $\mathcal{N}_0(.,.) = (N_0 \cdot, \cdot)$, respectively, on V_h, see (8.57), where $(.,.)$ denotes the $L^2(\Omega)$ inner product. Note that $H_0^T = H_0 > 0$ and that $\|\cdot\|_{\mathcal{H}_0}$ is equivalent to $\|\cdot\|_1$, by Poincaré-Freidrich's inequality.

If M_0 denotes any symmetric positive definite preconditioner for H_0, and $\text{Range}(R_0^T)$ denotes a coarse space $V_{h_0} \subset V_h$ for a *sufficiently small* coarse grid size h_0, then, given a parameter $\beta > 0$, the action of the inverse M^{-1}, of the additive preconditioner M of [XU8], on the matrix A has the form:

$$M^{-1}A = \beta\, M_0^{-1} A + R_0^T \left(R_0 A R_0^T\right)^{-1} R_0 A.$$

If P_0 denotes the $\mathcal{A}(.,.)$-oblique projection onto $\text{Range}(R_0^T) = V_{h_0}$:

$$\mathcal{A}(P_0 v_h, \phi_h) = \mathcal{A}(v_h, \phi_h), \quad \forall \phi_h \in \text{Range}(R_0^T),$$

then, the matrix representation of the oblique-projection P_0 will be:

$$\mathbf{P}_0 = R_0^T \left(R_0 A R_0^T\right)^{-1} R_0 A.$$

Note that $\mathbf{P}_0 \mathbf{P}_0 = \mathbf{P}_0$. However, since A is not symmetric positive definite, $\mathbf{P}_0$ will not be an orthogonal projection. The preconditioned matrix $M^{-1}A$ in [XU8] can be now be represented in terms of matrix $\mathbf{P}_0$ as:

$$M^{-1}A = \,\beta\, M_0^{-1} A + \mathbf{P}_0.$$

The following notation will be employed. We define a contraction factor:

$$\delta_0 = \sup_{u \in H_0^1(\Omega) \setminus \{0\}} \frac{\|(I - P_0)u\|_0}{\|u\|_{\mathcal{H}_0}}. \tag{8.62}$$

Since $\|\cdot\|_1$ is equivalent to $\|\cdot\|_{\mathcal{H}_0}$, Schatz's lemma yields a bound of the form $\delta_0 \le K_* h^\alpha$ provided (8.53) and its adjoint are uniquely solvable in $H^{1+\alpha}(\Omega)$ for $f(x) \in L_2(\Omega)$. We shall also let c_1 and c_2 denote parameters such that:

$$\begin{aligned} \|u\|_0 &\le c_1 \|u\|_{\mathcal{H}_0}, & \forall u \in H_0^1(\Omega) \\ \mathcal{N}_0(u,v) &\le c_2 \|u\|_0 \|v\|_{\mathcal{H}_0}, & \forall u, v \in H_0^1(\Omega). \end{aligned} \tag{8.63}$$

We shall also assume that matrix M_0 satisfies:

$$\lambda_m \le \frac{\mathbf{v}^T H_0 \mathbf{v}}{\mathbf{v}^T M_0 \mathbf{v}} \le \lambda_M, \tag{8.64}$$

with $0 < \lambda_m < \lambda_M$ where λ_m and λ_M are independent of h.

Remark 8.33. The parameters K_*, c_1, c_2 will be independent of h by standard results from elliptic regularity and finite element theory.

The following is a preliminary result.

Lemma 8.34. *Suppose that Schatz's lemma holds, and that* $c_1, c_2, \lambda_m, \lambda_M$ *and* δ_0 *are as defined before. Then, the following bounds will hold:*

1. $\|P_0 u\|_{\mathcal{H}_0}^2 \le 2\,(P_0 u, u)_{\mathcal{H}_0} + c_2^2\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2$.
2. $\|u\|_0^2 \le 4\,c_1^2\,(P_0 u, u)_{\mathcal{H}_0} + 2\,(c_1^2 c_2^2 + 1)\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2$.
3. $\|u\|_{\mathcal{H}_0}^2 \le 2\,\lambda_m^{-1} (M_0^{-1} A u, u)_{\mathcal{H}_0} + c_2^2 \left(\frac{\lambda_M}{\lambda_m}\right)^2 \|u\|_0^2$.

Proof. We follow [XU8]. To prove result 1, consider:

$$\begin{aligned} \|P_0 u\|_{\mathcal{H}_0}^2 &= \mathcal{H}_0(P_0 u, P_0 u) \\ &= \mathcal{H}_0(P_0 u, u) - \mathcal{H}_0(P_0 u, (I - P_0) u) \\ &= \mathcal{H}_0(P_0 u, u) - \mathcal{H}_0((I - P_0) u, P_0 u) \\ &= \mathcal{H}_0(P_0 u, u) - \mathcal{A}((I - P_0) u, P_0 u) + \mathcal{N}_0((I - P_0) u, P_0 u) \\ &= \mathcal{H}_0(P_0 u, u) + \mathcal{N}_0((I - P_0) u, P_0 u) \\ &\le \mathcal{H}_0(P_0 u, u) + c_2\,\|(I - P_0) u\|_0\,\|P_0 u\|_{\mathcal{H}_0} \\ &\le \mathcal{H}_0(P_0 u, u) + c_2\,\delta_0\,\|u\|_{\mathcal{H}_0}\,\|P_0 u\|_{\mathcal{H}_0} \\ &\le \mathcal{H}_0(P_0 u, u) + \tfrac{1}{2}\,c_2^2\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2 + \tfrac{1}{2}\,\|P_0 u\|_{\mathcal{H}_0}^2. \end{aligned}$$

In the last four lines above, the definition of P_0, the Schwartz inequality, the definition of δ_0, and the arithmetic-geometric inequality were employed. Subtracting the term $\frac{1}{2}\|P_0 u\|_{\mathcal{H}_0}^2$ and rescaling yields result 1.

To prove result 2, decompose $u = P_0 u + (I - P_0) u$ and estimate:

$$\begin{aligned} \|u\|_0^2 &\le 2\,\|P_0 u\|_0^2 + 2\,\|u - P_0 u\|_0^2 \\ &\le 2\,c_1^2\,\|P_0 u\|_{\mathcal{H}_0}^2 + 2\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2 \\ &\le 2\,c_1^2\,\left(2\,\mathcal{H}_0(P_0 u, u) + c_2^2\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2\right) + 2\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2 \\ &= 4\,c_1^2\,\mathcal{H}_0(P_0 u, u) + 2\,(c_1^2 c_2^2 + 1)\,\delta_0^2\,\|u\|_{\mathcal{H}_0}^2. \end{aligned}$$

The triangle inequality was employed in the first line, $\|\cdot\|_0 \le c_1 \|\cdot\|_{\mathcal{H}_0}$ and the definition of δ_0 were employed on the second line, while result 1 was substituted on the third line.

To prove result 3, consider:

$$\begin{aligned}
\|u\|_{\mathcal{H}_0}^2 &= \mathcal{H}_0(u,u) \\
&\le \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}H_0u,u) \\
&= \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) - \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}N_0u,u) \\
&= \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) - \lambda_m^{-1}(M_0^{-1}N_0u,H_0u)_0 \\
&= \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) - \lambda_m^{-1}(N_0u,M_0^{-1}H_0u)_0 \\
&= \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) - \lambda_m^{-1}\mathcal{N}_0(u,M_0^{-1}H_0u) \\
&\le \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) + c_2\lambda_m^{-1}\|u\|_0\|M_0^{-1}H_0u\|_{\mathcal{H}_0} \\
&\le \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) + c_2\lambda_m^{-1}\|u\|_0\lambda_M\|u\|_{\mathcal{H}_0} \\
&\le \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) + \tfrac{1}{2}c_2^2\lambda_m^{-2}\lambda_M^2\|u\|_0^2 + \tfrac{1}{2}\|u\|_{\mathcal{H}_0}^2 \\
&= \lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) + \tfrac{1}{2}c_2^2\left(\tfrac{\lambda_M}{\lambda_m}\right)^2\|u\|_0^2 + \tfrac{1}{2}\|u\|_{\mathcal{H}_0}^2.
\end{aligned}$$

Subtracting the term $\frac{1}{2}\|u\|_{\mathcal{H}_0}^2$ and rescaling yields result 3. Here, we have employed bounds for the Rayleigh quotient of $M_0^{-1}H_0$ in the $\mathcal{H}_0(.,.)$ inner product, that $\mathcal{N}_0(u,v) = (N_0u,v)_0$ and bounds on $\mathcal{N}_0(u,v)$, and the arithmetic-geometric mean inequality. □

Now consider a preconditioner M of the form:

$$M^{-1} \equiv \beta\, M_0^{-1} + R_0^T(R_0AR_0^T)^{-1}R_0,$$

where $\beta > 0$ is a parameter defined below. The following result provides a lower bound for the smallest eigenvalue of the symmetric part of $M^{-1}A$.

Proposition 8.35. *Suppose δ_0 satisfies $2c_2^2(c_1^2c_2^2+1)\delta_0^2\left(\frac{\lambda_M}{\lambda_m}\right)^2 \le \frac{1}{2}$. Then, there exists parameters $\mu_1 > 0$ and $\beta > 0$ independent of h such that:*

$$\beta\,\mathcal{H}_0(M_0^{-1}Au,u) + \mathcal{H}_0(P_0u,u) \ge \mu_1\mathcal{H}_0(u,u).$$

Proof. Combining result **3** from the preceding lemma:

$$\|u\|_{\mathcal{H}_0}^2 \le 2\lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) + c_2^2\left(\frac{\lambda_M}{\lambda_m}\right)^2\|u\|_0^2,$$

with bound **2** from the preceding lemma:

$$\|u\|_0^2 \le 4c_1^2\,\mathcal{H}_0(P_0u,u) + 2(c_1^2c_2^2+1)\delta_0^2\|u\|_{\mathcal{H}_0}^2,$$

yields the estimate:

$$\begin{aligned}\|u\|_{\mathcal{H}_0}^2 \le & \, 2\lambda_m^{-1}\mathcal{H}_0(M_0^{-1}Au,u) + 4c_1^2c_2^2 \left(\tfrac{\lambda_M}{\lambda_m}\right)^2 \mathcal{H}_0(P_0u,u) \\ & + 2c_2^2(c_1^2c_2^2+1)\delta_0^2 \left(\tfrac{\lambda_M}{\lambda_m}\right)^2 \|u\|_{\mathcal{H}_0}^2.\end{aligned}$$

When δ_0 is small enough so that $2c_2^2(c_1^2c_2^2+1)\delta_0^2 \left(\frac{\lambda_M}{\lambda_m}\right)^2 \le \frac{1}{2}$, we may subtract the term involving $\|u\|_{\mathcal{H}_0}^2$ on the right hand side. Substituting the parameters:

$$\mu_1 \equiv \frac{\lambda_m^2}{8c_1^2c_2^2\lambda_M^2}, \qquad \text{and} \qquad \beta \equiv \frac{\lambda_m}{2c_1^2c_2^2\lambda_M^2},$$

we obtain the bound:

$$\mu_1\, \mathcal{H}_0(u,u) \le \beta\, \mathcal{H}_0(M_0^{-1}Au,u) + \mathcal{H}_0(P_0u,u).$$

By construction μ_1 and β are independent of h. □

An upper bound is established for $\|M^{-1}A\|_{\mathcal{H}_0}$ in the following result.

Proposition 8.36. *There exists $\mu_2 > 0$ independent of h such that:*

$$\|\beta\, M_0^{-1}Au + P_0u\|_{\mathcal{H}_0} \le \mu_2\|u\|_{\mathcal{H}_0}, \quad \forall u \in V_h.$$

Proof. See [XU8]. First note that:

$$\begin{aligned}\|M_0^{-1}A\,u\|_{\mathcal{H}_0} &\le \|M_0^{-1}H_0u\|_{\mathcal{H}_0} + \|M_0^{-1}N_0u\|_{\mathcal{H}_0} \\ &\le \lambda_M\|u\|_{\mathcal{H}_0} + \|M_0^{-1}N_0u\|_{\mathcal{H}_0}.\end{aligned}$$

However:

$$\begin{aligned}\|M_0^{-1}N_0u\|_{\mathcal{H}_0}^2 &= (H_0M_0^{-1}N_0u, M_0^{-1}N_0u) \\ &= (N_0u, M_0^{-1}H_0M_0^{-1}N_0u) \\ &\le c_2\|u\|_{\mathcal{H}_0}\|M_0^{-1}H_0M_0^{-1}N_0u\|_0 \\ &\le c_2\lambda_M\|u\|_{\mathcal{H}_0}\|M_0^{-1}N_0u\|_0 \\ &\le c_1\,c_2\,\lambda_M\|u\|_{\mathcal{H}_0}\|M_0^{-1}N_0u\|_{\mathcal{H}_0}.\end{aligned}$$

It thus follows that:

$$\|M_0^{-1}N_0u\|_{\mathcal{H}_0} \le c_1c_2\lambda_M\|u\|_{\mathcal{H}_0}.$$

Combining these results gives:

$$\|M_0^{-1}Au\|_{\mathcal{H}_0} \le \lambda_M\|u\|_{\mathcal{H}_0} + c_1c_2\lambda_M\|u\|_{\mathcal{H}_0}.$$

Next, consider:

$$\begin{aligned}
\|P_0 u\|_{\mathcal{H}_0}^2 &= \mathcal{H}_0(P_0 u, P_0 u) \\
&= \mathcal{A}(P_0 u, P_0 u) - \mathcal{N}_0(P_0 u, P_0 u) \\
&= \mathcal{A}(u, P_0 u) - \mathcal{N}_0(P_0 u, P_0 u) \\
&= \mathcal{H}_0(u, P_0 u) + \mathcal{N}_0(u, P_0 u) - \mathcal{N}_0(P_0 u, P_0 u) \\
&= \mathcal{H}_0(u, P_0 u) + \mathcal{N}_0((I - P_0)u, P_0 u) \\
&\le \|u\|_{\mathcal{H}_0} \|P_0 u\|_{\mathcal{H}_0} + c_2 \|(I - P_0)u\|_0 \|P_0 u\|_{\mathcal{H}_0} \\
&\le \|u\|_{\mathcal{H}_0} \|P_0 u\|_{\mathcal{H}_0} + c_2 \delta_0 \|u\|_{\mathcal{H}_0} \|P_0 u\|_{\mathcal{H}_0}.
\end{aligned}$$

Canceling the common terms yields:

$$\|P_0 u\|_{\mathcal{H}_0} \le \|u\|_{\mathcal{H}_0} + c_2 \delta_0 \|u\|_{\mathcal{H}_0} = (1 + c_2 \delta_0) \|u\|_{\mathcal{H}_0}.$$

Combining bounds for $\|M_0^{-1} A u\|_{\mathcal{H}_0}$ and $\|P_0 u\|_{\mathcal{H}_0}$, we obtain:

$$\beta \|M_0^{-1} A u\|_{\mathcal{H}_0} + \|P_0 u\|_{\mathcal{H}_0} \le (1 + c_2 \delta_0 + \beta(1 + c_1 c_2)\lambda_M) \|u\|_{\mathcal{H}_0}.$$

Thus, $\mu_2 = (1 + c_2 \delta_0 + \beta(1 + c_1 c_2)\lambda_M)$. □

Remark 8.37. By construction, the upper and lower bounds μ_2 and μ_1 are independent of h. These bounds may be substituted in the GMRES algorithm to establish a rate of convergence independent of h:

$$\|A\mathbf{u}^{(k)} - \mathbf{f}\|_{\mathcal{H}_0} \le \left(1 - \frac{\mu_1^2}{\mu_2^2}\right)^k \|A\mathbf{u}^{(0)} - \mathbf{f}\|_{\mathcal{H}_0},$$

see [SA2].

Remark 8.38. When $a(x) = \varepsilon \ll 1$, the bilinear form $\mathcal{A}(.,.)$ will be coercive only if the mesh size h_0 satisfies:

$$h_0 < C\, \varepsilon^{1/2\alpha}.$$

When h_0 is very small, it will be prohibitively expensive to apply $\mathbf{P}_0$. As a result, the additive preconditioner based on a coarse space will be primarily suited for *diffusion dominated* problems.

9

Parabolic Equations

In this chapter, we describe domain decomposition methods for solving the linear system arising from an implicit discretization of a parabolic equation. If A denotes the discretization of the underlying elliptic operator, and $\tau > 0$ is the time step, and G is a mass or identity matrix, this yields the system:

$$(G + \alpha \, \tau \, A)\mathbf{u}^k = \tilde{\mathbf{f}}^k$$

at each time $t_k = k \, \tau$, where $\mathbf{u}^k$ denotes the discrete solution at time t_k. The condition number of $(G + \alpha \, \tau A)$ improves as $\tau \to 0^+$, and this facilitates various simplifications or improvements in domain decomposition solvers.

Firstly, if the time step satisfies a constraint of the form $\tau \leq C \, h_0^2$, where h_0 denotes the diameter of the subdomains, then *coarse space* correction may not be necessary in Schwarz and Schur complement methods, to maintain a rate of convergence independent of h and h_0. The resulting algorithms can be parallelized more easily. Secondly, it may be possible to formulate *non-iterative* solvers accurate to within truncation error, without altering the stability of the original scheme. Non-iterative solvers may employ $\mathbf{u}^{k-1}$ and apply an explicit method to predict the boundary values of $\mathbf{u}^k$ on subdomain interfaces and use implicit methods to update the solution in the interior of subdomains, or they may employ the boundary values of $\mathbf{u}^{k-1}$ to update the solution on overlapping subdomains with large overlap, and use a partition of unity to obtain $\mathbf{u}^k$, or they may employ domain decomposition operator splittings and update $\mathbf{u}^{k-1}$ to obtain $\mathbf{u}^k$ using splitting or generalized ADI schemes.

Our discussion in this chapter is organized as follows. Chap. 9.1 describes background on implicit schemes, truncation error, stability and convergence of discretizations. Chap. 9.2 describes iterative solvers, while Chap. 9.3 describes noniterative solvers. Chap. 9.4 describes the *parareal* method for solving a parabolic equation on a time interval $[0, T]$. It corresponds to a *multiple shooting* method on $[0, T]$, and is suited for applications to parabolic optimal control problems. Sample theoretical results are presented in Chap. 9.5.

9.1 Background

We consider the following parabolic equation:

$$\begin{cases} u_t + L\,u = f(x,t), & \text{in } \Omega \times (0,t) \\ u(x,0) \;\;\;= u_0(x), & \text{in } \Omega \\ u(x,t) \;\;\;= 0, & \text{on } \partial\Omega \times (0,t), \end{cases} \tag{9.1}$$

where $L\,u = -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)u$ denotes an underlying elliptic operator with $0 < a_0 \leq a(x)$. Except when specified otherwise, we shall assume that $\mathbf{b}(x) = \mathbf{0}$ and that the reaction coefficient $c(x) \geq 0$.

To obtain a spatial discretization of (9.1), let Ω_h denote a triangulation of Ω with grid size h and interior nodes $x_1, \ldots x_n$. If a finite element discretization is employed in space, let $u_h(x,t) \in V_h \subset H_0^1(\Omega)$ denote the finite element solution for each t. A Galerkin approximation of (9.1) will seek $u_h(x,t) \in V_h$:

$$\begin{cases} (u_{h,t}, v) + \mathcal{A}(u_h, v) = F(v), \qquad \forall v \in V_h \\ \qquad\qquad\qquad u_h(x,0) = I_h u_0(x), \end{cases}$$

where $(u,v) = \int_\Omega u\,v\,dx$ in the weak formulation above, and:

$$\begin{cases} \mathcal{A}(u,v) \equiv \int_\Omega \left(a(x)\nabla u \cdot \nabla v + (\mathbf{b}(x) \cdot \nabla u)\,v + c(x)\,u\,v\right)\,dx \\ \quad F(v) \equiv \int_\Omega f\,v\,dx. \end{cases}$$

This semi-discretization yields a stiff system of ordinary differential equations. Let $\mathbf{u}(t) \in \mathbb{R}^n$ denote a vector of nodal values of the discrete solution $u_h(x,t)$ with $(\mathbf{u})_i\,(t) = u_h(x_i,t)$ for $1 \leq i \leq n$. If $\{\phi_1, \ldots, \phi_n\}$ denotes a finite element nodal basis, define $G_{ij} = (\phi_i, \phi_j)$, $A_{ij} = \mathcal{A}(\phi_i, \phi_j)$ and $\mathbf{f}_i = (f, \phi_i)$. Then, the finite element semi-discretization will correspond to:

$$\begin{cases} G\,\mathbf{u}_t + A\,\mathbf{u} = \mathbf{f}(t) \\ \qquad\quad \mathbf{u}(0) = I_h u_0, \end{cases} \tag{9.2}$$

where $I_h u_0$ denotes the interpolation $(I_h u_0)_i = u_0(x_i)$ for $1 \leq i \leq n$.

Remark 9.1. If a finite difference discretization is employed in space, then matrix $G = I$ will be an identity matrix of size n, while $A\mathbf{u}$ will denote the finite difference discretization of the elliptic term $L\,u$. If $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \geq 0$ then matrix A will be symmetric positive definite for finite element and finite difference discretizations. In case $u(x,t) = z(x,t)$ on $\partial\Omega \times (0,t)$, then the forcing term $\mathbf{f}(t)$ must be replaced by $\mathbf{f}(t) - A_{IB}\,\mathbf{z}(t)$, where A_{IB} denotes the extended stiffness matrix involving boundary data and $\mathbf{z}(t) = I_{h,B}z(.,t)$ denotes interpolation of the boundary data onto nodes on $\partial\Omega$.

If $(\mathbf{u}_*(t))_i = u(x_i, t)$ denotes the restriction of the true solution $u(x,t)$ of (9.1) to the nodes $x_1, \ldots, x_n$, we define the truncation error $\mathbf{T}(t)$ as:

$$\begin{cases} G\,\mathbf{u}_{*\,t} + A\,\mathbf{u}_* = \mathbf{f}(t) + \mathbf{T}(t) \\ \qquad\qquad \mathbf{u}_*(0) = I_h\,u_0 + \mathbf{T}(0). \end{cases} \tag{9.3}$$

By construction $\mathbf{T}(0) = \mathbf{0}$. Subtracting (9.2) from (9.3), solving the resulting inhomogeneous equation for $\mathbf{e}(t) = \mathbf{u}_*(t) - \mathbf{u}(t)$ using Duhamel's principle (superposition), and estimating yields the following bounds [CO, AR3]:

$$\|\mathbf{e}(t)\| \;\leq\; \|\mathbf{T}(0)\|\,e^{c_1 t} + c_2\,t\,\left(e^{c_1\,t} - 1\right) \max_{0\leq s\leq t} \|\mathbf{T}(s)\|,$$

provided $\|e^{-G^{-1}\,A\,s}\| \leq e^{c_1\,s}$, for $c_1 > 0$ and $c_2 > 0$ independent of h. If $u(x,t)$ is sufficiently smooth, the truncation error will satisfy $\|\mathbf{T}(t)\| \leq Ch^2$.

To discretize (9.2) on a time interval $(0, t_*)$ let $\tau = (t_*/m)$ denote a time step, and let $t_k = k\,\tau$ for $0 \leq k \leq m$. An implicit or semi-implicit linear two step method [ST10, IS, SH, SH2, LA7]. will result in a system of linear equations of the following form at each time t_k:

$$\begin{cases} (G + \alpha\,\tau\,A)\,\mathbf{u}^{k+1} + C\,\mathbf{u}^k = \tilde{\mathbf{f}}^{k+1}, \quad \text{for } \; 0 \leq k \leq \;(m-1) \\ \qquad\qquad\qquad\qquad \mathbf{u}^0 = I_h u_0. \end{cases} \tag{9.4}$$

where $\mathbf{u}^k$ denotes the discrete solution at time $t_k = k\,\tau$. The θ-scheme, for instance, yields the following discretization for $0 \leq \theta \leq 1$:

$$(G + \tau\,\theta A)\,\mathbf{u}^{k+1} + (-G + \tau(1-\theta)A)\,\mathbf{u}^k = \tau\,\theta\,\mathbf{f}^{k+1} + \tau(1-\theta)\,\mathbf{f}^k.$$

We obtain the forward Euler method for $\theta = 0$, the Crank-Nicolson method for $\theta = \frac{1}{2}$, and the backward Euler method for $\theta = 1$. More general schemes may be found in [GE, SH, SH2, HA7, LA7, HA8].

9.1.1 Consistency

If $u(x,t)$ denotes the solution to (9.1) and $\left(\mathbf{u}_*^{(k)}\right)_i = u(x_i, t_k)$, we define the local truncation error $\mathbf{T}^{k+1}$ of scheme (9.4) at time t_{k+1} as:

$$\begin{cases} (G + \alpha\,\tau\,A)\,\mathbf{u}_*^{k+1} + C\,\mathbf{u}_*^k = \tilde{\mathbf{f}}^{k+1} + \mathbf{T}^{k+1} \quad \text{for } \; 0 \leq k \leq (m-1) \\ \qquad\qquad\qquad\qquad \mathbf{u}_*^0 = I_h u_0 + \mathbf{T}^0 \end{cases} \tag{9.5}$$

By construction $\mathbf{T}^0 = \mathbf{0}$. Discretization (9.4) will be said to be *consistent* if:

$$\tau^{-1}\,\|\mathbf{T}^k\| \to 0 \quad \text{as} \quad (h, \tau) \to (0, 0).$$

More generally, discretization (9.4) will be said to be accurate to order (q_1, q_2) if $\|\mathbf{T}^k\| \;\leq\; c\,\tau\;(h^{q_1} + \tau^{q_2})$, for some $c > 0$.

Remark 9.2. It can easily be verified that for the θ-scheme in time and a second order discretization in space, the local truncation error will satisfy:

$$\|\mathbf{T}^k\| \leq \begin{cases} C\,\tau\left(\tau^1 + h^2\right), & \text{if } \theta \neq 1/2 \\ C\,\tau\left(\tau^2 + h^2\right), & \text{if } \theta = 1/2. \end{cases}$$

Higher order discretizations of (9.1) may be constructed using more accurate spatial discretization $A\mathbf{u}$ of Lu, and higher order linear *multistep* discretization of $\mathbf{u}_t + A\mathbf{u} = \mathbf{f}$ in time [GE, SH, SH2, HA7, LA7, HA8]. A k step linear multistep method will typically require storage of the solution at k discrete times. The resulting scheme will be stable only if the eigenvalues of τA lie in the region of *stability* of the scheme, thereby imposing *constraints* on τ.

Remark 9.3. If a symmetric positive definite linear system is desirable, then a *semi-implicit* scheme may be employed. For instance, suppose $A = H + N$ where $H\mathbf{u}$ denotes the discretization of $-\nabla \cdot (a(x)\nabla u)$ and $N\mathbf{u}$ the discretization of $\mathbf{b}(x) \cdot \nabla u + c(x)u$. Then, the system of differential equations:

$$\mathbf{u}_t + H\mathbf{u} + N\mathbf{u} = \mathbf{f},$$

maybe discretized using an Adams-Moulton (implicit) scheme for the term $H\mathbf{u}$ and an Adams-Bashforth (explicit) scheme for the $N\mathbf{u}$ term. This will yield a linear multistep scheme, see [GE, SH, SH2, HA7, LA7, HA8], requiring storage of the solution at several discrete times. Care must be exercised to ensure that the resulting scheme is stable, as there will be constraints on admissible time steps τ. For instance, we may employ a backward Euler scheme for the $H\mathbf{u}$ term and a forward Euler scheme for the $N\mathbf{u}$ term resulting in:

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\tau} + H\mathbf{u}^{k+1} + N\mathbf{u}^k = \mathbf{f}^k, \quad \text{for } 0 \leq k \leq (m-1).$$

When H is symmetric positive definite, this scheme will be stable in the Euclidean norm $\|\cdot\|$ *provided*:

$$\|(I + \tau\,H)^{-1}(I - \tau\,N)\| \leq (1 + c_1\tau),$$

for some $c_1 > 0$ independent of τ and h.

9.1.2 Stability

Discretization (9.4) will be said to be stable in a norm $\|\cdot\|$ if its solution satisfies the following bound:

$$\|\mathbf{u}^{k+1}\| \leq (1 + c_1\tau)\|\mathbf{u}^k\| + c_2\|\tilde{\mathbf{f}}^k\|,$$

for arbitrary $\tilde{\mathbf{f}}_k$, where $c_1 > 0$ and $c_2 > 0$ are independent of h and τ.

Remark 9.4. If $\mathbf{b}(x) = \mathbf{0}$ and A is a symmetric positive definite matrix, then the θ-scheme will be stable in the Euclidean norm $\|\cdot\|$ and the mesh dependent norm $\|\cdot\|_{I+\alpha\tau A}$, see Chap. 9.5. If A is a diagonally dominant M-matrix, possibly nonsymmetric, then the θ-scheme will be stable in the maximum norm $\|\cdot\|_\infty$ provided:

$$\theta \leq \min_{1 \leq i \leq n} \frac{1}{(1-\theta) A_{ii}},$$

see Chap. 9.5. When $\mathbf{b}(x) \neq \mathbf{0}$, care must be exercised in the spatial discretization of $\mathbf{b}(x) \cdot \nabla u$, as a centered finite difference or traditional Galerkin discretization will be *unstable* [SO2, JO2]. Upwind finite difference or streamline diffusion discretizations may be employed if $a(x)$ is sufficiently small.

9.1.3 Lax Convergence Theorem

A discretization such as (9.4) will be said to be *convergent* if the norm of the error $\mathbf{e}^k = \mathbf{u}_*^k - \mathbf{u}^k$ goes to zero as $(h, \tau) \to (0, 0)$:

$$\max_{1 \leq k \leq m} \|\mathbf{e}^k\| \to 0 \text{ as } (h, \tau) \to (0, 0).$$

We have the following important convergence theorem due to Lax [RI].

Theorem 9.5. *If scheme (9.4) is consistent and stable, it will be convergent. In particular if a stable scheme is consistent of order* (q_1, q_2)*, then it will be convergent with the error satisfying:*

$$\|\mathbf{e}^k\| \leq c\, (h^{q_1} + \tau^{q_2}).$$

Proof. See [RI] or Chap. 9.5. □

9.2 Iterative Algorithms

At each discrete time t_k for $0 \leq k \leq (m-1)$, we must solve the linear system:

$$(G + \alpha\, \tau A)\mathbf{u}^{k+1} = \tilde{\mathbf{g}}^{k+1} = \tilde{\mathbf{f}}^{k+1} - C\mathbf{u}^k. \tag{9.6}$$

It will be sufficient to solve each system to truncation error, i.e., determine an approximate solution $\mathbf{w}^{k+1} \approx \mathbf{u}^{k+1}$ so that the residual is the same magnitude as the truncation error $O(\|\mathbf{T}^{k+1}\|)$. Most Schwarz and Schur complement preconditioners from preceding chapters can be employed, however, it may be possible to *omit coarse space* correction *provided* the time step τ is sufficiently small. Omitting coarse space correction will help to reduce computational costs and improve parallelizability of the solvers.

9.2.1 Schwarz Algorithms

Let $\Omega_1^*, \ldots, \Omega_p^*$ form an overlapping covering of Ω with shape regular subdomains of size h_0 having overlap $\beta\, h_0$. Let Ω_h denote a quasiuniform triangulation of Ω with grid size h, and for $1 \leq i \leq p$ let R_i denote the pointwise nodal restriction map onto interior nodes in Ω_i^* in the local ordering. Then, the action of the inverse of the additive Schwarz preconditioner M for $(G + \alpha\, \tau\, A)$, without coarse space correction, will have the matrix form:

$$M^{-1} = \sum_{i=1}^{p} R_i^T \left(R_i (G + \alpha\, \tau\, A) R_i^T \right)^{-1} R_i.$$

The *sequential* Schwarz preconditioner will yield more rapid convergence (with multicoloring for parallelizability). If A is nonsymmetric, then GMRES acceleration will be necessary. The following theoretical result indicates that if τ is sufficiently small, then coarse space correction can be omitted in Schwarz algorithms without adverse deterioration in convergence rate provided see [CA, CA3] and [KU3, KU5, KU6].

Lemma 9.6. *Let* $\mathbf{b}(x) = \mathbf{0}$ *and* $c(x) \geq 0$ *in (9.1), and let* $\Omega_1^*, \ldots, \Omega_p^*$ *have overlap* $\beta\, h_0$*. Then, the partition parameters* K_0 *and* K_1 *associated with the subspaces* $V_h(\Omega_i^*) \subset H_0^1(\Omega_i^*)$ *for* $1 \leq i \leq p$ *will satisfy:*

$$\begin{aligned} K_0 &\leq C(1 + \tau\, \|a\|_\infty\, \beta^{-2}\, h_0^{-2}) \\ K_1 &\leq C, \end{aligned}$$

without a coarse space, for some $C > 0$ *independent of* h*,* h_0*,* τ*,* $\|a\|_\infty$*, where:*

$$\|a\|_\infty = \max_{x \in \overline{\Omega}} |a(x)|.$$

Proof. See [CA, CA3] and Chap. 9.5. □

The preceding result suggests that provided $\tau\, h_0^{-2}$ is uniformly bounded, the *coarse space correction* term may be *omitted* without adverse deterioration in the convergence of additive or multiplicative Schwarz algorithms.

Remark 9.7. Since $\|\mathbf{u}^k - \mathbf{u}^{k-1}\|$ will formally be accurate to $O(\tau)$, we can employ $\mathbf{u}^{k-1}$ as a starting guess in the Schwarz iteration to solve:

$$(G + \alpha\, \tau\, A)\mathbf{u}^k = \tilde{\mathbf{g}}^k = \tilde{\mathbf{f}}^k - C\mathbf{u}^{k-1}.$$

The Schwarz iteration can be applied till the residual norm is $O(\tau(h^{q_1} + \tau^{q_2}))$. If a coarse space correction is included, it will speed up the convergence of

Schwarz algorithms, particularly if $a(x)$ has large jump discontinuities. In practice, it may be desirable to test the computational times with and without coarse space correction.

9.2.2 Schur Complement Algorithms

Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of Ω with common interface B. Traditional Schur complement algorithms can be employed to solve $(G + \alpha\,\tau\,A)\,\mathbf{u}^k = \tilde{\mathbf{g}}^k$. If we block partition each nodal vector as:

$$\mathbf{v} = \left(\mathbf{v}_I^T, \mathbf{v}_B^T\right)^T, \quad \text{where} \quad \mathbf{v}_I = \left(\mathbf{v}_I^{(1)^T}, \ldots, \mathbf{v}_I^{(p)^T}\right)^T,$$

then linear system (9.6) will have the following block structure:

$$\begin{bmatrix} G_{II} + \alpha\tau\, A_{II} & G_{IB} + \alpha\tau\, A_{IB} \\ G_{BI} + \alpha\tau\, A_{BI} & G_{BB} + \alpha\tau\, A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^k \\ \mathbf{u}_B^k \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{g}}_I^k \\ \tilde{\mathbf{g}}_B^k \end{bmatrix},$$

where $A_{II} = \text{blockdiag}(A_{II}^{(1)}, \ldots, A_{II}^{(p)})$. The Schur complement system is:

$$S(\tau)\,\mathbf{u}_B^k = \mathbf{g}_B^k,$$

where

$$\begin{aligned} S(\tau) &= (G_{BB} + \alpha\,\tau\,A_{BB}) - (G_{BI} + \alpha\tau A_{BI})(G_{II} + \alpha\tau A_{II})^{-1}(G_{IB} + \alpha\tau A_{IB}) \\ \mathbf{g}_B^k &= \tilde{\mathbf{g}}_B^k - (G_{BI} + \alpha\,\tau A_{BI})(G_{II} + \alpha\,\tau A_{II})^{-1}\tilde{\mathbf{g}}_I^k. \end{aligned}$$

Most of the Schur complement preconditioners from preceding chapters can be employed to precondition $S(\tau)$. For instance, if $\mathcal{G}_1, \ldots, \mathcal{G}_q$ are globs of B with corresponding pointwise nodal restriction maps $\mathcal{R}_i$ from nodes on B to nodes on $B^{(i)} = \partial\Omega_i \backslash \partial\Omega$, the action of the inverse of the Neumann-Neumann preconditioner without coarse space correction will be:

$$M^{-1} = \sum_{i=1}^{p} \mathcal{R}_i \left(S^{(i)}(\tau)\right)^{-1} \mathcal{R}_i,$$

where each $S^{(i)}(\tau)$ denotes a (non-singular) subdomain Schur complement matrix. This will yield a more parallelizable algorithm. As with Schwarz preconditioners, if a time-step constraint of the form $\tau \leq \, c\,h_0^2$ holds, here h_0 denotes the subdomain diameter, the above preconditioner should yield a condition number bound with poly-logarithmic dependence on the mesh parameters [FA14, FA15]. If $\mathcal{R}_0$ denotes a coarse space weighted restriction matrix and $S_0(\tau) = \mathcal{R}_0 S(\tau) \mathcal{R}_0^T$, then the coarse space correction term $\mathcal{R}_0 \left(S_0(\tau)\right)^{-1} \mathcal{R}_0$ can be added to the above Neumann-Neumann preconditioner.

9.3 Non-Iterative Algorithms

Given a discretization (9.4) of (9.1), a non-iterative algorithm determines an approximate solution of (9.4) *without iteration.* Such methods, motivated by the classical ADI and splitting methods, solve a *modified* discretization of (9.4), see [DO9, PE, BA11, LE14, KO, SA5, YA, ST11, GL7]:

$$\begin{cases} \tilde{H}\,\mathbf{w}^{k+1} + \tilde{C}\,\mathbf{w}^{k} = \tilde{\mathbf{f}}^{k+1}, & \text{for } 0 \le k \le (m-1) \\ \qquad\qquad \mathbf{w}^{0} = \mathbf{u}^{0}, \end{cases} \tag{9.7}$$

where $\tilde{H} \approx (G + \alpha\,\tau\,A)$ and $\tilde{C} \approx C$ are chosen so that the linear system $\tilde{H}\,\mathbf{w}^{k+1} = \tilde{\mathbf{f}}^{k+1} - \tilde{C}\,\mathbf{w}^{k}$ can be solved without iteration. Additionally, the modified discretization (9.7) must be *consistent* to the same order as the original discretization (9.4), and *stable.* Typically, however, some constraints will be required on τ to ensure stability of the modified scheme (9.7).

In this section, we loosely group non-iterative methods by modification of Schwarz or Schur complement methods, or based on domain decomposition operator splittings. For simplicity, we consider a finite difference discretization of (9.1), so that $G = I$, and employ a partition of unity.

Discrete Partition of Unity. Given subdomains $\Omega_1^*, \ldots, \Omega_p^*$ which form an overlapping decomposition of Ω, we define a discrete partition of unity subordinate to $\Omega_1^*, \ldots, \Omega_p^*$ as follows. Let n denote the number of interior nodes in Ω and n_i^* the number of nodes in Ω_i^*, with R_i denoting a nodal restriction matrix of size $n_i^* \times n$ which restricts a vector of nodal values on Ω to nodal values on Ω_i^*. We let $D^{(i)}$ denote a diagonal matrix of size n_i^*, with positive diagonal entries such that:

$$I = \sum_{i=1}^{p} R_i^T D^{(i)} R_i.$$

For instance, if $x_l^{(i)}$ denotes the l'th node in the local ordering of nodes within Ω_i^*, define $(D^{(i)})_{ll}$ as $1/\text{degree}(x_l^{(i)})$ where $\text{degree}(x_l^{(i)})$ denotes the number of subdomains Ω_j^* to which node $x_l^{(i)}$ belongs to, for $1 \le i \le p$.

Similarly, given *non-overlapping* subdomains $\Omega_1, \ldots, \Omega_p$, we shall employ a discrete partition of unity subordinate to $\overline{\Omega}_1, \ldots, \overline{\Omega}_p$ satisfying:

$$I = \sum_{i=1}^{p} \Phi_i,$$

where each $\Phi_i = R_{\overline{\Omega}_i}^T D^{(\overline{\Omega}_i)} R_{\overline{\Omega}_i}$ is a diagonal matrix of size n with non-negative diagonal entries, where $R_{\overline{\Omega}_i}$ denotes a nodal restriction matrix of size $n_i \times n$ which restricts a vector in Ω into nodal values on $\overline{\Omega}_i$ (with n_i nodes in $\overline{\Omega}_i$) and $D^{\overline{\Omega}_i}$ is a diagonal matrix of size n_i such that if $y_l^{(i)}$ denotes the l'th node in the local ordering within $\overline{\Omega}_i$, then $(D^{\overline{\Omega}_i})_{ll}$ is $1/\text{degree}(y_l^{(i)})$ where $\text{degree}(y_l^{(i)})$ denotes the number of subdomains $\overline{\Omega}_j^*$ to which node $y_l^{(i)}$ belongs to.

9.3.1 Non-Iterative Schwarz Algorithms

The first non-iterative algorithm we describe is due to [KU3, KU5, KU6, ME6]. At each time step, this algorithm determines an approximate solution to (9.6) by solving appropriately chosen problems on p *overlapping subdomains*. This algorithm is motivated by the property that the entries $(I + \alpha\,\tau\,A)^{-1}_{ij}$ of the discrete Green's function decay rapidly away from the diagonal, as $\tau \to 0^+$. This suggests using a partition of unity to decompose the forcing term into the subdomains, and to solve on overlapping subdomains with sufficiently large overlap, to obtain an approximate solution to a specified accuracy ϵ.

More specifically, let $\Omega_1, \ldots, \Omega_p$ form a nonoverlapping decomposition of Ω into subdomains of size h_0. Given an overlap parameter $\beta > 0$, for each subdomain define $\Omega_i^* \supset \Omega_i$ as an extended subdomain:

$$\Omega_i^* \equiv \{x \in \Omega : \, \mathrm{dist}(x, \Omega_i) < \beta\,h_0\}\,, \quad \text{for } 1 \le i \le p.$$

Using barrier functions and comparison functions as described in Chap. 15, it can be shown that for sufficiently small h, the entries $(I + \alpha\,\tau\,A)^{-1}_{ij}$ in the ith row of the discrete Green's function matrix decay rapidly with increasing distance between nodes x_i and x_j. For instance, if $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \ge 0$, given $0 < \epsilon \ll 1$, it can be shown that [KU3, KU5, KU6]:

$$\left|\left((I + \alpha\,\tau\,A)^{-1}\right)_{ij}\right| \le \epsilon, \text{ when } |x_i - x_j| \ge c_*\,\sqrt{\tau}\log(\epsilon^{-1}). \tag{9.8}$$

As a result, if $\mathbf{r}_i \in \mathbb{R}^n$ has support in $\overline{\Omega}_i$ then $\mathbf{w}_i = (I + \alpha\,\tau\,A)^{-1}\mathbf{r}_i$ will be of "small" magnitude at each node x_l for which $\mathrm{dist}(x_l, \overline{\Omega}_i) \ge c_*\,\sqrt{\tau}\log(\epsilon^{-1})$:

$$|(\mathbf{w}_i)_l| = |\sum_j \left((I + \alpha\,\tau A)^{-1}\right)_{lj} (\mathbf{r}_i)_j| \le \tilde{c}\,\epsilon\,\|\mathbf{r}_i\|,$$

for some $\tilde{c} > 0$, i.e., $(\mathbf{w}_i)_l$ will be $O(\epsilon)$. Thus, $\mathbf{w}_i = (I + \alpha\,\tau\,A)^{-1}\mathbf{r}_i$ may be approximated by $\mathbf{v}_i = R_i^T(I + \alpha\,\tau\,A_i)^{-1}R_i\,\mathbf{r}_i$ where $\mathrm{supp}(\mathbf{v}_i) \subset \Omega_i^*$ and $A_i \equiv R_i A R_i^T$ and $\beta\,h_0 \ge c_*\,\sqrt{\tau}\log(\epsilon^{-1})$. To approximate $(I + \alpha\,\tau\,A)^{-1}\tilde{\mathbf{g}}_k$, use the partition of unity to decompose $\tilde{\mathbf{g}}_k = \mathbf{r}_1 + \cdots + \mathbf{r}_p$ with $\mathbf{r}_i = \Phi_i\,\tilde{\mathbf{g}}_k$, where $\mathrm{supp}(\mathbf{r}_i) \subset \overline{\Omega}_i$, and apply the preceding approximation on each term:

$$(I + \alpha\,\tau\,A)^{-1}\tilde{\mathbf{g}}_k \approx \sum_{i=1}^{p} R_i^T(I + \alpha\,\tau\,A_i)^{-1}R_i\,\mathbf{r}_i = \sum_{i=1}^{p} R_i^T\mathbf{v}_i,$$

where we need to solve:

$$(I + \alpha\,\tau\,A_i)\mathbf{v}_i = R_i\mathbf{r}_i, \quad \text{for } 1 \le i \le p.$$

Below, we summarize the non-iterative algorithm of [KU3, KU5, KU6] for approximately solving $(I + \alpha\,\tau\,A)\,\mathbf{u}^k = \tilde{\mathbf{g}}_k = \tilde{\mathbf{f}}^k - C\mathbf{u}^{k-1}$ to accuracy $O(\epsilon)$. We let $\mathbf{w}^k \approx \mathbf{u}^k$ denote the non-iterative solution at time t_k.

Algorithm 9.3.1 *(Noniterative Algorithm of [KU3, KU5, KU6])*

```
Input $\mathbf{w}^0 = \mathbf{u}^0$, $\{\tilde{\mathbf{f}}^k\}$
1. For $k = 1, \ldots, m$ do:
2.     Compute $\tilde{\mathbf{g}}^k = \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1}$
3.     Update:
           $$\mathbf{w}^k = \sum_{i=1}^{p} R_i^T (I + \alpha\,\tau\,A_i)^{-1} R_i\,\Phi_i \tilde{\mathbf{g}}^k.$$
4. Endfor
Output: $\mathbf{w}^k \approx \mathbf{u}^k$
```

Parallel implementation of the above algorithm is described in [ME6]. The following error bound will hold, see [KU3, KU5, KU6].

Lemma 9.8. *If the extended subdomains Ω_i^* have overlap of:*

$$\beta\,h_0 = \mathcal{O}(\sqrt{\tau}\log(\epsilon^{-1})),$$

then the error will satisfy:

$$\|\mathbf{w}^k - \mathbf{u}^k\| \leq \mathcal{O}(\epsilon), \quad \text{for} \quad 1 \leq k \leq m.$$

Remark 9.9. If $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \geq 0$, the resulting scheme will be convergent for sufficiently small ϵ. If the original discretization is $\mathcal{O}(\tau^2 + h^2)$ accurate, then for the choice $\tau = h$ and $\epsilon = h^3$, the overlap should be approximately $\mathcal{O}(\sqrt{h}\log(h))$. While if the original discretization is $\mathcal{O}(\tau + h^2)$ accurate, then for the choice $\tau = h^2$ and $\epsilon = h^4$, the overlap should be $\mathcal{O}(h\log(h))$.

We next describe an *improved* version of the preceding non-iterative method, see [BL3, CH22]. This algorithm employs the solution $\mathbf{u}^{k-1}$ from the preceding time to compute an approximation to $\mathbf{u}^k$, and is motivated by the following *decay* result on the influence of the boundary data as $\tau \to 0^+$. The next result is stated for a finite difference discretization on Ω_i^*.

Lemma 9.10. *Suppose the following conditions hold.*

1. *Let $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \geq 0$, and let $\Omega_1^*, \ldots, \Omega_p^*$ have overlap $\beta\,h_0$.*
2. *Let $(I + \alpha\,\tau\,A)$ be an M-matrix.*
3. *Let $(I + \alpha\,\tau\,A)\mathbf{u}^k = \tilde{\mathbf{g}}^k$ and*

$$\begin{cases} ((I + \alpha\,\tau\,A)\mathbf{v}_i)_l = \left(\tilde{\mathbf{g}}^k\right)_l, & \text{for } x_l \in \Omega_i^* \\ (\mathbf{v}_i)_l = \left(\mathbf{u}^{k-1}\right)_l, & \text{for } x_l \in \partial\Omega_i^*. \end{cases}$$

Then, for sufficiently small h, the following bound will hold:

$$\max_{x_l \in \overline{\Omega}_i} |(\mathbf{u}^{(k)})_l - (\mathbf{v}_i)_l| \leq e^{-(\alpha\,\beta\,h_0/\sqrt{\tau})} \max_{x_l \in \partial\Omega_l^*} |(\mathbf{u}^k)_l - (\mathbf{u}^{k-1})_l|,$$

for some $\alpha > 0$ independent of h and τ.

Proof. See [LI7, BL3, CH22] and Chap. 15. □

Since $\max_{x_l \in \partial \Omega_l^*} |(\mathbf{u}^k)_l - (\mathbf{u}^{k-1})_l| = O(\tau)$, the preceding result can be applied to estimate the error $\mathbf{u}^k - \mathbf{v}_i$ within $\overline{\Omega}_i$ as follows:

$$\max_{x_l \in \overline{\Omega}_i} |(\mathbf{u}^k)_l - (\mathbf{v}_i)_l| \leq C\, \tau\, e^{-\alpha\, \beta\, h_0 / \sqrt{\tau}}.$$

So if an approximate solution is desired to accuracy ϵ in $\overline{\Omega}_i$, we must choose the overlap $\beta\, h_0$ depending on the time step τ so that:

$$e^{-(\alpha\, \beta\, h_0/\sqrt{\tau})}\, \tau \leq \epsilon \implies \beta\, h_0 \geq C\, \alpha \sqrt{\tau}\, \log(\tau/\epsilon).$$

Thus, if $\epsilon = \tau(h^2 + \tau^2) = O(\tau^3)$, we require $\beta\, h_0 \geq C\, \alpha\, \sqrt{\tau}\, \log(\tau^{-2})$. The algorithm of [BL3, CH22] combines the local solutions $\mathbf{v}_i$ on each $\overline{\Omega}_i$ using the discrete partition of unity $\Phi_1, \ldots, \Phi_p$ subordinate to $\{\overline{\Omega}_i\}$. Alg. 9.3.2 below summarizes the computation of the approximate solution of:

$$(I + \alpha\, \tau\, A)\, \mathbf{u}^k = \tilde{\mathbf{g}}^k = \tilde{\mathbf{f}}^k - C\, \mathbf{u}^{k-1},$$

employing the notation $\mathbf{w}^k \approx \mathbf{u}^k$ for the non-iterative solution.

Algorithm 9.3.2 *(Noniterative Algorithm of [BL3, CH22])*
Input $\mathbf{w}^0 = \mathbf{u}^0$, $\{\tilde{\mathbf{f}}^k\}$

```
1. For $k = 1, \ldots, m$ do:
2.     Compute $\tilde{\mathbf{g}}^k = \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1}$
3.     Compute the residual $\mathbf{r}^k = \tilde{\mathbf{g}}^k - (I + \alpha\, \tau\, A)\mathbf{w}^{k-1}$
4.     In parallel solve:
           $\mathbf{w}^k = \mathbf{w}^{k-1} + \sum_{i=1}^{p} \Phi_i\, R_i^T (I + \alpha\, \tau\, A_i)^{-1} R_i\, \mathbf{r}^k.$
5. Endfor
```

Output: $\mathbf{w}^k$

A *sequential* version of the above algorithm may also be employed.

Remark 9.11. In implementations, the contraction factor $e^{-\alpha\, \beta\, h_0/\sqrt{\tau}}$ can be estimated computationally. On each $\Omega_i^* \supset \Omega_i$, define a subdomain contraction factor κ_i as $\kappa_i \equiv \max_{x_l \in \overline{\Omega}_i} |(\mathbf{v}_i)_l|$, where $\mathbf{v}_i$ is the solution to:

$$\begin{cases} ((I + \alpha\, \tau\, A)\mathbf{v}_i)_l = 0, & \text{for } x_l \in \Omega_i^*, \\ (\mathbf{v}_i)_l = 1, & \text{for } x_l \in \partial \Omega_i^* \cap \Omega, \\ (\mathbf{v}_i)_l = 0, & \text{for } x_l \in \partial \Omega_i^* \cap \partial \Omega. \end{cases} \tag{9.9}$$

The global error reduction factor can then be estimated as:

$$e^{-\alpha\, \beta\, h_0/\sqrt{\tau}} \leq \max\{\kappa_1, \ldots, \kappa_p\} < 1,$$

see Chap. 15.

9.3.2 Non-Iterative Schur Complement Algorithms

Next, we describe non-iterative Schur complement based algorithms for (9.6). We describe algorithms of [KU3, KU5, KU6, DA4, DA5, ZH5], and alternative heuristic methods. Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of Ω with interface B. Corresponding to the nodes within these subregions, we block partition a nodal vector as $\mathbf{v} = \left(\mathbf{v}_I^T, \mathbf{v}_B^T\right)^T$ where $\mathbf{v}_I = \left(\mathbf{v}_I^{(1)^T}, \ldots, \mathbf{v}_I^{(p)^T}\right)^T$. This yields the following block structure for (9.6)

$$\begin{bmatrix} I + \alpha\tau\, A_{II} & \alpha\tau\, A_{IB} \\ \alpha\tau\, A_{BI} & I + \alpha\tau\, A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^k \\ \mathbf{u}_B^k \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{g}}_I^k \\ \tilde{\mathbf{g}}_B^k \end{bmatrix}, \tag{9.10}$$

where $A_{II} = \text{blockdiag}(A_{II}^{(1)}, \ldots, A_{II}^{(p)})$. The reduced system for $\mathbf{u}_B^k$ is:

$$S(\tau)\, \mathbf{u}_B^k = \mathbf{g}_B^k, \tag{9.11}$$

where the Schur complement matrix $S(\tau)$ and forcing $\mathbf{g}_B^k$ are:

$$\begin{cases} S(\tau) = (I + \alpha\,\tau\, A_{BB}) - \alpha^2\,\tau^2 A_{BI}(I + \alpha\,\tau\, A_{II})^{-1} A_{IB} \\ \mathbf{g}_B^k = \tilde{\mathbf{g}}_B^k - \alpha\,\tau\, A_{BI}(I + \alpha\,\tau A_{II})^{-1}\, \tilde{\mathbf{g}}_I^k. \end{cases}$$

An approximate solution to system (9.10) can be obtained *non-iteratively* by solving (9.11) approximately using a non-iterative method, yielding $\mathbf{w}_B^k \approx \mathbf{u}_B^k$. Using $\mathbf{w}_B^k \approx \mathbf{u}_B^k$, we may update $\mathbf{w}_I^k \approx \mathbf{u}_I^k$ by solving:

$$(I + \alpha\,\tau\, A_{II})\mathbf{w}_I^k = \tilde{\mathbf{g}}_I^k - \alpha\tau\, A_{IB}\mathbf{w}_B^k. \tag{9.12}$$

Since $(I + \alpha\,\tau\, A_{II})$ is block diagonal, this can be implemented in *parallel.* (It will be advantageous to expand $\mathbf{u}_I^k = \mathbf{u}_I^{k-1} + \mathbf{v}_I^k$ and $\mathbf{u}_B^k = \mathbf{u}_B^{k-1} + \mathbf{v}_B^k$, to form the residual equation and determine the updates $\mathbf{v}_I^k$ and $\mathbf{v}_B^k$).

Explicit-Implicit Algorithms. Explicit-implicit methods approximate either the Dirichlet data or Neumann flux on each subdomain interface $B^{(i)}$ using an *explicit* scheme, while the other components of $(\mathbf{u}_I^{k^T}, \mathbf{u}_B^{k^T})^T$ are approximated using an *implicit* scheme. For instance, the algorithm of [KU3, KU5, KU6] seeks an approximation $\mathbf{w}_B^k \approx \mathbf{u}_B^k$ on the interface B at time t_k by applying a conditionally stable *explicit* scheme with time step $\tau_e \ll \tau$ and having the same order temporal accuracy as the original implicit scheme, with time step $\tau_e = \tau/N_e$. An approximation $\mathbf{w}_I^k$ of $\mathbf{u}_I^k$ can then be obtained by solving (9.12). The resulting solver thus combines stable implicit and explicit discretizations. To reduce computational costs, the explicit scheme should only be applied to compute the solution at space-time grid points $(x_r, t_{k-1} + l\,\tau_e)$ for $l = 1, \ldots, N_e$ which lie in the numerical domain of dependence of nodes in B at time $t_k = t_{k-1} + N_e\,\tau_e$. Such a region will be the union of space-time cones emanating from (x_l, t_k) towards t_{k-1} for $x_l \in B$.

If the implicit scheme is 2nd order in τ (such as in Crank-Nicolson), and the explicit scheme is 1st order (as in forward Euler), we must choose $\tau_e = O(\tau^2)$ to preserve the original local truncation error.

An explicit-implicit scheme may also update the Neumann flux on each subdomain boundary [DA4, DA5, DA6]. We illustrate such a scheme in the two subdomain case. Suppose $\mathbf{u}_I^{k,(i)}$ and $\mathbf{u}_B^{k,(i)}$ denote restrictions of $\mathbf{u}_I^k$ and $\mathbf{u}_B^k$ to nodes in Ω_i and $B^{(i)}$, then $\tilde{\mathbf{g}}_B^{k,(i)} = \alpha\tau\, A_{BI}^{(i)}\mathbf{u}_I^{k,(i)} + (I + \alpha\tau A_{BB}^{(i)})\mathbf{u}_B^{k,(i)}$ will represent the local Neumann flux of $\mathbf{u}^k$ on $B^{(i)}$. If the flux $\tilde{\mathbf{g}}_B^{k,(1)}$ at time t_k is approximated by averaging the local fluxes of $\mathbf{u}^{k-1}$ on $B^{(i)}$ at time t_{k-1}, then we may approximate $\mathbf{w}_I^{k,(i)} \approx \mathbf{u}_I^{k,(i)}$ and $\mathbf{w}_B^{k,(i)} \approx \mathbf{u}_B^{k,(i)}$ as follows:

$$
\begin{aligned}
&\tilde{\mathbf{g}}_B^{k,(1)} \approx \tfrac{1}{2}\textstyle\sum_{i=1}^{2}(-1)^{i+1}\left(\tau\, A_{BI}^{(1)}\mathbf{w}_I^{k-1,(1)} + (I + \alpha\tau A_{BB}^{(1)})\mathbf{w}_B^{k-1,(1)}\right)\\
&\begin{bmatrix} I + \alpha\tau\, A_{II}^{(i)} & \alpha\tau\, A_{IB}^{(i)} \\ \alpha\tau\, A_{BI}^{(i)} & I + \alpha\tau\, A_{BB}^{(i)} \end{bmatrix}
\begin{bmatrix} \mathbf{w}_I^{k,(i)} \\ \mathbf{w}_B^{k,(i)} \end{bmatrix}
=
\begin{bmatrix} \tilde{\mathbf{g}}_I^{k,(i)} \\ (-1)^{i+1}\,\tilde{\mathbf{g}}_B^{k,(1)} \end{bmatrix}
\quad \text{for } 1 \le i \le 2.
\end{aligned}
$$

A different averaging operator is employed in [DA4, DA5, DA6]. However, the global error in this scheme deteriorates [DA6]. Alternative conditionally stable Schur complement predictor-corrector schemes are described in [ZH5].

Remark 9.12. Generally, caution must be exercised in employing power series expansions of matrix $S(\tau)$ to formulate modified schemes, as the following example illustrates. Note that $S(\tau)$ has the following formal expansion in τ:

$$
S(\tau) = I + \alpha\,\tau\, A_{BB} - \alpha^2\tau^2 A_{BI}A_{IB} + \mathcal{O}(\tau^3). \tag{9.13}
$$

However, for finite difference discretizations, the diagonal entries in A_{BB} will be $O(h^{-2})$ while $A_{BI}A_{IB}$ have entries $O(h^{-4})$, making formal truncation not meaningful. If we erroneously "truncate" $S(\tau) \approx (I + \alpha\,\tau\, A_{BB})$, we may obtain the following formal non-iterative "approximation" $\mathbf{w}_B^k \approx \mathbf{u}_B^k$:

$$
(I + \alpha\,\tau\, A_{BB})\mathbf{w}_B^k = \mathbf{g}_B^k + \alpha^2\,\tau^2 A_{BI}(I + \alpha\,\tau\, A_{II})^{-1}A_{IB}\,\mathbf{w}_B^{k-1}.
$$

The resulting Scheme will be stable, however, its local truncation error will be quite large, due to the erroneous truncation. To verify stability, note that:

$$
0 \le S(\tau) = (I + \alpha\,\tau\, A_{BB}) - \alpha^2\,\tau^2 A_{IB}^T(I + \alpha\,\tau\, A_{II})^{-1}A_{IB} \le (I + \alpha\,\tau\, A_{BB})
$$

yielding that $(I + \alpha\,\tau\, A_{BB})^{-1} \le S(\tau)^{-1}$ in terms of quadratic forms. Define $F \equiv \alpha\tau\, A_{IB}^T(I + \alpha\,\tau\, A_{II})^{-1}$. Then the block LU factorization of $(I + \alpha\,\tau\, A)$:

$$
(I + \alpha\,\tau\, A) = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}\begin{bmatrix} I + \alpha\,\tau\, A_{II} & 0 \\ 0 & S(\tau) \end{bmatrix}\begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^T,
$$

will yield the following relations between the quadratic forms:

$$\begin{aligned}
(I + \alpha\,\tau\,A)^{-1} &= \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^{-T} \begin{bmatrix} (I + \alpha\,\tau\,A_{II})^{-1} & 0 \\ 0 & S(\tau)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^{-1} \\
&\geq \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^{-T} \begin{bmatrix} (I + \alpha\,\tau\,A_{II})^{-1} & 0 \\ 0 & (I + \alpha\,\tau\,A_{BB})^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^{-1} \equiv M^{-1}.
\end{aligned}$$

Here matrix M^{-1} defines the block matrix approximation of $(I + \alpha\,\tau\,A)^{-1}$. This will ensure stability of the truncated Schur complement scheme provided the original scheme is stable, however, this scheme is expected to be inaccurate. This scheme corresponds to one symmetrized block Gauss-Seidel iteration for (9.10) with starting guess from the previous time step.

Schur Complement-Operator Splitting Method. We shall next outline a *heuristic* non-iterative algorithm for solving the Schur complement system (9.11) approximately, using the generalized ADI algorithm. The reader is referred to the next section for a more general discussion of operator splittings and the generalized ADI algorithm [DO9, PE, DO10, DO12]. The algorithm we describe will be based on a *splitting* of the Schur complement matrix $S(\tau)$, motivated by the following algebraic identity:

$$\begin{aligned}
S(\tau) &= I + \alpha\,\tau\,A_{BB} - \alpha^2\tau^2 A_{BI}(I + \alpha\,\tau\,A_{II})^{-1}A_{IB} \\
&= I + \alpha\,\tau\left(A_{BB} - A_{BI}(\tfrac{I}{\alpha\tau} + A_{II})^{-1}A_{IB}\right) \\
&\equiv I + \alpha\,\tau\,\overline{S}(\tau),
\end{aligned}$$

where $\overline{S}(\tau) \equiv A_{BB} - A_{BI}(\frac{I}{\alpha\,\tau} + A_{II})^{-1}A_{IB}$. Matrix $\overline{S}(\tau)$ denotes the Schur complement associated with the symmetric *positive definite* matrix $\overline{A}(\tau)$:

$$\overline{A}(\tau) = \begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{bmatrix} + \frac{1}{\alpha\,\tau}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

Employing the subassembly identity for the Schur complement matrix, based on the non-overlapping decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, and decomposing the interface B into the globs $B^{(1)}, \ldots, B^{(p)}$, where $B^{(i)} = \partial\Omega_i \cap \Omega$, and using the nodal restriction matrix $\mathcal{R}_i$ which maps a vector of nodal values on B to nodal values on $B^{(i)}$, we obtain a matrix splitting of $\overline{S}(\tau)$ as:

$$\begin{cases} \overline{S}(\tau) = \sum_{i=1}^{p} \mathcal{R}_i^T \left(A_{BB}^{(i)} - A_{BI}^{(i)}(\frac{I}{\alpha\,\tau} + A_{II}^{(i)})^{-1}A_{IB}^{(i)}\right)\mathcal{R}_i \\ \qquad\;\; \equiv \sum_{i=1}^{p} \mathcal{R}_i^T \overline{S}^{(i)}(\tau)\mathcal{R}_i, \end{cases} \tag{9.14}$$

where $\overline{S}^{(i)}(\tau) \equiv \left(A_{BB}^{(i)} - A_{BI}^{(i)}(\frac{I}{\alpha\,\tau} + A_{II}^{(i)})^{-1}A_{IB}^{(i)}\right)$. Each matrix $\overline{S}^{(i)}(\tau)$ will be symmetric and *positive definite*, while $H_i \equiv \mathcal{R}_i^T \overline{S}^{(i)}(\tau)\mathcal{R}_i$ will be symmetric *positive semi-definite.* In terms of the matrices H_i, the splitting (9.14) becomes $\overline{S}(\tau) = (H_1 + \cdots + H_p)$, while the Schur complement system (9.11) becomes:

$$\left(I + \alpha\,\tau\,\overline{S}(\tau)\right)\,\mathbf{u}_B^k = \left(I + \alpha\,\tau\,(H_1 + \cdots + H_p)\right)\mathbf{u}_B^k \;=\; \mathbf{g}_B^k.$$

The generalized ADI algorithm for *approximately* solving the above system, applies one block Gauss-Seidel iteration to $(I + \alpha\,\tau\,(H_1 + \cdots + H_p))\;\mathbf{u}_B^k = \mathbf{g}_B^k$ using $\mathbf{w}_B^{k-1} \approx \mathbf{u}_B^{k-1}$ as a starting guess, as described next in matrix form.

Re-arranging the block terms in $(I + \alpha\,\tau\,(H_1 + \cdots + H_p))\;\mathbf{u}_B^k = \mathbf{g}_B^k$, will trivially yield the following equivalent block linear system:

$$\begin{bmatrix} (I + \alpha\,\tau\,H_1) & \alpha\tau H_2 & \cdots & \alpha\tau H_p \\ \alpha\tau H_1 & (I + \alpha\tau H_2) & \cdots & \alpha\tau H_p \\ \vdots & \vdots & \ldots & \vdots \\ \alpha\tau H_1 & \alpha\tau H_2 & \cdots & (I + \alpha\tau H_p) \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_p \end{bmatrix} = \begin{bmatrix} \mathbf{g}_B^k \\ \mathbf{g}_B^k \\ \vdots \\ \mathbf{g}_B^k \end{bmatrix}, \qquad (9.15)$$

with solution $\mathbf{v}_1 = \cdots = \mathbf{v}_p = \mathbf{u}_B^k$. The generalized ADI method corresponds to one sweep of a block Gauss-Seidel iteration to solve the above block system, using $\mathbf{v}_1 = \cdots = \mathbf{v}_p = \mathbf{w}_B^{k-1}$ as a starting guess, and defining $\mathbf{w}_B^k \equiv \mathbf{v}_p$ after the sweep as the approximation to $\mathbf{u}_B^k$.

Algorithm 9.3.3 *(Schur Complement-ADI Algorithm to Solve (9.11))*
Input: $\mathbf{w}_B^{(k-1)}$, $\mathbf{g}_B^k$

```
1. Solve for $\mathbf{v}_1$:
   $$\left(I + \alpha\,\tau\mathcal{R}_1^T\overline{S}^{(1)}(\tau)\mathcal{R}_1\right)\mathbf{v}_1 = \mathbf{g}_B^k - \sum_{j=2}^{p}\alpha\,\tau\,\left(\mathcal{R}_j^T\overline{S}^{(j)}(\tau)\mathcal{R}_j\right)\mathbf{w}_B^{k-1}.$$
2. For $i = 2, \cdots, p$, solve for $\mathbf{v}_i$:
   $$\left(I + \alpha\,\tau\,\mathcal{R}_i^T\overline{S}^{(i)}(\tau)\mathcal{R}_i\right)\mathbf{v}_i = \mathbf{v}_{i-1} + \alpha\,\tau\,\left(\mathcal{R}_i^T\overline{S}^{(i)}\mathcal{R}_i\right)\mathbf{w}_B^{k-1}.$$
3. Endfor
Output: $\mathbf{w}_B^k \equiv \mathbf{v}_p$.
```

Remark 9.13. Each linear system of the form $(I + \alpha\,\tau\mathcal{R}_i^T\overline{S}^{(i)}\mathcal{R}_i)\mathbf{x}_i = \mathbf{r}_i$ arising in the generalized ADI algorithm can be solved at the cost of solving one Neumann problem on Ω_i as indicated below. With the exception of the submatrix of $(I + \alpha\,\tau\mathcal{R}_i^T\overline{S}^{(i)}\mathcal{R}_i)$ corresponding to the nodes on $B^{(i)}$, this matrix will have the same entries as the identity matrix. This yields $(\mathbf{x}_i)_l = (\mathbf{r}_i)_l$ for all nodes $x_l \in (B \backslash B^{(i)})$. To determine $\mathcal{R}_i\mathbf{x}_i$ we need to solve:

$$\begin{bmatrix} I + \alpha\,\tau\,A_{II}^{(i)} & \alpha\,\tau\,A_{IB}^{(i)} \\ \alpha\,\tau\,A_{BI}^{(i)} & I + \alpha\,\tau\,A_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_I^{(i)} \\ \mathcal{R}_i\mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathcal{R}_i\mathbf{r}_i \end{bmatrix}.$$

This can be verified using standard identities for Schur complement matrices.

Remark 9.14. Rigorous results are not known on how the truncation error of the preceding non-iterative scheme depends on τ, h and h_0, nor on its conditional stability. Heuristically, we expect the truncation error to deteriorate as the subdomain size $h_0 \to 0$. As with all non-iterative solvers, care must be exercised to test the scheme for stability and accuracy in each application. Alternative fractional step algorithms may also be employed to solve (9.11).

9.3.3 Non-Iterative Operator Splitting Methods

Operator splitting methods [DO9, PE, BA11, LE14, KO, SA5, YA, ST11], [GL7], such as fractional step and generalized ADI (Alternating Directions Implicit) methods are classical methods for obtaining an approximate solution to an evolution equation. Given an evolution equation $u_t + L\,u = f$ with initial value $u(0) = u_0$, these methods employ a *splitting* of the operator L:

$$L = L_1 + \cdots + L_q.$$

Using the splitting, these methods seek an *approximate solution* to $u_t+Lu = f$ by solving evolution equations of the form $w_t + L_i\,w = f_i$ for different f_i. If the operators L_i in the splitting can be chosen so that $w_t + L_i\,w = f_i$ are computationally "simpler" to solve than the original problem, then such operator splitting methodology may offer computational advantages.

Traditional splittings [DO9, PE] are based on *separation of variables.* For instance, the traditional splitting based on a separation of variables for parabolic equation (9.1) with $L\,u \equiv -\left(u_{x_1x_1} + u_{x_2x_2}\right)$ and $\Omega \subset \mathbb{R}^2$ is:

$$L\,u = L_1\,u + L_2\,u, \quad \text{where} \quad L_i\,u \equiv -u_{x_ix_i} \;\; \text{for} \;\; i = 1, 2.$$

An implicit discretization of (9.1) will yield the linear system:

$$(I + \alpha\,\tau\,A)\mathbf{u}^k + C\mathbf{u}^{k-1} + \tilde{\mathbf{f}}^k, \tag{9.16}$$

at each discrete time $t_k = k\,\tau$, where A denotes the discretization of the elliptic operator L. The operator splitting $L = L_1 + L_2$ will yield a matrix splitting $A = A_1 + A_2$, where A_i corresponds to a discretization of L_i. For a traditional finite difference discretization of L on a uniform grid on Ω, matrix A_i will be *tridiagonal* for an appropriate ordering of the nodes. In this case, the parabolic equation $w_t + L_i\,w = f_i$ will yield a tridiagonal system with coefficient matrix $(I + \alpha\,\tau\,A_i)$, and can thus be solved very efficiently. However, these computational advantages cannot be realized when the grid is *non-uniform* or when the underlying elliptic operator is *not separable.*

Our discussion will focus on *domain decomposition* operator *splittings* based on a *partition of unity* [VA, VA2, LA3, LA4, MA34]. Given an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ of Ω, we let $\chi_1(x), \ldots, \chi_p(x)$ denote a partition of unity subordinate to the subdomains. Then, an elliptic operator:

$$L\,u \;=\; -\nabla\cdot(a(x)\nabla u)+\mathbf{b}(x)\cdot\nabla u+c(x)u, \tag{9.17}$$

can be split as $L\,u = L_1\,u+\cdots+L_p\,u$, where for $1\le i\le p$:

$$L_i\,u \;\equiv\; -\nabla\cdot(\chi_i(x)\,a(x)\nabla u)+\chi_i(x)\,\mathbf{b}(x)\cdot\nabla u+\chi_i(x)\,c(x)u. \tag{9.18}$$

By construction, $L_i\,u$ will have support in Ω_i^* and $L\,u = L_1\,u+\cdots+L_p\,u$, since $\chi_1(x)+\cdots+\chi_p(x)=1$. Furthermore, when $\mathbf{b}(x)=\mathbf{0}$ and $c(x)\ge 0$, each L_i and A_i will be *semi-coercive.* Unlike splittings based on separation of variables, however, the truncation error of domain decomposition splittings can deteriorate as the size of the subdomains $h_0\to 0$.

Fractional Step Methods. This operator splitting method can be motivated by considering a system of differential equations [BA11, YA, ST11]:

$$\mathbf{u}_t + A\mathbf{u} = \mathbf{f}(t), \quad \mathbf{u}(0)=\mathbf{u}_0, \tag{9.19}$$

where matrix A is of size n and $\mathbf{u}(t)\in\mathbb{R}^n$. Using Duhamel's principle, the exact solution to this inhomogeneous system of ordinary differential equations has the following representation involving matrix exponentials:

$$\mathbf{u}(t) \;=\; e^{-A\,t}\,\mathbf{u}_0+\int_0^t e^{-A(t-s)}\mathbf{f}(s)\,ds, \quad\text{where}\quad e^{-A\,t}=\sum_{i=0}^{\infty}\frac{(-t)^iA^i}{i!}.$$

If matrix A can be split as $A=A_1+\cdots+A_q$, then *provided* the matrices A_i *commute*, i.e., $A_iA_j=A_jA_i$ for each pair i,j, it will hold that:

$$e^{-A\,t}=e^{-(A_1+\cdots+A_q)\,t}=e^{-A_1\,t}\cdots e^{-A_q\,t}.$$

If it is simpler to compute the action of $e^{-A_i\,t}$ than the action of $e^{-A\,t}$, then this representation will yield a sequential algorithm for solving (9.19).

Generally, the matrices A_i in a splitting will not commute. However, if τ denotes a time step and $t=m\,\tau$, it will hold that $e^{-m\,A\,\tau}=e^{-A\,\tau}\cdots e^{-A\,\tau}$, since each term in $m\,\tau A=\tau A+\cdots+\tau A$ commutes. If $\tau\ll 1$ is small, we may heuristically *approximate* $e^{-\tau\,A}$ as follows:

$$e^{-\tau\,A}=e^{-\tau(A_1+\cdots+A_q)}=e^{-\tau A_1}\cdots e^{-\tau A_p}+o(\tau^2).$$

This can be verified by substituting $A=A_1+\cdots+A_q$ into the formal power series expansion for $e^{-A\,\tau}$ and grouping terms. The first order fractional step method is motivated by this property. From a matrix viewpoint, an implicit discretization of (9.19) will yield a linear system of the form:

$$(I+\alpha\,\tau\,A)\,\mathbf{u}^k+C\mathbf{u}^{k-1}=\tilde{\mathbf{f}}^k, \tag{9.20}$$

at each discrete time t_k. The first order fractional step method approximates $(I+\alpha\,\tau\,A)$ by a product of matrices of the form $(I+\alpha\,\tau\,A_i)$:

$$(I+\alpha\,\tau\,A)=(I+\alpha\,\tau\,A_1)\cdots(I+\alpha\,\tau\,A_p)+O(\tau^2), \tag{9.21}$$

which can be verified by formally multiplying all of the terms. The truncation error in (9.21) will be $O(\tau^2)$ at each time step, and due to accumulation of errors, the global error will be 1st order. Substituting (9.21) into (9.20) yields a modified discretization for $\mathbf{w}^{k-1} \approx \mathbf{u}^{k-1}$ and $\mathbf{w}^k \approx \mathbf{u}^k$:

$$(I + \alpha\,\tau\, A_1) \cdots (I + \alpha\,\tau\, A_q)\,\mathbf{w}^k = \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1}. \tag{9.22}$$

An algorithm for determining $\mathbf{w}^k$ is summarized next.

Algorithm 9.3.4 *(First Order Fractional Step Algorithm to Solve (9.22))*
Input: $\mathbf{w}^{k-1}$, $\mathbf{r}_1 = \tilde{\mathbf{f}}^k - C\,\mathbf{w}^{k-1}$

```
1. For $i = 1, \ldots, q$ solve:
       $(I + \alpha\,\tau\, A_i)\mathbf{z}_i = \mathbf{r}_i,$ and define $\mathbf{r}_{i+1} = \mathbf{z}_i.$
   Endfor
Output: $\mathbf{w}^k \equiv \mathbf{z}_q$
```

Remark 9.15. If L and L_i are defined by (9.17) and (9.18) respectively, then the entries of matrix A_i will be zero for nodes outside Ω_i^*, i.e., $(A_i)_{lj} = 0$ if x_l or x_j do not lie in Ω_i^*. As a result, the solution to $(I + \alpha\,\tau\, A_i)\mathbf{z}_i = \mathbf{r}_i$ can be computed at the cost of solving a subdomain problem:

$$\left((I + \alpha\,\tau\, A_i)^{-1}\mathbf{r}\right)_l = \begin{cases} (\mathbf{r})_l, & \text{if } x_l \notin \Omega_i^* \\ \left((R_i(I + \alpha\,\tau\, A_i)R_i^T)^{-1}R_i\mathbf{r}\right)_l, & \text{if } x_l \in \Omega_i^*. \end{cases}$$

where R_i denotes a restriction map from Ω into Ω_i^*.

Approximation (9.22) will be locally 2nd order accurate in τ, however, due to accumulation of errors [RI] the scheme will be 1st order accurate globally, provided the modified scheme is stable. The following result concerns the stability of the 1st order fractional step method.

Lemma 9.16. *Suppose each A_i is symmetric and positive semi-definite, with $\|C\| \le 1$ in the Euclidean norm $\|.\|$. Then, the 1st order fractional step scheme will be stable with:*

$$\|\mathbf{w}^k\| \le \|\mathbf{w}^{k-1}\| + c\,\|\tilde{\mathbf{f}}^k\|,$$

for some $c > 0$ independent of τ and h.

Proof. Since $\mathbf{w}^k = (I + \alpha\,\tau\, A_q)^{-1} \cdots (I + \alpha\,\tau\, A_1)^{-1}\left(\tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1}\right)$ and since $\|C\| \le 1$ in the Euclidean norm, we only need verify that:

$$\|\,(I + \alpha\,\tau\, A_q)^{-1} \cdots (I + \alpha\,\tau\, A_1)^{-1}\,\| \le 1.$$

However, this will follow since each of the terms $(I + \alpha\,\tau\, A_i)$ are symmetric positive definite with eigenvalues greater than 1, so that the Euclidean norms of their inverses will be bounded by one. □

The truncation error of the 1st order fractional step scheme is derived next.

Lemma 9.17. *The solution* $\mathbf{w}^k$ *of the 1st order fractional step scheme:*

$$(I + \alpha\,\tau\,A_1)\cdots(I + \alpha\,\tau\,A_q)\mathbf{w}^k = \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1},$$

will satisfy:

$$\begin{aligned}&(I + \alpha\,\tau\,A)\mathbf{w}^k + \textstyle\sum_{m=2}^{q} \left(\alpha^m \tau^m \sum_{1\le\sigma_1<\cdots<\sigma_m\le q} A_{\sigma_1}\cdots A_{\sigma_m}\right)\mathbf{w}^k \\ &\quad = \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1}.\end{aligned} \tag{9.23}$$

The local truncation error $\mathbf{T}_{\text{mod}}$ *of (9.22) will have the following terms in addition to the terms in the original local truncation error* $\mathbf{T}_{\text{orig}}$ *of (9.20):*

$$\mathbf{T}_{\text{mod}} = \mathbf{T}_{\text{orig}} + \sum_{m=2}^{q}\left(\alpha^m\tau^m \sum_{1\le\sigma_1<\cdots<\sigma_m\le q} A_{\sigma_1}\cdots A_{\sigma_m}\,\mathbf{u}_*^k\right), \tag{9.24}$$

where $\mathbf{u}_*^k$ *denotes the restriction of the exact solution of the parabolic equation to the grid points at time* $k\,\tau$*.*

Proof. See [DO9, PE, BA11, ST11]. Evaluating the product yields:

$$\begin{aligned}&(I + \alpha\,\tau\,A_1)\cdots(I + \alpha\,\tau\,A_q) \\ &\quad = I + \alpha\,\tau\,A + \sum_{m=2}^{q}\left(\alpha^m\tau^m \sum_{1\le\sigma_1<\cdots<\sigma_m\le q} A_{\sigma_1}\cdots A_{\sigma_m}\right).\end{aligned}$$

The proof immediately follows by replacing $(I + \alpha\tau A)$ in the original discretization by $(I + \alpha\tau A_1)\cdots(I + \alpha\tau A_q)$ and substituting the preceding. □

Remark 9.18. The preceding result shows that the local truncation error of the 1st order fractional step scheme is $O(\tau^2)$, resulting in a global error of $O(\tau)$ due to accumulation of errors, see [RI]. Thus, 1st order fractional step methods will only be suitable for globally 1st order schemes in time, such as backward Euler, but not for globally 2nd order schemes such as Crank-Nicolson.

To preserve the accuracy of globally 2nd order schemes, a 2nd order fractional step approximation can be employed using *Strang splitting* [ST11]. We illustrate this for the matrix splitting $A = A_1 + A_2$. Strang splitting approximates $e^{-\alpha\,\tau\,A}$ by a third order accurate approximation in τ as follows:

$$e^{-\alpha\,\tau\,A} = e^{-\frac{\alpha\,\tau}{2}A_1} e^{-\alpha\,\tau\,A_2} e^{-\frac{\alpha\,\tau}{2}A_1} + O(\tau^3). \tag{9.25}$$

This can be verified by substituting power series expansions of each exponential, observing that the matrices may not commute. Each term $e^{-\gamma\,\tau A_i}$ can be approximated using the Crank-Nicolson method to yield the following 3rd order approximation in τ, where the order of the terms is not important:

$$e^{-\gamma\tau A_i} = \left(I + \frac{\gamma\,\tau}{2}A_i\right)^{-1}\left(I - \frac{\gamma\,\tau}{2}A_i\right) + O(\tau^3).$$

Substituting such approximations into each of the terms of (9.25), will yield a locally 3rd order approximation of $e^{-\tau\,A} = (I + \frac{\tau}{2}\,A)^{-1}(I - \frac{\tau}{2}\,A) + O(\tau^3)$.

Lemma 9.19. *Consider the update in the Crank-Nicolson discretization of (9.19):*

$$\mathbf{u}^k = (I + \frac{\tau}{2}A)^{-1}(I - \frac{\tau}{2}A)\mathbf{u}^{k-1} + \frac{\tau}{2}\,(I + \frac{\tau}{2}A)^{-1}(\mathbf{f}^k + \mathbf{f}^{k-1}),$$

where $A = A_1 + A_2$ and A_1, A_2 are symmetric positive semi-definite matrices. Then, the following approximations will be $O(\tau^3)$ and unconditionally stable:

$$\begin{aligned} (I + \tfrac{\tau}{2}A)^{-1}(I - \tfrac{\tau}{2}A) &= H_1\,H_2\,H_1 + O(\tau^3) \\ (I + \tfrac{\tau}{2}A)^{-1}\tfrac{\tau}{2}\,(\mathbf{f}^k + \mathbf{f}^{k-1}) &= \tfrac{\tau}{2}\,(I + \tfrac{\tau}{2}A_2)^{-1}(I + \tfrac{\tau}{2}A_1)^{-1}\,(\mathbf{f}^k + \mathbf{f}^{k-1}) + O(\tau^3), \end{aligned}$$

where $H_1 = (I + \frac{\tau}{4}A_1)^{-1}(I - \frac{\tau}{4}A_1)$ and $H_2 = (I + \frac{\tau}{2}A_2)^{-1}(I - \frac{\tau}{2}A_2)$.

Proof. We leave the proof to the reader. □

Remark 9.20. Strang splittings require more linear systems to be solved than first order fractional step methods. It can be formulated for multiple splittings $A = A_1 + \cdots + A_q$ by recursive use of two matrix splittings, where $A_i = A_i^T \geq 0$:

$$\begin{aligned} (I + \tfrac{\tau}{2}A)^{-1}(I - \tfrac{\tau}{2}A) &= H_1 \cdots H_{q-1}H_qH_{q-1}\cdots H_1 + O(\tau^3) \\ (I + \tfrac{\tau}{2}A)^{-1}\tfrac{\tau}{2}\,(\mathbf{f}^k + \mathbf{f}^{k-1}) &= \tfrac{\tau}{2}\,(I + \tfrac{\tau}{2}A_q)^{-1}\cdots(I + \tfrac{\tau}{2}A_1)^{-1} \\ &\qquad (\mathbf{f}^k + \mathbf{f}^{k-1}) + O(\tau^3), \end{aligned}$$

where $H_q \equiv (I + \frac{\tau}{2}A_q)^{-1}(I - \frac{\tau}{2}A_q)$ and $H_i \equiv (I + \frac{\tau}{4}A_i)^{-1}(I - \frac{\tau}{4}A_i)$ for $i \neq q$. This will be unconditionally stable, but in practice, the generalized *alternating directions implicit* method, as described next, will be preferable. It generates smaller truncation errors and requires fewer systems to be solved. However, it may not be unconditionally stable without additional assumptions.

Generalized Alternating Directions Implicit (ADI) Method. This classical operator splitting method [DO9, PE, DO10, DO12] *formally* yields 3rd order local truncation error in τ, and can be applied to approximate any implicit time stepped scheme. However, the modified scheme may only be conditionally stable. Our discussion will consider the following discretization:

$$(I + \alpha\,\tau\,A)\,\mathbf{u}^k = \tilde{\mathbf{g}}^k \equiv \tilde{\mathbf{f}}^k - C\mathbf{u}^{k-1}, \tag{9.26}$$

with $A = A_1 + \ldots + A_q$, where A_i are symmetric positive semi-definite matrices.

To motivate the generalized ADI method, we substitute the matrix splitting $A = A_1 + \ldots + A_q$ into (9.26) and rearrange terms to obtain the following block linear system whose block rows are each equivalent to (9.26) for the choice $\mathbf{v}_1 = \cdots = \mathbf{v}_q = \mathbf{u}^k$:

$$\begin{bmatrix} (I + \alpha\tau A_1) & \alpha\tau A_2 & \cdots & \alpha\tau A_q \\ \alpha\tau A_1 & (I + \alpha\tau A_2) & \cdots & \alpha\tau A_q \\ \vdots & \vdots & \ldots & \vdots \\ \alpha\tau A_1 & \alpha\tau A_2 & \cdots & (I + \alpha\tau A_q) \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_q \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{g}}^k \\ \tilde{\mathbf{g}}^k \\ \vdots \\ \tilde{\mathbf{g}}^k \end{bmatrix}. \qquad (9.27)$$

The generalized ADI method can be obtained by applying one sweep of a block Gauss-Seidel iteration to solve the above system, using $\mathbf{v}_1 = \cdots = \mathbf{v}_q = \mathbf{w}^{k-1}$ as a starting guess, and defining $\mathbf{w}^k \equiv \mathbf{v}_q$ as the approximate solution.

The block Gauss-Seidel iterates with starting guess $\mathbf{w}^{k-1}$ will satisfy:

$$\begin{cases} (I + \alpha\tau A_1)\,\mathbf{v}_1 + \sum_{j=2}^{q} \alpha\tau A_j \mathbf{w}^{k-1} = \tilde{\mathbf{g}}^k \\ \vdots \\ \sum_{j=1}^{i-1} \alpha\tau A_j \mathbf{v}_j + (I + \alpha\tau A_i)\,\mathbf{v}_i + \sum_{j=i+1}^{q} \alpha\tau A_j \mathbf{w}^{k-1} = \tilde{\mathbf{g}}^k \\ \vdots \\ \sum_{j=1}^{q-1} \alpha\tau A_j \mathbf{v}_j + (I + \alpha\tau A_q)\,\mathbf{v}_q = \tilde{\mathbf{g}}^k. \end{cases}$$

Moving known quantities on the left hand side to the right, subtracting each block equation from its preceding, and solving yields the following algorithm for *approximately* solving $(I + \alpha\tau A)\mathbf{w}^k = \tilde{\mathbf{g}}^k$ using starting guess $\mathbf{w}^{k-1}$.

Algorithm 9.3.5 *(Generalized ADI Algorithm to Solve (9.26))*
Input: $\mathbf{w}^{k-1} \approx \mathbf{u}^{k-1}$

```
1. Solve for $\mathbf{v}_1$:
       $(I + \alpha\tau A_1)\,\mathbf{v}_1 = \tilde{\mathbf{g}}^k - \sum_{j=2}^{q} \alpha\tau A_j \mathbf{w}^{k-1}.$
2. For $i = 2, \cdots, q$, solve for $\mathbf{v}_i$:
       $(I + \alpha\tau A_i)\,\mathbf{v}_i = \mathbf{v}_{i-1} + \alpha\,\tau\,A_i \mathbf{w}^{k-1}.$
3. Endfor
```

Output: $\mathbf{w}^k \equiv \mathbf{v}_q \approx \mathbf{u}^k$.

Remark 9.21. The generalized ADI algorithm requires the solution of q linear systems, each with a coefficient matrix of the form $(I + \alpha\tau A_i)$ of size n. If the entries of A_i are zero for nodes outside Ω_i^*, then each such linear system can be reduced to a smaller linear system involving only the unknowns in Ω_i^*. If Ω_i^* is the union of disjoint subdomains of the same color, then problems on disjoint subregions can be solved in *parallel*, see Fig. 9.1.

Stability of the generalized ADI method is analyzed in [DO12].

Lemma 9.22. *If the matrices A_i are symmetric, positive semi-definite for $i = 1, \cdots, q$, if C is symmetric, and if A_i and C commute pairwise, then the generalized ADI scheme will be unconditionally stable. In the non-commuting case, if $q = 2$ and if the matrices A_i are symmetric positive semi-definite, and if $\| \cdot \|$ is any norm in which the original scheme (9.26) is stable, then the generalized ADI scheme will be stable in the norm $|||u||| \equiv \|(I + \alpha\tau A_2)u\|$. For $q \geq 3$, if the matrices A_i are positive semi-definite, then the generalized ADI scheme will be conditionally stable.*

Remark 9.23. In the non-commuting case, examples are known for $q \geq 3$ of positive semi-definite splittings for which the generalized ADI method can loose unconditional stability [DO12]. However, the above stability results are a bit pessimistic and instability is only rarely encountered in practice.

The truncation error of the generalized ADI method is described next.

Lemma 9.24. *Let $\mathbf{T}_{orig}$ denote the truncation error of the original scheme:*

$$(I + \alpha\tau A)\, \mathbf{u}^k + C\mathbf{u}^{k-1} = \tilde{\mathbf{f}}^k.$$

The solution $\mathbf{w}^k$ of the generalized ADI method will solve:

$$\begin{array}{c}(I + \alpha\tau A)\, \mathbf{w}^k + \sum_{m=2}^{q} \left(\alpha^m \tau^m \sum_{1 \leq \sigma_1 < \cdots < \sigma_m \leq q} A_{\sigma_1} \cdots A_{\sigma_m} \left(\mathbf{w}^k - \mathbf{w}^{k-1} \right) \right) \\ + C\mathbf{w}^{k-1} = \tilde{\mathbf{f}}^k,\end{array} \tag{9.28}$$

thereby introducing additional terms in the local truncation error:

$$\mathbf{T}_{ADI} = \mathbf{T}_{orig} + \sum_{m=2}^{q} \left(\alpha^m \tau^m \sum_{1 \leq \sigma_1 < \cdots < \sigma_m \leq q} A_{\sigma_1} \cdots A_{\sigma_m} \left(\mathbf{u}_*^k - \mathbf{u}_*^{k-1} \right) \right), \tag{9.29}$$

where $\mathbf{u}_*^k$ denotes the restriction of the exact solution to the parabolic equation to the grid points at time t_k.

Proof. See [DO12]. Step 2 of the generalized ADI method yields that:

$$(I + \alpha\tau A_i)\left(\mathbf{v}_i - \mathbf{w}^{k-1}\right) = \mathbf{v}_{i-1} - \mathbf{w}^{k-1}, \quad \text{for } 2 \leq i \leq q.$$

Recursively applying this, we deduce that:

$$(I + \alpha\tau A_2) \cdots (I + \alpha\tau A_q)\left(\mathbf{v}_q - \mathbf{w}^{k-1}\right) = (\mathbf{v}_1 - \mathbf{w}^{k-1}). \tag{9.30}$$

Now, step 1 of the generalized ADI algorithm yields that:

$$\begin{aligned}(I+\alpha\tau A_1)(\mathbf{v}_1-\mathbf{w}^{k-1}) &= \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1} - (I+\alpha\tau A_1)\mathbf{w}^{k-1} - \textstyle\sum_{j=2}^{q} \alpha\tau A_j \mathbf{w}^{k-1} \\ &= \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1} - (I+\alpha\tau A)\mathbf{w}^{k-1}.\end{aligned} \tag{9.31}$$

Multiplying equation (9.30) by $(I+\alpha\tau A_1)$ and substituting (9.31) for the resulting right hand side yields:

$$(I+\alpha\tau A_1)\cdots(I+\alpha\tau A_q)\,(\mathbf{v}_q-\mathbf{w}^{k-1}) = \tilde{\mathbf{f}}^k - C\mathbf{w}^{k-1} - (I+\alpha\tau A)\mathbf{w}^{k-1}. \tag{9.32}$$

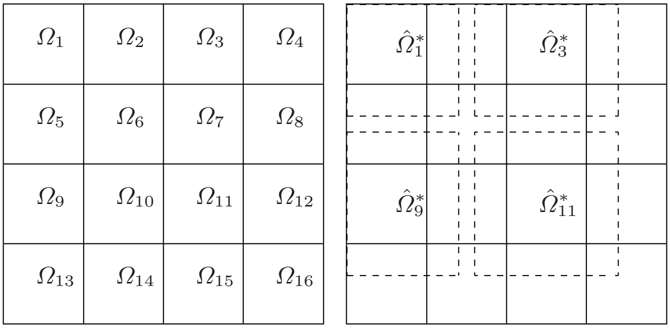
Using that $\mathbf{w}^k=\mathbf{v}_q$ and noting that:

$$\begin{aligned}&(I+\alpha\tau A_1)\cdots(I+\alpha\tau A_q) \\ &\quad = I+\alpha\tau A + \sum_{m=2}^{q} \alpha^m\tau^m \left(\sum_{1\le\sigma_1<\cdots<\sigma_m\le q} A_{\sigma_1}\cdots A_{\sigma_m} \right),\end{aligned}$$

yields equation (9.28), which is the desired result. □

Remark 9.25. The preceding truncation error term is estimated in Lemma 9.27 for domain decomposition operator splittings.

Domain Decomposition Operator Splittings. Here, we describe properties of domain decomposition splittings [VA, VA2, DR5, LA3, LA4, MA34]. To construct a partition of unity, we shall employ the following notation. Let $\Omega_1,\ldots,\Omega_p$ denote a *non-overlapping* decomposition of Ω into p subdomains of size h_0. We shall construct an overlapping covering $\hat{\Omega}_1^*,\ldots,\hat{\Omega}_p^*$ of Ω having overlap $\beta\,h_0$, by enlarging each subdomain Ω_k to $\hat{\Omega}_k^*$ to include all points in Ω within a distance $\beta\,h_0>0$ of Ω_k as in Fig. 9.1. We then group the overlapping subdomains into a small number $q\ll p$, of *colors* so that any two subdomains of the same color are disjoint, see Fig. 9.1. We denote the colored subdomains as $\Omega_1^*,\ldots,\Omega_q^*$ where each Ω_i^* is the union of several disjoint subdomains $\hat{\Omega}_l^*$. Multi-coloring reduces the number of operators in a splitting.



Non-overlapping subdomains A colored subdomain Ω_1^*

Fig. 9.1. Multicolored subdomains

Once an overlapping collection or multicolored subdomains have been constructed, a piecewise smooth partition of unity $\chi_1(x), \ldots, \chi_q(x)$ subordinate to $\Omega_1^*, \ldots, \Omega_q^*$ can be constructed as follows:

1. For $x \in \Omega_k^*$, let $\omega_k(x)$ denote the distance of x to the boundary $\partial\Omega_k^* \cap \Omega$:
$$\omega_k(x) = \begin{cases} \text{dist}(x, \partial\Omega_k^* \cap \Omega), & \text{for } x \in \Omega_k^* \\ 0, & \text{for } x \notin \Omega_k^*. \end{cases}$$
By construction $0 \leq \omega_k(x)$ will be continuous and *zero* outside $\overline{\Omega}_k^*$.
2. Define $\chi_k(x)$ by normalizing the $\omega_k(x)$ so that its sum equals 1:
$$\chi_k(x) \equiv \frac{\omega_k(x)}{\sum_{j=1}^q \omega_j(x)}, \quad \text{for } 1 \leq k \leq q.$$

The functions $\chi_1(x), \ldots, \chi_q(x)$ will be continuous and piecewise smooth.

Remark 9.26. Alternatively, any other choice of sufficiently smooth functions $\omega_k(x)$ positive in Ω_k^* and vanishing outside $\overline{\Omega_k}^*$ can be employed.

The preceding partition of unity functions $\chi_k(x)$ will satisfy:
$$\begin{cases} 0 \leq \chi_k(x) \leq 1, & \text{for } 1 \leq k \leq q \\ \text{supp}(\chi_k(x)) \subset \overline{\Omega}_k^*, & \text{for } 1 \leq k \leq q \\ \chi_1(x) + \cdots \chi_q(x) = 1, & \text{in } \overline{\Omega}. \end{cases}$$

For computational purposes, we have considered partition of unity functions $\chi_k(x)$ which are continuous and piecewise smooth. Smoother partition of unity functions (such as in $C^\infty(\Omega)$) can also be constructed, see [ST9].

Given a partition of unity $\chi_1(x), \ldots, \chi_q(x)$ subordinate to $\Omega_1^*, \ldots, \Omega_q^*$, a domain decomposition *splitting* $Lu = L_1 u + \cdots + L_q u$ of an elliptic operator $Lu = -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\, u$ can be constructed as follows:
$$\begin{cases} L_k u = -\nabla \cdot (a_k(x, y)\nabla u) + \mathbf{b}_k(x) \cdot \nabla u + c_k(x) u(x), \quad \text{where} \\ a_k(x) \equiv \chi_k(x)\, a(x) \geq 0, \quad \mathbf{b}_k(x) \equiv \chi_k(x)\, \mathbf{b}(x), \quad c_k(x) \equiv \chi_k(x)\, c(x). \end{cases}$$

Given a discretization $A\,\mathbf{u}$ of $L\,u$, we formally define $A_i\,\mathbf{u}$ as the discretization of $L_i\,u$, so that $A = A_1 + \cdots + A_q$, with $(A_i)_{lj} = 0$ for nodes $x_l, x_j \in \Omega \backslash \overline{\Omega}_i^*$. If $c(x) \geq 0$ and $\mathbf{b}(x) = 0$, then by construction each L_k and A_k will be self-adjoint and semi-coercive, and zero outside $\overline{\Omega}_i^*$. They can be employed in fractional step and generalized alternating directions implicit methods.

The following result estimates the truncation error of the generalized ADI method with domain decomposition splitting [VA, VA2, MA34].

Lemma 9.27. *Let $u(x,t) \subset C^1(\overline{\Omega} \times [0,t])$ denote the solution of (9.1) with $\partial_t u(.,t) \in C_0^{2q}(\Omega)$, and let $a(x) \in C^{2q}(\Omega)$. Then, the truncation error $\mathbf{T}_{ADI}^k$ of the generalized ADI method with domain decomposition splitting will satisfy:*

$$\|\mathbf{T}_{ADI}^k\| \leq \|\mathbf{T}_{orig}^k\| + K(\tau, \beta\, h_0, u)\, \tau^3,$$

where $\mathbf{T}_{orig}$ is the truncation error of the original scheme (9.44) and:

$$K(\tau, \beta\, h_0, u) \equiv \left(\frac{1}{(\beta\, h_0)^3} + \frac{\tau}{(\beta\, h_0)^5} + \cdots + \frac{\tau^{q-2}}{(\beta\, h_0)^{2q-1}} \right) \sup_{[0,t]} \|\partial_t u\|_{C^{2q}(\Omega)},$$

estimates the truncation error due to non-iterative approximate solver. If the ADI scheme is stable, then the error $\mathbf{e}_{ADI}^k = \mathbf{w}^k - \mathbf{u}_*(k\,\tau)$ will satisfy:

$$\|\mathbf{e}_{ADI}^k\| \leq \|\mathbf{e}_{orig}^k\| + K(\tau, \beta\, h_0, u)\, \tau^2. \tag{9.33}$$

Proof. See [VA, VA2, MA34]. □

Remark 9.28. The preceding result indicates that the accuracy of the ADI-domain decomposition splitting method can deteriorate as $\beta\, h_0 \to 0^+$ or if the exact solution $u(.,.)$ is not sufficiently smooth. However, for sufficiently smooth $u(.,.)$ and fixed overlap, the additional error due to the non-iterative solver will be $O(\tau^2)$ globally. Alternative non-iterative domain decomposition solvers are described in [DR5, DA4, DA5, LA3, LA4, ZH5].

9.4 Parareal-Multiple Shooting Method

The parareal method is a *parallel-in-time* iterative method for solving a dissipative evolution equation based on a *decomposition* of its *time* domain [LI3]. Given a dissipative equation $u_t + L\,u = f$ posed on a time interval $[0,T]$ with initial value $u(0) = u_0$, the parareal method decomposes the interval into p sub-intervals $[T_{i-1}, T_i]$ for $1 \leq i \leq p$ with $0 = T_0 < T_1 < \cdots < T_p = T$ and determines the solution at the times T_i for $1 \leq i \leq p$ using a *multiple-shooting* technique which solves the evolution equation on each interval in *parallel.* To speed up the multiple shooting iteration, the residual equations are "preconditioned" by solving a "coarse" time-grid discretization of the evolution equation using a time-step $|T_i - T_{i-1}|$, see [LI3, MA6, CH20]. The resulting algorithm is parallel, with coarse granularity, and suited for application to time dependent optimal control problems, see [MA7].

We shall describe the parareal method for solving parabolic equation (9.1) on $\Omega \times (0,T)$ discretized by an unconditionally stable θ-scheme in time and a finite difference discretization in space. From a matrix viewpoint, the resulting algorithm can be viewed as a Schur complement algorithm [GA7, AS] with an appropriately chosen "coarse time-grid" preconditioner, however, the truncated algorithm can also be regarded as a predictor-corrector time-discretization of the underlying parabolic equation, see [LI3, MA6, FA4]. We shall decompose $[0,T]$ into p uniform sub-intervals $[T_{i-1}, T_i]$ for $1 \le i \le p$:

$$T_i = i\,\Delta T \quad \text{for } \ 0 \le i \le p, \quad \text{and } \ \Delta T \equiv \frac{T}{p}. \tag{9.34}$$

We partition each interval $[T_{i-1}, T_i]$ into m sub-intervals with step size τ:

$$\tau \equiv \frac{\Delta T}{m} = \frac{T}{p\,m} \quad \text{and} \quad t_l = l\,\tau \quad \text{for } \ 0 \le l \le p\,m. \tag{9.35}$$

We express the θ-scheme (9.45) for (9.1) compactly as:

$$-F\,\mathbf{u}^{k-1} + \mathbf{u}^k = \tilde{g}^k, \qquad \text{for } \ 1 \le k \le p\,m, \ \text{ with } \ \mathbf{u}^0 = I_h u_0, \tag{9.36}$$

where matrix F of size n and $\tilde{\mathbf{g}}^k \in \mathbb{R}^n$ are defined by:

$$F = (I + \theta\,\tau A)^{-1}(I - (1-\theta)\,\tau\,A) \quad \text{and} \quad \tilde{\mathbf{g}}^k \equiv (I + \theta\,\tau\,A)^{-1}\tilde{\mathbf{f}}^k, \tag{9.37}$$

and n denotes the number of nodal unknowns in Ω. This corresponds to:

$$\begin{bmatrix} I & & & 0 \\ -F & I & & \\ & \ddots & \ddots & \\ 0 & & -F & I \end{bmatrix} \begin{bmatrix} \mathbf{u}^0 \\ \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^{p\,m} \end{bmatrix} = \begin{bmatrix} I_h u_0 \\ \tilde{\mathbf{g}}^1 \\ \vdots \\ \tilde{\mathbf{g}}^{p\,m} \end{bmatrix}, \tag{9.38}$$

which is a block lower bi-diagonal system of equations for $\mathbf{u}^0, \ldots, \mathbf{u}^{p\,m}$. Since it is lower block diagonal, it can be solved by marching in time. At each time step $t_k = k\,\tau$, an application of F requires solving a linear system with coefficient matrix $(I + \theta\,\tau\,A)$. For unconditional stability, we choose $\frac{1}{2} \le \theta \le 1$.

Given a temporal decomposition $[T_{i-1}, T_i]$ of $[0,T]$ for $1 \le i \le p$, we partition the unknowns in $\{\mathbf{u}^k\}_{k=0}^{p\,m}$ as *interior* and *boundary* nodal vectors, where the "boundary" denotes the union of all endpoints of the time intervals $[T_{i-1}, T_i]$ for $1 \le i \le p$. We shall employ the following block nodal vectors:

$$\mathbf{u}_I^{(i)} = \begin{bmatrix} \mathbf{u}^{(i-1)m+1} \\ \vdots \\ \vdots \\ \mathbf{u}^{i\,m-1} \end{bmatrix} \ \text{for } \ 1 \le i \le p, \ \ \mathbf{u}_I = \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \vdots \\ \vdots \\ \mathbf{u}_I^{(p)} \end{bmatrix}, \ \text{ and } \ \mathbf{u}_B = \begin{bmatrix} \mathbf{u}^0 \\ \mathbf{u}^m \\ \vdots \\ \mathbf{u}^{p\,m} \end{bmatrix}.$$

Similarly, we define nodal vectors $\tilde{\mathbf{g}}_I^{(i)}$, $\tilde{\mathbf{g}}_I$ and $\tilde{\mathbf{g}}_B$ for the forcing term $\tilde{\mathbf{g}}$ in (9.38). By construction $\mathbf{u}_I^{(i)} \in \mathbb{R}^{(m-1)n}$, $\mathbf{u}_I \in \mathbb{R}^{p(m-1)n}$ and $\mathbf{u}_B \in \mathbb{R}^{n(p+1)}$, and similarly for the forcing terms. Based on the block partition $\mathbf{u}_I$ and $\mathbf{u}_B$, we shall block partition the evolution block matrix system (9.38) as:

$$\begin{bmatrix} E_{II} & E_{IB} \\ E_{BI} & E_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{g}}_I \\ \tilde{\mathbf{g}}_B \end{bmatrix}, \tag{9.39}$$

where the block matrices E_{II}, E_{IB}, E_{BI} and E_{BB} are defined as:

$$E_{II} \equiv \begin{bmatrix} E_{II}^{(1)} & & 0 \\ & \ddots & \\ 0 & & E_{II}^{(p)} \end{bmatrix}, \; E_{IB} = \begin{bmatrix} E_{IB}^{(1)} \\ \vdots \\ \vdots \\ E_{IB}^{(p)} \end{bmatrix},$$

$$E_{BI} = \begin{bmatrix} E_{BI}^{(1)^T} \\ \vdots \\ \vdots \\ E_{BI}^{(p)^T} \end{bmatrix}^T, \; E_{BB} = \begin{bmatrix} I & & 0 \\ & \ddots & \\ 0 & & I \end{bmatrix},$$

where each of the above submatrices have the following block structure:

$$E_{II}^{(i)} = \begin{bmatrix} I & & & 0 \\ -F & I & & \\ & & \ddots & \\ 0 & & -F & I \end{bmatrix}, \; E_{IB}^{(i)} = \begin{bmatrix} X_{i0} & \cdots & X_{ip} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \; E_{BI}^{(i)} = \begin{bmatrix} 0 & \cdots & 0 & Y_{0i} \\ 0 & \cdots & 0 & Y_{1i} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & Y_{pi} \end{bmatrix},$$

where $X_{il} \equiv -\delta_{i\,(l+1)}\,F$ for $0 \le l \le p$ and $Y_{li} = -\delta_{l\,i}\,F$ for $0 \le l \le p$. Here δ_{ij} denotes the Kronecker index function $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. Matrices $E_{II}^{(i)}$, $E_{IB}^{(i)}$ and $E_{BI}^{(i)}$ are block $(m-1) \times (m-1)$, $(m-1) \times (p+1)$ and $(p+1) \times (m-1)$ matrices, respectively, with each block of size $n \times n$.

The parareal algorithm solves the discrete evolution equations (9.39) by solving the reduced Schur complement system for $\mathbf{u}_B$ using a preconditioner S_0 (which shall be described shortly). Since $\mathbf{u}_B$ denotes the block vector consisting of the unknown solution at the times $T_0, \ldots, T_p$, given $\mathbf{u}_B$ we can determine the solution $\mathbf{u}_I$ at the interior of the time intervals $[T_{i-1}, T_i]$ by solving for $\mathbf{u}_I = E_{II}^{-1}\,(\tilde{\mathbf{g}}_I - E_{IB}\mathbf{u}_B)$ in (9.39). Substituting $\mathbf{u}_I$ into the second block row in (9.39) yields the Schur complement system for $\mathbf{u}_B$:

$$S\,\mathbf{u}_B = \mathbf{g}_B \quad \text{where} \begin{cases} S \equiv (E_{BB} - E_{BI}E_{II}^{-1}E_{IB}) \\ \mathbf{g}_B \equiv (\tilde{\mathbf{g}}_B - E_{BI}E_{II}^{-1}\tilde{\mathbf{g}}_I). \end{cases} \tag{9.40}$$

The following is an explicit expression for $S = (E_{BB} - E_{BI}E_{II}^{-1}E_{IB})$.

Lemma 9.29. *Let the matrices E_{II}, E_{IB}, E_{BI} and E_{BB} be as defined earlier. Then, the Schur complement matrix will have the following expression:*

$$S = \begin{bmatrix} I & & & 0 \\ -F_\tau & I & & \\ & \ddots & \ddots & \\ 0 & & -F_\tau & I \end{bmatrix}, \quad \text{with} \quad F_\tau \equiv F^m, \tag{9.41}$$

where S is a block $(p+1) \times (p+1)$ matrix with blocks of size n.

Proof. Employ the following expression for S:

$$\begin{bmatrix} \mathbf{0} \\ S\,\mathbf{u}_B \end{bmatrix} = \begin{bmatrix} E_{II} & E_{IB} \\ E_{BI} & E_{BB} \end{bmatrix} \begin{bmatrix} -E_{II}^{-1}\,E_{IB}\,\mathbf{u}_B \\ \mathbf{u}_B \end{bmatrix},$$

$$\text{where} \quad E_{II}^{-1} = \begin{bmatrix} E_{II}^{(1)^{-1}} & & 0 \\ & \ddots & \\ 0 & & E_{II}^{(p)^{-1}} \end{bmatrix},$$

and substitute for E_{BB}, E_{IB}, E_{BI}, and E_{II}^{-1} using the block matrix identity:

$$E_{II}^{(i)} = \begin{bmatrix} I & & & 0 \\ -F & I & & \\ & \ddots & \ddots & \\ 0 & & -F & I \end{bmatrix} \Longrightarrow E_{II}^{(i)^{-1}} = \begin{bmatrix} I & & & 0 \\ F & I & & \\ \vdots & & \ddots & \ddots \\ F^{m-2} & \cdots & F & I \end{bmatrix}.$$

The desired result follows directly. □

The matrix F^m represents a discrete time approximation of the matrix exponential $e^{-m\,\tau\,A}$. Since $m\,\tau = \Delta T$, we may *heuristically* approximate $e^{-\Delta T\,A}$ by $F_{\Delta T}$ obtained by employing a stable θ-scheme with time-step ΔT:

$$F_{\Delta T} \equiv (I + \theta\,\Delta T\,A)^{-1}\,(I - (1-\theta)\,\Delta T\,A) \approx e^{-\Delta T\,A} \approx F_\tau = F^m.$$

Substituting this into (9.41) yields the *parareal preconditioner* S_0 for S:

$$S_0 \equiv \begin{bmatrix} I & & & 0 \\ -F_{\Delta T} & I & & \\ & \ddots & \ddots & \\ 0 & & -F_{\Delta T} & I \end{bmatrix}.$$

The error amplification matrix $(I - S_0^{-1}S)$ will be a *contraction* in the Euclidean norm for certain θ, and an *unaccelerated* iteration to solve $S\,\mathbf{u}_B = \mathbf{g}_B$ with preconditioner S_0 will be convergent. The rate will be independent of τ and ΔT. Below, we summarize the matrix version of the parareal algorithm.

Algorithm 9.4.1 *(Matrix Version of the Parareal Algorithm to Solve (9.39))*
Input: $\tilde{\mathbf{g}}_I$, $\tilde{\mathbf{g}}_B$ *and starting guess* $\mathbf{v}_B$

```
1. For $i = 1, \dots p$ in parallel solve:
       $E_{II}^{(i)}\, \mathbf{w}_I^{(i)} = \tilde{\mathbf{g}}_I^{(i)}$
2. Endfor
3. Compute $\mathbf{g}_B = (\tilde{\mathbf{g}}_B - E_{BI}\mathbf{w}_I)$
4. For $k = 0, 1, \dots$ until convergence do
5.     Compute $\mathbf{r}_B \leftarrow (\mathbf{g}_B - S\,\mathbf{v}_B)$
6.     Solve $S_0\,\mathbf{x}_B = \mathbf{r}_B$
7.     Update $\mathbf{v}_B \leftarrow \mathbf{v}_B + \mathbf{x}_B$
8. Endfor
9. Solve $E_{II}\mathbf{v}_I = \tilde{\mathbf{g}}_I - E_{IB}\,\mathbf{v}_B$
```

Output: $\mathbf{v}_I$, $\mathbf{v}_B$

Remark 9.30. The notation $\mathbf{v}_B \leftarrow \mathbf{v}_B + \mathbf{r}_B$ means that we determine a *new* update $\mathbf{v}_B$ by adding $\mathbf{x}_B$ to the *old* $\mathbf{v}_B$. In step 1, solving $E_{II}^{(i)}\mathbf{w}_I^{(i)} = \tilde{\mathbf{g}}_I^{(i)}$ corresponds to solving the discretized parabolic equation on $[T_{i-1}, T_i]$. Since $E_{II}^{(i)}$ is block lower bi-diagonal, the computational time will be proportional to $(m-1)$ solves of a linear system with coefficient matrix $(I + \theta\,\tau\,A)$. In step 5, the time for computing $S\,\mathbf{v}_B$ will be proportional to the time for computing $E_{II}^{-1}\,(E_{IB}\mathbf{v}_B)$. Since E_{II} is block diagonal with p blocks $E_{II}^{(i)}$, this can be implemented on p parallel processors in a time proportional to $(m-1)$ solves of systems with coefficient matrix $(I + \theta\,\tau\,A)$. In step 6, the solution of $S_0\mathbf{x}_B = \mathbf{r}_B$ will be sequential and the computational time will be proportional to the cost of solving p linear systems with coefficient matrix $(I + \theta\,\Delta T\,A)$. Each solve of $(I + \theta\,\Delta T\,A)$ may be parellelized using spatial domain decomposition.

The following result estimates $\|I - S_0^{-1}S\|$ in the Euclidean norm [LI3].

Lemma 9.31. *Consider the θ-scheme discretization of (9.1) for $\frac{1}{2} \le \theta \le 1$. Let $A = A^T > 0$, $\Delta T = (T/p)$, $\tau = (T/p\,m)$, with $F_{\Delta T}$, F and F_τ defined by:*

$$\begin{cases} F_{\Delta T} \equiv (I + \theta\,\Delta T\,A)^{-1}\,(I - (1-\theta)\,\Delta T\,A) \\ F \quad\;\; = (I + \theta\,\tau\,A)^{-1}\,(I - (1-\theta)\,\tau\,A) \\ F_\tau \quad \equiv F^m. \end{cases}$$

Let $Q^T\,A\,Q =$ $\mathrm{diag}(\lambda_1, \dots, \lambda_n)$ denote the spectral decomposition of matrix A, where $Q^TQ = I$. Then, the matrices $F_{\Delta T}$ and F_τ will be diagonalized by Q with $\mathrm{diag}(\mu_1, \dots, \mu_n) \equiv Q^T F_\tau Q$ *and* $\mathrm{diag}(\alpha_1, \dots, \alpha_n) \equiv Q^T F_{\Delta T} Q$. *Furthermore, the following bounds will hold for the matrices S and S_0 defined earlier:*

$$\|I - S_0^{-1}S\| \le \max_{1 \le i \le n} \left(\frac{1 - |\alpha_i|^p}{1 - |\alpha_i|} \right) |\mu_i - \alpha_i|, \qquad (9.42)$$

Proof. The Euclidean norm of $\|I - S_0^{-1}S\|$ will correspond to the square root of the maximal eigenvalue of $(I - S_0^{-1}S)^T(I - S_0^{-1}S)$. To estimate this, we employ the following matrix identity for S_0^{-1} to obtain:

$$S_0^{-1} = \begin{bmatrix} I & & & 0 \\ F_{\Delta T} & I & & \\ \vdots & \ddots & \ddots & \\ F_{\Delta T}^p & \cdots & F_{\Delta T} & I \end{bmatrix} \Longrightarrow \left(I - S_0^{-1}S\right) = D \begin{bmatrix} 0 & & & 0 \\ F_{\Delta T}^0 & 0 & & \\ \vdots & \ddots & \ddots & \\ F_{\Delta T}^{p-1} & \cdots & F_{\Delta T}^0 & 0 \end{bmatrix},$$

where $D \equiv \text{blockdiag}(F_\tau - F_{\Delta T}, \ldots, F_\tau - F_{\Delta T})$. Here, we have used that since I and A commute, matrices $F_\tau = F_\tau^T > 0$ and $F_{\Delta T} = F_{\Delta T}^T > 0$ also commute. Forming $(I - S_0^{-1}S)^T(I - S_0^{-1}S)$ yields:

$$(I - S_0^{-1}S)^T(I - S_0^{-1}S) = D^2 \begin{bmatrix} X_{p-1} & \cdots & F_{\Delta T}^{p-2}X_1 & F_{\Delta T}^{p-1}X_0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ X_1 F_{\Delta T}^{p-2} & \cdots & X_1 & F_{\Delta T}X_0 & \vdots \\ X_0 F_{\Delta T}^{p-1} & \cdots & X_0 F_{\Delta T} & X_0 & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix},$$

where $X_l \equiv \sum_{i=0}^{l} F_{\Delta T}^i$. Since A is diagonalized by matrix Q, each submatrix of $(I - S_0^{-1}S)^T(I - S_0^{-1}S)$ will also be diagonalized by Q. So if we define $\mathcal{Q} = \text{blockdiag}(Q, \ldots, Q)$, then $\mathcal{Q}^T(I - S_0^{-1}S)^T(I - S_0^{-1}S)\mathcal{Q}$ will have diagonal submatrices. The maximal eigenvalue of $(I - S_0^{-1}S)^T(I - S_0^{-1}S)$ will equal the maximal eigenvalue of $\mathcal{Q}^T(I - S_0^{-1}S)^T\mathcal{Q}\,\mathcal{Q}^T(I - S_0^{-1}S)\mathcal{Q}$, and this can be estimated using $\|\mathcal{Q}^T(I - S_0^{-1}S)^T\mathcal{Q}\|_\infty \, \|\mathcal{Q}^T(I - S_0^{-1}S)\mathcal{Q}\|_\infty$. Using maximal absolute row sums, and the identity $(1 + \alpha + \alpha^2 + \cdots + \alpha^{p-1}) = (1 - \alpha^p)/(1 - \alpha)$ for $\alpha \neq 1$, yields the desired result. □

The following Corollary shows that $\|I - S_0^{-1}S\| \leq \frac{1}{2}$ for $\theta = 1$.

Corollary 9.32. *Let $\theta = 1$ and $m \geq 2$. Then, it will hold that $\|I - S_0^{-1}S\| \leq \frac{1}{2}$.*

Proof. Let $0 < x_i \equiv \tau\,\lambda_i < \infty$ denote the i'th eigenvalue of $\tau\,A$. Since A is diagonalized by Q, matrices $F_{\Delta T} = (I + m\,\tau\,A)^{-1}$ and $F_\tau = (I + \tau\,A)^{-m}$ will also be diagonalized by Q. The eigenvalues α_i and μ_i of $F_{\Delta T}$ and F_τ, will be:

$$\alpha_i = \frac{1}{(1 + m\,x_i)} \quad \text{and} \quad \mu_i = \frac{1}{(1 + x_i)^m}, \quad \text{for} \quad 1 \leq i \leq p.$$

Substituting this into the expression (9.42) and simplifying yields:

$$\left(\frac{1 - |\alpha_i|^p}{1 - |\alpha_i|}\right) |\mu_i - \alpha_i| = \left(\frac{(1 + m\,x_i)^p - 1}{m\,x_i\,(1 + m\,x_i)^{p-1}}\right) \left(\frac{(1 + x_i)^m - (1 + m\,x_i)}{(1 + m\,x_i)\,(1 + x_i)^m}\right).$$

Employing the binomial expansion, rearranging terms, and simplifying yields:

$$\begin{aligned}
\left(\frac{1-|\alpha_i|^p}{1-|\alpha_i|}\right)|\mu_i-\alpha_i| &= \left(\frac{\sum_{l=1}^{p}\binom{p}{l}(m\,x_i)^l}{m\,x_i\,(1+m\,x_i)^{p-1}}\right)\left(\frac{\sum_{k=2}^{m}\binom{m}{k}x_i^k}{(1+m\,x_i)\,(1+x_i)^m}\right)\\
&= \left(\frac{\sum_{l=1}^{p}\binom{p}{l}(m\,x_i)^l}{(1+m\,x_i)^{p}}\right)\left(\frac{\sum_{k=2}^{m}\binom{m}{k}x_i^k}{m\,x_i\,(1+x_i)^m}\right)\\
&= \left(\frac{\sum_{l=1}^{p}\binom{p}{l}(m\,x_i)^l}{(1+m\,x_i)^{p}}\right)\left(\frac{\sum_{k=2}^{m}\binom{m}{k}x_i^{k-1}}{m\,(1+x_i)^m}\right)\\
&= \left(\frac{\sum_{l=1}^{p}\binom{p}{l}(m\,x_i)^l}{\sum_{l=0}^{p}\binom{p}{l}(m\,x_i)^l}\right)\left(\frac{\sum_{k=1}^{m-1}\binom{m}{k+1}x_i^{k-1}}{m\,\sum_{k=0}^{m}\binom{m}{k}x_i^k}\right).
\end{aligned}$$

Since $x_i > 0$, we note that the quotient in the first bracket will be bounded by 1 by comparing coefficients of x_i^l. Similarly, comparing the coefficients of x_i^l in the quotient in the second bracket yields the following upper bound:

$$\max_{1\le k\le m-1}\frac{\binom{m}{k+1}}{m\binom{m}{k}} = \max_{1\le k\le m-1}\frac{(m-k)}{m\,(k+1)} \le \frac{1}{2}.$$

This yields the desired result. □

Remark 9.33. The eigenvalues λ_i of matrix A ranges from $O(1)$ to $O(h^{-2})$, so that the eigenvalues $x_i = \tau\,\lambda_i$ of $\tau\,A$ will range from $O(\tau)$ to $O(\tau\,h^{-2})$. If $\tau = O(h)$ this range will be from $O(h)$ to $O(h^{-1})$. In particular, when $\lambda_i = O(1)$, the contraction factor $(1-|\alpha_i|^p)\,|\alpha_i-\mu_i|/(1-|\alpha_i|) = O(\Delta T\,\tau)$.

Remark 9.34. In Alg. 9.4.1, steps 1, step 5 and step 6 are the most expensive in terms of computational costs. Steps 1 and 5 require the solution of linear systems with coefficient matrices $E_{II}^{(i)}$ in parallel for $1 \le i \le p$, and when p is large, the total memory requirements can also be significant if implemented on a shared memory architecture. Step 6 requires the solution of the block lower bi-diagonal system with coefficient matrix S_0, with storage requirement for $\mathbf{v}_B$ or $\mathbf{r}_B$ proportional to $n\,p$.

Remark 9.35. The theoretical bounds in Lemma 9.31 and Coro. 9.32 can be improved to $\|(I - S_0^{-1}S)\| \leq 0.294$, we refer the reader to [LI3, GA7, SC2] for more general analysis and notation. For Alg. 9.4.1, it is *heuristically* expected that the error propagation matrix $(I - S_0^{-1}S)$ will still be *contractive* for $\frac{1}{2} \leq \theta \leq 1$. However, for $\frac{1}{2} \leq \theta < 1$, the contraction factors may depend on τ, h, ΔT, p. More generally, the implementation of the parareal coarse preconditioner S_0 can use a different choice of θ for $F_{\Delta T}$ than for F (or F_τ).

9.5 Theoretical Results

In this section, we describe some background results on the stability and convergence of discretizations of parabolic equations. Following that, we estimate convergence bounds for a Schwarz preconditioner without a coarse space correction term. We consider a discretization of the parabolic equation:

$$\begin{cases} u_t + Lu = f, & \text{in } \Omega \times (0,t), \\ u(x,0) = u_0(x), & \text{in } \Omega, \\ u(x,t) = 0, & \text{on } \partial\Omega \times (0,t), \end{cases} \tag{9.43}$$

where $L\,u \equiv -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u$ has smooth coefficients satisfying $a(x) \geq a_0 > 0$ and $c(x) \geq 0$ in Ω. In most of our applications, we shall assume that $\mathbf{b}(x) = 0$, in which case L will be self-adjoint and coercive. However, several results are also described for $\mathbf{b}(x) \neq 0$.

9.5.1 Stability and Convergence of Time-Stepping Schemes

Consider a discretization of (9.43) having the following form:

$$(I + \alpha\,\tau\,A)\,\mathbf{u}^k + C\,\mathbf{u}^{k-1} = \tilde{\mathbf{f}}^k, \tag{9.44}$$

where $\tau > 0$ denotes the time step and matrix A denotes a *finite difference* discretization of L, where C is a matrix, $\tilde{\mathbf{f}}^k$ a forcing term, and $\alpha > 0$. To be specific, we shall consider the θ-scheme for $0 \leq \theta \leq 1$:

$$(I + \theta\,\tau A)\,\mathbf{u}^k - (I - (1-\theta)\tau A)\,\mathbf{u}^{k-1} = \theta\,\tau\,\mathbf{f}^k + (1-\theta)\,\tau\,\mathbf{f}^{k-1}. \tag{9.45}$$

In this case $\alpha = \theta$, $C = -\,(I - (1-\theta)\tau A)$ and $\tilde{\mathbf{f}}^k = \theta\,\tau\,\mathbf{f}^k + (1-\theta)\,\tau\,\mathbf{f}^{k-1}$. The θ-scheme yields the forward Euler method when $\theta = 0$, the Crank-Nicolson method when $\theta = 1/2$ and the backward Euler method when $\theta = 1$.

Definition. We define the *local* truncation error $\mathbf{T}^k_{orig}$ of (9.44) as:

$$\mathbf{T}^k_{orig} \equiv (I + \alpha\tau A)\,\mathbf{u}^k_* + C\mathbf{u}^{k-1}_* - \tilde{\mathbf{f}}^k, \tag{9.46}$$

at time $t_k = k\,\tau$, where $\mathbf{u}^k_*$ denotes the restriction of the exact solution $u(x,t)$ of (9.43) to the spatial grid at time t_k.

Remark 9.36. For instance, if $\mathbf{b}(x) = \mathbf{0}$ and A is a 2nd order accurate spatial discretization of L, then the θ-scheme in time will yield a local truncation error of $O\left(\tau(\tau + h^2)\right)$ if $\theta \neq \frac{1}{2}$ and $O\left(\tau^2(\tau^2 + h^2)\right)$ if $\theta = \frac{1}{2}$.

Definition. Discretization (9.44) is said to be stable in a norm $\|\cdot\|$ if:

$$\|\mathbf{u}^k\| \;\leq\; (1 + c_1\,\tau)\,\|\mathbf{u}^{k-1}\| + c_2\,\|\tilde{\mathbf{f}}^k\|, \tag{9.47}$$

holds with c_1, $c_2 > 0$ independent of τ and h for arbitrary $\tilde{\mathbf{f}}^k$. If this holds without restrictions on τ, the discretization is *unconditionally* stable. If this holds only with restrictions on τ, it is said to be *conditionally* stable.

The next result summarizes the Euclidean norm stability of the θ-scheme for (9.43) when $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \geq 0$. In this case, matrix $A = A^T > 0$.

Lemma 9.37. *Consider discretization (9.45) with $A = A^T > 0$, and let $\|\cdot\|$ denote the Euclidean norm. Then, the following results will hold:*

1. *For $\frac{1}{2} \leq \theta \leq 1$, the θ-scheme will be unconditionally stable in $\|\cdot\|$.*
2. *For $0 \leq \theta < \frac{1}{2}$, the θ-scheme will be stable in $\|\cdot\|$ provided:*

$$\tau \leq \frac{1}{(1-\theta)\,\lambda_1},$$

where $0 < \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of A.

Proof. The solution $\mathbf{u}^k$ in the θ-scheme has the representation:

$$\mathbf{u}^k \;=\; (I + \theta\,\tau\,A)^{-1}\,(I - (1-\theta)\,\tau\,A)\;\mathbf{u}^{k-1} + (I + \theta\,\tau\,A)^{-1}\;\tilde{\mathbf{f}}^k.$$

Since $A = A^T > 0$ is of size n, let $A = Q\Lambda Q^T$ be its eigendecomposition where Q is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Substituting that $(I + \theta\,\tau\,A) = Q(I + \theta\,\tau\,\Lambda)Q^T$ for $\theta \in [0,1]$ and using that $\|Q\| = \|Q^T\| = 1$ in the Euclidean norm $\|\cdot\|$ yields:

$$\|\mathbf{u}^k\| \leq \|(I + \theta\,\tau\,\Lambda)^{-1}(I - (1-\theta)\,\tau\,\Lambda)\|\;\|\mathbf{u}^{k-1}\| + \|(I + \theta\,\tau\,\Lambda)^{-1}\|\;\|\tilde{\mathbf{f}}^k\|.$$

The bound $\|(I + \theta\,\tau\,\Lambda)^{-1}\| \leq 1$ will hold trivially for $\theta \geq 0$. To estimate $\|(I + \theta\,\tau\,\Lambda)^{-1}(I - (1-\theta)\,\tau\,\Lambda)\|$ substitute that Λ is a diagonal matrix with diagonal entries $0 < \lambda_1 \leq \cdots \leq \lambda_n$. The desired result will follow trivially from the scalar requirements:

$$\left|\frac{1 - (1-\theta)\,\tau\,\lambda_i}{1 + \theta\,\tau\,\lambda_i}\right| \;\leq\; 1, \qquad \text{for} \quad 1 \leq i \leq n,$$

using that $1 + \theta\,\tau\,\lambda_i \geq 1$. $\square$

Remark 9.38. The stability bounds in the preceding result will also be valid in the mesh dependent norm $\|\cdot\|_{I+\alpha\,\tau\,A}$ for $0 \leq \alpha$. To verify this, suppose that F is a matrix which *commutes* with $(I+\alpha\,\tau\,A)$, such that F and $(I+\alpha\,\tau\,A)$ are simultaneously diagonalized by an orthogonal matrix Q. Let $F = QD_F Q^T$ and $(I+\alpha\,\tau\,A) = Q(I+\alpha\,\tau\,\Lambda)Q^T$. Then, the following will hold:

$$\begin{aligned} \|F\|^2_{I+\alpha\,\tau\,A} = \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\|F\mathbf{v}\|^2_{I+\alpha\,\tau\,A}}{\|\mathbf{v}\|^2_{I+\alpha\,\tau\,A}} &= \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^T F^T (I+\alpha\,\tau\,A) F\mathbf{v}}{\mathbf{v}^T (I+\alpha\,\tau\,A)\mathbf{v}} \\ &= \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^T D_F^T (I+\alpha\,\tau\Lambda) D_F\mathbf{v}}{\mathbf{v}^T (I+\alpha\,\tau\,\Lambda)\mathbf{v}} \\ &= \sup_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^T D_F^T D_F\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \;=\; \|F\|^2, \end{aligned}$$

since the factors $(1+\alpha\,\tau\,\lambda_i)$ which occur in the numerator and denominator of the above quotient cancel. Choosing $F = (I+\theta\,\tau\,A)^{-1}(I-(1-\theta)\theta\tau A)$ and $F = (I+\theta\,\tau\,A)^{-1}$ yields stability bounds in $\|\cdot\|_{I+\alpha\,\tau\,A}$.

In the following, we shall describe a few Euclidean norm stability results for implicit discretizations of *non-self adjoint* parabolic equations ($\mathbf{b}(x) \neq 0$) and for hyperbolic equations ($a(x) = 0$ and $c(x) = 0$). We shall decompose $A = H + N$ with $H = H^*$ denoting the Hermitian part of A and $N = -N^*$ the skew-Hermitian part of A, i.e., $H = \frac{1}{2}(A+A^*)$ and $N = \frac{1}{2}(A-A^*)$, where $X^* = \overline{X}^T$ denotes the complex adjoint of X, i.e., the complex conjugate of the transpose. Matrices H and N will be unitarily diagonalizable (though not simultaneously). If $H = H^* \geq 0$, then the eigenvalues of A will have non-negative real part (based on its field of values), while the eigenvalues of N will be pure imaginary. In the non-self adjoint case, the following Euclidean norm stability result will hold for the backward Euler scheme.

Lemma 9.39. *Consider discretization (9.45) for $\theta = 1$ with $A = H + N$, where matrix $H = H^* > 0$ and $N^* = -N$. Then, the backward Euler scheme will be unconditionally stable in the (complex) Euclidean norm $\|\cdot\|$.*

Proof. We substitute $A = H + N$ into the backward Euler scheme to obtain:

$$(I+\tau\,H+\tau\,N)\,\mathbf{u}^k = \mathbf{u}^{k-1} + \tau\,\mathbf{f}^k,$$

at each discrete time $t_k = k\,\tau$. Matrix $(I+\tau\,H)$ will be Hermitian positive definite so that $(I+\tau\,H)^{1/2}$ is well defined. We may thus express:

$$(I+\tau\,H)+\tau\,N \;=\; (I+\tau\,H)^{1/2}\,(I+\tau\,\tilde{N})\,(I+\tau\,H)^{1/2},$$

where $\tilde{N} \equiv (I+\tau\,H)^{-1/2}N(I+\tau\,H)^{-1/2}$ is skew-Hermitian. Matrices $(I+\tau\,H)$ and $(I+\tau\,\tilde{N})$ will be unitarily diagonalizable (although, not simultaneously), with eigenvalues of magnitude greater that 1, yielding $\|(I+\tau\,H)^{-1/2}\| \leq 1$ and $\|(I+\tau\,\tilde{N})^{-1}\| \leq 1$, so that $\|(I+\tau\,A)^{-1}\| \leq 1$. We shall thus obtain:

$$\begin{aligned}\|\mathbf{u}^k\| &\le \|(I+\tau\,A)^{-1}\mathbf{u}^{k-1}\| + \tau\,\|(I+\tau\,A)^{-1}\mathbf{f}^k\| \\ &\le \|\mathbf{u}^{k-1}\| + \tau\,\|\mathbf{f}^k\|.\end{aligned}$$

This verifies the unconditional stability of the backward Euler scheme. □

Remark 9.40. More generally, if $A = H + N = W\,\Lambda\,W^{-1}$ is *diagonalizable* (where W may not be unitary), with eigenvalues $\lambda_i = (\Lambda)_{ii}$, then:

$$(I+\theta\,\tau\,A)^{-1}(I-(1-\theta)\,\tau\,A) = W(I+\theta\,\tau\,\Lambda)^{-1}(I-(1-\theta)\,\tau\,\Lambda)\,W^{-1},$$

since I, A, and the other terms will be *simultaneously* diagonalized by W. If $(1-(1-\theta)\,\tau\,\mathrm{Re}(\lambda_i)) \ge 0$, then the eigenvalues will satisfy:

$$|(1-(1-\theta)\,\tau\,\lambda_i)/(1+\theta\,\tau\,\lambda_i)| \le 1, \quad \text{provided} \quad \frac{1}{2} \le \theta \le 1.$$

In this case $\|(I+\theta\,\tau\,A)^{-1}(I-(1-\theta)\,\tau\,A)\| \le \|W\|\,\|W^{-1}\|$. This will not guarantee Euclidean norm stability of (9.45) since W may not be unitary. However, when $H = 0$ (for instance, if $a(x) = 0$ and $c(x) = 0$), then W will be unitary (and $(I+\tau\,N)$ will be diagonalized by W), yielding that the θ-scheme for the hyperbolic equation will be unconditionally stable for $\frac{1}{2} \le \theta \le 1$.

We next consider the *maximum norm* stability of θ-schemes when A is an M-matrix (and $A^T \ne A$). We shall employ the following preliminary results.

Lemma 9.41. *Let A be a diagonally dominant M-matrix [VA9, SA2]. Then:*

$$\|\,(I+\alpha\tau A)^{-1}\,\|_\infty \le 1,$$

for any $\alpha > 0$, $\tau > 0$.

Proof. It will be sufficient to show that for any vector $\mathbf{w} \in R^n$:

$$\|\mathbf{w}\|_\infty \le \|\,(I+\alpha\tau A)\,\mathbf{w}\|_\infty.$$

To show this, choose i such that: $|(\mathbf{w})_i| = \|\mathbf{w}\|_\infty$, and without loss of generality, assume that $(\mathbf{w})_i > 0$ (otherwise, replace $\mathbf{w}$ by $-\mathbf{w}$ and repeat the argument). It will thus also hold that $-|(\mathbf{w})_j| \ge -(\mathbf{w})_i$. Now, since A is an M-matrix, it will hold that $A_{ii} > 0$ and $A_{ij} \le 0$ for $j \ne i$. Applying the preceding properties at index i yields:

$$\begin{aligned}(I+\alpha\tau A\mathbf{w})_i &= (\mathbf{w})_i + \alpha\tau \textstyle\sum_j A_{ij}(\mathbf{w})_j \\ &= (\mathbf{w})_i + \alpha\tau A_{ii}(\mathbf{w})_i + \alpha\tau \textstyle\sum_{j\ne i} A_{ij}(\mathbf{w})_j \\ &= (\mathbf{w})_i + \alpha\tau A_{ii}(\mathbf{w})_i - \alpha\tau \textstyle\sum_{j\ne i} |A_{ij}|\,(\mathbf{w})_j \\ &\ge (\mathbf{w})_i + \alpha\tau A_{ii}(\mathbf{w})_i - \alpha\tau \textstyle\sum_{j\ne i} |A_{ij}|(\mathbf{w})_i \\ &= (\mathbf{w})_i + \alpha\tau \left(A_{ii} - \textstyle\sum_{j\ne i} |A_{ij}|\right)(\mathbf{w})_i \\ &\ge (\mathbf{w})_i,\end{aligned}$$

where the last line holds by the diagonal dominance of A. This yields the desired result since $\|(I+\alpha\tau A)\mathbf{w}\|_\infty \ge \|\mathbf{w}\|_\infty$. □

We next estimate the maximum norm of $(I - \alpha\tau A)$ with constraints on τ.

Lemma 9.42. *Let A be a diagonally dominant M-matrix of size n. Then:*

$$\|I - \alpha\tau A\|_\infty \le 1,$$

provided $\alpha > 0$ and $\tau > 0$ satisfies:

$$\tau \le \min_{1\le i\le n} \frac{1}{\alpha\, A_{ii}}.$$

Proof. We use the property that:

$$\begin{aligned}
\|I - \alpha\tau A\|_\infty &= \max_{1\le i\le n} \left(|(I-\alpha\tau A)_{ii}| + \textstyle\sum_{j\ne i} |(I-\alpha\tau A)_{ij}|\right) \\
&= \max_{1\le i\le n} \left(|1-\alpha\tau A_{ii}| + \textstyle\sum_{j\ne i} |\alpha\tau A_{ij}|\right) \\
&= \max_{1\le i\le n} \left(|1-\alpha\tau A_{ii}| - \alpha\,\tau \textstyle\sum_{j\ne i} A_{ij}\right) \\
&= \max_{1\le i\le n} \left(1-\alpha\tau A_{ii} - \alpha\,\tau \textstyle\sum_{j\ne i} A_{ij}\right) \\
&= \max_{1\le i\le n} \left(1-\alpha\,\tau \textstyle\sum_{j} A_{ij}\right) \ \le 1.
\end{aligned}$$

Here, we have used that $(1-\alpha\tau A_{ii}) \ge 0$ to obtain the 4th equality. The last line follows since $\sum_j A_{ij} \ge 0$. □

As a corollary of the preceding two results, we obtain sufficient conditions for the maximum norm stability of the θ-scheme.

Lemma 9.43. *Let A be a diagonally dominant M-matrix of size n. Then:*

1. *If $0 \le \theta < 1$, the θ-scheme will be stable in $\|\cdot\|_\infty$ provided:*

$$\tau \le \min_{1\le i\le n} \frac{1}{(1-\theta)\, A_{ii}}. \tag{9.48}$$

2. *If $\theta = 1$, the backward Euler scheme will be unconditionally stable in $\|\cdot\|_\infty$.*

Proof. We shall employ the matrix representation of the θ-scheme:

$$\mathbf{u}^k \;=\; (I + \theta\,\tau\,A)^{-1}\,(I - (1-\theta)\tau\,A)\;\mathbf{u}^{k-1} + (I+\theta\,\tau\,A)^{-1}\;\tilde{\mathbf{f}}^k,$$

and estimate $\|\mathbf{u}^k\|_\infty$ using the triangle inequality:

$$\begin{aligned}
\|\mathbf{u}^k\|_\infty \le{}& \|\,(I+\theta\,\tau\,A)^{-1}\,\|_\infty\,\|I-(1-\theta)\,\tau\,A\|_\infty\,\|\mathbf{u}^{k-1}\|_\infty \\
&+ \|\,(I+\theta\,\tau\,A)^{-1}\,\|_\infty\,\|\tilde{\mathbf{f}}^k\|_\infty.
\end{aligned}$$

For $0 \le \theta < 1$, an application of Lemma 9.41 and Lemma 9.42 yields:

$$\|\mathbf{u}^k\|_\infty \le \|\mathbf{u}^{k-1}\|_\infty + \|\tilde{\mathbf{f}}^k\|_\infty, \tag{9.49}$$

provided (9.48) holds. For $\theta = 1$, the term $I-(1-\theta)\,\tau\,A = I$ and Lemma 9.41 yields bound (9.49). □

Lax's Convergence Theorem. The discretization (9.44) of the parabolic equation (9.43) will be convergent in a norm $\|\cdot\|$, if the error satisfies:

$$\|\mathbf{u}_*^k - \mathbf{u}^k\| \to 0 \quad \text{as} \quad (h, \tau) \to (0, 0),$$

where $\mathbf{u}_*^k$ denotes the restriction of the exact solution $u(x,t)$ of (9.43) to the grid points of $\mathcal{T}_h(\Omega)$ at time $t_k = k\,\tau$. The following result provides sufficient conditions for convergence, see [RI].

Lemma 9.44. *Suppose the following conditions hold.*

1. *Let discretization (9.44) be stable in a norm $\|\cdot\|$, satisfying (9.47).*
2. *Let discretization (9.44) be consistent, satisfying the following:*

$$\|\mathbf{T}_{orig}^k\| \le c_3(u)\,\tau(h^{q_1} + \tau^{q_2}) \;\text{ and }\; \|\mathbf{e}^0\| \le c_3(u)\,(h^{q_1} + \tau^{q_2}),$$

where $c_3(u) > 0$ is independent of h and τ (but dependent on u).

3. *Given $t_* > 0$, let $k_* = (t_*/\tau)$.*

Then, the error $\mathbf{e}^k \equiv \mathbf{u}_*^k - \mathbf{u}^k$ will satisfy:

$$\|\mathbf{e}^k\| \le c_4(u)\,t_*\,e^{c_1\,t_*}\,(h^{q_1} + \tau^{q_2})\,, \qquad \text{for} \quad 0 \le k \le k_*,$$

where $c_4(u) = (1 + c_2\,c_3(u)) > 0$ is independent of h and τ.

Proof. We describe the proof of sufficiency, see [RI] for necessary conditions. By definition of the local truncation error, $\mathbf{u}_*^k$ will satisfy:

$$(I + \alpha\,\tau\,A)\,\mathbf{u}_*^k + C\,\mathbf{u}_*^{k-1} = \tilde{\mathbf{f}}^k + \mathbf{T}_{orig}^k,$$

at time t_k. On the other hand, the discrete solution $\mathbf{u}^k$ will satisfy:

$$(I + \alpha\,\tau\,A)\,\mathbf{u}^k + C\,\mathbf{u}^{k-1} = \tilde{\mathbf{f}}^k.$$

Subtracting these two equations yields:

$$(I + \alpha\,\tau\,A)\,(\mathbf{u}_*^k - \mathbf{u}^k) + C\,(\mathbf{u}_*^k - \mathbf{u}^k) = \mathbf{T}_{orig}^k, \quad \text{for } 1 \le k \le k_*,$$

where typically $(\mathbf{u}_*^0 - \mathbf{u}^0) = \mathbf{0}$. Defining $\mathbf{e}^k \equiv (\mathbf{u}_*^k - \mathbf{u}^k)$ as the error vector and employing stability of the scheme in the norm $\|\cdot\|$, we obtain that:

$$\|\mathbf{e}^k\| \le (1 + c_1\,\tau)\,\|\mathbf{e}^{k-1}\| + c_2\,\|\mathbf{T}_{orig}^k\|, \quad \text{for } 1 \le k \le k_*.$$

Applying this bound recursively and using that $(1 + c_1\,\tau) \le e^{c_1\,\tau}$ yields:

$$\begin{aligned}
\|\mathbf{e}^k\| &\le e^{c_1\,k\,\tau}\,\|\mathbf{e}^0\| + c_2 \textstyle\sum_{i=1}^k e^{c_1\,(k-i)\,\tau} \|\mathbf{T}_{orig}^{k-i}\| \\
&\le e^{c_1\,k\,\tau}\|\mathbf{e}^0\| + c_2 \textstyle\sum_{i=1}^k e^{c_1\,k\,\tau} \|\mathbf{T}_{orig}^{k-i}\| \\
&\le e^{c_1\,k\,\tau}\|\mathbf{e}^0\| + c_2\,e^{c_1\,k\,\tau}\,k_*\,\max_i \|\mathbf{T}_{orig}^{k-i}\| \\
&\le e^{c_1\,k\,\tau}\|\mathbf{e}^0\| + c_2\,c_3(u)\,t_*\,e^{c_1\,t_*}\,(h^{q_1} + \tau^{q_2})\,,
\end{aligned}$$

where we have also used that $k_* \tau = t_*$. Since the initial error $\mathbf{e}^0$ satisfies $\|\mathbf{e}^0\| \leq c_3(u) \, (h^{q_1} + \tau^{q_2})$, we obtain:

$$\|\mathbf{e}^k\| \leq c_4(u) \, t_* \, e^{c_1 \, t_*} \, (h^{q_1} + \tau^{q_2}) \, ,$$

where $c_4(u) \equiv (1 + c_2 \, c_3(u))$ is independent of h and τ (but depends on the higher order derivatives of $u(x,t)$). □

9.5.2 Convergence Bounds for Time-Stepped Iterative Algorithms

We conclude this section by describing two convergence results on iterative algorithms for solving an implicit discretization of parabolic equation (9.1). The first result estimates the convergence rate of a Schwarz iterative algorithm *without coarse space* correction for solving a symmetric positive definite system of the form $(M + \tau \, A) \, \mathbf{u}^k = \tilde{\mathbf{g}}^k$, obtained by applying an implicit scheme in time and a finite element discretization in space, when $\mathbf{b}(x) = \mathbf{0}$, $c(x) \geq 0$. Here M is the finite element mass matrix. The second result analyzes the symmetric positive definite preconditioner $(I + \tau \, H)$ for solving a *non-symmetric* system $(I + \tau \, H + \tau \, N) \, \mathbf{u}^k = \tilde{\mathbf{g}}^k$, obtained by an implicit scheme in time and a finite difference discretization in space, when $\mathbf{b}(x) \neq \mathbf{0}$ and $c(x) \geq 0$.

Let $\Omega_1, \ldots, \Omega_p$ be a non-overlapping decomposition of Ω with subdomains of size h_0, and let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping decomposition of Ω with overlap $\beta \, h_0$, obtained by extending the non-overlapping subdomains.

Lemma 9.45. *Suppose the following conditions hold.*

1. *Let $\mathbf{b}(x) = \mathbf{0}$, $c(x) \geq 0$ and $a(x) = a_j$ in $\Omega_j \subset \Omega_j^*$ for $1 \leq j \leq p$.*
2. *Let V_h denote the continuous piecewise linear finite element space on a quasi-uniform triangulation $\mathcal{T}_h(\Omega)$ of Ω. Let $V_i \equiv V_h \cap H_0^1(\Omega_i^*)$.*
3. *Let $\mathcal{H}(.,.)$ and $\mathcal{M}(.,.)$ denote the following bilinear forms:*

$$\begin{cases} \mathcal{H}(u,v) & \equiv \int_\Omega a(x) \nabla u \cdot \nabla v \, dx, \quad \textit{for } u, v \in H_0^1(\Omega) \\ \mathcal{M}(u,v) & \equiv \int_\Omega (1 + \tau \, c(x)) u \, v \, dx, \quad \textit{for } u, v \in H_0^1(\Omega) \\ \mathcal{M}_{\Omega_i^*}(u,v) & \equiv \int_{\Omega_i^*} (1 + \tau \, c(x)) u \, v \, dx, \quad \textit{for } u, v \in H_0^1(\Omega). \end{cases}$$

Then, given $v_h \in V_h$, there exists $v_i \in V_i$ satisfying:

$$v_h = v_1 + \cdots + v_p,$$

and a parameter $C > 0$ independent of h, τ and h_0, such that:

$$\begin{array}{c} \sum_{i=1}^p \left(\mathcal{M}(v_i, v_i) + \tau \, \mathcal{H}(v_i, v_i) \right) \\ \leq C (1 + \tau \, \|a\|_\infty \, \beta^2 \, h_0^{-2}) \, \left(\mathcal{M}(v_h, v_h) + \tau \mathcal{H}(v_h, v_h) \right) . \end{array}$$

Proof. See [CA, CA3]. Given $v_h \in V_h$, we shall employ Lemma 2.72 from Chap. 2.5.3 and define $v_i = I_h \left(\chi_i(x) \, v_h(x) \right)$ where each $\chi_i(\cdot)$ is a partition of unity function subordinate to Ω_i^*, and I_h is the nodal interpolation map. If g_0 denotes the maximum number of subdomains Ω_j^* adjacent to a subdomain Ω_i^*, then equation (2.47) from Lemma 2.72 yields the bound:

$$\begin{aligned} \textstyle\sum_{i=1}^p \mathcal{H}(v_i, v_i) &\le 2g_0\, \mathcal{H}(v_h, v_h) + 2\, g_0\, C\beta^{-2} h_0^{-2} \textstyle\sum_{j=1}^p a_j \|v_h\|^2_{L^2(\Omega_j)} \\ &\le 2g_0\, \mathcal{H}(v_h, v_h) + 2\, g_0\, C\beta^{-2} h_0^{-2} \|a\|_\infty \|v_h\|^2_{L^2(\Omega)} \\ &\le 2g_0\, \mathcal{H}(v_h, v_h) + 2\, g_0\, C\beta^{-2} h_0^{-2} \|a\|_\infty \mathcal{M}(v_h, v_h). \end{aligned} \tag{9.50}$$

We shall next estimate $\|v_i\|^2_{L^2(\Omega)}$ where $v_i = I_h(\chi_i\, v_h)$ has support in Ω_i^*. We shall employ the property that on each element κ, the elemental mass matrix M_κ is *well conditioned* for a quasi-uniform triangulation, i.e., if I_κ denotes the identity matrix on κ and $\Omega \subset \mathbb{R}^d$, then there exists $0 < \gamma_1 < \gamma_2$ independent of h such that $\gamma_1\, h^d\, I_\kappa \le M_\kappa \le \gamma_2\, h^d\, I_\kappa$ holds in the sense of quadratic forms. Let R_κ denote the nodal restriction map onto nodes in κ, and let $\mathbf{v}$ and $\mathbf{v}_i$ denote the global nodal vectors associated with v_h and v_i, respectively. Then, using the well conditioned property yields that:

$$\begin{aligned} \|v_i\|^2_{L^2(\Omega)} = \textstyle\sum_{\kappa\in\Omega_i^*} \mathbf{v}_i^T R_\kappa^T M_\kappa R_\kappa \mathbf{v}_i &\le \gamma_2\, h^d \textstyle\sum_{\kappa\in\Omega_i^*} \mathbf{v}_i^T R_\kappa^T I_\kappa R_\kappa \mathbf{v}_i \\ &\le \gamma_2\, h^d \textstyle\sum_{\kappa\in\Omega_i^*} \mathbf{v}^T R_\kappa^T I_\kappa R_\kappa \mathbf{v}, \end{aligned}$$

since each entry of $\mathbf{v}_i$ is an entry of $\mathbf{v}$ multiplied by a factor $0 \le \chi_i(\cdot) \le 1$.

Again applying the well conditionedness of M_κ yields:

$$\|v_i\|^2_{L^2(\Omega)} \le \gamma_2\, \gamma_1^{-1} \sum_{\kappa\in\Omega_i^*} \mathbf{v}^T R_\kappa^T M_\kappa R_\kappa \mathbf{v} = \gamma_2\, \gamma_1^{-1} \|v\|^2_{L^2(\Omega_i^*)}.$$

Letting $\gamma = (\gamma_2/\gamma_1)$ we obtain:

$$\begin{aligned} \tfrac{1}{1+\tau\|c\|_\infty}\, \mathcal{M}(v_i, v_i) &\le \gamma\, \|v_h\|^2_{L^2(\Omega_i^*)} \\ &\le \gamma\, \mathcal{M}_{\Omega_i^*}(v_h, v_h). \end{aligned}$$

Assuming $\tau < 1$ and summing the preceding bounds for $1 \le i \le p$ yields:

$$\sum_{i=1}^p \mathcal{M}(v_i, v_i) \le g_0\, \gamma(1 + \|c\|_\infty)\mathcal{M}(v_h, v_h). \tag{9.51}$$

Combining (9.50) and (9.51) yields:

$$\begin{aligned} &\textstyle\sum_{i=1}^p \left(\mathcal{M}(v_i, v_i) + \tau\mathcal{H}(v_i, v_i)\right) \\ &\le 2g_0\, \tau\, \mathcal{H}(v_h, v_h) + \left(g_0\, \gamma(1 + \|c\|_\infty) + 2g_0\, C\, \tau\, \|a\|_\infty\, \beta^{-2} h_0^{-2}\right) \mathcal{M}(v_h, v_h), \end{aligned}$$

from which the desired result follows. □

Remark 9.46. Thus, the partition parameter K_0 from Chap. 2.3 satisfies:

$$K_0 \le C\,(1 + \tau\, \|a\|_\infty\, \beta^{-2}\, h_0^{-2}).$$

Since the partition parameter K_1 (see Chap. 2.3) is bounded independent of h, h_0, τ and $a(.)$, we may combine both bounds to obtain that the convergence rate of Schwarz algorithms, without coarse space correction, depends primarily on $\tau \, \|a\|_\infty \, \beta^{-2} \, h_0^{-2}$ for implicit discretizations of parabolic equations. Thus, provided $\tau \, \leq \, c \, \beta^2 \, h_0^2$, its convergence will be uniform.

We next consider the solution of a *non-symmetric* time-stepped system.

Lemma 9.47. *Suppose the following assumptions hold.*

1. *Let $A = H + N$, with $H^T = H > 0$ and N non-symmetric.*
2. *For $0 \leq \tau$ let $M = (I + \tau \, H)$ denote a preconditioner for $(I + \tau A)$.*
3. *Let $\tau \, \sigma_{\max}(N) \, \leq \, \delta_0 \, < \, 1$, where $\sigma_{\max}(N) \, = \, \|N\|$ denotes the maximum singular value of N and $\| \cdot \|$ the Euclidean norm.*

Then, the following bounds will hold.

1. $\|I - M^{-1} \, (I + \tau A) \, \| \leq \delta_0$.
2. *The GMRES iterates $\mathbf{z}^{(k)}$ when solving $M^{-1}(I + \tau A)\mathbf{u} = \mathbf{f}$ will satisfy:*

$$\|\mathbf{u} - \mathbf{z}^{(k)}\| \leq \left(1 - \left(\frac{1 - \delta_0}{1 + \delta_0} \right)^2 \right)^k \|\mathbf{u} - \mathbf{z}^{(0)}\|.$$

Proof. Since $(I + \tau H)^{-1}(I + \tau H + \tau N) = I + \tau \, (I + \tau H)^{-1} N$, we obtain:

$$\begin{aligned} \|I - (I + \tau H)^{-1}(I + \tau H + \tau N)\| &= \tau \, \|(I + \tau H)^{-1} N\| \\ &\leq \tau \, \|(I + \tau H)^{-1}\| \, \|N\| \\ &\leq \tau \|N\| \\ &= \delta_0, \end{aligned}$$

since $\|(I + \tau H)^{-1}\| \leq 1$. It thus holds that $\|M^{-1}(I + \tau \, A)\| \leq \, 1 + \delta_0$. Next, since $\|I - M^{-1} \, (I + \tau A) \, \| \leq \delta_0$, it will hold that:

$$\left(\mathbf{z}^* \mathbf{z} - \mathbf{z}^* M^{-1}(I + \tau \, A)\mathbf{z} \right) \, \leq \, \delta_0 \, \mathbf{z}^* \mathbf{z},$$

so that:

$$(1 - \delta_0) \, \mathbf{z}^* \mathbf{z} \, \leq \, \mathbf{z}^* M^{-1}(I + \tau \, A)\mathbf{z},$$

which shows that $\lambda_{\min} \left(\frac{1}{2}(M^{-1}(I + \tau \, A) + (I + \tau \, A)^* M^{-*}) \right) \geq (1 - \delta_0) > 0$. The bounds for the GMRES iterates follows immediately. □

Remark 9.48. For a finite difference discretization, $H\mathbf{u}$ will correspond to a discretization of $-\nabla \cdot (a(x) \nabla u)$ and $N\mathbf{u}$ the discretization of $\mathbf{b}(x) \cdot \nabla u + c(x) u$. In this case, it will hold that $H^T = H > 0$, and $\|N\| = \sigma_{\max}(N) \leq c_1 \, h^{-1}$ for some $c_1 > 0$ independent of h. Thus, if $\tau \, \leq \, c \, h$, then $(I + \tau \, H)$ will be an effective preconditioner for $(I + \tau \, A)$.

Remark 9.49. In applications, the symmetric positive definite preconditioner $(I + \tau H)$ may be replaced by an appropriate symmetric domain decomposition preconditioner, provided the time step τ satisfies $c_1 \, \tau \, h^{-1} = \delta_0 < 1$.

10

Saddle Point Problems

Saddle point problems are associated with *constrained optimization* problems. The search for the minimum $\mathbf{u}$ of an objective functional $J(\mathbf{v})$ in $\mathbb{R}^n$ subject to m constraints $B(\mathbf{v}) - \mathbf{g} = \mathbf{0}$ can be transformed into a search for the *saddle point* $(\mathbf{u}, \boldsymbol{\mu})$ (i.e., a critical point which is not a local optimum) of the associated Lagrangian functional $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) = J(\mathbf{v}) + \boldsymbol{\eta}^T (B(\mathbf{v}) - \mathbf{g})$. Here, the introduced variables $\boldsymbol{\eta} \in \mathbb{R}^m$ are referred to as Lagrange multipliers, and the search for the saddle point of $\mathcal{L}(.,.)$ is unconstrained.

The problem that we shall consider will seek to minimize the quadratic objective functional $J(\mathbf{v}) \equiv \frac{1}{2}\mathbf{v}^T A\mathbf{v} - \mathbf{v}^T\mathbf{f}$ defined in $\mathbb{R}^n$, subject to $m < n$ *linear constraints* $B\mathbf{v} = \mathbf{g}$, where $A = A^T$ is positive semi-definite matrix of size n and B is a full rank matrix of size $m \times n$. The Lagrangian functional $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) = J(\mathbf{v}) + \boldsymbol{\eta}^T (B\mathbf{v} - \mathbf{g})$ is associated with this constrained minimization problem, where $\boldsymbol{\eta} \in \mathbb{R}^m$ denotes the Lagrange multipliers. The first derivative test for a critical point $(\mathbf{u}, \boldsymbol{\mu})$ of $\mathcal{L}(.,.)$ yields the *indefinite* linear system:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \tag{10.1}$$

which is referred to as a *saddle point* linear system. By construction, $\mathbf{u}$ will minimize $J(\mathbf{v})$ within the constraint set satisfying $B\mathbf{v} = \mathbf{g}$.

Our focus in this chapter will be on *iterative algorithms* for solving (10.1). Chap. 10.1 describes various properties of saddle point systems. Chap. 10.2 introduces the *duality* formulation and *Uzawa's algorithm.* Chap. 10.3 describes the *penalty and regularization* method for obtaining an approximate solution. Chap. 10.4 describes *projection methods.* Chap. 10.5 describes *block matrix* preconditioners and Krylov algorithms. Applications to Navier-Stokes equations, mixed formulations of elliptic equations, and to optimal control problems, are described in Chaps. 10.6, 10.7 and 10.8, respectively. For a more detailed discussion of saddle point problems, readers are referred to [CI4, GI3, BE12].

10.1 Properties of Saddle Point Systems

In this section, we shall describe properties of the saddle point system:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{10.2}$$

where $A = A^T$ is a positive semi-definite matrix of size n, and B is a full rank matrix of size $m \times n$ with $m < n$, with $\mathbf{u}$, $\mathbf{f} \in \mathbb{R}^n$ and $\boldsymbol{\mu}$, $\mathbf{g} \in \mathbb{R}^m$. We discuss the solvability of saddle point systems, and the distribution of its eigenvalues. See [CI4, GI3, BE12] for a more detailed discussion of saddle point problems.

10.1.1 Constrained Minimization Formulation

The following result [CI4, GI3] describes how linear system (10.2) arises as a necessary condition for determining the minimum of the functional $J(\cdot)$:

$$J(\mathbf{v}) \equiv \frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}, \tag{10.3}$$

within the constraint set $B\mathbf{v} = \mathbf{g}$. The additional variables $\boldsymbol{\mu} \in \mathbb{R}^m$ in (10.2), referred to as Lagrange multipliers, arise when enforcing first order conditions for the minimization of $J(\cdot)$ within the constraint set $B\mathbf{v} = \mathbf{g}$. We define $\mathcal{K}_{\mathbf{g}} \subset \mathbb{R}^n$ and $\mathcal{K}_0 \subset \mathbb{R}^n$ as the following convex *constraint sets*:

$$\begin{aligned} \mathcal{K}_{\mathbf{g}} &= \{\mathbf{v} \in \mathbb{R}^n : B\mathbf{v} = \mathbf{g}\} \\ \mathcal{K}_0 &= \{\mathbf{v} \in \mathbb{R}^n : B\mathbf{v} = \mathbf{0}\} \;=\; \text{Kernel}(B), \end{aligned} \tag{10.4}$$

defined by the linear constraints $B\mathbf{v} = \mathbf{g}$ and $B\mathbf{v} = \mathbf{0}$, respectively.

Lemma 10.1. *Suppose the following conditions hold.*

1. *Let* $\mathbf{u}$ *denote the solution of the constrained minimization problem:*

$$J(\mathbf{u}) = \min_{\mathbf{v} \in \mathcal{K}_{\mathbf{g}}} J(\mathbf{v}) \tag{10.5}$$

2. *Let* B *be a matrix of full rank* m.

Then, there will exist a vector $\boldsymbol{\mu} \in \mathbb{R}^m$ *such that:*

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{10.6}$$

Proof. To verify the first block row of (10.6) consider the curve $\mathbf{x}(t) = \mathbf{u} + t\,\mathbf{v}$ for $t \in \mathbb{R}$, for arbitrary $\mathbf{v} \in \mathcal{K}_0$. By construction $\mathbf{x}(t) \in \mathcal{K}_{\mathbf{g}}$ and passes through $\mathbf{u}$ when $t = 0$. Since the minimum of $J(\cdot)$ within $\mathcal{K}_{\mathbf{g}}$ is attained at $\mathbf{u}$ and since

$\mathbf{x}(0) = \mathbf{u}$ with $\mathbf{x}(t) \subset \mathcal{K}_{\mathbf{g}}$, the function $J\left(\mathbf{x}(t)\right)$ must attain its minimum when $t = 0$. Applying the derivative test and substituting that $\left.\frac{d\mathbf{x}(t)}{dt}\right|_{t=0} = \mathbf{v}$ yields:

$$\begin{aligned}
\left.\frac{dJ(\mathbf{x}(t))}{dt}\right|_{t=0} = 0, \;\; \forall \mathbf{v} \in \mathcal{K}_0 &\Leftrightarrow \nabla J\left(\mathbf{u}\right) \cdot \mathbf{v} = 0, \;\; \forall \mathbf{v} \in \mathcal{K}_0 \\
&\Leftrightarrow \nabla J\left(\mathbf{u}\right) \perp \text{Kernel}(B) \\
&\Leftrightarrow \nabla J\left(\mathbf{u}\right) \in \text{Kernel}(B)^{\perp} \\
&\Leftrightarrow \nabla J\left(\mathbf{u}\right) \in \text{Range}(B^T).
\end{aligned}$$

Since we may represent any vector in Range(B^T) in the form $-B^T\boldsymbol{\mu}$ for some $\boldsymbol{\mu} \in \mathbb{R}^m$ (the negative sign here is for convenience), we obtain that:

$$\nabla J\left(\mathbf{u}\right) = A\mathbf{u} - \mathbf{f} = -B^T\boldsymbol{\mu}, \qquad \text{for some } \boldsymbol{\mu} \in \mathbb{R}^m,$$

which yields the first block row of (4.22). The second block row of (10.6) holds since $\mathbf{u} \in \mathcal{K}_{\mathbf{g}}$, yielding $B\mathbf{u} = \mathbf{g}$. □

Definition 10.2. *The components μ_i of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$ are referred to as Lagrange multipliers. A functional $\mathcal{L}(\mathbf{v}, \boldsymbol{\lambda})$, referred to as a Lagrangian function, is associated with (10.2):*

$$\mathcal{L}(\mathbf{v}, \boldsymbol{\lambda}) \equiv \left(\frac{1}{2}\mathbf{v}^T A\mathbf{v} - \mathbf{v}^T\mathbf{f}\right) + \boldsymbol{\lambda}^T\left(B\mathbf{v} - \mathbf{g}\right). \tag{10.7}$$

It is easily verified that linear system (10.2) arises from the first order derivative test for a *critical point* $(\mathbf{u}, \boldsymbol{\mu})$ of $\mathcal{L}(\cdot, \cdot)$:

$$\begin{cases}
\left.\frac{\partial\mathcal{L}}{\partial\mathbf{v}}\right|_{(\mathbf{u},\boldsymbol{\mu})} = A\,\mathbf{u} + B^T\boldsymbol{\mu} - \mathbf{f} = \mathbf{0} \\
\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda}}\right|_{(\mathbf{u},\boldsymbol{\mu})} = B\mathbf{u} - \mathbf{g} \qquad\qquad = \mathbf{0}.
\end{cases}$$

In § 5.2, it is shown that $(\mathbf{u}, \boldsymbol{\mu})$ corresponds to a *saddle point* of $\mathcal{L}(\mathbf{v}, \boldsymbol{\lambda})$, i.e.,

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\lambda}) \le \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) \le \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}), \qquad \forall\, \mathbf{v} \in \mathbb{R}^n,\, \boldsymbol{\lambda} \in \mathbb{R}^m.$$

This property motivates the name saddle point system for (10.2).

10.1.2 Well-Posedness of the Saddle Point System

We shall now consider the solvability and well posedness of system (10.2), and provide estimates for the Euclidean norm of its solution. For simplicity, we shall assume that matrix A is positive definite and that matrix B has full rank. When matrix A is positive semidefinite, system (10.2) will still be uniquely solvable, provided A is positive definite within the subspace $\mathcal{K}_0 = \text{Kernel}(B)$, see [GI3]. In this case, the solution to (10.2) may be sought using an augmented Lagrangian approach [GL7] described in Chap. 10.2, the projection approach from Chap. 10.4, or the FETI approach [FA14] from Chap. 4.

Lemma 10.3. *Suppose the following conditions hold.*

1. *Let A be symmetric positive definite of size n.*
2. *Let matrix B of size $m \times n$ have full rank m, with $m < n$.*
3. *Let $S \equiv (BA^{-1}B^T)$ denote a Schur complement matrix.*

Then system (10.2) will be nonsingular with solution $(\mathbf{u}, \boldsymbol{\mu})$ satisfying:

$$\begin{cases} \|\mathbf{u}\| \le \left(\|A^{-1}\mathbf{f}\| + \|A^{-1}B^TS^{-1}BA^{-1}\mathbf{f}\| + \|A^{-1}B^TS^{-1}\mathbf{g}\|\right) \\ \|\boldsymbol{\mu}\| \le \left(\|S^{-1}BA^{-1}\mathbf{f}\| + \|S^{-1}\mathbf{g}\|\right), \end{cases}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Proof. Since A is nonsingular, solving for $\mathbf{u}$ in terms of $\boldsymbol{\mu}$ using the first block row of (10.2) yields $\mathbf{u} = A^{-1}\left(\mathbf{f} - B^T\boldsymbol{\mu}\right)$. Substituting this expression for $\mathbf{u}$ into the second block row of (10.2), yields $S\,\boldsymbol{\mu} = (BA^{-1}\mathbf{f} - \mathbf{g})$. Since A is nonsingular and B is of full rank, S will be *nonsingular* (see Lemma 10.11 for bounds on the eigenvalues of S), yielding the bound:

$$\|\boldsymbol{\mu}\| \le \left(\|S^{-1}B\,A^{-1}\mathbf{f}\| + \|S^{-1}\mathbf{g}\|\right).$$

Substituting the expression for $\boldsymbol{\mu}$ into $\mathbf{u} = A^{-1}\left(\mathbf{f} - B^T\boldsymbol{\mu}\right)$ and estimating:

$$\|\mathbf{u}\| \le \left(\|A^{-1}\mathbf{f}\| + \|A^{-1}B^TS^{-1}BA^{-1}\mathbf{f}\| + \|A^{-1}B^TS^{-1}\mathbf{g}\|\right).$$

This completes the proof. □

Remark 10.4. In our applications, A^{-1}, S^{-1}, $A^{-1}B^TS^{-1}BA^{-1}$, $A^{-1}B^TS^{-1}$ and $S^{-1}BA^{-1}$ will have bounds independent of h in the Euclidean norm.

Remark 10.5. System (10.6) will be solvable even if A is *singular*, provided A is *coercive* within the subspace $\mathcal{K}_0$ and matrix B has rank m. Under these assumptions, the solution to (10.6) can be determined in three steps. In the first step, solve the 2nd block row of (10.6) to determine any $\mathbf{w} \in \mathbb{R}^n$ such that $B\,\mathbf{w} = \mathbf{g}$. For instance, we may seek $\mathbf{w} = B^T\boldsymbol{\gamma}$ for $\boldsymbol{\gamma} \in \mathbb{R}^m$ and solve $(BB^T)\boldsymbol{\gamma} = \mathbf{g}$, where (BB^T) will have full rank. By construction, it will hold that $\mathbf{u}_0 \equiv (\mathbf{u} - \mathbf{w}) \in \mathcal{K}_0$. In the second step, using the first block row of (10.6), we seek $\mathbf{u}_0 \in \mathcal{K}_0$ satisfying:

$$\begin{cases} \mathbf{v}_0^T\,A\,\mathbf{u}_0 = \mathbf{v}_0^T\left(\mathbf{f} - A\,\mathbf{w} - B^T\boldsymbol{\mu}\right), & \forall\,\mathbf{v}_0 \in \mathcal{K}_0 \\ \qquad\qquad = \mathbf{v}_0^T\left(\mathbf{f} - A\,\mathbf{w}\right), & \forall\,\mathbf{v}_0 \in \mathcal{K}_0, \end{cases}$$

where $\mathbf{v}_0^TB^T\boldsymbol{\mu} = (B\mathbf{v}_0)^T\boldsymbol{\mu} = 0$. The problem to determine $\mathbf{u}_0 \in \mathcal{K}_0$ will be well posed, since A is *coercive* within $\mathcal{K}_0$. Furthermore, the residual will satisfy $\mathbf{v}_0^T\left(\mathbf{f} - A\left(\mathbf{w} + \mathbf{u}_0\right)\right) = 0$ for all $\mathbf{v}_0 \in \mathcal{K}_0$, yielding that $\left(\mathbf{f} - A\left(\mathbf{w} + \mathbf{u}_0\right)\right) \in \mathcal{K}_0^\perp$. Since $\mathcal{K}_0 = \text{Kernel}(B)$, the fundamental theorem of linear algebra will yield that $\left(\mathbf{f} - A\left(\mathbf{w} + \mathbf{u}_0\right)\right) \in \text{Range}(B^T)$. Thus, there must be some $\boldsymbol{\mu} \in \mathbb{R}^m$ such that $\left(\mathbf{f} - A\left(\mathbf{w} + \mathbf{u}_0\right)\right) = -B^T\boldsymbol{\mu}$. To determine $\boldsymbol{\mu} \in \mathbb{R}^m$, in the third step multiply the first block row of (10.6) by B^T and solve $(BB^T)\boldsymbol{\mu} = B^T(\mathbf{f} - A\mathbf{u})$ for $\boldsymbol{\mu}$. By construction, $(\mathbf{w} + \mathbf{u}_0, \boldsymbol{\mu})$ will solve (10.6).

In various applications, it will be of interest to employ different norms for $\mathbf{u}$ and $\boldsymbol{\mu}$ and to employ estimates in such norms [GI3]. It will be convenient to formulate saddle point problem (10.6) weakly using bilinear forms. Accordingly, let $\mathcal{U}$ and $\mathcal{Q}$ denote finite dimensional Hilbert spaces of dimension n and m respectively, with inner products $(.,.)_{\mathcal{U}}$ and $(.,.)_{\mathcal{Q}}$, and norms $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{Q}}$, respectively. We shall let $\mathcal{U}'$ and $\mathcal{Q}'$ denote the dual spaces of $\mathcal{U}$ and $\mathcal{Q}$, respectively, and employ $\langle .,.\rangle_{\mathcal{U}}$ and $\langle .,.\rangle_{\mathcal{Q}}$ to denote their respective duality pairings. The following notation shall be employed.

- Given a symmetric bilinear form $\mathcal{A}(u,v)$ defined for $u,v \in \mathcal{U}$, we define $A : \mathcal{U} \to \mathcal{U}'$ as a linear map associated with it:

  $$\langle Au, v\rangle_{\mathcal{U}} = \mathcal{A}(u,v), \quad \forall u, v \in \mathcal{U}.$$

  If $\mathbf{u}$ and $\mathbf{v}$ denote vector representations of u and v in $\mathcal{U}$, relative to some chosen basis, then we may represent $\mathcal{A}(u,v) \equiv \mathbf{v}^T A\mathbf{u}$ with $A = A^T$.
- Similarly, given a bilinear form $\mathcal{B}(u,\mu)$ for $u \in \mathcal{U}$ and $\mu \in \mathcal{Q}$, we let $B : \mathcal{U} \to \mathcal{Q}'$ denote the linear map associated it:

  $$\langle Bu, \mu\rangle_{\mathcal{Q}} = \mathcal{B}(u,\mu), \quad \forall u \in \mathcal{U}, \ \ \mu \in \mathcal{Q}.$$

  If $\mathbf{u}$ and $\boldsymbol{\mu}$ denote vector representations of $u \in \mathcal{U}$ and $\mu \in \mathcal{Q}$, then B will be a matrix of size $m \times n$ such that $\mathcal{B}(u,\mu) \equiv \boldsymbol{\mu}^T B\mathbf{u}$, in the chosen basis.
- Given $f \in \mathcal{U}'$ and $g \in \mathcal{Q}'$ let $\langle u, f\rangle_{\mathcal{U}} = \mathbf{u}^T\mathbf{f}$ and $\langle \mu, g\rangle_{\mathcal{Q}} = \boldsymbol{\mu}^T\mathbf{g}$ denote their vector representations.

The weak form of saddle point problem (10.6) will seek $u \in \mathcal{U}$ and $\mu \in \mathcal{Q}$:

$$\begin{cases} \mathcal{A}(u,v) + \mathcal{B}(v,\mu) = \langle f, v\rangle_{\mathcal{U}}, & \forall v \in \mathcal{U} \\ \qquad\qquad\quad \mathcal{B}(u,\eta) = \langle g, \eta\rangle_{\mathcal{Q}}, & \forall \eta \in \mathcal{Q}. \end{cases} \tag{10.8}$$

The solvability of (10.8) will require the *coercivity* of $\mathcal{A}(.,.)$ within the subspace $\mathcal{K}_0 = \{v \in \mathcal{U} \, : \, \mathcal{B}(v,\eta) = 0, \ \ \forall v \in \mathcal{U}\}$ and an *inf-sup* condition for $\mathcal{B}(.,.)$.

Definition 10.6. *We say that the bilinear form $\mathcal{B}(\cdot,\cdot)$ satisfies an* **inf-sup** *condition, if there exists a constant $\beta > 0$ such that:*

$$\inf_{\mu \in \mathcal{Q}\setminus\{0\}} \ \sup_{u \in \mathcal{U}\setminus\{0\}} \frac{\mathcal{B}(u,\mu)}{\|u\|_{\mathcal{U}} \, \|\mu\|_{\mathcal{Q}}} \geq \beta. \tag{10.9}$$

The *inf-sup* condition is trivially equivalent to the requirement:

$$\sup_{u \in \mathcal{U}\setminus\{0\}} \frac{\mathcal{B}(u,\mu)}{\|u\|_{\mathcal{U}}} \geq \beta \, \|\mu\|_{\mathcal{Q}}, \quad \forall \mu \in \mathcal{Q}. \tag{10.10}$$

For the *inf-sup* condition to hold, the map $B : \mathcal{U} \to \mathcal{Q}'$ must be *surjective.*

Remark 10.7. If $\mathcal{U} = R^n$ and $\mathcal{Q} = R^m$ are endowed with standard Euclidean inner products, and $\mathcal{B}(\mathbf{u}, \boldsymbol{\mu}) = \boldsymbol{\mu}^T B\mathbf{u}$, then the *inf-sup* condition requires $\text{Range}(B) = \mathcal{Q}$, i.e., B must have full rank $m < n$. Furthermore, it requires:

$$\beta = \sigma_1(B) > 0,$$

where $\sigma_1(B)$ denotes the smallest singular value of B.

Remark 10.8. Given a map $B : \mathcal{U} \to \mathcal{Q}'$, let $\mathcal{K}_0 = \text{Kernel}(B)$ denote:

$$\mathcal{K}_0 = \{u \in \mathcal{U} : B\,u = 0\}.$$

Then, the restricted mapping $B : \mathcal{K}_0^{\perp} \subset \mathcal{U} \to \mathcal{Q}'$ will be *one to one*. The *inf-sup* condition shows that B will also be *onto*, see [GI3], since:

$$\sup_{u\,\in\,\mathcal{K}_0^{\perp}} \frac{\mathcal{B}(u,\mu)}{\|u\|_{\mathcal{U}}} \;\geq\; \beta\,\|\mu\|_{\mathcal{Q}},$$

otherwise there would be $\mu \in \mathcal{Q}$ for which the right hand side above is zero. Furthermore, the map B will satisfy:

$$\|B\,u\|_{\mathcal{Q}'} \;\geq\; \beta\,\|u\|_{\mathcal{U}}, \qquad \text{for} \;\; u \,\in\, \mathcal{K}_0^{\perp},$$

so that the pseudoinverse of B satisfies $\|B^{\dagger}\| \leq \frac{1}{\beta}$.

The following result derives norm bounds for the solution of (10.8).

Lemma 10.9. *Suppose the following conditions hold:*

1. *Let the symmetric bilinear form $\mathcal{A}(.,.)$ be coercive in $\mathcal{K}_0$ with:*

$$\alpha_0\,\|v\|_{\mathcal{U}}^2 \leq \mathcal{A}(v,v), \quad \forall v \in \mathcal{K}_0,$$

   for some $\alpha_0 > 0$, where $\mathcal{K}_0 = \{v \in \mathcal{U} \,:\, \mathcal{B}(v,\eta) = 0, \;\; \forall v \in \mathcal{U}\}$.
2. *Let $\mathcal{A}(.,.)$ be bounded with $|\mathcal{A}(u,v)| \;\leq\; \alpha_1\,\|u\|_{\mathcal{U}}\,\|v\|_{\mathcal{U}}$ for some $\alpha_1 > 0$.*
3. *Let the inf-sup condition hold for $\mathcal{B}(\cdot,\cdot)$ with constant $\beta > 0$.*
4. *Let $c > 0$ be such that $\|Bu\|_{\mathcal{Q}'} \;\leq\; c\,\|u\|_{\mathcal{U}}$, for all $u \in \mathcal{U}$.*

Then, the solution $(u,\mu) \in \mathcal{U} \times \mathcal{Q}$ of (10.8) will satisfy the bounds:

$$\begin{cases} \|u\|_{\mathcal{U}} \;\leq\; \frac{1}{\alpha_0}\|f\|_{\mathcal{U}'} + \frac{1}{\beta}\left(1 + \frac{\alpha_1}{\alpha_0}\right)\|g\|_{\mathcal{Q}'} \\ \|\mu\|_{\mathcal{Q}} \leq \frac{1}{\beta}\left(1 + \frac{\alpha_1}{\alpha_0}\right)\|f\|_{\mathcal{U}'} + \frac{\alpha_1}{\beta^2}\left(1 + \frac{\alpha_1}{\alpha_0}\right)\|g\|_{\mathcal{Q}'}. \end{cases}$$

Proof. Employ the three steps from Remark 10.5 and estimate the terms. □

10.1.3 Eigenvalues of the Saddle Point Matrix

The distribution of eigenvalues of the saddle point matrix L:

$$L = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$$

can be estimated using energy arguments [RU5]. The convergence rate of Krylov space algorithms to solve (10.2) will depend on the intervals $[a, b] \cup [c, d]$ containing the eigenvalues of L, where $a < b < 0$ and $0 < c < d$, see [RU5]. In Chap. 10.5, block matrix preconditioners are described for L.

Theorem 10.10. *Let λ be an eigenvalue of L with eigenvector $\left(\mathbf{u}^T, \boldsymbol{\mu}^T\right)^T$:*

$$\begin{cases} A\mathbf{u} + B^T\boldsymbol{\mu} = \lambda\, \mathbf{u} \\ B\mathbf{u} \qquad\quad\; = \lambda\, \boldsymbol{\mu}. \end{cases} \tag{10.11}$$

Furthermore, suppose that the following conditions hold:

1. $0 < \gamma_1 \le \cdots \le \gamma_n$ denote the eigenvalues of the $n \times n$ matrix $A = A^T \ge 0$.
2. Let $0 < \sigma_1 \le \cdots \le \sigma_m$ denote the singular values of the $m \times n$ matrix B.

Then any eigenvalue λ of L must lie in the union of the intervals:

$$\lambda \in \left[\frac{1}{2}(\gamma_1 - \sqrt{\gamma_1^2 + 4\sigma_m^2}), \frac{1}{2}(\gamma_n - \sqrt{\gamma_n^2 + 4\sigma_1^2})\right] \cup \left[\gamma_1, \frac{1}{2}(\gamma_n + \sqrt{\gamma_n^2 + 4\sigma_m^2})\right],$$

where $\frac{1}{2}(\gamma_n - \sqrt{\gamma_n^2 + 4\sigma_1^2}) < 0$ and $0 < \gamma_1$.

Proof. See [RU5]. Since matrix L is real and symmetric, its eigenvalues and eigenvectors will be real. Taking an inner product of the first row of (10.11) with $\mathbf{u}$, and the second row with $\boldsymbol{\mu}$ will yield the following expressions:

$$\begin{cases} \mathbf{u}^T A\mathbf{u} + \mathbf{u}^T B^T\boldsymbol{\mu} = \lambda\, \mathbf{u}^T\mathbf{u} \\ \qquad\qquad\; \boldsymbol{\mu}^T B\mathbf{u} = \lambda\, \boldsymbol{\mu}^T\boldsymbol{\mu}. \end{cases} \tag{10.12}$$

Since L is non-singular, we obtain $\lambda \ne 0$. Thus, the second row of (10.11) yields $\boldsymbol{\mu} = \frac{1}{\lambda} B\,\mathbf{u}$, and substituting this into the first block row of (10.12) and multiplying the resulting equation by λ and rearranging terms yields the following quadratic equation for λ:

$$\lambda^2\,(\mathbf{u}^T\mathbf{u}) \;-\; \lambda\,(\mathbf{u}^T A\mathbf{u}) \;-\; (\mathbf{u}^T B^T B\mathbf{u}) \;=\; 0. \tag{10.13}$$

Since $\lambda \ne 0$, it can easily be verified that $\mathbf{u} \ne \mathbf{0}$. Upper and lower bounds can be obtained for λ by substituting the following bounds into (10.13):

$$\begin{cases} \gamma_1\, \mathbf{u}^T\mathbf{u} \le \quad \mathbf{u}^T A\mathbf{u} \quad \le \gamma_n\, \mathbf{u}^T\mathbf{u} \\ \sigma_1\, \mathbf{u}^T\mathbf{u} \le \mathbf{u}^T B^T B\mathbf{u} \le \sigma_m\, \mathbf{u}^T\mathbf{u} \end{cases}$$

and solving for λ, as described next.

(i). To derive a *lower* bound for λ, when $\lambda > 0$, we substitute into (10.13) that $(\mathbf{u}^T B^T B\mathbf{u}) \geq 0$ and $(\mathbf{u}^T A\mathbf{u}) \geq \gamma_1 (\mathbf{u}^T \mathbf{u})$ to obtain:

$$\begin{aligned} 0 \leq \lambda^2 (\mathbf{u}^T \mathbf{u}) - \lambda (\mathbf{u}^T A\mathbf{u}) &\leq \lambda^2 (\mathbf{u}^T \mathbf{u}) - \lambda \gamma_1 (\mathbf{u}^T \mathbf{u}) \\ &= \lambda (\mathbf{u}^T \mathbf{u}) (\lambda - \gamma_1) . \end{aligned}$$

Since $\lambda (\mathbf{u}^T \mathbf{u}) > 0$ this requires $\lambda \geq \gamma_1$.

(ii). To derive an *upper* bound for λ when $\lambda > 0$, we substitute into (10.13) that $(\mathbf{u}^T A\mathbf{u}) \leq \gamma_n (\mathbf{u}^T \mathbf{u})$ and $(\mathbf{u}^T B^T B\mathbf{u}) \leq \sigma_m^2 (\mathbf{u}^T \mathbf{u})$ to obtain:

$$\begin{aligned} 0 = \lambda^2(\mathbf{u}^T \mathbf{u}) - \lambda(\mathbf{u}^T A\mathbf{u}) - (\mathbf{u}^T B^T B\mathbf{u}) &\leq \lambda^2(\mathbf{u}^T \mathbf{u}) - \lambda \gamma_n(\mathbf{u}^T \mathbf{u}) - \sigma_m^2(\mathbf{u}^T \mathbf{u}) \\ &\leq \left(\lambda^2 - \gamma_n \lambda - \sigma_m^2\right) (\mathbf{u}^T \mathbf{u}). \end{aligned}$$

Since $r_1 = \frac{1}{2}\left(\gamma_n - \sqrt{\gamma_n^2 + 4\sigma_m^2}\right) < 0$ and $r_2 = \frac{1}{2}\left(\gamma_1 + \sqrt{\gamma_n^2 + 4\sigma_m^2}\right)$ are the roots of the quadratic $\left(\lambda^2 - \gamma_n \lambda - \sigma_m^2\right) = (\lambda - r_1)(\lambda - r_2)$, we obtain that this polynomial is *non-positive* in the interval (r_1, r_2). Since $\lambda > 0$, we require:

$$\lambda \leq \frac{1}{2}\left(\gamma_n + \sqrt{\gamma_n^2 + 4\sigma_m^2}\right).$$

(iii). To derive a *lower* bound for λ when $\lambda < 0$, we substitute into (10.13) that $-\lambda (\mathbf{u}^T A\mathbf{u}) \geq -\lambda \gamma_1 (\mathbf{u}^T \mathbf{u})$ and $-(\mathbf{u}^T B^T B\mathbf{u}) \geq -\sigma_m^2 (\mathbf{u}^T \mathbf{u})$ to obtain:

$$\begin{aligned} 0 = \lambda^2(\mathbf{u}^T \mathbf{u}) - \lambda(\mathbf{u}^T A\mathbf{u}) - (\mathbf{u}^T B^T B\mathbf{u}) &\geq \lambda^2(\mathbf{u}^T \mathbf{u}) - \gamma_1 \lambda(\mathbf{u}^T \mathbf{u}) - \sigma_m^2(\mathbf{u}^T \mathbf{u}) \\ &= \left(\lambda^2 - \gamma_1 \lambda - \sigma_m^2\right) (\mathbf{u}^T \mathbf{u}). \end{aligned}$$

Since $r_1 = \frac{1}{2}\left(\gamma_1 - \sqrt{\gamma_1^2 + 4\sigma_m^2}\right) < 0$ and $r_2 = \frac{1}{2}\left(\gamma_1 + \sqrt{\gamma_1^2 + 4\sigma_m^2}\right)$ are the roots of the quadratic $(\lambda^2 - \gamma_1 \lambda - \sigma_m^2) = (\lambda - r_1)(\lambda - r_2)$, we obtain that this polynomial is *non-positive* for $\lambda \in (r_1, r_2)$ which yields the requirement:

$$\frac{1}{2}\left(\gamma_1 - \sqrt{\gamma_1^2 + 4\sigma_m^2}\right) \leq \lambda.$$

(iv). To derive an *upper* bound for λ when $\lambda < 0$, we substitute into (10.13) that $-\lambda (\mathbf{u}^T A\mathbf{u}) \leq -\lambda \gamma_n (\mathbf{u}^T \mathbf{u})$ and $-(\mathbf{u}^T B^T B\mathbf{u}) \leq -\sigma_1^2 (\mathbf{u}^T \mathbf{u})$ to obtain:

$$\begin{aligned} 0 = \lambda^2(\mathbf{u}^T \mathbf{u}) - \lambda(\mathbf{u}^T A\mathbf{u}) - (\mathbf{u}^T B^T B\mathbf{u}) &\leq \lambda^2(\mathbf{u}^T \mathbf{u}) - \gamma_n \lambda(\mathbf{u}^T \mathbf{u}) - \sigma_1^2(\mathbf{u}^T \mathbf{u}) \\ &= \left(\lambda^2 - \gamma_n \lambda - \sigma_1^2\right) (\mathbf{u}^T \mathbf{u}). \end{aligned}$$

Since $r_1 = \frac{1}{2}\left(\gamma_n - \sqrt{\gamma_n^2 + 4\sigma_1^2}\right) < 0$ and $r_2 = \frac{1}{2}\left(\gamma_n + \sqrt{\gamma_n^2 + 4\sigma_1^2}\right)$ are the roots of the quadratic $(\lambda^2 - \gamma_n \lambda - \sigma_1^2) = (\lambda - r_1)(\lambda - r_2)$, we obtain that this polynomial is *non-negative* for $\lambda < r_1$or $\lambda > r_2$. Since $\lambda < 0$ this yields:

$$\lambda \leq \frac{1}{2}\left(\gamma_n - \sqrt{\gamma_n^2 + 4\sigma_1^2}\right),$$

which yields the desired bound. □

10.1.4 Condition Number of the Schur Complement

Duality and Krylov based algorithms for (10.2) require preconditioners for the Schur complement matrix $S = (BA^{-1}B^T)$. Here, we discuss its conditioning.

Lemma 10.11. *Suppose the following conditions hold:*

1. *Let bilinear form $\mathcal{A}(.,.)$ be symmetric and coercive with:*
$$\alpha_0 \|v\|_{\mathcal{U}}^2 \le \mathcal{A}(v,v) \le \alpha_1 \|v\|_{\mathcal{U}}^2, \quad \forall v \in \mathcal{U},$$
for some $\alpha_1 > \alpha_0 > 0$.
2. *Let the inf-sup condition hold for $\mathcal{B}(\cdot,\cdot)$ with constant $\beta > 0$:*
$$\sup_{u \in \mathcal{U}\setminus\{0\}} \frac{\mathcal{B}(u,\mu)}{\|u\|_{\mathcal{U}}} \ge \beta \|\mu\|_{\mathcal{Q}}, \quad \forall \mu \in \mathcal{Q}.$$
3. *Let $c > 0$ be such that:*
$$\|Bu\|_{\mathcal{Q}'} \le c \|u\|_{\mathcal{U}}, \quad \forall u \in \mathcal{U}.$$

Then, the Schur complement $S = (BA^{-1}B^T)$ will satisfy:

$$\frac{\beta^2}{\alpha_1} \|\mu\|_{\mathcal{Q}}^2 \le \left\langle BA^{-1}B^T\mu, \mu \right\rangle_{\mathcal{Q}} \le \frac{c^2}{\alpha_0} \|\mu\|_{\mathcal{Q}}^2, \quad \forall \mu \in \mathcal{Q}. \tag{10.14}$$

Proof. We first verify (10.14). Since bilinear form $\mathcal{A}(.,.)$ is symmetric and coercive, it defines an inner product on $\mathcal{U}$, so that $A^{-1} : \mathcal{U}' \to \mathcal{U}$ is bounded. Employing this yields the following equivalent expression for $\langle S\mu, \mu\rangle_{\mathcal{Q}}$:

$$\begin{aligned} \left\langle BA^{-1}B^T\mu, \mu \right\rangle_{\mathcal{Q}} &= \left\langle A^{-1}B^T\mu, B^T\mu \right\rangle_{\mathcal{U}} \\ &= \left\langle AA^{-1}B^T\mu, A^{-1}B^T\mu \right\rangle_{\mathcal{U}} \\ &= \sup_{v \in \mathcal{U}\setminus\{0\}} \frac{\left\langle AA^{-1}B^T\mu, v \right\rangle_{\mathcal{U}}^2}{\langle Av, v\rangle_{\mathcal{U}}} \\ &= \sup_{v \in \mathcal{U}\setminus\{0\}} \frac{\left\langle B^T\mu, v \right\rangle_{\mathcal{U}}^2}{\langle Av, v\rangle_{\mathcal{U}}}. \end{aligned}$$

To obtain a lower bound, apply the *inf-sup* condition:

$$\langle S\mu, \mu\rangle_{\mathcal{Q}} = \sup_{v \in \mathcal{U}\setminus\{0\}} \frac{\left\langle B^T\mu, v \right\rangle_{\mathcal{U}}^2}{\langle Av, v\rangle_{\mathcal{U}}} \ge \frac{\beta^2}{\alpha_1} \|\mu\|_{\mathcal{Q}}^2.$$

To obtain an upper bound, the third assumption can be employed:

$$\langle S\mu, \mu\rangle_{\mathcal{Q}} = \sup_{v \in \mathcal{U}'\setminus\{0\}} \frac{\langle \mu, Bv\rangle_{\mathcal{U}}^2}{\langle Av, v\rangle_{\mathcal{U}}} \le \frac{c^2 \|\mu\|_{\mathcal{Q}}^2 \|v\|_{\mathcal{U}}^2}{\|v\|_A^2} \le \frac{c^2}{\alpha_0} \|\mu\|_{\mathcal{Q}}^2.$$

The preceding two bounds verify (10.14). □

Remark 10.12. The parameters β, c, α_0, α_1 can be estimated in applications using results from finite element theory, yielding bound (10.14) for cond(S).

10.2 Algorithms Based on Duality

Duality formulations are motivated by a *geometric* characterization of saddle points [CI4, GI3]. *Heuristically*, if $(\mathbf{u}, \boldsymbol{\mu})$ is a saddle point of $\mathcal{L}(.,.)$, then the "restricted" functional $\mathcal{L}(\mathbf{v}, \boldsymbol{\mu})$, when considered as a function of $\mathbf{v} \in \mathbb{R}^n$, will attain its minimum at $\mathbf{u}$, while the restricted functional $\mathcal{L}(\mathbf{u}, \boldsymbol{\eta})$, when considered as a function of $\boldsymbol{\eta} \in \mathbb{R}^m$, will attain its maximum at $\boldsymbol{\mu}$. The saddle point $(\mathbf{u}, \boldsymbol{\mu})$ will thus be invariant when $\mathcal{L}(.,.)$ is optimized within such "sections" (planes). This suggests an alternate characterization of a saddle point. Given $\boldsymbol{\eta} \in \mathbb{R}^m$, let $\mathbf{v}_{\boldsymbol{\eta}} \in \mathbb{R}^n$ denote the minimum of the restricted functional $\mathcal{L}(\cdot, \boldsymbol{\eta})$. Then $\mathbf{v}_{\boldsymbol{\eta}}$ can be determined by solving an *unconstrained* minimization problem in $\mathbb{R}^n$. As $\boldsymbol{\eta} \in \mathbb{R}^m$ is varied, it can be shown that the functional $D(\boldsymbol{\eta}) \equiv \mathcal{L}(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$, referred to as the *dual functional*, will attain its maximal value at $\boldsymbol{\eta} = \boldsymbol{\mu}$. Thus, a saddle point of $\mathcal{L}(.,.)$ can be sought by maximization of the dual functional $D(\boldsymbol{\eta}) = \mathcal{L}(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$, which will yield $\boldsymbol{\eta} = \boldsymbol{\mu}$ as the maximum and $(\mathbf{u}, \boldsymbol{\mu}) = (\mathbf{v}_{\boldsymbol{\mu}}, \boldsymbol{\mu})$ as the saddle point.

To elaborate the details, recall our constrained minimization problem:

$$J(\mathbf{u}) = \min_{\mathbf{v} \in \mathcal{K}_{\mathbf{g}}} J(\mathbf{v}) \tag{10.15}$$

where $J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}$ and $\mathcal{K}_{\mathbf{g}} = \{\mathbf{v} : B\mathbf{v} = \mathbf{g}\}$. We defined $\mathcal{L}(.,.)$ as:

$$\mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \equiv J(\mathbf{v}) + \boldsymbol{\eta}^T (B\mathbf{v} - \mathbf{g}), \tag{10.16}$$

where $\boldsymbol{\eta} \in \mathbb{R}^m$ denotes the vector of Lagrange multipliers. By construction, the first derivative test for a critical point of $\mathcal{L}(.,.)$ yields system (10.2). In most of the algorithms we describe, we shall assume that $A = A^T > 0$.

Definition 10.13. *A point* $(\mathbf{u}, \boldsymbol{\mu})$ *is said to be a saddle point of* $\mathcal{L}(.,.)$ *if:*

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\eta}) \le \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) \le \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}), \qquad \forall \mathbf{v} \in \mathbb{R}^n,\ \boldsymbol{\eta} \in \mathbb{R}^m. \tag{10.17}$$

We associate the following two functionals with the Lagrangian $\mathcal{L}(.,.)$.

Definition 10.14. *We define a dual functional* $D(\cdot)$ *and a functional* $E(\cdot)$*:*

$$\begin{aligned} D(\boldsymbol{\eta}) &\equiv \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}), \qquad \forall \boldsymbol{\eta} \in \mathbb{R}^m \\ E(\mathbf{v}) &\equiv \sup_{\boldsymbol{\eta}} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}), \qquad \forall \mathbf{v} \in \mathbb{R}^n, \end{aligned} \tag{10.18}$$

where it is easily verified that:

$$E(\mathbf{v}) = \begin{cases} +\infty, & \text{if} \quad \mathbf{v} \notin \mathcal{K}_{\mathbf{g}} \\ J(\mathbf{v}), & \text{if} \quad \mathbf{v} \in \mathcal{K}_{\mathbf{g}}, \end{cases}$$

while $-\infty < D(\boldsymbol{\eta}) < \infty$. *Since* $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta})$ *is quadratic in* $\mathbf{v}$ *and* $A = A^T > 0$, *for each* $\boldsymbol{\eta} \in \mathbb{R}^m$ *the infimum will be attained for some* $\mathbf{v}_{\boldsymbol{\eta}} \in \mathbb{R}^n$ *such that:*

$$\mathcal{L}(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta}) = \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}).$$

By construction, for arbitrary $\mathbf{v} \in \mathbb{R}^n$ and $\boldsymbol{\eta} \in \mathbb{R}^m$ it will hold that:

$$D(\boldsymbol{\eta}) \leq \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \leq E(\mathbf{v}).$$

Since $E(\mathbf{v}) < \infty$ for $\mathbf{v} \in \mathcal{K}_{\mathbf{g}}$, minimizing over $\mathbf{v}$ and maximizing over $\boldsymbol{\eta}$ yields:

$$D(\boldsymbol{\eta}) = \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \leq \inf_{\mathbf{v}} E(\mathbf{v}) \Longrightarrow \sup_{\boldsymbol{\eta}} D(\boldsymbol{\eta}) \leq \inf_{\mathbf{v}} E(\mathbf{v}) \, < \, \infty. \tag{10.19}$$

If $D(\boldsymbol{\mu}) = \sup D(\cdot) = \inf E(\cdot) = E(\mathbf{u})$, then $(\mathbf{u}, \boldsymbol{\mu})$ will be a saddle point.

Lemma 10.15. *Let A be symmetric positive definite and B be of full rank. Then $(\mathbf{u}, \boldsymbol{\mu})$ will be a saddle point of $\mathcal{L}(.,.)$ iff:*

$$\min_{\mathbf{v} \in \mathbb{R}^n} E(\mathbf{v}) = \min_{\mathbf{v} \in \mathbb{R}^n} \left(\sup_{\boldsymbol{\eta} \in \mathbb{R}^m} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \right) = \max_{\boldsymbol{\eta} \in \mathbb{R}^m} \left(\inf_{\mathbf{v} \in \mathbb{R}^n} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \right) = \max_{\boldsymbol{\eta} \in \mathbb{R}^m} D(\boldsymbol{\eta}). \tag{10.20}$$

Proof. We shall first show that condition (10.20) will hold at a saddle point. Accordingly, suppose that $(\mathbf{u}, \boldsymbol{\mu})$ is a saddle point:

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\eta}) \leq \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) \leq \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}), \quad \forall \mathbf{v} \in \mathbb{R}^n \text{ and } \boldsymbol{\eta} \in \mathbb{R}^m.$$

Maximizing over $\boldsymbol{\eta}$ and minimizing over $\mathbf{v}$ yields:

$$E(\mathbf{u}) \leq \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) \leq D(\boldsymbol{\mu}).$$

Using property (10.19) at the saddle point yields:

$$D(\boldsymbol{\mu}) \leq \sup_{\boldsymbol{\eta}} D(\boldsymbol{\eta}) \leq \inf_{\mathbf{v}} E(\mathbf{v}) \leq E(\mathbf{u}) \leq \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) \leq D(\boldsymbol{\mu}),$$

so it follows that $D(\boldsymbol{\mu}) = E(\mathbf{u})$ at the saddle point, yielding (10.20).

Next, we consider the converse. Suppose (10.20) holds. Let the minimal value of $E(\cdot)$ be attained at $\mathbf{u}$:

$$E(\mathbf{u}) = \min_{\mathbf{v} \in \mathbb{R}^n} E(\mathbf{v}) = \min_{\mathbf{v} \in \mathbb{R}^n} \left(\sup_{\boldsymbol{\eta} \in \mathbb{R}^m} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \right).$$

Similarly, let the maximal value of $D(\cdot)$ be attained at $\boldsymbol{\mu}$:

$$D(\boldsymbol{\mu}) = \max_{\boldsymbol{\eta} \in \mathbb{R}^m} D(\boldsymbol{\eta}) = \max_{\boldsymbol{\eta} \in \mathbb{R}^m} \left(\inf_{\mathbf{v} \in \mathbb{R}^n} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) \right).$$

Then, condition (10.20) is equivalent to requiring requiring $E(\mathbf{u}) = D(\boldsymbol{\mu})$. Substituting the definitions of $E(\cdot)$ and $D(\cdot)$ yields:

$$E(\mathbf{u}) = \sup_{\boldsymbol{\eta}} \mathcal{L}(\mathbf{u}, \boldsymbol{\eta}) \geq \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) \geq \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}) = D(\boldsymbol{\mu}).$$

Since $E(\mathbf{u}) = D(\boldsymbol{\mu})$, we obtain that $E(\mathbf{u}) = \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) = D(\boldsymbol{\mu})$. Thus:

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\eta}) \leq \sup_{\boldsymbol{\eta}} \mathcal{L}(\mathbf{u}, \boldsymbol{\eta}) = E(\mathbf{u}) = \mathcal{L}(\mathbf{u}, \boldsymbol{\mu}) = D(\boldsymbol{\mu}) = \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}) \leq \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}),$$

for arbitrary $\mathbf{v}$ and $\boldsymbol{\eta}$, and $(\mathbf{u}, \boldsymbol{\mu})$ is a saddle point of $\mathcal{L}(.,.)$. □

Uzawa's Algorithm. Based on (10.20), the Lagrange multiplier $\boldsymbol{\mu}$ at a saddle point $(\mathbf{u}, \boldsymbol{\mu})$ of $\mathcal{L}(.,.)$ can be sought by maximization of the dual function $D(\cdot)$. Uzawa's method is a *gradient ascent* algorithm with a fixed step size $\tau > 0$ for maximizing $D(\cdot)$, see [AR7, CI4, GI3]. Given an iterate $\boldsymbol{\mu}^{(k-1)} \in \mathbb{R}^m$, Uzawa's method computes an update $\boldsymbol{\mu}^{(k)} \in \mathbb{R}^m$ as follows:

$$\boldsymbol{\mu}^{(k)} = \boldsymbol{\mu}^{(k-1)} + \tau \left[\frac{\partial D}{\partial \boldsymbol{\eta}} \right] \Bigg|_{\boldsymbol{\mu}^{(k-1)}}, \tag{10.21}$$

where the gradient:

$$\left[\frac{\partial D}{\partial \boldsymbol{\eta}} \right] \equiv \left[\frac{\partial D}{\partial \eta_1}, \ldots, \frac{\partial D}{\partial \eta_m} \right]^T \in \mathbb{R}^m.$$

Given $\boldsymbol{\mu}^{(k)}$, Uzawa's method constructs an approximation $\mathbf{u}^{(k+1)}$ of $\mathbf{u}$ as the argument which minimizes $\mathcal{L}(\mathbf{v}, \boldsymbol{\mu}^{(k)})$ for $\mathbf{v} \in \mathbb{R}^n$:

$$D(\boldsymbol{\mu}^{(k)}) = \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}^{(k)}) = \mathcal{L}(\mathbf{u}^{(k+1)}, \boldsymbol{\mu}^{(k)}),$$

where index $(k+1)$ has been used instead of k for convenience. An explicit expression for $\frac{\partial D}{\partial \boldsymbol{\eta}}(\boldsymbol{\mu}^{(k)})$ can be obtained as described below.

Lemma 10.16. *Given $\boldsymbol{\mu}^{(k)}$ let $\mathbf{u}^{(k+1)}$ denote the minimum:*

$$D(\boldsymbol{\mu}^{(k)}) = \mathcal{L}(\mathbf{u}^{(k+1)}, \boldsymbol{\mu}^{(k)}) = \inf_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}^{(k)}),$$

where $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) = \left(\frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}\right) + \boldsymbol{\eta}^T (B\mathbf{v} - \mathbf{g})$. Then, the following will hold:

1. *The update $\mathbf{u}^{(k+1)}$ will solve:*

$$A\mathbf{u}^{(k+1)} = \mathbf{f} - B^T \boldsymbol{\mu}^{(k)}. \tag{10.22}$$

2. *The following expression will hold for $\frac{\partial D}{\partial \boldsymbol{\eta}}$:*

$$\left[\frac{\partial D}{\partial \boldsymbol{\eta}} \right] \Bigg|_{\boldsymbol{\mu}^{(k)}} = B\mathbf{u}^{(k+1)} - \mathbf{g}. \tag{10.23}$$

Proof. Given $\boldsymbol{\eta} \in \mathbb{R}^m$, let $\mathbf{v}_{\boldsymbol{\eta}} \in \mathbb{R}^n$ denote the minimum:

$$\mathcal{L}(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta}) = \inf_{\mathbf{v} \in \mathbb{R}^n} \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}).$$

Since $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) = (\frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T) + \boldsymbol{\eta}^T (B\mathbf{v} - \mathbf{g})$, an application of the first order derivative test for the minimum yields the following linear system for $\mathbf{v}_{\boldsymbol{\eta}}$:

$$A\,\mathbf{v}_{\boldsymbol{\eta}} = \mathbf{f} - B^T \boldsymbol{\eta}.$$

Since $\mathbf{v}_{\boldsymbol{\eta}}$ minimizes $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta})$ for $\mathbf{v} \in \mathbb{R}^n$ it will hold that:

$$\left[\frac{\partial \mathcal{L}}{\partial \mathbf{v}}\right]\Big|_{(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta})} = \left[\frac{\partial \mathcal{L}}{\partial v_1}, \ldots, \frac{\partial \mathcal{L}}{\partial v_n}\right]\Big|_{(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta})} = 0.$$

Applying the chain rule using $D(\boldsymbol{\eta}) = \mathcal{L}(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$ will yield:

$$\left[\frac{\partial D}{\partial \boldsymbol{\eta}}\right]^T = \left[\frac{\partial \mathcal{L}}{\partial \mathbf{v}}\right]^T \left[\frac{\partial \mathbf{v}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\eta}}\right] + \left[\frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}}\right]^T = \left[\frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}}\right]^T,$$

since $\left[\frac{\partial \mathcal{L}}{\partial \mathbf{v}}\right] = 0$ at $(\mathbf{v}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$. Since $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}} = B\mathbf{v} - \mathbf{g}$, the desired result now follows by substituting $\mathbf{v} = \mathbf{u}^{(k+1)}$ and $\boldsymbol{\eta} = \boldsymbol{\mu}^{(k)}$. □

Uzawa's algorithm can now be summarized by substituting (10.23) for $(\partial D/\partial \eta)$ into (10.21) and employing (10.22) to determine $\mathbf{u}^{(k+1)}$ given $\boldsymbol{\mu}^{(k)}$.

Algorithm 10.2.1 *(Uzawa's Algorithm for Solving (10.2))*
Given $\mathbf{u}^{(0)}$, $\boldsymbol{\mu}^{(0)}$*:*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     Solve: $A\mathbf{u}^{(k+1)} = \mathbf{f} - B^T \boldsymbol{\mu}^{(k)}$
3.     Update: $\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} + \tau \left(B\mathbf{u}^{(k+1)} - \mathbf{g}\right)$
4. Endfor
```

Remark 10.17. Substituting $\mathbf{u}^{(k+1)} = A^{-1}(\mathbf{f} - B^T \boldsymbol{\mu}^{(k)})$ from step 2 above, into the expression for the update $\boldsymbol{\mu}^{(k+1)}$in step 3, will yield:

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} + \tau \left(BA^{-1}\mathbf{f} - \mathbf{g} - S\boldsymbol{\mu}^{(k)}\right), \tag{10.24}$$

where $S = (BA^{-1}B^T)$ denotes the Schur complement. This corresponds to an *unaccelerated* Richardson method to solve $S\,\boldsymbol{\mu} = (BA^{-1}\mathbf{f} - \mathbf{g})$ for $\boldsymbol{\mu}$. To ensure convergence, τ must satisfy $0 < \tau < \frac{1}{\lambda_{\max}(S)}$.

Remark 10.18. From a matrix viewpoint, Uzawa's algorithm to solve (10.2) corresponds to an unaccelerated matrix splitting iteration:

$$\begin{bmatrix} \mathbf{u}^{(k+1)} \\ \boldsymbol{\mu}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} + \begin{bmatrix} A & 0 \\ B & -(I/\tau) \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} - \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} \right).$$

Thus, its rate of convergence will depend on the spectral radius of the error propagation matrix, which computation yields as:

$$\rho \begin{bmatrix} 0 & -A^{-1}B^T \\ 0 & I - \tau S \end{bmatrix} = \rho(I - \tau S), \quad \text{where } S = (BA^{-1}B^T). \tag{10.25}$$

Other variants of Uzawa's algorithm will be indicated later.

The following result concerns an optimal choice of *fixed* parameter τ.

Proposition 10.19. *Suppose the following conditions hold:*

1. *$A = A^T > 0$ and B be of full rank $m < n$.*
2. *Let $\lambda_1 = \lambda_{min}(S)$ and $\lambda_m = \lambda_{max}(S)$ where $S = BA^{-1}B^T$.*

Then, the optimal fixed step size τ_* will satisfy:

$$\tau_* = \frac{2}{\lambda_1 + \lambda_m},$$

and the error $\mathbf{e}^{(k)} \equiv (\boldsymbol{\mu} - \boldsymbol{\mu}^{(k)})$ in the iterates will satisfy:

$$\|\mathbf{e}^{(k)}\| \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{e}^{(0)}\|,$$

where $(\kappa = \lambda_m/\lambda_1)$ denotes the condition number of matrix S.

Proof. The following error contraction estimate will hold for expression (10.24):

$$\|\mathbf{e}^{(k)}\| \le (\rho(I - \tau\, S))^k \; \|\mathbf{e}^{(0)}\|,$$

in the Euclidean norm $\|\cdot\|$. The optimal choice of parameter τ_* must satisfy:

$$\rho\left(I - \tau_* S\right) = \min_{\tau} \rho\left(I - \tau\, S\right) = \min_{\tau} \left\{\left|1 - \tau\lambda_1\right|, \left|1 - \tau\lambda_m\right|\right\},$$

which yields the expression:

$$1 - \tau_*\lambda_1 = \tau_*\lambda_m - 1 \Longrightarrow \tau_* = \frac{2}{\lambda_1 + \lambda_m}.$$

For the above choice of parameter τ, we obtain:

$$\rho\left(I - \tau_* S\right) = \min\left\{\left|1 - \frac{2}{\lambda_1 + \lambda_m}\lambda_1\right|, \left|1 - \frac{2}{\lambda_1 + \lambda_m}\lambda_m\right|\right\} = \frac{\kappa - 1}{\kappa + 1},$$

where $\kappa = (\lambda_m/\lambda_1)$. □

Remark 10.20. When the Schur complement $S = (BA^{-1}B^T)$ is *ill conditioned*, Uzawa's algorithm will converge slowly. In this case, a *preconditioner* S_0 must be employed for S. Step 3 of Uzawa's algorithm can be updated as follows:

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} + S_0^{-1}\left(B\mathbf{u}^{(k+1)} - \mathbf{g}\right),$$

where τ has been absorbed into the preconditioner. Preconditioners S_0 for S are described in § 5.6 and § 5.7 for different applications.

Inexact Versions of Uzawa's Algorithm. In applications, it may be expensive to solve the system $A\mathbf{u}^{(k+1)} = \left(\mathbf{f} - B^T\boldsymbol{\mu}^{(k)}\right)$ exactly. If an inexact *iterative* solver is used, it will result in an *inner* iteration. Care must be exercised to ensure that the modified iterates converge. Two alternative approaches are possible [VE, QU7, GL7, BA18, EL5, BR16].

- If a fixed preconditioner $A_0 = A_0^T$ is employed for $A = A^T$, then the update in step 2 of Uzawa's algorithm can be *modified* as follows [BR16]:
$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + A_0^{-1}\left(\mathbf{f} - A\mathbf{u}^{(k)} - B^T\boldsymbol{\mu}^{(k)}\right),$$
where A_0^{-1} has formally replaced A^{-1}. This iteration will be *linear.*
- If the stopping criterion for the inexact iterative solver is based on the magnitude of the *residual* vector $\mathbf{r}^{(k)}$, then step 2 of Uzawa's algorithm may involve a varying number of *inner* iterations to update $\mathbf{u}^{(k+1)}$:
$$A\mathbf{u}^{(k+1)} = \left(\mathbf{f} - B^T\boldsymbol{\mu}^{(k)}\right) + \mathbf{r}^{(k)}.$$
To ensure convergence of the modified Uzawa iterates, analysis in [EL5] suggests the stopping criterion $\|\mathbf{r}^{(k)}\| \le \tau\, \|B\mathbf{u}^{(k+1)} - \mathbf{g}\|$ for some $\tau < 1$

Below, we list the *inexact* Uzawa algorithm to solve (10.2), incorporating a fixed preconditioner A_0 for A. Since the convergence rate of Uzawa's algorithm will deteriorate if the Schur complement S is ill-conditioned, the following algorithm additionally incorporates a preconditioner S_0 for S, see [BR16].

Algorithm 10.2.2 *(Preconditioned Inexact Uzawa Algorithm)*
Given $\mathbf{u}^{(0)}$ *and* $\boldsymbol{\mu}^{(0)}$.

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + A_0^{-1}\left(\mathbf{f} - A\mathbf{u}^{(k)} - B^T\boldsymbol{\mu}^{(k)}\right)$.
3.     $\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} + S_0^{-1}(B\mathbf{u}^{(k+1)} - \mathbf{g})$
4. Endfor
```

Here, the step size $\tau > 0$ has been absorbed into S_0.

Remark 10.21. From a matrix viewpoint, the above preconditioned inexact Uzawa iteration corresponds to an *unaccelerated* iteration to solve (10.2) based on the following matrix splitting:

$$\begin{bmatrix} \mathbf{u}^{(k+1)} \\ \boldsymbol{\mu}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} + \begin{bmatrix} A_0 & 0 \\ B & -S_0 \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} - \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} \right). \quad (10.26)$$

Note that this matrix splitting expression will not be valid for a fixed A_0 if the update in step 2 has the form $A\mathbf{u}^{(k+1)} = \left(\mathbf{f} - B^T\boldsymbol{\mu}^{(k)}\right) + \mathbf{r}^{(k)}$.

Remark 10.22. The rate of convergence of iteration (10.26) will depend on the spectral radius of its error propagation matrix, which computation yields as:

$$\rho \begin{bmatrix} I - A_0^{-1}A & -A_0^{-1}B^T \\ S_0^{-1}B(I - A_0^{-1}A) & I - S_0^{-1}BA_0^{-1}B^T \end{bmatrix}. \quad (10.27)$$

This spectral radius is estimated in [BR16], as stated next.

Lemma 10.23. *Suppose the following assumptions hold.*

1. *Let there be* $0 \leq \gamma < 1$ *satisfying:*

$$(1-\gamma)\,(\boldsymbol{\eta}^T S_0 \boldsymbol{\eta}) \,\leq\, (\boldsymbol{\eta}^T S \boldsymbol{\eta}) \,\leq\, (\boldsymbol{\eta}^T S_0 \boldsymbol{\eta}), \quad \forall \boldsymbol{\eta} \in \mathbb{R}^m.$$

2. *Let there be* $0 \leq \delta < 1$ *satisfying:*

$$(1-\delta)\,(\mathbf{v}^T A_0 \mathbf{v}) \,\leq\, (\mathbf{v}^T A \mathbf{v}) \,\leq\, (\mathbf{v}^T A_0 \mathbf{v}), \quad \forall \mathbf{v} \in \mathbb{R}^n.$$

Then, the following bound will hold for $\mathbf{e}^{(k)} = (\mathbf{u} - \mathbf{u}^{(k)}, \boldsymbol{\mu} - \boldsymbol{\mu}^{(k)})$*:*

$$\|\mathbf{e}^{(k)}\| \,\leq\, \rho^k \,\|\mathbf{e}^{(0)}\|,$$

for the iterates in Algorithm 10.2.2, where

$$\rho \equiv \frac{\gamma(1-\delta) + \sqrt{\gamma^2(1-\delta)^2 + 4\delta}}{2}.$$

Proof. See [BR16]. It can be noted that $\rho \leq 1 - (1/2)(1-\gamma)(1-\delta)$, so that this inexact Uzawa iteration will converge provided $\delta < 1$ and $\gamma < 1$. □

Remark 10.24. Provided $A_0^{-1}\left(\mathbf{f} - A\mathbf{u}^{(k)} - B^T\boldsymbol{\mu}^{(k)}\right)$ corresponds to a descent direction of $\mathcal{L}(\cdot, \boldsymbol{\mu}^{(k)})$, and provided S_0 is symmetric positive definite, it can be verified that the iterates in the inexact Uzawa algorithm satisfy:

$$\begin{aligned} \mathcal{L}(\mathbf{u}^{(k+1)}, \boldsymbol{\mu}^{(k)}) \quad &\leq \mathcal{L}(\mathbf{u}^{(k)}, \boldsymbol{\mu}^{(k)}), \qquad \text{and} \\ \mathcal{L}(\mathbf{u}^{(k+1)}, \boldsymbol{\mu}^{(k+1)}) &\geq \mathcal{L}(\mathbf{u}^{(k+1)}, \boldsymbol{\mu}^{(k)}), \end{aligned}$$

due to the alternating descent (minimization) and ascent (maximization) of $\mathcal{L}(.,.)$. At the saddle point $(\mathbf{u}, \boldsymbol{\mu})$, these iterates will remain stationary.

Remark 10.25. In certain applications matrix B may not be of full rank. In this case Kernel(B^T) will not be trivial and system (10.2) will be *consistent* only if $\mathbf{g} \perp$ Kernel(B^T). If this holds, then Uzawa's iterates for $A_0 = A$ will be well defined provided $S_0 = (I/\tau)$.

The *Arrow-Hurwicz* algorithm [AR7, TE, GI3] corresponds to a special case of the inexact Uzawa method, in which, given $\boldsymbol{\mu}^{(k)}$, one step of a *gradient descent* method with a step size ω is applied to approximate the minimum of $\mathcal{L}(\mathbf{v}, \boldsymbol{\mu}^{(k)})$ for $\mathbf{v} \in \mathbb{R}^n$. This corresponds to approximately solving the system:

$$A\,\mathbf{u}^{(k+1)} = \left(\mathbf{f} - B^T\boldsymbol{\mu}^{(k)}\right).$$

Define $\phi_k(\mathbf{v}) = \mathcal{L}(\mathbf{v}, \boldsymbol{\mu}^{(k)}) = \frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T\left(\mathbf{f} - B^T\boldsymbol{\mu}^{(k)}\right)$ as the functional to be minimized. Then, the direction $\mathbf{d}^{(k)}$ of steepest descent of $\phi_k(\cdot)$ at $\mathbf{u}^{(k)}$ is:

$$\mathbf{d}^{(k)} = -\nabla\phi_k(\mathbf{u}^{(k)}) = \left(\mathbf{f} - B^T\boldsymbol{\mu}^{(k)} - A\mathbf{u}^{(k)}\right).$$

The gradient descent update with step size ω will satisfy:

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \omega \left(\mathbf{f} - B^T \boldsymbol{\mu}^{(k)} - A\mathbf{u}^{(k)} \right),$$

and the resulting algorithm is summarized below.

Algorithm 10.2.3 *(Arrow-Hurwicz Algorithm)*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \omega \left( \mathbf{f} - A\mathbf{u}^{(k)} - B^T \boldsymbol{\mu}^{(k)} \right)$
3.     $\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} + \tau \left( B\mathbf{u}^{(k+1)} - \mathbf{g} \right)$
4. Endfor
```

From a matrix viewpoint, the Arrow-Hurwicz algorithm corresponds to an inexact Uzawa algorithm with $A_0 = (I/\omega)$ and $S_0 = (I/\tau)$.

Lemma 10.26. *The Arrow-Hurwicz algorithm has block matrix form:*

$$\begin{bmatrix} \mathbf{u}^{(k+1)} \\ \boldsymbol{\mu}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} + \begin{bmatrix} (I/\omega) & 0 \\ B & -(I/\tau) \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} - \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} \right).$$

If $\omega = (1/\lambda_{\max}(A))$ and $\tau = (1/\lambda_{\max}(S))$ then the convergence factor ρ of the Arrow-Hurwicz algorithm will satisfy:

$$\rho \leq \frac{\gamma(1-\delta) + \sqrt{\gamma^2(1-\delta)^2 + 4\delta}}{2},$$

for $\gamma = (\lambda_{\min}(A)/\lambda_{\max}(A))$ and $\delta = (\lambda_{\min}(S)/\lambda_{\max}(S))$.

Proof. Follows by Lemma 10.23 using $A_0 = \lambda_{\max}(A)I$ and $S_0 = \lambda_{\max}(S)I$, see [BR16]. □

Augmented Lagrangian Method. The variants of Uzawa's algorithm that we have described are applicable when A is symmetric and *positive definite.* When A is *singular*, saddle point problem (10.2) may still be well posed, provided B is of full rank and A is coercive in the subspace $\mathcal{K}_0 = \text{Kernel}(B)$, see Lemma 10.9. In this case, minimizing $\mathcal{L}(\mathbf{v}, \boldsymbol{\mu})$ for $\mathbf{v} \in \mathbb{R}^n$ will not yield a unique minimum. However, if the *augmented Lagrangian* method is employed to construct an equivalent reformulation of the saddle point system (10.6), then matrix A will be replaced by a *non-singular* matrix [GL7], and Uzawa's algorithm can be applied to solve the augmented saddle point system.

The original constrained minimization problem sought to minimize the functional $J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}$ within the constraint set $\mathcal{K}_{\mathbf{g}}$. If we replace the objective functional $J(\mathbf{v})$ by $J_\rho(\mathbf{v}) \equiv J(\mathbf{v}) + \rho \, \|B\mathbf{v} - \mathbf{g}\|_W^2$, where $\rho > 0$ is a parameter and $W = W^T > 0$ is a weight matrix of size m, then the minima of $J_\rho(\mathbf{v})$ and $J(\mathbf{v})$ will coincide within the constraint set $\mathcal{K}_{\mathbf{g}}$, since since $\rho \, \|B\mathbf{v} - \mathbf{g}\|_W^2$ vanishes within it. However, in the augmented saddle point system, matrix A will be replaced by the non-singular matrix $A + \rho \, B^T W B$.

The augmented Lagrangian $\mathcal{L}_\rho(\mathbf{v}, \boldsymbol{\eta})$ associated with the minimization of $J_\rho(\mathbf{v})$ within $\mathcal{K}_{\mathbf{g}}$ is defined as:

$$\begin{cases} \mathcal{L}_\rho(\mathbf{v}, \boldsymbol{\eta}) \equiv J(\mathbf{v}) + \frac{\rho}{2} \|B\mathbf{v} - \mathbf{g}\|_W^2 + \boldsymbol{\eta}^T (B\mathbf{v} - \mathbf{g}) \\ \qquad\qquad\quad = \mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) + \frac{\rho}{2} \|B\mathbf{v} - \mathbf{g}\|_W^2, \end{cases} \tag{10.28}$$

where $\|B\mathbf{v} - \mathbf{g}\|_W^2 = (B\mathbf{v} - \mathbf{g})^T W (B\mathbf{v} - \mathbf{g})$ vanishes within $\mathcal{K}_{\mathbf{g}}$. The following properties can easily be verified.

- As $\rho \to 0$, the augmented Lagrangian $\mathcal{L}_\rho(.,.) \to \mathcal{L}(.,.)$.
- For $\mathbf{v} \in \mathcal{K}_{\mathbf{g}} = \{\mathbf{v} \,:\, B\,\mathbf{v} = \mathbf{g}\}$, it will hold that: $\mathcal{L}(\mathbf{v}, \boldsymbol{\eta}) = \mathcal{L}_\rho(\mathbf{v}, \boldsymbol{\eta})$.
- $(\mathbf{u}, \boldsymbol{\mu})$ is a saddle point of $\mathcal{L}(.,.)$ *iff* it is a saddle point of $\mathcal{L}_\rho(.,.)$.

These properties suggest that the saddle point of $\mathcal{L}(.,.)$ may be sought by determining the saddle point of $\mathcal{L}_\rho(.,.)$. Applying the first derivative test to determine the saddle point of $\mathcal{L}_\rho(.,.)$ yields the following linear system:

$$\begin{bmatrix} A + \rho\, B^T W B & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{f} + \rho\, B^T W \mathbf{g} \\ \mathbf{g} \end{bmatrix}. \tag{10.29}$$

Since A is *coercive* within $\mathcal{K}_0$ and since $\rho\, B^T W B$ is coercive within $\mathcal{K}_0^\perp$ (when B has full rank and since $\mathcal{K}_0 = \text{Kernel}(B)$), matrix $A + \rho\, B^T W B$ will be positive definite for $\rho > 0$. Thus, Uzawa's algorithms can be employed to solve (10.29). The choice of parameter $\rho > 0$ is considered in [GL7].

10.3 Penalty and Regularization Methods

Penalty and regularization methods are related techniques for approximating a constrained optimization problem or its saddle point formulation [CI4, GI3]. Both methods formulate a family of computationally simpler problems which depend on a small parameter $\epsilon > 0$, such that as $\epsilon \to 0^+$, the penalty or regularized solutions converge to the solution of the original problem.

We shall first describe the *penalty* (or "barrier" function) method. Given a *constrained* minimization problem $(\mathbf{P})$, the penalty method constructs a family of *unconstrained* minimization problems $(\mathbf{P}_\epsilon)$ for $\epsilon > 0$, whose solutions converge to the constrained minimum as $\epsilon \to 0^+$. Let $(\mathbf{P})$ denote the constrained minimization problem:

$$(\mathbf{P}) \qquad\qquad J(\mathbf{u}) = \min_{\mathbf{v} \in \mathcal{K}_{\mathbf{g}}} J(\mathbf{v}), \tag{10.30}$$

where $J(v) = \frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}$, and $\mathcal{K}_{\mathbf{g}} = \{\mathbf{v} \in \mathbb{R}^n \,:\, B\mathbf{v} = \mathbf{g}\}$. The penalty method employs a *nonnegative* penalty function $\psi(\cdot)$ which *vanishes* in $\mathcal{K}_{\mathbf{g}}$:

$$\psi(\mathbf{v}) = \frac{1}{2} \|B\mathbf{v} - \mathbf{g}\|_{D^{-1}}^2 = \frac{1}{2} \left(B\mathbf{v} - \mathbf{g}\right)^T D^{-1} \left(B\mathbf{v} - \mathbf{g}\right), \tag{10.31}$$

where D^{-1} denotes a symmetric positive definite matrix of size m. Using the penalty function, a penalized objective function $J_\epsilon(\mathbf{v})$ is defined for $\epsilon > 0$ as:

$$J_\epsilon(\mathbf{v}) = J(\mathbf{v}) + \frac{1}{\epsilon}\psi(\mathbf{v}) = \left(\frac{1}{2}\mathbf{v}^T A\mathbf{v} - \mathbf{v}^T\mathbf{f}\right) + \frac{1}{2\epsilon}\left(B\mathbf{v} - \mathbf{g}\right)^T D^{-1}\left(B\mathbf{v} - \mathbf{g}\right), \tag{10.32}$$

so that as $\epsilon \to 0^+$ the penalty term $(1/\epsilon)\psi(\mathbf{v})$ dominates $J(\mathbf{v})$ except when the constraints $B\mathbf{v} = \mathbf{g}$ are satisfied.

The penalty method seeks the *unconstrained* minimum of $J_\epsilon(\cdot)$:

$$(\mathbf{P}_\epsilon) \qquad\qquad J_\epsilon(\mathbf{u}_\epsilon) = \min_{\mathbf{v}\in\mathbb{R}^n} J_\epsilon(\mathbf{v}). \tag{10.33}$$

Heuristically, we expect the minimum $\mathbf{u}_\epsilon$ of $J_\epsilon(\cdot)$ to satisfy the constraints as $\epsilon \to 0^+$, since the penalty term $\frac{1}{\epsilon}\,\psi(\mathbf{u}_\epsilon)$ will dominate $J(\mathbf{u}_\epsilon)$ otherwise. Applying the first derivative test to $J_\epsilon(\mathbf{v})$ yields the linear system:

$$\left(A + \frac{1}{\epsilon}B^T D^{-1} B\right)\mathbf{u}_\epsilon = \mathbf{f} + \frac{1}{\epsilon}B^T D^{-1}\mathbf{g}. \tag{10.34}$$

Its solution $\mathbf{u}_\epsilon$ can be shown to converge to the solution $\mathbf{u}$ of $(\mathbf{P})$.

Proposition 10.27. *Let A be a symmetric positive definite matrix of size n and let D^{-1} be symmetric positive definite of size m. If $\mathbf{u}_\epsilon$ and $\mathbf{u}$ denote the solutions to problems $(\mathbf{P})_\epsilon$ and $(\mathbf{P})$, respectively, then the following will hold:*

$$\|\mathbf{u}_\epsilon - \mathbf{u}\| \le c\,\epsilon,$$

for some $c > 0$ independent of ϵ (but dependent on A, B, D, $\mathbf{f}$ and $\mathbf{g}$).

Proof. See [CI4] and Lemma 10.29. □

The advantage of the penalty method is that it replaces a *constrained* minimization problem by an *unconstrained* minimization problem. However, the linear system (10.34) for $\mathbf{u}_\epsilon$ can become highly ill-conditioned as $\epsilon \to 0^+$:

$$\text{cond}\left(A + \frac{1}{\epsilon}B^T D^{-1} B\right) \le O\left(\frac{\lambda_{\max}(A) + \frac{1}{\epsilon}\,\lambda_{\max}(B^T D^{-1} B)}{\lambda_{\min}(A)}\right).$$

As a result care must be exercised when solving for $\mathbf{u}_\epsilon$. Alternatively, $\mathbf{u}_\epsilon$ may be obtained by solving a related *regularized* saddle point system (10.35).

The *regularization* method [GI3] is closely related to the penalty method. Given a small parameter $\epsilon > 0$, the regularization method perturbs the saddle point system (10.2) by introducing a perturbation term $-\epsilon\, D\,\boldsymbol{\mu}_\epsilon$ in the second block row, resulting in the following block system:

$$\begin{bmatrix} A & B^T \\ B & -\epsilon\, D \end{bmatrix}\begin{bmatrix} \mathbf{w}_\epsilon \\ \boldsymbol{\mu}_\epsilon \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{10.35}$$

Here D is a symmetric positive definite matrix of size m. System (10.35) can easily be verified to be well posed. Indeed, using $\mathbf{w}_\epsilon = A^{-1}\left(\mathbf{f} - B^T\boldsymbol{\mu}_\epsilon^T\right)$ and substituting this into the 2nd block row of (10.35) yields the reduced system:

$$\left(BA^{-1}B^T + \epsilon\, D\right)\boldsymbol{\mu}_\epsilon = \left(BA^{-1}\mathbf{f} - \mathbf{g}\right).$$

Since D and $BA^{-1}B^T$ are symmetric positive definite, the Schur complement $(BA^{-1}B^T + \epsilon\, D)$ will also be symmetric and positive definite, for $\epsilon > 0$. Thus, system (10.35) will be solvable when $A = A^T > 0$ and B has full rank.

The regularized system (10.35) can be solved by *block elimination* of $\boldsymbol{\mu}_\epsilon$. Indeed, the second block row of (10.35) can be solved for $\boldsymbol{\mu}_\epsilon$ resulting in:

$$\boldsymbol{\mu}_\epsilon = \frac{1}{\epsilon}\, D^{-1}\left(B\mathbf{w}_\epsilon - \mathbf{g}\right).$$

Substituting this expression for $\boldsymbol{\mu}_\epsilon$ into the first block row of (10.35) yields:

$$\left(A + \frac{1}{\epsilon}\, B^T D^{-1} B\right)\mathbf{w}_\epsilon = \mathbf{f} + \frac{1}{\epsilon}\, B^T D^{-1}\mathbf{g}. \tag{10.36}$$

This system is identical to the *penalty* system (10.34), and it follows that component $\mathbf{w}_\epsilon$ in the regularized saddle point problem (10.35) is identical to the penalty solution $\mathbf{u}_\epsilon$ in (10.34).

Remark 10.28. The solution to (10.35) can be sought iteratively obtained by formally modifying step 3 of Uzawa's method:

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} + \tau\left(B\,\mathbf{u}^{(k+1)} - \epsilon\, D\boldsymbol{\mu}^{(k)}\right),$$

leaving all other steps the same. If matrix $(BA^{-1}B^T + \epsilon\, D)$ is ill-conditioned, we may replace τ by M^{-1}, where M is a preconditioner for $(BA^{-1}B^T + \epsilon\, D)$.

The next result estimates the error $(\mathbf{w}_\epsilon - \mathbf{u}, \boldsymbol{\mu}_\epsilon - \boldsymbol{\mu})$ between the solution to the regularized problem (10.35) and to the saddle point system (10.2).

Lemma 10.29. *Let A and D be symmetric positive definite matrices of size n and m, respectively. Let $(\mathbf{w}_\epsilon, \boldsymbol{\mu}_\epsilon)$ solve the regularized system (10.35) and $(\mathbf{u}, \boldsymbol{\mu})$ solve (10.2). Then, there exists $c > 0$ independent of ϵ such that:*

$$\|\mathbf{w}_\epsilon - \mathbf{u}\| \le c\,\epsilon.$$

Proof. We subtract (10.35) from (10.2) obtaining:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}\begin{bmatrix} \mathbf{u} - \mathbf{w}_\epsilon \\ \boldsymbol{\mu} - \boldsymbol{\mu}_\epsilon \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \epsilon\, D\boldsymbol{\mu}_\epsilon \end{bmatrix}. \tag{10.37}$$

The well posedness of (10.35) in the Euclidean norm yields the estimate:

$$\|\boldsymbol{\mu}_\epsilon\| \le \left(\|(S + \epsilon\, D)^{-1}BA^{-1}\mathbf{f}\| + \|(S + \epsilon\, D)^{-1}\,\mathbf{g}\|\right),$$

where $S = (BA^{-1}B^T)$. Using the well posedness of (10.2) from Lemma 10.3, and substituting the preceding bound yields the estimate:

$$\begin{cases} \|\mathbf{u} - \mathbf{w}_\epsilon\| \le \epsilon \, \|A^{-1}B^T S^{-1} D(S + \epsilon\, D)^{-1}(BA^{-1}\mathbf{f} - \mathbf{g})\| \\ \qquad\qquad\quad \le c\, \epsilon, \end{cases}$$

where $c \equiv \|A^{-1}B^T S^{-1} D(S + \epsilon\, D)^{-1}(BA^{-1}\mathbf{f} - \mathbf{g})\|$. In our applications, $c > 0$ is expected to have bounds independent of h and ϵ. □

10.4 Projection Methods

Projection methods to solve saddle point system (10.2) are motivated by an *orthogonal decomposition* property which system (10.2) inherits, that enables computing its solution using *orthogonal projections* [CH27, CH28, CI4, TE]. The use of projections can be motivated by considering system (10.2) when $A = I$ is of size n, matrix B of size $m \times n$ has rank m and $\mathbf{g} = \mathbf{0} \in \mathbb{R}^m$. In this case, system (10.2) reduces to:

$$\begin{cases} \mathbf{u} + B^T \boldsymbol{\mu} = \mathbf{f}, \\ \qquad B\mathbf{u} = \mathbf{0}. \end{cases} \tag{10.38}$$

The second block equation $B\,\mathbf{u} = \mathbf{0}$ in (10.38) requires $\mathbf{u} \in \mathcal{K}_0 = \text{Kernel}(B)$. Since $\text{Range}(B^T) = \text{Kernel}(B)^\perp$ it follows that $B^T\boldsymbol{\mu} \in \text{Range}(B^T) = \mathcal{K}_0^\perp$. Thus, $\mathbf{u} + B^T\boldsymbol{\mu} = \mathbf{f}$ corresponds to an Euclidean orthogonal decomposition of $\mathbf{f} \in \mathbb{R}^n$ with $\mathbf{u} \in \mathcal{K}_0$ and $B^T\boldsymbol{\mu} \in \mathcal{K}_0^\perp$. To determine $\boldsymbol{\mu}$, multiply the first block row of (10.38) by B and use that $B\mathbf{u} = \mathbf{0}$ to obtain $\left(BB^T\right)\boldsymbol{\mu} = B\,\mathbf{f}$. This system will be non-singular since B has full rank $m < n$, yielding:

$$\boldsymbol{\mu} \;=\; \left(BB^T\right)^{-1} B\mathbf{f} \;\text{ and }\; \mathbf{u} \;=\; P_{\mathcal{K}_0}\mathbf{f} = \left(I - B^T(BB^T)^{-1}B\right)\mathbf{f},$$

where $P_{\mathcal{K}_0}\mathbf{f}$ denotes the Euclidean orthogonal projection of $\mathbf{f}$ onto $\mathcal{K}_0$.

The preceding procedure is applicable only when $A = I$ and $\mathbf{g} = \mathbf{0}$. More generally, saddle point system (10.2) with $A = A^T > 0$ and $\mathbf{g} \neq \mathbf{0}$ can be reduced to the case $\mathbf{g} = \mathbf{0}$ in a preliminary step, and the computation of the component of $\mathbf{u}$ in $\mathcal{K}_0$ will involve an A-orthogonal projection onto $\mathcal{K}_0$, while $\boldsymbol{\mu}$ can be computed using an Euclidean orthogonal projection (or by an alternative approach in the case Schwarz projections). The following preliminary result, applicable when $\mathbf{g} = \mathbf{0}$, describes why the problem to determine $\mathbf{u}$ in system (10.2) is *positive definite* within the subspace $\mathcal{K}_0$.

Lemma 10.30. *Consider system (10.2) in which $A = A^T > 0$ and matrix B of size $m \times n$ has rank m, and $\mathbf{f} \in \mathbb{R}^n$, $\mathbf{g} = \mathbf{0} \in \mathbb{R}^m$:*

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}. \tag{10.39}$$

Then, the following properties will hold.

- $\mathbf{w} \in \mathcal{K}_0 = \text{Kernel}(B)$ *and will satisfy:*

$$\mathbf{v}^T A\mathbf{w} = \mathbf{v}^T \mathbf{f}, \quad \forall \mathbf{v} \in \mathcal{K}_0. \tag{10.40}$$

- *If* $\mathbf{f} = A\,\mathbf{u} + B^T\,\boldsymbol{\eta}$ *for some* $\mathbf{u} \in \mathbb{R}^n$ *and* $\boldsymbol{\eta} \in \mathbb{R}^m$*, then* $\mathbf{w} = P^A_{\mathcal{K}_0}\mathbf{u}$ *will be an A-orthogonal projection of* $\mathbf{u}$ *onto* $\mathcal{K}_0$*.*

In particular, $\mathbf{w}$ *will not depend on* $\boldsymbol{\eta}$*, if* $\mathbf{f} = A\,\mathbf{u} + B^T\,\boldsymbol{\eta}$*.*

Proof. The second block row $B\,\mathbf{w} = \mathbf{0}$ of (10.39) yields that $\mathbf{w} \in \mathcal{K}_0$. Choose $\mathbf{v} \in \mathcal{K}_0$ and compute the inner product of $\left(\mathbf{v}^T, \mathbf{0}^T\right)^T$ with (10.39) to obtain:

$$\mathbf{v}^T A\mathbf{w} + \mathbf{v}^T B^T \boldsymbol{\gamma} = \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} = \mathbf{v}^T \mathbf{f}. \tag{10.41}$$

Since $\mathbf{v}^T B^T = \mathbf{0}$ for $\mathbf{v} \in \mathcal{K}_0$, this reduces to $\mathbf{v}^T A\mathbf{w} = \mathbf{v}^T \mathbf{f}$. Since $A = A^T > 0$, the problem (10.40) to determine $\mathbf{w} \in \mathcal{K}_0$ is coercive within $\mathcal{K}_0$ and solvable. When $\mathbf{f} = A\,\mathbf{u} + B^T\,\boldsymbol{\eta}$, the inner product of $\mathbf{v} \in \mathcal{K}_0$ with $\mathbf{f}$ will satisfy:

$$\mathbf{v}^T \mathbf{f} = \mathbf{v}^T A\mathbf{u} + \mathbf{v}^T B^T \boldsymbol{\eta} = \mathbf{v}^T A\mathbf{u}, \quad \forall \mathbf{v} \in \mathcal{K}_0,$$

since $\mathbf{v}^T B^T = \mathbf{0}$ for $\mathbf{v} \in \mathcal{K}_0$. Substituting this into (10.40) will yield:

$$\mathbf{v}^T A\mathbf{w} \; = \; \mathbf{v}^T A\mathbf{u}, \qquad \forall\, \mathbf{v} \in \mathcal{K}_0.$$

Thus $\mathbf{w} = P^A_{\mathcal{K}_0}\mathbf{u}$ corresponds to an A-orthogonal projection of $\mathbf{u}$ onto $\mathcal{K}_0$. □

Thus, if $\mathbf{g} = \mathbf{0}$, then $\mathbf{w}$ in (10.39) can be determined by solving (10.40) within $\mathcal{K}_0$. Since a basis for $\mathcal{K}_0$ can be computationally expensive to construct, we shall describe *projection* algorithms which will compute approximations of $\mathbf{w}$ iteratively, without an explicit basis for $\mathcal{K}_0$. The following result describes how to reduce the general case of saddle point problem (10.2) to the case with $\mathbf{g} = \mathbf{0}$ and furthermore, how $\boldsymbol{\mu}$ can be determined.

Lemma 10.31. *Consider system (10.2) with* $A = A^T > 0$ *and* B *of rank* m*:*

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{10.42}$$

Suppose $\mathbf{u}_* \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^n$ *and* $\boldsymbol{\gamma} \in \mathbb{R}^m$ *are computed as follows.*

1. *Let* $\mathbf{u}_* = B^T(BB^T)^{-1}B\mathbf{g}$.
2. *Let* $\mathbf{w} \in \mathcal{K}_0$ *solve* $\mathbf{v}^T A\mathbf{w} \; = \; \mathbf{v}^T\,(\mathbf{f} - A\mathbf{u}_*)$*, for all* $\mathbf{v} \in \mathbb{R}^n$.
3. *Let* $\boldsymbol{\gamma} \; = \; (BB^T)^{-1}B\,(\mathbf{f} - A\mathbf{u}_* - A\mathbf{w})\,.$

Then, $\mathbf{u} = (\mathbf{w}_* + \mathbf{w})$ *and* $\boldsymbol{\mu} = \boldsymbol{\gamma}$*.*

Proof. Firstly, since B is $m \times n$ with rank m, matrix (BB^T) is invertible. Thus, $\mathbf{u}_* = B^T(BB^T)^{-1}\mathbf{g}$ is well defined and $B\mathbf{u}_* = BB^T(BB^T)^{-1}\mathbf{g} = \mathbf{g}$.

Secondly, since $\mathbf{f} = A\mathbf{u} + B^T\boldsymbol{\mu}$, we obtain that $\mathbf{w} \in \mathcal{K}_0$ solves:

$$\mathbf{v}^T A\mathbf{w} = \mathbf{v}^T \left(A\mathbf{u} + B^T\boldsymbol{\mu} - A\mathbf{u}_*\right) \quad \forall \mathbf{v} \in \mathcal{K}_0.$$

Applying Lemma 10.30 to the above expression yields $\mathbf{w} = P^A_{\mathcal{K}_0}(\mathbf{u} - \mathbf{u}_*)$. By construction of $\mathbf{u}_*$, we obtain $B(\mathbf{u} - \mathbf{u}_*) = \mathbf{0}$, yielding that $(\mathbf{u} - \mathbf{u}_*) \in \mathcal{K}_0$. Uniqueness of the projection within $\mathcal{K}_0$ yields $\mathbf{w} = (\mathbf{u} - \mathbf{u}_*)$.

Thirdly, once $\mathbf{u}_*$ and $\mathbf{w}$ have been determined, the first block row of (10.42) yields the overdetermined system $B^T\boldsymbol{\mu} = (\mathbf{f} - A\mathbf{u}_* - A\mathbf{w})$. This system will be consistent since $\mathbf{v}^T(\mathbf{f} - A\mathbf{u}_* - A\mathbf{w}) = 0$ for $\mathbf{v} \in \mathcal{K}_0$, by step 2. Multiplying both sides by B yields $\boldsymbol{\mu} = (BB^T)^{-1}B(\mathbf{f} - A\mathbf{u})$. □

We now summarize the general projection algorithm to solve (10.2).

Algorithm 10.4.1 *(General Projection Algorithm to Solve (10.2))*

1. *Determine* $\mathbf{u}_*$ *such that* $B\mathbf{u}_* = \mathbf{g}$*:*

$$\mathbf{u}_* = B^T\left(BB^T\right)^{-1}\mathbf{g}$$

2. *Determine* $\mathbf{w} \in \mathcal{K}_0$ *satisfying:*

$$\mathbf{v}^T A\,\mathbf{w} = \mathbf{v}^T(\mathbf{f} - A\mathbf{u}_*), \quad \forall \mathbf{v} \in \mathcal{K}_0 \tag{10.43}$$

3. *Determine* $\boldsymbol{\mu}$ *such that* $B^T\boldsymbol{\mu} = (\mathbf{f} - A\,\mathbf{u}_* - A\,\mathbf{w})$*:*

$$\boldsymbol{\mu} = \left(BB^T\right)^{-1} B\,(\mathbf{f} - A\,\mathbf{u}_* - A\,\mathbf{w})$$

Output: $(\mathbf{u} = \mathbf{u}_* + \mathbf{w}, \boldsymbol{\mu})$

The *first* step in the preceding algorithm involves computing $\mathbf{u}_*$. This requires solving a linear system with the coefficient matrix (BB^T) of size m. Depending on m and the application, efficient solvers may be available. However, for Stokes and mixed formulations of elliptic equations, we shall indicate an alternate method for computing $\mathbf{u}_*$ satisfying $B\mathbf{u}_* = \mathbf{g}$, using domain decomposition methods. Typically, the *second* step is the most computationally expensive. We shall describe a projected gradient and projection Schwarz algorithm for this step. Step *three* has a similar computational cost as step one. In specific applications, we shall indicate alternatives for computing $\boldsymbol{\mu}$, based on domain decomposition.

10.4.1 Projected Gradient Descent

The projected gradient algorithm [CI4] can be employed to solve (10.43) for $\mathbf{w} \in \mathcal{K}_0$ in step 2, without constructing a basis for $\mathcal{K}_0$. We define:

$$\Phi(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T A\mathbf{v} - \mathbf{v}^T\left(\mathbf{f} - A\mathbf{u}_*\right). \tag{10.44}$$

Since $\mathcal{K}_0 = \text{Kernel}(B)$ is a subspace of $\mathbb{R}^n$ and $A = A^T > 0$, solving (10.43) corresponds to a first order derivative test for minimizing $\Phi(\mathbf{v})$ within $\mathcal{K}_0$:

$$\mathbf{v}^T\left(A\mathbf{w} + A\mathbf{u}_* - \mathbf{f}\right) = 0 \quad\Leftrightarrow\quad \mathbf{v}^T \cdot \nabla\Phi(\mathbf{w}) = 0, \;\; \forall \mathbf{v} \in \mathcal{K}_0.$$

Thus (10.43) is equivalent to the constrained minimization problem:

$$\Phi\left(\mathbf{w}\right) = \min_{\mathbf{v}\in\mathcal{K}_0} \Phi\left(\mathbf{v}\right). \tag{10.45}$$

At the constrained minimum $\mathbf{w} \in \mathcal{K}_0$ of $\Phi(\cdot)$, applying the stationarity of the directional derivative of $\Phi(\cdot)$ within $\mathcal{K}_0$ will yield the equivalences:

$$\begin{aligned} \mathbf{v}^T \cdot \nabla\Phi(\mathbf{w}) = 0, \quad \forall \mathbf{v} \in \mathcal{K}_0 &\Leftrightarrow P_{\mathcal{K}_0}\nabla\Phi(\mathbf{w}) = \mathbf{0} \\ &\Leftrightarrow \mathbf{w} = P_{\mathcal{K}_0}\left(\mathbf{w} - \tau\,\nabla\Phi(\mathbf{w})\right), \end{aligned} \tag{10.46}$$

since $P_{\mathcal{K}_0}\mathbf{w} = \mathbf{w}$ for $\mathbf{w} \in \mathcal{K}_0$, where $P_{\mathcal{K}_0} = I - B^T\left(BB^T\right)^{-1}B$ is the Euclidean orthogonal projection onto $\mathcal{K}_0$ and $\tau > 0$ is a fixed parameter. The equation $\mathbf{w} = P_{\mathcal{K}_0}\left(\mathbf{w} - \tau\,\nabla\Phi(\mathbf{w})\right)$ in (10.46) expresses that $\mathbf{w}$ is a *fixed point* of:

$$T(\mathbf{v}) \equiv P_{\mathcal{K}_0}\left(\mathbf{v} - \tau\,\nabla\Phi(\mathbf{v})\right). \tag{10.47}$$

Substituting that $\nabla\Phi(\mathbf{v}) = A\mathbf{v} - (\mathbf{f} - A\,\mathbf{u}_*)$ in (10.47) yields:

$$T(\mathbf{v}) \equiv P_{\mathcal{K}_0}\left(\mathbf{v} - \tau A\mathbf{v} + \tau(\mathbf{f} - A\mathbf{u}_*)\right). \tag{10.48}$$

For suitable $\tau > 0$, the map $T(\cdot)$ will be a *contraction* and Picard iteration:

$$\mathbf{v}^{(k+1)} = P_{\mathcal{K}_0}\left(\mathbf{v}^{(k)} - \tau A\mathbf{v}^{(k)} + \tau(\mathbf{f} - A\,\mathbf{u}_*)\right), \tag{10.49}$$

can be shown to converge geometrically to the unique fixed point $\mathbf{w}$ of $T(\cdot)$. This is described in the following result.

Lemma 10.32. *Suppose the following conditions hold:*

1. *Let* $\tau = 2/\left(\lambda_{\min}(A) + \lambda_{\max}(A)\right)$.
2. *Given* $\mathbf{v}^{(0)}$ *define the Picard iterates:*

$$\mathbf{v}^{(k+1)} = P_{\mathcal{K}_0}\left(\mathbf{v}^{(k)} - \tau A\mathbf{v}^{(k)} + \tau(\mathbf{f} - A\,\mathbf{u}_*)\right). \tag{10.50}$$

Then the iterates $\mathbf{v}^{(k)}$ *will converge to the solution* $\mathbf{w}$ *of (10.43):*

$$\|\mathbf{v}^{(k)} - \mathbf{w}\| \le \left(\frac{\kappa(A) - 1}{\kappa(A) + 1}\right)^k \|\mathbf{v}^{(0)} - \mathbf{w}\|,$$

where $\|\cdot\|$ *denotes the Euclidean norm,* $\lambda(A)$ *an eigenvalue of* A *and* $\kappa(A)$ *the spectral condition number of* A.

Proof. See [CI4]. Since $\|P_{\mathcal{K}_0}\| \le 1$ we estimate:

$$\|T(\mathbf{v}) - T(\mathbf{w})\| = \|P_{\mathcal{K}_0}\left((\mathbf{v}-\mathbf{w}) - \tau\, A(\mathbf{v}-\mathbf{w})\right)\| \le \|\left(I - \tau\, A\right)(\mathbf{v}-\mathbf{w})\|.$$

Since $\|I - \tau\, A\| \le \max\{|1-\tau\,\lambda_{\min}(A)|, |1-\tau\,\lambda_{\max}(A)\}$, substituting the optimal fixed choice of parameter τ:

$$\tau_* = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)},$$

yields the contraction factor:

$$\|T(\mathbf{v}) - T(\mathbf{w})\| \le \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)\|\mathbf{v}-\mathbf{w}\|, \qquad \text{where } \kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

This completes the proof. □

Algorithm 10.4.2 *(Projected Gradient Algorithm to Solve (10.43))*
Let $\mathbf{v}^{(0)}$ denote a starting guess

```
1. For $k = 0, 1, \ldots$ until convergence do:
       $\mathbf{v}^{(k+1)} = P_{\mathcal{K}_0}\left(\mathbf{v}^{(k)} - \tau A\mathbf{v}^{(k)} + \tau(\mathbf{f} - A\,\mathbf{u}_*)\right)$
2. Endfor
Output: $\mathbf{v}^{(k)}$
```

Remark 10.33. The action of the projection $P_{\mathcal{K}_0} = I - B^T\left(BB^T\right)^{-1}B$ on a vector can be computed at a cost proportional to the cost of solving a linear system with coefficient matrix $\left(BB^T\right)$. Matrix $\left(BB^T\right)$ will be sparse if B is sparse, and direct or iterative methods may be employed.

Remark 10.34. The projected gradient algorithm with parameter τ:

$$\tau = 2/\left(\lambda_{\min}(A) + \lambda_{\max}(A)\right),$$

will converge *slowly* when matrix A is *ill-conditioned.*

Since $\mathcal{K}_0 \subset \mathbb{R}^n$ is a *subspace*, the projected gradient algorithm can be accelerated based on the following alternate expression for $\mathbf{w}$:

$$A\mathbf{w} + B^T\boldsymbol{\mu} \,=\, \mathbf{f} - A\mathbf{u}_*. \tag{10.51}$$

Multiply (10.51) by B and solve for $\boldsymbol{\mu}$ to obtain:

$$\boldsymbol{\mu} = \left(BB^T\right)^{-1} B\left(\mathbf{f} - A\,\mathbf{u}_* - A\mathbf{w}\right).$$

Substituting the preceding into (10.51) yields the formal expression:

$$\left(I - B^T(BB^T)^{-1}B\right)\, A\,\mathbf{w} = \left(I - B^T(BB^T)^{-1}B\right)\, (\mathbf{f} - A\mathbf{u}_*)\,.$$

Here $\left(I - B^T(BB^T)^{-1}B\right) = P_{\mathcal{K}_0}$ denotes the Euclidean orthogonal projection onto $\mathcal{K}_0$. Since $\mathbf{w} \in \mathcal{K}_0$, we may seek $\mathbf{w} = P_{\mathcal{K}_0}\mathbf{w}$ by solving the following consistent symmetric positive semidefinite system:

$$\left(P_{\mathcal{K}_0} A P_{\mathcal{K}_0}\right) \mathbf{w} = P_{\mathcal{K}_0} (\mathbf{f} - A\mathbf{u}_*)\,. \tag{10.52}$$

Matrix $\left(P_{\mathcal{K}_0} A P_{\mathcal{K}_0}\right)$ is symmetric and positive semidefinite, and the *conjugate gradient* method can be applied to solve this *singular* but *consistent* system. The following should be noted when solving the singular system (10.52).

- The initial iterate $\mathbf{v}^{(0)}$ must be chosen from $\mathcal{K}_0$ (or *projected* onto $\mathcal{K}_0$ using the orthogonal projection $P_{\mathcal{K}_0} = I - B^T\left(BB^T\right)^{-1} B$..)
- For $\mathbf{v} \in \mathcal{K}_0$ the matrix-vector product $\left(P_{\mathcal{K}_0} A P_{\mathcal{K}_0}\right)\mathbf{v}$ simplifies to $\left(P_{\mathcal{K}_0} A\right)\mathbf{v}$. Thus, the action of $P_{\mathcal{K}_0}$ need only be computed *once* per iteration.
- A preconditioner A_0^{-1} can be applied to solve the system:

  $$A_0^{-1}\left(P_{\mathcal{K}_0} A P_{\mathcal{K}_0}\right)\mathbf{w} = A_0^{-1} P_{\mathcal{K}_0} (\mathbf{f} - A\mathbf{u}_*)\,,$$

  using the inner product generated by $P_{\mathcal{K}_0} A P_{\mathcal{K}_0}$.

By construction, all iterates will lie in $\mathcal{K}_0$.

10.4.2 Schwarz Projection Algorithms

An alternative approach to solve (10.43) is to use a divide and conquer approach based on the projection formulation of *Schwarz* methods, see Chap. 2.2 and Chap. 2.3. If $\mathcal{K}_0^{(0)}, \mathcal{K}_0^{(1)}, \ldots, \mathcal{K}_0^{(p)}$ denote subspaces of $\mathcal{K}_0 = \text{Kernel}(B)$, then Schwarz algorithms can be formulated to solve (10.43) using A-orthogonal projections onto these subspaces. In applications, if these projections are computed by solving smaller saddle point problems, then not only can $\mathbf{w} \in \mathcal{K}_0$ be determined, but also the Lagrange multiplier $\boldsymbol{\mu}$. We assume the following.
Assumptions.

1. For $0 \le i \le p$, let $\mathcal{K}_0^{(i)} \subset \mathcal{K}_0$ denote subspaces of $\mathcal{K}_0$ of dimension k_i:

   $$\mathcal{K}_0 = \mathcal{K}_0^{(0)} + \mathcal{K}_0^{(1)} + \cdots + \mathcal{K}_0^{(p)}.$$

   Thus, it must also hold that $k_0 + k_1 + \cdots + k_p \ge \dim\left(\mathcal{K}_0\right) = (n - m)$.
2. For $0 \le i \le p$ let $\mathcal{U}^{(i)} \subset R^n$ denote a subspace of dimension n_i such that:

   $$\mathcal{U}^{(i)} = \text{Range}(U_i),$$

   where U_i is a matrix of size $n \times n_i$ whose columns form a basis for $\mathcal{U}^{(i)}$.

3. For $0 \leq i \leq p$ let $\mathcal{Q}^{(i)} \subset R^m$ denote a subspace of dimension m_i such that:
$$\mathcal{Q}^{(i)} = \text{Range}(Q_i),$$
where Q_i is a matrix of size $m \times m_i$ whose columns form a basis for $\mathcal{Q}^{(i)}$.
4. For $0 \leq i \leq p$ let $\mathcal{K}_0^{(i)} \subset \mathcal{U}^{(i)}$ and satisfy:
$$\mathcal{K}_0^{(i)} = \left\{ \mathbf{v} \in \mathcal{U}^{(i)} \, : \, \mathbf{q}^T B \mathbf{v} = 0, \;\; \forall \mathbf{q} \in \mathcal{Q}^{(i)} \right\}.$$
Thus, we implicitly require that $k_i = (n_i - m_i)$.

Employing A-orthogonal projections onto each subspace $\mathcal{K}_0^{(i)}$, it will be possible to formulate implicitly preconditioned additive and multiplicative Schwarz algorithms to solve (10.43), without employing a basis for $\mathcal{K}_0$.

Definition 10.35. *We define $P^A_{\mathcal{K}_0^{(i)}}$ as the A-orthogonal projection onto subspace $\mathcal{K}_0^{(i)}$. Given $\mathbf{w} \in \mathbb{R}^n$ the projection $P^A_{\mathcal{K}_0^{(i)}} \mathbf{w} \in \mathcal{K}_0^{(i)}$ will satisfy:*
$$\mathbf{v}^T A \, P^A_{\mathcal{K}_0^{(i)}} \mathbf{w} = \mathbf{v}^T A \, \mathbf{w}, \qquad \forall \, \mathbf{v} \in \mathcal{K}_0^{(i)}. \tag{10.53}$$

To obtain a matrix representation, let $P^A_{\mathcal{K}_0^{(i)}} \mathbf{w} = U_i \, \mathbf{w}_i \in \mathcal{K}_0^{(i)}$ for $\mathbf{w}_i \in \mathbb{R}^{n_i}$:
$$P^A_{\mathcal{K}_0^{(i)}} \mathbf{w} = U_i \, \mathbf{w}_i.$$

Then, the requirement $\mathbf{v}^T A \, U_i \, \mathbf{w}_i = \mathbf{v}^T A \, \mathbf{w}$, for all $\mathbf{v} \in \mathcal{K}_0^{(i)}$ can be formulated as a saddle point system, as described in the following.

Applying Lemma 10.30 with local spaces $\text{Range}(U_i)$ and $\text{Range}(Q_i)$ yields:
$$\begin{cases} U_i^T A U_i \, \mathbf{w}_i + U_i^T B^T Q_i \, \boldsymbol{\gamma}_i = U_i^T A \, \mathbf{w} \\ Q_i^T B U_i \, \mathbf{w}_i \qquad\qquad\qquad\;\; = 0, \end{cases} \tag{10.54}$$
where $\boldsymbol{\gamma}_i \in \mathbb{R}^{m_i}$ denotes a vector of local Lagrange multiplier variables which enforce the local constraints $Q_i^T B \, U_i \, \mathbf{v}_i = 0$. Defining submatrices:
$$A_i = U_i^T A U_i \quad \text{and} \quad B_i = Q_i^T B U_i \tag{10.55}$$
of size $n_i \times n_i$ and $m_i \times n_i$ respectively, system (10.54) can be expressed as:
$$\begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \boldsymbol{\gamma}_i \end{bmatrix} = \begin{bmatrix} U_i^T A \, \mathbf{w} \\ \mathbf{0} \end{bmatrix}, \tag{10.56}$$
so that:
$$P^A_{\mathcal{K}_0^{(i)}} \mathbf{w} = U_i \begin{bmatrix} I \\ 0 \end{bmatrix}^T \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T A \, \mathbf{w} \\ \mathbf{0} \end{bmatrix}. \tag{10.57}$$
This yields a matrix representation of the A-orthogonal projection $P^A_{\mathcal{K}_0^{(i)}}$.

Additive Schwarz Algorithm. The *additive Schwarz* algorithm to solve problem (10.43) is based on the following equation equivalent to (10.43):

$$P^A \mathbf{w} = \mathbf{r}, \tag{10.58}$$

where P^A denotes the additive Schwarz preconditioned matrix, defined by:

$$P^A \equiv \sum_{i=0}^{p} P^A_{\mathcal{K}_0^{(i)}},$$

while vector $\mathbf{r} = P^A \mathbf{w}$ can be computed explicitly, even though $\mathbf{w}$ is unknown, by replacing $U_i^T A\mathbf{w} = U_i^T (\mathbf{f} - A\,\mathbf{u}_*)$:

$$\mathbf{r} = \sum_{i=0}^{p} P^A_{\mathcal{K}_0^{(i)}} \mathbf{w} = \sum_{i=0}^{p} U_i \begin{bmatrix} I \\ 0 \end{bmatrix}^T \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T (\mathbf{f} - A\,\mathbf{u}_*) \\ \mathbf{0} \end{bmatrix}. \tag{10.59}$$

By construction, the additive Schwarz preconditioned matrix P^A will be *self adjoint* in the A-inner product, with $\mathbf{r} \in \mathcal{K}_0$, since it is a sum of self adjoint projections onto subspaces of $\mathcal{K}_0$. Furthermore, results from § 2.3 and § 2.3 show that P^A will be *coercive* given our assumptions. We may thus determine $\mathbf{w}$ as the solution to (10.58) using the *conjugate gradient* method with inner product $\langle \mathbf{v}, \mathbf{w} \rangle_A \equiv \mathbf{v}^T A\mathbf{w}$. Provided the initial iterate $\mathbf{v}^{(0)} \in \mathcal{K}_0$, all subsequent iterates will lie in $\mathcal{K}_0$. Condition number bounds for $\mathrm{cond}(P^A)$ can be estimated in terms of the partition parameters for $\mathcal{K}_0^{(0)}, \ldots, \mathcal{K}_0^{(p)}$.

Multiplicative Schwarz Algorithm. We next list the unaccelerated multiplicative Schwarz algorithm to solve (10.43).

Algorithm 10.4.3 *(Multiplicative Schwarz Algorithm to Solve (10.43))*
Let $\mathbf{v}^{(0)} \in \mathcal{K}_0$ *be a starting iterate.*

```
1. For $l = 0, 1, \ldots$ until convergence do:
2.     For $i = 0, 1, \ldots, p$ do:
           $\mathbf{v}^{(l+\frac{i+1}{p+1})} = \mathbf{v}^{(l+\frac{i}{p+1})} + U_i \begin{bmatrix} I \\ 0 \end{bmatrix}^T \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T (\mathbf{f} - A\,\mathbf{u}_* - A\mathbf{v}^{(l+\frac{i}{p+1})}) \\ \mathbf{0} \end{bmatrix}.$
3.     Endfor
4. Endfor
```

Output: $\mathbf{v}^{(l)}$

Remark 10.36. The iterates $\mathbf{v}^{(k)}$ can be shown to converge *geometrically* to the solution $\mathbf{w}$ of (10.43). The rate of convergence will depend on the partition parameters associated with the subspaces $\mathcal{K}_0^{(0)}, \mathcal{K}_0^{(1)}, \ldots, \mathcal{K}_0^{(p)}$ of $\mathcal{K}_0$, as analyzed in Chap. 2.3. In applications to the Stokes equation and mixed formulations of elliptic equations, the Lagrange multiplier variables in the smaller saddle point problems, will also approximate the Lagrange multiplier variables $\boldsymbol{\mu}$ in (10.43), and converge as the iterates $\mathbf{v}^{(k)}$ converge to $\mathbf{w}$.

10.5 Krylov Space and Block Matrix Methods

In this section, we describe preconditioned Krylov methods for solving (10.2). Since saddle point system (10.2) is symmetric *indefinite*, the CG algorithm cannot be employed to solve it. However, the MINRES (or the conjugate residual) method [PA3, CH19, RU5, SA2, HA3] will be applicable, with the same storage requirements as the CG algorithm. Here, we shall describe:

- The Schur complement method for (10.2) using CG acceleration.
- A symmetric positive definite reformulation of (10.2) for use with CG.
- Block diagonal preconditioner for (10.2) using MINRES acceleration.
- Block triangular preconditioner for (10.2) using GMRES acceleration.
- Saddle point preconditioner for (10.2) using GMRES acceleration.

Each of the above methods will employ a preconditioner A_0 for A, and a preconditioner S_0 for the Schur complement $S = (BA^{-1}B^T)$.

10.5.1 Schur Complement Method to Solve (10.2)

The Schur complement method to solve (10.2) is based on the elimination of $\mathbf{u} = A^{-1}\left(\mathbf{f} - B^T\boldsymbol{\mu}\right)$. Substituting this into the second block row of (10.2) and rearranging terms yields the following reduced system for $\boldsymbol{\mu}$:

$$S\,\boldsymbol{\mu} = (BA^{-1}\mathbf{f} - \mathbf{g}), \quad \text{where} \quad S \equiv \left(BA^{-1}B^T\right). \tag{10.60}$$

The Schur complement method solves system (10.60) for $\boldsymbol{\mu}$ using a PCG algorithm with preconditioner S_0, and $\mathbf{u} = A^{-1}\left(\mathbf{f} - B^T\boldsymbol{\mu}\right)\mathbf{u}$ is subsequently determined. Although similar to a duality method, $\boldsymbol{\mu}$ is determined first, and the methodology requires a solver for matrix A.

Algorithm 10.5.1 *(Schur Complement Algorithm to Solve (10.2))*

```
1. Solve: $S\,\boldsymbol{\mu} = (BA^{-1}\mathbf{f} - \mathbf{g})$ with preconditioner $S_0$
2. Solve: $A\,\mathbf{u} = (\mathbf{f} - B^T\boldsymbol{\mu})$
```

Output: $(\mathbf{u}, \boldsymbol{\mu})$

Remark 10.37. Matrix $S = (BA^{-1}B^T)$ need not be assembled. Instead, when B is of full rank and $A = A^T > 0$, matrix S will be symmetric and positive definite, and $S\,\boldsymbol{\mu} = (BA^{-1}\mathbf{f} - \mathbf{g})$ can be solved by a PCG method. Each matrix-vector product with S will require products with B^T, solving a linear system of the form $A\mathbf{v} = \mathbf{r}$, and a product with B.

The Schur complement method can also be motivated by the factorization:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & -S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix},$$

when $A = A^T > 0$ and B is of full rank. Its inverse will be:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}^{-1} = \begin{bmatrix} I & -A^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & -S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -BA^{-1} & I \end{bmatrix}.$$

Preconditioners can be obtained by approximation of matrices A and S.

10.5.2 A Symmetric Positive Definite Reformulation of (10.2)

When $A = A^T > 0$, the symmetric *indefinite* saddle point system (10.2) can be transformed into a symmetric *positive definite* system, see [BR8]. The reformulation of [BR8] is not based on the normal equations, and employs a preconditioner A_0 of A. The transformed system can be solved by a PCG method, but its implementation requires computing the action of A_0 (which may not be available for various preconditioners), see Remark 10.39.

To construct a positive definite reformulation of (10.2), let $A_0 = A_0^T > 0$ be a preconditioner for matrix A, satisfying:

$$\alpha_0 \left(\mathbf{v}^T A \mathbf{v}\right) \le \left(\mathbf{v}^T A_0 \mathbf{v}\right) \le \alpha_1 \left(\mathbf{v}^T A \mathbf{v}\right), \quad \forall \mathbf{v} \in \mathbb{R}^n, \tag{10.61}$$

where $0 < \alpha_0 \le \alpha_1 < 1$. It can be noted that given any positive definite symmetric preconditioner $\tilde{A}$ for A, it will be sufficient to define A_0 as:

$$A_0 = \delta\, \tilde{A},$$

for some $\delta < \lambda_{\min}\left(\tilde{A}^{-1} A\right)$, and this will yield $\alpha_1 < 1$. This will require estimating the minimal eigenvalue of $\tilde{A}^{-1}A$ by the Lanczos method [GO4].

To reformulate (10.2) as a positive definite system, given a preconditioner A_0 for A satisfying (10.61), apply the following block transformation:

$$\begin{bmatrix} A_0^{-1} & 0 \\ BA_0^{-1} & -I \end{bmatrix} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A_0^{-1} & 0 \\ BA_0^{-1} & -I \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{10.62}$$

This yields the system:

$$\begin{bmatrix} A_0^{-1} A & A_0^{-1} B^T \\ BA_0^{-1} A - B & BA_0^{-1} B^T \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A_0^{-1}\mathbf{f} \\ BA_0^{-1}\mathbf{f} - \mathbf{g} \end{bmatrix}, \tag{10.63}$$

which we shall write more compactly as:

$$\mathcal{M} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A_0^{-1}\mathbf{f} \\ BA_0^{-1}\mathbf{f} - \mathbf{g} \end{bmatrix}, \tag{10.64}$$

where $\mathcal{M}$ is defined as the following block matrix:

$$\mathcal{M} \equiv \begin{bmatrix} A_0^{-1} A & A_0^{-1} B^T \\ BA_0^{-1} A - B & BA_0^{-1} B^T \end{bmatrix}. \tag{10.65}$$

Matrix $\mathcal{M}$ is not symmetric (in the standard Euclidean sense). However, it will be shown to be symmetric in the following inner product $\langle ., \rangle$:

$$\left\langle \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{pmatrix} \right\rangle \equiv \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} A - A_0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} = \mathbf{u}^T \left(A - A_0\right) \mathbf{v} + \boldsymbol{\mu}^T \boldsymbol{\lambda}. \tag{10.66}$$

Expression (10.66) defines an inner product since by assumption on the choice of A_0 matrix $(A - A_0)$ is positive definite symmetric. To verify that the block matrix $\mathcal{M}$ is *symmetric* in the inner product $\langle ., . \rangle$ note that:

$$\begin{aligned} \left\langle \mathcal{M} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} \right\rangle &= \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} (A - A_0)A_0^{-1}A & (AA_0^{-1} - I)\, B^T \\ B(A_0^{-1}A - I) & BA_0^{-1}B^T \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} \\ &= \left\langle \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}, \mathcal{M} \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} \right\rangle. \end{aligned} \tag{10.67}$$

Heuristically, we note that as $A_0 \to A$:

$$\begin{bmatrix} (A - A_0)A_0^{-1}A & (AA_0^{-1} - I)\, B^T \\ B(A_0^{-1}A - I) & BA_0^{-1}B^T \end{bmatrix} \to \begin{bmatrix} A - A_0 & 0 \\ 0 & BA^{-1}B^T \end{bmatrix},$$

which is *positive definite*. As a result, when $A_0 \to A$, we heuristically expect:

$$\left\langle \mathcal{M} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} \right\rangle \to \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} A - A_0 & 0 \\ 0 & BA^{-1}B^T \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} = \left\langle \mathcal{M}_* \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{bmatrix} \right\rangle,$$

where matrix $\mathcal{M}_*$ is a symmetric positive definite matrix (in the Euclidean and $\langle ., . \rangle$ inner products) defined by:

$$\mathcal{M}_* \equiv \begin{bmatrix} I & 0 \\ 0 & BA^{-1}B^T \end{bmatrix}. \tag{10.68}$$

Rigorous bounds for the eigenvalues of $\mathcal{M}_*^{-1}\mathcal{M}$ are stated below.

Lemma 10.38. *Suppose the following conditions hold.*

1. *Let matrix A_0 satisfy (10.61).*
2. *Let $\mathcal{M}$ and $\mathcal{M}_*$ be as defined by (10.65) and (10.68), respectively.*
3. *Let $\lambda_0 \equiv \left(1 + \frac{\alpha}{2} + \sqrt{\alpha + \frac{\alpha^2}{4}}\right)^{-1}$ and $\lambda_1 = \dfrac{1 + \sqrt{\alpha}}{1 - \alpha}$, where $\alpha \equiv 1 - \alpha_0$.*

Then the following bound will hold.

$$\lambda_0 \left\langle \mathcal{M}_* \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right\rangle \le \left\langle \mathcal{M} \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{pmatrix} \right\rangle \le \lambda_1 \left\langle \mathcal{M}_* \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \boldsymbol{\lambda} \end{pmatrix} \right\rangle. \tag{10.69}$$

Proof. See [BR8]. □

The parameter α in the preceding lemma can easily be verified to depend on $1/\text{cond}(A_0, A)$. Additionally, the condition number of $\mathcal{M}_*$ depends solely on the condition number of $S = BA^{-1}B^T$. As a result, the preceding lemma shows that system (10.64) can be solved by a conjugate gradient algorithm using the inner product $\langle ., . \rangle$, with preconditioner $\mathcal{M}_*$ if S is ill conditioned, or without preconditioner if $\mathcal{M}_*$ is well conditioned. The algorithm of [BR8] solves (10.64) without preconditioning. We summarize the resulting algorithm in the following, noting that it requires computing the action of A_0.

Algorithm 10.5.2 *(Algorithm of [BR8] to Solve (10.2))*

```
1. Compute: $A_0^{-1}\mathbf{f}$ and $BA_0^{-1}\mathbf{f} - \mathbf{g}$.
2. Solve using CG with $\langle ., . \rangle$ inner product:
   $$\mathcal{M} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A_0^{-1}\mathbf{f} \\ BA_0^{-1}\mathbf{f} - \mathbf{g} \end{bmatrix}$$
Output: $(\mathbf{u}, \boldsymbol{\mu})$
```

Remark 10.39. Multiplying equation (10.63) by blockdiag$(A - A_0, I)$ yields:

$$\begin{bmatrix} AA_0^{-1}A - A & (AA_0^{-1} - I)B^T \\ B(A_0^{-1}A - I) & BA_0^{-1}B^T \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} (AA_0^{-1} - I)\mathbf{f} \\ BA_0^{-1}\mathbf{f} - \mathbf{g} \end{bmatrix}. \tag{10.70}$$

This system is symmetric. By Lemma10.38, its coefficient matrix will also be positive definite and spectrally equivalent to blockdiag(A, S). The above system may thus be solved by PCG using a preconditioner blockdiag(A_0, S_0). Solving the above system does not require computing the action of A_0.

Remark 10.40. The normal equations associated with (10.2) has the form:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}^T \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}^T \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}.$$

This yields the symmetric positive definite system:

$$\begin{bmatrix} A^2 + B^TB & AB^T \\ BA & BB^T \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A\mathbf{f} + B^T\mathbf{g} \\ B\mathbf{g} \end{bmatrix}.$$

However, this squares the condition number. To reduce the condition number, a preconditioner may be applied before the normal equations are formed.

Remark 10.41. Yet another positive definite reformulation of (10.2) can be obtained by multiplying (10.2) on the left as follows:

$$\begin{bmatrix} I & 0 \\ 2BA^{-1} & -I \end{bmatrix} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 2BA^{-1} & -I \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix},$$

which yields:

$$\begin{bmatrix} A & B^T \\ B & 2BA^{-1}B^T \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 2BA^{-1}\mathbf{f} - \mathbf{g} \end{bmatrix}.$$

The transformed coefficient matrix will have the block factorization:

$$\begin{aligned} \begin{bmatrix} A & B^T \\ B & 2BA^{-1}B^T \end{bmatrix} &= \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ 0 & S \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & 2S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix}, \end{aligned}$$

where $S = (BA^{-1}B^T)$. The transformed coefficient matrix is symmetric. Additionally, an application of Sylvester's law of inertia [GO4] shows that the coefficient matrix is positive definite when $A = A^T > 0$ and B has full rank. However, the resulting formulation requires the action of A^{-1}.

10.5.3 Block Diagonal Preconditioner for (10.2)

When saddle point system (10.2) is solved using the MINRES algorithm, its rate of convergence will depend on the distribution of the eigenvalues of the coefficient matrix [PA3, CH19, RU5, SA2]. Preconditioning can help improve its rate of convergence, and here, we shall describe a block diagonal preconditioner for a saddle point system [EL8, EL2, KL2, EL3, SI, EL4, EL9, ZU].

We shall denote the saddle point matrix in (10.2) as L and denote its symmetric positive definite block diagonal preconditioner as L_0:

$$L = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \quad \text{and} \quad L_0 = \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix}, \tag{10.71}$$

where $A_0 = A_0^T > 0$ denotes a preconditioner for A, while $S_0 = S_0^T > 0$ denotes a preconditioner for the Schur complement matrix $S = BA^{-1}B^T$. The following result describes the distribution of eigenvalues of the (left) preconditioned matrix $L_0^{-1}L$ in the *special case* when $A_0 = A$.

Lemma 10.42. *Suppose the following conditions hold:*

1. *Let $A_0 = A$ be a symmetric positive definite matrix of size n, and let B be an $m \times n$ matrix of full rank m, where $m < n$.*
2. *Let $S_0 = S_0^T > 0$ denote a preconditioner for the Schur complement matrix $S = BA^{-1}B^T$, with γ_i denoting the i'th eigenvalue of $S_0^{-1}S$:*

$$S\,\mathbf{q}_i = \gamma_i\, S_0\, \mathbf{q}_i,$$

with corresponding eigenvector $\mathbf{q}_i \in \mathbb{R}^m$.

Then, an eigenvalue λ of the above preconditioned matrix $L_0^{-1}L$ will lie in:

$$\lambda \in I^{(-)} \cup I^{(+)},$$

where

$$\begin{aligned} I^{(-)} &\equiv \qquad [\frac{1-\sqrt{1+4\gamma_{max}}}{2}, \frac{1-\sqrt{1+4\gamma_{min}}}{2}] \subset (-\infty, 0) \\ I^{(+)} &\equiv \{1\} \cup [\frac{1+\sqrt{1+4\gamma_{min}}}{2}, \frac{1+\sqrt{1+4\gamma_{max}}}{2}] \subset [1, \infty). \end{aligned} \tag{10.72}$$

Proof. We follow the proof in [EL8, EL2, KL2, EL3, SI]. Let λ be an eigenvalue of $L_0^{-1}L$ corresponding to eigenvector $\left(\mathbf{u}^T, \boldsymbol{\mu}^T\right)^T$:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \lambda \begin{bmatrix} A & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}.$$

This yields the equations:

$$\begin{cases} (1-\lambda)\, A\mathbf{u} + B^T\boldsymbol{\mu} = \mathbf{0} \\ B\mathbf{u} \qquad\qquad\quad = \lambda\, S_0\boldsymbol{\mu}. \end{cases} \tag{10.73}$$

If $\lambda = 1$, then the first block row yields that $\boldsymbol{\mu} = \mathbf{0}$ since B^T has full rank. Substituting this into the second block row yields that $B\mathbf{u} = \mathbf{0}$, i.e., $\mathbf{u} \in \mathcal{K}_0$. Since $\dim(\mathcal{K}_0) = n - m$, this yields that $\lambda = 1$ will be an eigenvalue of $L_0^{-1}L$ of multiplicity $(n-m)$. If $\{\mathbf{w}_1, \ldots, \mathbf{w}_{n-m}\}$ forms a basis for $\mathcal{K}_0$, the $(n-m)$ independent eigenvectors of $L_0^{-1}L$ will be $(\mathbf{w}_i^T, \mathbf{0}^T)^T$ for $1 \le i \le (n-m)$.

To determine the remaining $2m$ eigenvalues of $L_0^{-1}L$, let $\lambda \neq 1$. In this case, the first block row of (10.73) yields $\mathbf{u} = (\lambda - 1)^{-1} A^{-1} B^T \boldsymbol{\mu}$. Substituting this into the second block row yields:

$$\left(BA^{-1}B^T\right) \boldsymbol{\mu} = \lambda \, (\lambda - 1) \, S_0 \, \boldsymbol{\mu}.$$

If γ_i is an eigenvalue of $S_0^{-1}S$ corresponding to eigenvector $\mathbf{q}_i$:

$$S \, \mathbf{q}_i = \gamma_i \, S_0 \, \mathbf{q}_i, \qquad \text{for } i = 1, \ldots, m,$$

then it must hold that $\gamma_i = \lambda \, (\lambda - 1)$. Solving this quadratic equation for λ using each of the m eigenvalues γ_i of $S_0^{-1}S$ yields two roots λ_{+i} and λ_{-i} as:

$$\lambda_{\pm i} = \frac{1 \pm \sqrt{1 + 4\gamma_i}}{2}, \qquad \text{for } i = 1, \ldots, m,$$

with the associated eigenvectors:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{(\lambda_{\pm i} - 1)} A^{-1} B^T \mathbf{q}_i \\ \mathbf{q}_i \end{bmatrix} = \lambda_{\pm i} \begin{bmatrix} A & 0 \\ 0 & S_0 \end{bmatrix} \begin{bmatrix} \frac{1}{(\lambda_{\pm i} - 1)} A^{-1} B^T \mathbf{q}_i \\ \mathbf{q}_i \end{bmatrix}.$$

Thus, we have determined all the eigenvalues and eigenvectors of $L_0^{-1}L$. Bound (10.72) follows immediately. □

Remark 10.43. The analysis in Lemma 10.42 applies only when $A_0 = A$. In practice, it will be preferable to employ a block diagonal preconditioner $L_0 = \text{blockdiag}(A_0, S_0)$ where $A_0 \neq A$. In this case, determining the exact eigenvalues of $L_0^{-1}L$ will be complicated. However, we may still obtain *bounds* for distribution of eigenvalues of $L_0^{-1}L$ as follows [RU5]. Let $A_0 = A_0^T > 0$ and $S_0 = S_0^T > 0$, so that $L_0^{1/2}$ is well defined. It is easily verified that the eigenvalues of $L_0^{-1}L$ are identical to the eigenvalues of $\tilde{L} \equiv (L_0^{-1/2} L L_0^{-1/2})$:

$$\tilde{L} \equiv L_0^{-1/2} L L_0^{-1/2} = \begin{bmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & 0 \end{bmatrix},$$

where $\tilde{A} = (A_0^{-1/2} A A_0^{-1/2})$ and $\tilde{B} = (S_0^{-1/2} B A_0^{-1/2})$. The eigenvalues of $\tilde{L}$ can be estimated by applying Lemma 10.10 (with A replaced by $\tilde{A}$ and B replaced by $\tilde{B}$). When $A_0 \asymp A$, the eigenvalues γ_i of $\tilde{A} = (A_0^{-1/2} A A_0^{-1/2})$ will lie in an interval independent of the mesh parameters. Since the singular values σ_j of $\tilde{B}$ correspond to the positive square roots of the eigenvalues of $\tilde{B}\tilde{B}^T = S_0^{-1/2}(BA_0^{-1}B^T)S_0^{-1/2}$, when $A_0 \asymp A$, we will obtain $(BA_0^{-1}B^T) \asymp S$. If additionally $S \asymp S_0$, we will obtain $S_0^{-1/2}(BA_0^{-1}B^T)S_0^{-1/2} \asymp I$. As a result, the singular values σ_j of $\tilde{B}$ will also be independent of the mesh parameters.

Motivated by the preceding, the preconditioner $L_0 = \mathrm{blockdiag}(A_0, S_0)$ may be employed to precondition (10.2) using a Krylov space algorithm such as MINRES [PA3, CH19, RU5] or the *conjugate residual* method [SA2]. These algorithms require the coefficient matrix to be *symmetric*. Since the preconditioned matrix $L_0^{-1} L$ will unfortunately not be symmetric in the Euclidean inner product, though it is easily verified to be symmetric in the inner product $\langle ., . \rangle_{L_0}$ generated by the positive definite symmetric preconditioner L_0:

$$\langle (\mathbf{u}, \boldsymbol{\mu}), (\mathbf{v}, \boldsymbol{\lambda}) \rangle_{L_0} \equiv \mathbf{v}^T A_0 \mathbf{u} + \boldsymbol{\lambda}^T S_0 \boldsymbol{\mu}. \tag{10.74}$$

As a result, the MINRES algorithm may be employed to solve:

$$\begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix}^{-1} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} A_0 & 0 \\ 0 & S_0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{10.75}$$

with $\langle ., . \rangle_{L_0}$ replacing the Euclidean inner product. See [PA3, CH19, RU5] or [SA2] for a listing of the algorithm.

Remark 10.44. An alternate preconditioning method was described for (10.2) in [RU5]. Let $M = \mathrm{blockdiag}(R, Q)$ denote a (possibly nonsymmetric) block diagonal matrix with diagonal blocks R and Q of size n and m, respectively. Then, system (10.2) can be transformed into the symmetric indefinite system:

$$\left(M^{-1} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} M^{-T} \right) \left(M^T \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} \right) = M^{-1} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}.$$

We shall denote the transformed system as:

$$\tilde{L} \begin{bmatrix} \tilde{\mathbf{u}} \\ \tilde{\boldsymbol{\mu}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}} \\ \tilde{\mathbf{g}} \end{bmatrix}, \tag{10.76}$$

where

$$\tilde{L} = \begin{bmatrix} R^{-1} A R^{-T} & R^{-1} B^T Q^{-T} \\ Q^{-1} B R^{-T} & 0 \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{u}} \\ \tilde{\boldsymbol{\mu}} \end{bmatrix} = \begin{bmatrix} R^T \mathbf{u} \\ Q^T \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{f}} \\ \tilde{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} R^{-1} \mathbf{f} \\ Q^{-1} \mathbf{g} \end{bmatrix}. \tag{10.77}$$

System (10.77) may be solved by the MINRES algorithm, from which we may obtain $\mathbf{u} = R^{-T} \tilde{\mathbf{u}}$ and $\boldsymbol{\mu} = Q^{-T} \tilde{\boldsymbol{\mu}}$. Its convergence rate will depend on the eigenvalues of $\tilde{L}$, which depends on the eigenvalues of $R^{-1} A R^{-T}$ and the singular values of $R^{-1} B^T Q^{-T}$ by Lemma 10.10. Heuristically, R must be chosen such that:

$$R^{-1} A R^{-T} \asymp I \;\; \Leftrightarrow \;\; A \asymp R R^T.$$

Requiring $R^{-1} B^T Q^{-T}$ to be a unitary matrix will be equivalent to requiring:

$$Q^{-1} B R^{-T} R^{-1} B^T Q^{-T} \asymp I \;\; \Leftrightarrow \;\; Q^{-1} S Q^{-T} \asymp I \;\; \Leftrightarrow \;\; S \asymp Q Q^T,$$

since $R^{-T} R^{-1} \asymp A^{-1}$. Matrices R and Q may be obtained by *incomplete* factorization of A and S, respectively. Since S is typically not assembled, heuristic approximations can be employed for Q, based on the structure of S.

10.5.4 Block Triangular Preconditioner for (10.2)

If a block triangular matrix is employed to precondition the saddle point system (10.2), symmetry will be lost. However, it may be noted that:

- For suitably chosen diagonal blocks, the eigenvalues of the resulting preconditioned system can be chosen to be all *positive.* By comparison, the preconditioned system resulting from a symmetric positive definite block diagonal preconditioner is *indefinite.*
- Block triangular preconditioners can be inverted at almost the computational cost as block diagonal preconditioners.

Due to non-symmetry of the preconditioned system, a *GMRES* method will be required. In this case, the eigenvalues of the preconditioned matrix may not be the sole factor determining the convergence rate of the algorithm [SA2].

We shall consider block triangular preconditioners of the following form:

$$M_1 = \begin{bmatrix} A_0 & 0 \\ B & -S_0 \end{bmatrix} \quad \text{or} \quad M_2 = \begin{bmatrix} A_0 & B^T \\ 0 & -S_0 \end{bmatrix}, \tag{10.78}$$

where A_0 and S_0 are symmetric positive definite matrices of size n and m, respectively. The following preliminary result describes the distribution of eigenvalues of the preconditioned matrix $M_1^{-1}L$ in the case when $A_0 = A$, where L denotes the coefficient matrix in (10.2).

Lemma 10.45. *Suppose the following conditions hold:*

1. *Let A be a symmetric positive definite matrix of size n and let B be full rank matrix of size $m \times n$ with $m < n$.*
2. *Let M_1 denote the block lower triangular preconditioner defined in (10.78) with $A_0 = A$, and a symmetric positive definite matrix S_0 of size m.*
3. *Let γ_i denote the i'th eigenvalue of $S_0^{-1}S$:*

$$S\,\mathbf{q}_i = \gamma_i\, S_0\mathbf{q}_i, \tag{10.79}$$

with corresponding eigenvector $\mathbf{q}_i \in \mathbb{R}^m$.

Then, if λ is an eigenvalue of $M_1^{-1}L$:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \lambda \begin{bmatrix} A & 0 \\ B & -S_0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix},$$

it will lie in the set:

$$\lambda \in (\{1\} \cup [\gamma_{min}, \gamma_{max}]) \subset (0, \infty),$$

where $\gamma_{min} = \min\{\gamma_1, \ldots, \gamma_m\}$ and $\gamma_{max} = \max\{\gamma_1, \ldots, \gamma_m\}$.

Proof. The proof is similar to the block diagonal case [EL8, EL2, KL2, EL3, SI]. The equation corresponding to the generalized eigenvalue problem is:

$$\begin{cases} (1-\lambda)\; A\mathbf{u} + B^T\boldsymbol{\mu} = \mathbf{0} \\ (1-\lambda)\; B\mathbf{u} \qquad\quad = -\lambda\, S_0\boldsymbol{\mu}. \end{cases}$$

The first block row yields that if $\lambda = 1$, then $\boldsymbol{\mu} = \mathbf{0}$, since B^T has rank m. When $\lambda = 1$ and $\boldsymbol{\mu} = \mathbf{0}$, the second block row is satisfied. As a result, $\lambda = 1$ is an eigenvalue of $M_1^{-1}L$ of multiplicity n. Indeed, if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ form a basis for $\mathbb{R}^n$, then $\left(\mathbf{u}_i^T, \mathbf{0}^T\right)^T$ for $i = 1, \ldots, n$ will form a basis for the eigenspace corresponding to $\lambda = 1$. To determine the remaining m eigenvalues of $M_1^{-1}L$, suppose $\lambda \neq 1$. The first block row yields $\mathbf{u} = (\lambda - 1)^{-1}\, A^{-1}\, B^T\boldsymbol{\mu}$. Substituting this into the second block row yields:

$$S\,\boldsymbol{\mu} \;=\; \lambda\, S_0\, \boldsymbol{\mu},$$

where $S = BA^{-1}B^T$. Thus λ corresponds to a generalized eigenvalue of $S\,\mathbf{q} = \gamma\, S_0\mathbf{q}$. Given m linearly independent eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_m$ satisfying $S\,\mathbf{q}_i = \gamma_i\, S_0\, \mathbf{q}_i$, it will hold that $\mathbf{u}_i = (\gamma_i - 1)^{-1}\, A^{-1}\, B^T\, \mathbf{q}_i$. Thus $\lambda = \gamma_i$ will be an eigenvalue of $M_1^{-1}L$ corresponding to the eigenvector $\left(\mathbf{u}_i^T, \mathbf{q}_i^T\right)^T$ for $i = 1, \ldots, m$. Thus, the eigenvalues of $M_1^{-1}L$ lie in the set $\lambda \in \left(\{1\} \cup [\gamma_{min}, \gamma_{max}]\right)$. □

Remark 10.46. The eigenvalues of $M_2^{-1}L$ will also lie in the same interval when $A_0 = A$. However, the eigenvectors will differ from those for $M_1^{-1}L$.

Remark 10.47. The preconditioner M_1 defined by (10.78) is identical to that employed in the block matrix form (10.26) of the inexact preconditioned Uzawa algorithm [BR16] described in Chap. 10.2. As a result, by Lemma 10.23, if matrices A_0 and S_0 satisfy the following for some $0 \leq \gamma < 1$ and $0 \leq \delta < 1$:

$$\begin{cases} (1-\gamma)\, \boldsymbol{\eta}^T S_0 \boldsymbol{\eta} \leq \boldsymbol{\eta}^T S \boldsymbol{\eta} \leq \boldsymbol{\eta}^T S_0 \boldsymbol{\eta}, & \forall \boldsymbol{\eta} \in \mathbb{R}^m \\ (1-\delta)\, \mathbf{v}^T A_0 \mathbf{v} \leq \mathbf{v}^T A \mathbf{v} \leq \mathbf{v}^T A_0 \mathbf{v}, & \forall \mathbf{v} \in \mathbb{R}^m, \end{cases}$$

then the eigenvalues of $M_1^{-1}L$ will have positive real part. See also [ZU].

Remark 10.48. A result in [KL2] shows that the (right) preconditioned matrix LM_2^{-1} is *symmetrizable* with *positive eigenvalues*, under appropriate restrictions on $A_0 = A_0^T > 0$ and $S_0 = S_0^T > 0$. Additional results on the spectra of $M_1^{-1}L$ and LM_2^{-1} may be found in [EL8, EL2, EL3, SI, EL4, EL9, ZU].

10.5.5 A Saddle Point Preconditioner for (10.2)

In certain applications, it may be advantageous to precondition system (10.2) using another saddle point matrix [AL2], see Chap. 10.7. The efficacy of such a preconditioner will depend on the computational cost of solving the preconditioned system. We consider a saddle point preconditioner L_0 of the form:

$$L_0 = \begin{bmatrix} A_0 & B^T \\ B & 0 \end{bmatrix} \quad \text{where} \quad L = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where A_0 denotes a symmetric positive definite preconditioner for A. Below, we list an *unaccelerated* splitting algorithm for solving (10.2).

Algorithm 10.5.3 *(Unaccelerated Splitting Algorithm)*

```
1. Let $\mathbf{u}^{(0)}$ and $\boldsymbol{\mu}^{(0)}$ be starting iterates
2. For $k = 0, 1, \ldots$ until convergence do:
       $$\begin{bmatrix} \mathbf{u}^{(k+1)} \\ \boldsymbol{\mu}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} + \begin{bmatrix} A_0 & B^T \\ B & 0 \end{bmatrix}^{-1} \left( \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} - \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{bmatrix} \right) \qquad (10.80)$$
3. Endfor
Output: $\left( \mathbf{u}^{(k)^T}, \boldsymbol{\mu}^{(k)^T} \right)^T$
```

If the linear system with coefficient matrix L_0 is solved using a Schur complement algorithm, each iteration will involve an *inner iteration*. The following result describes the distribution of eigenvalues of $L_0^{-1} L$.

Lemma 10.49. *Suppose the following conditions hold.*

1. *Let A and A_0 be symmetric positive definite matrices of size n and let B be a full rank matrix of size $m \times n$ with $m < n$.*
2. *Let γ_i denote a generalized eigenvalue of the system:*

$$A\,\mathbf{w}_i = \gamma_i\, A_0\, \mathbf{w}_i,$$

corresponding to eigenvector $\mathbf{w}_i \in \mathbb{R}^n$.

3. *Let λ be a generalized eigenvalue of:*

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix} = \lambda \begin{bmatrix} A_0 & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\mu} \end{bmatrix}. \qquad (10.81)$$

Then, the eigenvalue λ will be real and lie in the interval:

$$\lambda \in (\{1\} \cup [\gamma_{\min}, \gamma_{\max}]),$$

where $\gamma_{\min} = \min\{\gamma_1, \ldots, \gamma_n\}$ and $\gamma_{\max} = \max\{\gamma_1, \ldots, \gamma_n\}$.

Proof. We shall follow the proof in [AL2]. We rewrite (10.81) as:

$$\begin{cases} (A - \lambda\, A_0)\, \mathbf{u} + (1 - \lambda)\, B^T \boldsymbol{\mu} = \mathbf{0} \\ (1 - \lambda)\, B\, \mathbf{u} \qquad\qquad\qquad = \mathbf{0}. \end{cases} \tag{10.82}$$

Suppose $\lambda = 1$. In this case system (10.82) reduces to:

$$(A - \ A_0)\, \mathbf{u} = \mathbf{0},$$

for arbitrary $\boldsymbol{\mu} \in \mathbb{R}^m$. There will be a *nonzero* solution $\mathbf{u}$ to the above equation *iff* $\gamma = 1$ is an eigenvalue of $A_0^{-1}A$. Regardless, the component $\boldsymbol{\mu} \in \mathbb{R}^m$ can be chosen arbitrarily, so that $\lambda = 1$ will be an eigenvector of $L_0^{-1}L$ of multiplicity $m+r$, where r denotes the multiplicity of $\gamma = 1$ as an eigenvalue of $A_0^{-1}A$. We set $r = 0$ if $\gamma = 1$ is not an eigenvalue of $A_0^{-1}A$. Corresponding eigenvectors can be obtained easily in the form $(\mathbf{0}^T, \boldsymbol{\mu}_i^T)^T$ or $(\mathbf{u}_i^T, \mathbf{0}^T)^T$.

Next, suppose that $\lambda \neq 1$, possibly complex. In this case, the second block row of (10.82) reduces to $B\,\mathbf{u} = \mathbf{0}$, so that $\mathbf{u} \in \mathcal{K}_0 = \text{Kernel}(B) \subset \mathbb{C}^n$. Taking the complex inner product of the first block row of (10.82) with $\mathbf{u} \in \mathcal{K}_0 \subset \mathbb{C}^n$ and rearranging terms yields:

$$\lambda = \frac{\mathbf{u}^H A \mathbf{u}}{\mathbf{u}^H A_0 \mathbf{u}}.$$

Bounds for the generalized Rayleigh quotient yields $\gamma_{\min} \leq \lambda \leq \gamma_{\max}$. It can be verified that an eigenvalue $\lambda \neq 1$ will correspond to the value of $\mathbf{u}^H A\mathbf{u}/\mathbf{u}^H A_0 \mathbf{u}$ at critical points $\mathbf{u}$ within $\mathcal{K}_0$. Corresponding eigenvectors can be obtained by solving for $\mathbf{u} \in \mathcal{K}_0$ and $\boldsymbol{\mu}$. We omit further details. □

Remark 10.50. The preceding lemma shows that $\rho(I - L_0^{-1}L) = \rho(I - A_0^{-1}A)$. As a result, if matrix A_0 is chosen such that $\rho\left(I - A_0^{-1}A\right) < 1$, then the *unaccelerated* iteration (10.80) will converge with an error contraction factor given by $\rho\left(I - A_0^{-1}A\right)$. In practice, A_0 must also be chosen so that saddle point matrix L_0 is easily inverted, which requires that both A_0 and $(BA_0^{-1}B^T)$ be invertible at low cost.

Remark 10.51. If matrix A_0 is spectrally equivalent to A then $BA_0^{-1}B^T$ will be spectrally equivalent to $BA^{-1}B^T$. Indeed, if $\lambda\left(A_0^{-1}A\right) \in [\beta_1, \beta_2]$ then $\lambda\left((BA_0^{-1}B^T)^{-1}\,(BA^{-1}B^T)\right) \in [\beta_2^{-1}, \beta_1^{-1}]$ since:

$$\frac{\boldsymbol{\mu}^T BA^{-1}B^T\boldsymbol{\mu}}{\boldsymbol{\mu}^T BA_0^{-1}B^T\boldsymbol{\mu}} = \frac{\mathbf{y}^T A^{-1}\mathbf{y}}{\mathbf{y}^T A_0^{-1}\mathbf{y}}, \qquad \text{for} \quad \mathbf{y} = B^T\boldsymbol{\mu},$$

and since $\lambda\left(A_0 A^{-1}\right) \in [\beta_2^{-1}, \beta_1^{-1}]$ when $\lambda\left(A_0^{-1}A\right) \in [\beta_1, \beta_2]$.

10.6 Applications to the Stokes and Navier-Stokes Equations

In this section, we describe applications of saddle point iterative methods to solve discretizations of the *incompressible* Stokes and Navier-Stokes equations. Computational methods for the incompressible Navier-Stokes equations are described in [CH27, CH28, PE4, TE, GI3, BR33, CH29, CA28, BE19]. Although we shall focus primarily on the *steady state* Stokes equation, we shall indicate extensions to *implicit* time discretizations of linearizations of the Navier-Stokes equations. Our discussion will include the following.

- Background on the Stokes and Navier-Stokes equations.
- Properties of matrices A and B for the *steady state Stokes* equation.
- Applications of Uzawa, penalty, projection and block matrix algorithms for the *steady state Stokes* equation.
- Applications to the Stokes-Oseen (*linearized Navier-Stokes*) problem.
- Applications to the *time dependent* Stokes and Navier-Stokes equations.

We consider finite element and finite difference discretizations of the Stokes equations [TE, GI3, BR33, CH27, CH28]. Domain decomposition applications to fluid flow problems are described in [CA12, CA16, CA11].

10.6.1 Background

The incompressible Navier-Stokes equations on a domain $\Omega \subset \mathbb{R}^d$ are:

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \nu\, \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\, \mathbf{u} + \nabla p = \mathbf{f}, & \text{in } \Omega \times (0,t) \\ \nabla \cdot \mathbf{u} = 0, & \text{in } \Omega \times (0,t), \end{cases} \tag{10.83}$$

where $\mathbf{u}(x,t)$ and $p(x,t)$ denote the unknown velocity and pressure of the fluid, while $\mathbf{f}(x,t)$ denotes a forcing term. The first equation expresses conservation of momentum, while the second equation expresses incompressibility of the fluid, see [CH29]. The parameter $\nu > 0$ represents the viscosity of the flow on the domain Ω, and is the reciprocal of the Reynolds number. The Navier-Stokes equation is of parabolic character, and appropriate *boundary* and *initial* conditions must be prescribed. Typical boundary conditions are:

$$\begin{cases} \mathbf{u} = \mathbf{g}_D, & \text{on } \partial\Omega_D \times (0,t), \text{ Dirichlet type} \\ -\nu\, \dfrac{\partial \mathbf{u}}{\partial \mathbf{n}} + p\, \mathbf{n} = \mathbf{g}_N, & \text{on } \partial\Omega_N \times (0,t), \text{ Neumann type} \end{cases} \tag{10.84}$$

where $\mathbf{n}$ represents the unit exterior normal to the boundary segment $\partial\Omega$. Initial conditions will specify $\mathbf{u}(x,0) = \mathbf{u}_0(x)$ on Ω.

Remark 10.52. Since The momentum equation in (10.83) will admit multiple solutions for the pressure, since $\nabla\,(p(x,t) + c) = \nabla p(x,t)$ for a constant c. Imposing Dirichlet boundary conditions $\mathbf{u} = \mathbf{g}_D$ on $\partial\Omega$ will not eliminate this nonuniqueness in the pressure. However, Neumann boundary conditions $-\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} + p\,\mathbf{n} = \mathbf{g}_N$ on a segment $\partial\Omega_N \neq \emptyset$ will yield a unique pressure.

An important approximation of the Navier-Stokes equation occurs when the term $(\mathbf{u}\cdot\nabla)\,\mathbf{u}$ is relatively *small* in relation to the other terms, typically when ν is large. The Stokes problem is obtained by omitting this nonlinear term. The steady state *Stokes* equation with Dirichlet boundary conditions is:

$$\begin{cases} -\nu\,\Delta\mathbf{u} + \nabla p = \mathbf{f}, & \text{in } \Omega \\ \nabla\cdot\mathbf{u} = 0, & \text{in } \Omega \\ \mathbf{u} = \mathbf{g}_D, & \text{on } \partial\Omega. \end{cases} \tag{10.85}$$

An alternative approximation of the steady state Navier-Stokes equations arises when it is linearized about a given velocity field $\mathbf{w}(x)$ using a fixed point linearization. This results in the *Stokes-Oseen* problem:

$$\begin{cases} -\nu\,\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\,\mathbf{w} + (\mathbf{w}\cdot\nabla)\,\mathbf{u} + \nabla p = \mathbf{f}, & \text{in } \Omega \\ \nabla\cdot\mathbf{u} = 0, & \text{in } \Omega \\ \mathbf{u} = \mathbf{g}_D, & \text{on } \partial\Omega. \end{cases} \tag{10.86}$$

where we have considered the Dirichlet problem for simplicity. In applications, for further simplification, we shall shift the term $(\mathbf{u}\cdot\nabla)\,\mathbf{w}$ (zeroth order in the derivatives of $\mathbf{u}$) to the right hand side as in a Picard iteration.

Discretization of (10.85). Let $\mathbf{g}_D(x) = \mathbf{0}$ and $\Omega \subset \mathbb{R}^d$. A finite element discretization of (10.85) can be obtained by Galerkin approximation of its *weak form*, see [TE, GI3, BR33]. Define $\mathbf{V}_D \equiv (V_D)^d$ and $Q_D = L^2(\Omega)/\mathbb{R}$ (functions in $L^2(\Omega)$ with mean value zero), where $V_D \equiv H_0^1(\Omega)$.Then, the weak form of (10.85) is obtained by multiplying the momentum equation in (10.85) by $\mathbf{v}(x) \in \mathbf{V}_D$ and integrating by parts over Ω, and multiplying the incompressibility equation by $-q(x) \in Q_D$ and integrating over Ω. The resulting weak formulation seeks $\mathbf{u}(x) \in \mathbf{V}_D$ and $p(x) \in Q_D$ satisfying:

$$\begin{cases} \mathcal{A}(\mathbf{u},\mathbf{v}) + \mathcal{B}(\mathbf{v},p) = (\mathbf{f},\mathbf{v}), & \forall \mathbf{v}\in\mathbf{V}_D \\ \mathcal{B}(\mathbf{u},q) = 0, & \forall q \in Q_D, \end{cases} \tag{10.87}$$

where $\mathcal{A}(.,.) : \mathbf{V}_D \times \mathbf{V}_D \to \mathbb{R}$ and $\mathcal{B}(.,.) : \mathbf{V}_D \times Q_D \to \mathbb{R}$ are bilinear forms:

$$\begin{cases} \mathcal{A}(\mathbf{u},\mathbf{v}) \equiv \nu \sum_{i=1}^d \int_\Omega (\nabla u_i \cdot \nabla v_i)\,dx \\ \mathcal{B}(\mathbf{v},q) \;\equiv -\int_\Omega q\,(\nabla\cdot\mathbf{v})\,dx. \end{cases} \tag{10.88}$$

for $\mathbf{u}(x) = (u_1(x),\ldots,u_d(x)) \in \mathbf{V}_D$ and $\mathbf{v} = (v_1(x),\ldots,v_d(x)) \in \mathbf{V}_D$, and $p(x)$, $q(x) \in Q_D$. The term:

$$\mathcal{B}(\mathbf{v},p) = -\int_\Omega p(x)\,(\nabla\cdot\mathbf{v}(x))\,dx = \int_\Omega \nabla p(x)\cdot\mathbf{v}(x)\,dx,$$

using integration by parts, since $\mathbf{v}(x)$ has *zero* boundary values.

Given a quasiuniform triangulation Ω_h of Ω with elements of size h, let $\mathbf{V}_h \subset \mathbf{V}_D$ and $Q_h \subset Q_D$ denote finite element spaces for the velocity and pressure, with $\mathbf{V}_h = (V_h)^d$ where $V_h \subset V_D$. A finite element discretization of (10.87) is obtained by seeking $\mathbf{u}_h(x) \in V_D$ and $p_h(x) \in Q_h$ such that:

$$\begin{cases} \mathcal{A}(\mathbf{u}_h, \mathbf{v}_h) + \mathcal{B}(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h), & \forall \mathbf{v}_h \in \mathbf{V}_h \\ \mathcal{B}(\mathbf{u}_h, q_h) \qquad\qquad\quad\;\; = 0, & \forall q_h \in Q_h. \end{cases} \tag{10.89}$$

A linear system corresponding to the above discretization can be obtained as follows. Let $\{\boldsymbol{\psi}_1(x), \ldots, \boldsymbol{\psi}_n(x)\}$ denote a basis for $\mathbf{V}_h$ and $\{q_1(x), \ldots, q_m(x)\}$ a basis for $Q_h \subset L^2(\Omega)$. Expand $\mathbf{u}_h(x)$ and $p_h(x)$ using this basis:

$$\begin{aligned} \mathbf{u}_h(x) &= \textstyle\sum_{i=1}^{n} (\mathbf{u}_h)_i \, \boldsymbol{\psi}_i(x) \\ p_h(x) &= \textstyle\sum_{i=1}^{m} (\mathbf{p}_h)_i \, q_i(x), \end{aligned}$$

where with some abuse of notation, we have used $\mathbf{u}_h(x)$ to denote a finite element function and $\mathbf{u}_h$ to denote its vector representation relative to the given basis. Substituting $\mathbf{v}_h(x) = \boldsymbol{\psi}_i(x)$ for $i = 1, \ldots, n$ and $q_h(x) = q_i(x)$ for $i = 1, \ldots, m$ into the above yields the following *saddle point* linear system:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_h \\ \mathbf{p}_h \end{bmatrix} = \begin{bmatrix} \mathbf{f}_h \\ \mathbf{0} \end{bmatrix}, \tag{10.90}$$

where matrices A and B, and vector $\mathbf{f}_h$ are defined by:

$$(A)_{ij} = \mathcal{A}(\boldsymbol{\psi}_i, \boldsymbol{\psi}_j), \quad (B)_{ij} = \mathcal{B}(\boldsymbol{\psi}_j, q_i) \quad (\mathbf{f}_h)_i = (\mathbf{f}, \boldsymbol{\psi}_i).$$

Importantly, for a suitable ordering of the basis of $\mathbf{V}_h$, matrix A will be block diagonal. Let $\{\phi_1, \ldots, \phi_k\}$ form a nodal basis for V_h where $n = k\, d$. Define:

$$\begin{cases} \boldsymbol{\psi}_i(x) \qquad\quad\; = (\phi_i(x), \ldots, 0)^T, & \text{for } 1 \le i \le k \\ \vdots & \\ \boldsymbol{\psi}_{(d-1)k+i}(x) = (0, \ldots, \phi_i(x))^T, & \text{for } 1 \le i \le k. \end{cases}$$

In this case matrix A will have the block diagonal form:

$$A = \begin{bmatrix} A^{(1)} & & \\ & \ddots & \\ & & A^{(d)} \end{bmatrix}, \quad \text{where } \left(A^{(l)}\right)_{ij} = \nu \int_{\Omega} (\nabla \phi_i \cdot \nabla \phi_j) \, dx.$$

By construction, matrix A will be *symmetric positive definite* where each diagonal block $A^{(l)}$ corresponds to a finite element discretization of $-\nu\, \Delta$. Matrix B^T will correspond to a discretization of the gradient operator, while B will be a discretization of the negative of the divergence operator. To ensure that discretization (10.89) of (10.87) is *stable* and that saddle point system (10.90) is solvable, the finite element spaces $\mathbf{V}_h$ and Q_h must be compatibly chosen. The following result states sufficient conditions for a stable discretization.

Lemma 10.53. *Suppose the following conditions hold.*

1. *Coercivity: Let $\alpha > 0$ be independent of h such that:*
$$\mathcal{A}(\mathbf{v}_h, \mathbf{v}_h) \ge \alpha \, \|\mathbf{v}_h\|_{\mathbf{V}_D}^2, \quad \forall \mathbf{v}_h \in \mathcal{K}_0^h, \tag{10.91}$$
where $\mathcal{K}_0^h = \{\mathbf{v}_h \in \mathbf{V}_h \, : \, \mathcal{B}(\mathbf{v}_h, q_h) = 0, \;\; \forall q_h \in Q_h\}$.
2. *Uniform inf-sup condition: Let $\beta > 0$ be independent of h such that:*
$$\sup_{\mathbf{v}_h \in \mathbf{V}_h \setminus \{\mathbf{0}\}} \frac{\mathcal{B}(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{\mathbf{V}_D}} \ge \beta \, \|q_h\|_{Q_D}, \quad \forall q_h \in Q_h. \tag{10.92}$$

Then, discretization (10.89) of (10.87) will be stable with the error bound:
$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{V}_D} + \|p - p_h\|_{Q_D} \le C \left(\inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}_h\|_{\mathbf{V}_D} + \inf_{q_h \in Q_h} \|p - q_h\|_{Q_D} \right),$$
for some $C > 0$ independent of h, but dependent on $\alpha > 0$ and $\beta > 0$.

Proof. See [TE, GI3, BR33]. □

Remark 10.54. The *coercivity* condition (10.91) will hold trivially for the Stokes equation since $\mathcal{A}(\mathbf{v}, \mathbf{v}) = \nu \, |\mathbf{v}|_{1,\Omega}^2$ is equivalent to the Sobolev norm $\|\mathbf{v}\|_{1,\Omega}^2$. As a result, the stability of a discretization of Stokes equation will depend primarily on the uniform inf-sup condition (10.92) holding for the given choice of spaces $\mathbf{V}_h$ and Q_h. From a matrix viewpoint, the uniform inf-sup condition (10.92) is equivalent to the requirement that given $\mathbf{q}_h \in \mathbb{R}^m$ (satisfying $\mathbf{1}^T \mathbf{q}_h = 0$, for the Dirichlet problem), there exists $\mathbf{v}_h \in \mathbb{R}^m$:
$$B \, \mathbf{v}_h = \mathbf{q}_h \quad \text{with} \quad \|\mathbf{v}_h\|_A \le \beta^{-1} \, \|\mathbf{q}_h\|_M, \tag{10.93}$$
for $\beta > 0$ independent of h, where M denotes the mass matrix:
$$M_{ij} \equiv \int_\Omega q_i(x) \, q_j(x) \, dx \quad \text{with} \quad \mathbf{q}_h^T M \mathbf{q}_h = \|q_h\|_{0,\Omega}^2. \tag{10.94}$$
Here, we have used that $\|\mathbf{v}_h\|_{\mathbf{V}_h}$ is equivalent to $\|\mathbf{v}_h\|_A$. Thus, given a finite element space Q_h, the finite element space $\mathbf{V}_h$ must be "large enough" to ensure that (10.93) holds. See [GI3, BR33] for examples of such spaces.

Remark 10.55. For *Dirichlet* boundary conditions, if we choose $Q_D = L^2(\Omega)$, and $q_h(x) = 1$ belongs to Q_h, then it can be verified that matrix B^T will satisfy $B^T \, \mathbf{1} = \mathbf{0}$, where $\mathbf{1} = (1, \ldots, 1)^T$. This will hold because $\nabla \, q_h(x) = 0$, or equivalently $\mathcal{B}(\mathbf{v}_h, q_h) = 0$ for all $\mathbf{v}_h \in \mathbf{V}_h$, when $q_h(x) = 1$. Thus, under these assumptions matrix B will not have full rank, and the saddle point matrix in (10.90) will be *singular*. However, system (10.90) will be consistent since $\mathbf{1}^T \mathbf{0} = 0$. A full rank matrix B can be constructed, if desired, by choosing the pressure spaces Q_h and Q_D to consist only of functions having *mean value zero*, i.e., $Q_h \subset Q_D = L^2(\Omega)/\mathbb{R}$. However, this will be cumbersome and we shall instead work with a singular coefficient matrix, and modify algorithms from preceding sections appropriately.

Properties of A and B. Matrices A and B in the system (10.90) will satisfy the following *properties* for a finite element discretization of (10.85). Matrix $A = \text{blockdiag}\left(A^{(1)}, \ldots, A^{(d)}\right)$ of size n will be *symmetric positive definite* and *sparse*. Each diagonal block $A^{(l)}$ will correspond to a discretization of $-\nu\,\Delta$. Thus, by standard finite element theory [ST14, CI2, JO2], there will exist $\gamma_0 > 0$ and $\gamma_1 > 0$ independent of h such that:

$$\gamma_0\, h^d \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \gamma_1 h^{d-2}.$$

The condition number of matrix A will satisfy $\text{cond}(A) \leq C\,h^{-2}$ for some $C > 0$ independent of h. By construction, matrix B of size $m \times n$ will be *sparse*, with $m < n$. We shall assume that it satisfies the uniform inf-sup condition (10.93). The following additional property will hold for matrix B.

Lemma 10.56. *The matrix B arising in the discretization (10.90) of Stokes equation (10.85) will satisfy:*

$$\left|\mathbf{q}_h^T B \mathbf{v}_h\right| \leq C_*\, \|\mathbf{v}_h\|_A\, \|\mathbf{q}_h\|_M, \tag{10.95}$$

for some $C_* > 0$ independent of h, where M denotes the mass matrix (10.94) for the pressure satisfying $\mathbf{q}_h^T M \mathbf{q}_h = \|q_h\|_{0,\Omega}^2$.

Proof. Let $\mathbf{v}_h(x)$ and $q_h(x)$ denote finite element functions with nodal vectors $\mathbf{v}_h \in \mathbb{R}^n$ and $\mathbf{q}_h \in \mathbb{R}^m$, respectively. Then, bound (10.95) will hold since:

$$\begin{aligned}
\left|\mathbf{q}_h^T B \mathbf{v}_h\right| = \left|\mathcal{B}(\mathbf{v}_h, q_h)\right| &= \left|\int_\Omega q_h(x)\; (\nabla \cdot \mathbf{v}_h(x))\; dx\right| \\
&\leq \|\nabla \cdot \mathbf{v}_h\|_{0,\Omega}\, \|q_h\|_{0,\Omega} \\
&\leq C_1\, \|\mathbf{v}_h\|_{1,\Omega}\, \|q_h\|_{0,\Omega} \\
&\leq C_*\, \|\mathbf{v}_h\|_A\, \|q_h\|_{0,\Omega},
\end{aligned}$$

for some $C_* > 0$ independent of h. □

An application of bound (10.95) and the inf-sup condition (10.93) will yield the following spectral bounds for the Schur complement matrix $S = BA^{-1}B^T$.

Lemma 10.57. *Let $\mathcal{A}(.,.)$ and $\mathcal{B}(.,.)$ satisfy the inf-sup condition (10.92) and bound (10.95). Then, there will exist $0 < \alpha_0 < \alpha_1$ independent of h, such that:*

$$\alpha_0\, \beta^2\, \left(\mathbf{q}_h^T M \mathbf{q}_h\right) \;\leq\; \mathbf{q}_h^T \left(BA^{-1}B^T\right) \mathbf{q}_h \;\leq\; \alpha_1\, C_*^2\, \left(\mathbf{q}_h^T M \mathbf{q}_h\right), \tag{10.96}$$

where M denotes the mass matrix (10.94) for the pressure space Q_h (for simplicity, we assume $Q_h \subset L^2(\Omega)/\mathbb{R}$ for Dirichlet conditions on $\partial\Omega$).

Proof. Follows immediately by an application of Lemma 10.11. □

Remark 10.58. If Dirichlet boundary conditions are imposed on $\partial\Omega$ and the pressure space satisfies $Q_h \subset L^2(\Omega)$ (without the zero mean value requirement), then matrix B^T will be singular (with $B^T\mathbf{1} = \mathbf{0}$) and bound (10.96) will be valid only for $\mathbf{q}_h \in \mathbb{R}^m$ satisfying $\mathbf{1}^T\mathbf{q}_h = 0$.

Since the mass matrix M defined by (10.94) is *well conditioned*, having diagonal entries $M_{ii} = O(h^d)$, it will follow that the Schur complement matrix $S = (BA^{-1}B^T)$ is also well conditioned (apart from the singularity $S\,\mathbf{1} = \mathbf{0}$ when Dirichlet boundary conditions are imposed).

Remark 10.59. Bounds for the *singular values* of matrix B^T can be obtained, as the square root of the eigenvalues of BB^T. To obtain such an estimate, consider $BA^{-1}B^T$ and substitute bounds for the eigenvalues of matrix A^{-1}. Since $\gamma_0\, h^d \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le \gamma_1 h^{d-2}$, for $0 < \alpha_0 < \alpha_1$ it will hold that:

$$\gamma_0^{-1}\, h^{-d} \ge \lambda_{\max}(A^{-1}) \ge \lambda_{\min}(A^{-1}) \ge \gamma_1^{-1} h^{2-d}.$$

Substituting this into $S = BA^{-1}B^T$ will yield:

$$\gamma_0^{-1}\, h^{-d}\, \mathbf{q}_h^T \left(BB^T\right) \mathbf{q}_h \ge \mathbf{q}_h^T \left(BA^{-1}B^T\right) \mathbf{q}_h \ge \gamma_1^{-1} h^{2-d}\, \mathbf{q}_h^T \left(BB^T\right) \mathbf{q}_h.$$

Since $BA^{-1}B^T$ is spectrally equivalent to the mass matrix M, where M defined by (10.94) is spectrally equivalent to $h^d\, I$, we obtain:

$$c_0\, h^{-d}\, \mathbf{q}_h^T \left(BB^T\right) \mathbf{q}_h \ge h^d I \ge c h^{2-d}\, \mathbf{q}_h^T \left(BB^T\right) \mathbf{q}_h. \tag{10.97}$$

This yields $c\, h^d \le \sigma_i(B^T) \le C\, h^{d-1}$ for some $0 < c < C$ independent of h. However, $0 = \sigma_1 < \sigma_2 = c\, h^d \le \sigma_i(B) \le C\, h^{d-1}$ if Q_h contains constants.

10.6.2 Algorithms for the Steady State Stokes Equation

In the following, we comment on the algorithms from Chaps. 10.2 to 10.5 to solve discretizations of Stokes equations with Dirichlet boundary conditions.

Uzawa type algorithms. As described in Chap. 10.2, different versions of Uzawa's algorithm correspond to different choices of preconditioners A_0 for A and S_0 for $S = BA^{-1}B^T$ in the preconditioned inexact Uzawa algorithm.

Algorithm 10.6.1 *(Preconditioned Inexact Uzawa Algorithm)*
Given $\mathbf{u}^{(0)}$, $\mathbf{p}^{(0)}$ *and* $\mathbf{g} = \mathbf{0}$

```
For $k = 0, 1, \ldots$ until convergence do:
    $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + A_0^{-1}\left(\mathbf{f} - A\mathbf{u}^{(k)} - B^T\mathbf{p}^{(k)}\right).$
    $\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + S_0^{-1}(B\mathbf{u}^{(k+1)} - \mathbf{g})$
Endfor
```

Recall that $A = \text{blockdiag}(A^{(1)}, \ldots, A^{(d)})$ is block diagonal in Stokes applications, with d diagonal blocks when $\Omega \subset \mathbb{R}^d$, where $A^{(i)}$ corresponds to a discretization of $-\nu\,\Delta$ with Dirichlet boundary conditions. As a result, we may employ any suitable preconditioner $A_0^{(l)}$ for $A^{(l)}$ (for instance domain decomposition or multigrid), and define $A_0 = \text{blockdiag}(A_0^{(1)}, \ldots, A_0^{(d)})$. If $\text{cond}(A_0^{(l)}, A^{(l)})$ is independent of h for each l, then $\text{cond}(A_0, A)$ will also be independent of h.

Since the Schur complement $S = BA^{-1}B^T$ is spectrally equivalent to $h^d I$ (when $Q_h \subset L^2(\Omega)/\mathbb{R}$), we may choose $S_0 = c\, h^d I$ for some $c > 0$. In particular, if matrices A_0 and S_0 are scaled as in Lemma 10.23:

$$\begin{aligned} (1-\gamma)\, \mathbf{q}^T S_0 \mathbf{q} \, &\leq \mathbf{q}^T S \mathbf{q} \, \leq \mathbf{q}^T S_0 \mathbf{q}, \quad && \forall \mathbf{q} \in \mathbb{R}^m \\ (1-\delta)\, \mathbf{v}^T A_0 \mathbf{v} \, &\leq \mathbf{v}^T A \mathbf{v} \, \leq \mathbf{v}^T A_0 \mathbf{v}, \quad && \forall \mathbf{v} \in \mathbb{R}^n, \end{aligned} \tag{10.98}$$

for some $0 \leq \gamma < 1$ and $0 \leq \delta < 1$, then the preconditioned inexact Uzawa algorithm will converge at a rate independent of h. If $B^T \mathbf{1} = \mathbf{0}$ and $\mathbf{1}^T \mathbf{p}^{(0)} = 0$, then all subsequent iterates $\mathbf{p}^{(k)}$ will satisfy $\mathbf{1}^T \mathbf{p}^{(k)} = 0$ provided $\mathbf{1}^T S_0^{-1} = \mathbf{0}$.

Penalty and Regularization Methods. An approximate solution $(\mathbf{u}_\epsilon, \mathbf{p}_\epsilon)$ to (10.90) can also be obtained by solving the regularized saddle point system (10.35) from Chap. 10.3. The parameter ϵ can be chosen to be the same order as the *discretization error*, with $D = S$ (or $D = h^d I$), and Uzawa's algorithm can be suitably modified to solve (10.35). However, regularization does not provide particular advantages in this application [GI3].

Projection Methods. In applications to the stationary Stokes equation, the projected gradient algorithm from Chap. 10.4 will converge slowly since matrix A is ill-conditioned with $\text{cond}(A) = O(h^{-2})$. The projected conjugate gradient method with a preconditioner $A_0 = A_0^T > 0$ should converge more rapidly for an effective choice of preconditioner, however, rigorous studies are not known. Both algorithms require applying the Euclidean projection $P_{\mathcal{K}_0} = I - B^T \left(BB^T\right)^{-1} B$. Matrix (BB^T) corresponds to a discretization of $-\Delta$ and although BB^T will be sparse, fast solvers may be available only for special geometries and grids. Parallelization may also pose a challenge.

In the following, we describe Schwarz projection algorithms to solve (10.90), see [FO, LI6, PA2, CA34, PA12, CO5]. Let $\Omega_1, \ldots, \Omega_l$ denote a decomposition of Ω into nonoverlapping subdomains of size h_0. Additionally, let $\Omega_1^*, \ldots, \Omega_l^*$ denote an associated overlapping decomposition with subdomains $\Omega_i^* \equiv \{x \in \Omega \,:\, \text{dist}(x, \Omega_i) < \beta\, h_0\}$, having overlap $\beta\, h_0$. We shall assume that the elements of the triangulation Ω_h of Ω align with the subdomains Ω_i and Ω_i^*. Given finite element spaces $\mathbf{V}_h \subset \left(H_0^1(\Omega)\right)^d$ and $Q_h \subset L^2(\Omega)$ we define *subdomain* velocity and pressure spaces $\mathbf{V}_i$ and Q_i on Ω_i^* as follows:

$$\begin{cases} \mathbf{V}_i = \mathbf{V}_h \cap \left(H_0^1(\Omega_i^*)\right)^d, & \text{for } 1 \leq i \leq p \\ \mathcal{Q}_i = Q_h \cap L^2(\Omega_i^*), & \text{for } 1 \leq i \leq p. \end{cases}$$

We shall employ the following notation for *discrete divergence free* subspaces:

$$\begin{aligned} \mathcal{K}_0^h &= \{\mathbf{v}_h \in \mathbf{V}_h \,:\, \mathcal{B}(\mathbf{v}_h, q_h) = 0, \;\; \forall q_h \in Q_h\} \\ \mathcal{K}_0^i &= \{\mathbf{v} \in \mathbf{V}_i \,:\, \mathcal{B}(\mathbf{v}, q_h) = 0, \;\; \forall q_h \in Q_i\}. \end{aligned} \tag{10.99}$$

All Schwarz projection iterates will be required to remain within the constraint set $\mathcal{K}_0^h$. For most traditional choices of finite element spaces, the local discrete divergence free space $\mathcal{K}_0^i$ will be a subset of $\mathcal{K}_0^h$. However, without additional assumptions, this property may not hold if a *coarse space* is employed.

To define a *coarse space*, let $\Omega_1, \ldots, \Omega_l$ form a *coarse triangulation* $\tau_{h_0}(\Omega)$ of Ω, and let $\mathbf{V}_0$ and Q_0 denote *coarse* velocity and pressure spaces:

$$\begin{cases} \mathbf{V}_0 = \mathbf{V}_{h_0} \cap \left(H_0^1(\Omega)\right)^d \\ \mathcal{Q}_0 = Q_{h_0} \cap L^2(\Omega) \end{cases}$$

defined on the triangulation $\tau_{h_0}(\Omega)$. If the coarse grid divergence free space $\mathcal{K}_0^0$ is a subspace of the fine grid divergence free space $\mathcal{K}_0^h$, then the coarse space can be employed. In the special case that $\mathcal{K}_0^h \subset \mathcal{K}_0$ and $\mathcal{K}_0^0 \subset \mathcal{K}_0$, where $\mathcal{K}_0$ denotes the continuous divergence free space:

$$\mathcal{K}_0 \equiv \left\{ \mathbf{v} \in \left(H_0^1(\Omega)\right)^d \,:\, \nabla \cdot \mathbf{v} = 0 \right\}, \tag{10.100}$$

then it can be verified that $\mathcal{K}_0^0 \subset \mathcal{K}_0^h$.

To implement the Schwarz algorithm, let U_i denote a matrix of size $n \times n_i$ whose columns form a basis for the space of nodal vectors associated with $\mathbf{V}_i$. Similarly, let Q_i denote a matrix of size $m \times m_i$ whose columns form a basis for the space of nodal vectors associated with $\mathcal{Q}_i$. We define A_i and B_i as:

$$A_i = U_i^T A U_i \quad \text{and} \quad B_i = Q_i^T B U_i, \tag{10.101}$$

of size $n_i \times n_i$ and $m_i \times n_i$ respectively. By construction, it will hold that:

$$B_i \mathbf{w}_i = \mathbf{0} \implies U_i \mathbf{w}_i \subset \mathcal{K}_0^h, \quad \text{for } 0 \le i \le l. \tag{10.102}$$

We assume that $B^T \mathbf{1} = \mathbf{0}$ and that $B_i^T \mathbf{1}_i = \mathbf{0}$ for some $\mathbf{1}_i \in \mathbb{R}^{m_i}$. To avoid discontinuous pressures across Ω_i^*, we shall employ a restriction matrix R_i of size $m_i \times m_i$, such that $(R_i \boldsymbol{\mu}_i)_j = (\boldsymbol{\mu}_i)_j$ when node j is in Ω_i^* and $(R_i \boldsymbol{\mu}_i)_j = 0$ when node j is on $\partial\Omega_i^*$. We define $R_0 = 0$. We list the multiplicative Schwarz algorithm to solve the Dirichlet problem (10.90).

Algorithm 10.6.2 *(Multiplicative Schwarz Algorithm to Solve (10.90))*
Let $\mathbf{v}^{(0)} = \mathbf{0}$ *and* $\mathbf{p}^{(0)} = \mathbf{0}$ *denote starting iterates*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $i = 0, 1, \ldots, l$ do:
           $$\begin{bmatrix} \mathbf{w}_i^{(k)} \\ \boldsymbol{\mu}_i^{(k)} \end{bmatrix} = \begin{bmatrix} U_i & 0 \\ 0 & Q_i R_i \end{bmatrix} \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T (\mathbf{f} - A \mathbf{v}^{(k+\frac{i}{l+1})} - B^T \mathbf{p}^{(k+\frac{i}{l+1})}) \\ \mathbf{0} \end{bmatrix}$$
           Define $\gamma_i = \left( \mathbf{1}^T (\mathbf{p}^{(k+\frac{i}{l+1})} + \boldsymbol{\mu}_i^{(k)}) / \mathbf{1}^T Q_i R_i \mathbf{1}_i \right)$ and update:
           $$\begin{bmatrix} \mathbf{v}^{(k+\frac{i+1}{l+1})} \\ \mathbf{p}^{(k+\frac{i+1}{l+1})} \end{bmatrix} = \begin{bmatrix} \mathbf{v}^{(k+\frac{i}{l+1})} \\ \mathbf{p}^{(k+\frac{i}{l+1})} \end{bmatrix} + \begin{bmatrix} \mathbf{w}_i^{(k)} \\ \boldsymbol{\mu}_i^{(k)} - \gamma_i Q_i R_i \mathbf{1}_i \end{bmatrix}$$
3.     Endfor
4. Endfor
```

Output: $\mathbf{v}^{(k)}$ *and* $\mathbf{p}^{(k)}$

Remark 10.60. Several modifications have been introduced to obtain the above algorithm from the multiplicative Schwarz algorithm listed Chap. 10.4. First, since $\mathbf{g} = \mathbf{0}$, we have omitted step 1, setting $\mathbf{u}_* = \mathbf{0}$. Next, steps 2 and 3 to determine $\mathbf{w} \in \mathcal{K}_0^h$ and $\boldsymbol{\mu} = \mathbf{p}$ have been combined by computing the local pressures $\boldsymbol{\mu}_i^{(k)}$ to determine $\mathbf{p}$ as indicated above. By construction, all velocity iterates will be discrete divergence free, i.e., $\mathbf{v}^{(k+\frac{i}{l+1})} \in \mathcal{K}_0^h$, while the pressure iterates $\mathbf{p}^{(k+\frac{i+1}{l+1})}$ will have mean value zero. For sufficiently large overlap, the iterates $\mathbf{v}^{(k)}$ in the above algorithm will converge geometrically to the discrete solution $\mathbf{u}_h$ at a rate independent h, see [LI6, PA2, CA34]. The pressure $\mathbf{p}^{(k)}$ will similarly converge geometrically to $\mathbf{p}_h$. If a coarse space is included, then this rate of convergence is expected to be robust as $h_0 \to 0$.

Remark 10.61. When $i \neq 0$, each fractional iteration in the multiplicative Schwarz algorithm above corresponds to a discrete version of the iteration:

$$\begin{cases} -\nu\,\Delta \mathbf{u}^{(k+\dfrac{i+1}{l+1})} + \nabla p^{(k+\dfrac{i+1}{l+1})} = \mathbf{f}, & \text{in } \Omega_i^* \\ -\nabla \cdot \mathbf{u}^{(k+\dfrac{i+1}{l+1})} = 0, & \text{in } \Omega_i^* \\ \mathbf{u}^{(k+\dfrac{i+1}{l+1})} = \mathbf{u}^{(k+\dfrac{i}{l+1})}, & \text{on } \partial\Omega_i^*, \end{cases}$$

for $1 \leq i \leq l$ and $k = 0, 1, \ldots$. Here, the velocity and pressure are defined as $\mathbf{u}^{(k+\frac{i+1}{l+1})} = \mathbf{u}^{(k+\frac{i}{l+1})}$ and $p^{(k+\frac{i+1}{l+1})} = p^{(k+\frac{i}{l+1})}$ on $\Omega \setminus \Omega_i^*$. The pressure is also normalized to have mean value zero on Ω. The above can be derived from the discrete version by expressing the equations satisfied by the updated velocity and pressure in a weak form. The multiplicative Schwarz algorithm is sequential, but can be parallelized by *multicoloring* the subdomains.

Remark 10.62. The additive Schwarz algorithm from Chap. 10.4 may also be employed to solve (10.90). It is highly parallel:

- We can skip step 1, since $\mathbf{g} = \mathbf{0}$, setting $\mathbf{u}_* = \mathbf{0}$.
- In step 2, the velocity vector $\mathbf{w} = \mathbf{u}_h$ can be determined by solving the equation $P^A \mathbf{w} = \mathbf{r}$ using a CG algorithm with the A-inner product.
- In step 3, once the velocity $\mathbf{u}_h$ has been determined, subdomain pressures $\boldsymbol{\mu}_i$ can be computed for $i = 1, \ldots, l$ as follows:

$$\boldsymbol{\mu}_i = Q_i \begin{bmatrix} 0 \\ I \end{bmatrix}^T \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T(\mathbf{f} - A\mathbf{u}_h) \\ \mathbf{0} \end{bmatrix}.$$

  A global pressure $\mathbf{p}_h$ can be constructed from the local pressures $\boldsymbol{\mu}_i$ such that $\mathbf{p}_h = \boldsymbol{\mu}_i + \gamma_i\, Q_i \mathbf{1}_i$ on each subdomain Ω_i^*. The parameters γ_i can be determined sequentially so that the subdomain pressures have the same mean value as the pressures on adjacent subdomains on the regions of overlap, and so that the global pressure has mean value zero [MA31].

The additive Schwarz is heuristically expected to converge independent of h. If a coarse space is included, then its rate of convergence is expected to remain robust as $h_0 \to 0$. Theoretical bounds are not known.

Block Matrix Methods. The block matrix methods described in Chap. 10.5 are very effective for the iterative solution of (10.90).

- Since $S = BA^{-1}B^T$ is well conditioned for the stationary Stokes equation, the *Schur complement* system can be solved using the CG method without preconditioning. Computing the action of S requires the action of A^{-1}, and an inner iteration may be employed with preconditioner A_0. However, an inexact Uzawa or block matrix preconditioned Krylov space algorithm will be preferable [VE, QU7, BA18, EL5, BR16].
- The *positive definite reformulation* of [BR8] can also be employed to solve (10.90) using CG acceleration. Any *suitably scaled* preconditioner A_0 for A can be employed, satisfying (10.61). Since S is well conditioned, the resulting algorithm will converge at a rate independent of h.
- *Block diagonal* preconditioning is effective for solving stationary Stokes with *MINRES* acceleration [EL8, EL2, KL2, EL3, SI, ZU]. *Block triangular* preconditioning can be even more effective, but *GMRES* acceleration is required [EL4, EL9, ZU]. If preconditioner A_0 is spectrally equivalent to A and $S_0 = c\,h^d I$ is the preconditioner for S, then the resulting rate of convergence will be independent of h. For block triangular preconditioners, the matrices A_0 and S_0 must be suitably scaled as indicated in (10.98).
- *Saddle point preconditioners* do not offer particular advantages in applications to steady state Stokes, since matrix A is ill-conditioned.

If $A_0 \asymp A$ and $S_0 \asymp S$, the rate of convergence will be independent of h.

10.6.3 The Stokes-Oseen Problem

The Stokes-Oseen problem (10.86) arises when a solution to the stationary Navier-Stokes equations is sought based on the linearization $(\mathbf{u}\cdot\nabla)\mathbf{w}+(\mathbf{w}\cdot\nabla)\mathbf{u}$ of the nonlinear term $(\mathbf{u}\cdot\nabla)\,\mathbf{u}$ about a velocity field $\mathbf{w}(x)$. We shall further simplify the zeroth order term $(\mathbf{u}\cdot\nabla)\mathbf{w}$ by shifting it to the right hand side in a Picard iteration. A discretization of (10.86) without the $(\mathbf{u}\cdot\nabla)\mathbf{w}$ term will yield the following *nonsymmetric* linear system:

$$\begin{bmatrix} \nu\,A + N & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_h \\ \mathbf{p}_h \end{bmatrix} = \begin{bmatrix} \mathbf{f}_h \\ \mathbf{0} \end{bmatrix}, \tag{10.103}$$

where matrices B and B^T are as in the discretization of Stokes equation, $A\mathbf{u}_h$ and $N\mathbf{u}_h$ correspond to discretizations of the diffusion term $-\Delta\mathbf{u}$ and advection term $(\mathbf{w}\cdot\nabla)\,\mathbf{u}$, respectively. By construction, matrices A and N will be *block diagonal* with d identical diagonal blocks when $\Omega \subset \mathbb{R}^d$. We shall assume that matrix N is *skew-symmetric*, i.e., $N^T = -N$, satisfying:

$$\|A^{-1/2}NA^{-1/2}\| \le \gamma, \tag{10.104}$$

for some $\gamma > 0$ independent of h. When the parameter ν is small, a streamline diffusion or upwind scheme may need to be employed in constructing N, to ensure stability of the discretization.

Remark 10.63. Since $A^{-1/2}NA^{-1/2}$ is *unitarily diagonalizable* (because N is skew-Hermitian and A is Hermitian), to show that $\|A^{-1/2}NA^{-1/2}\| \leq \gamma$, it will be sufficient to show that:

$$(\mathbf{v}^H N\mathbf{v}) \leq \gamma\, (\mathbf{v}^H A\mathbf{v}), \quad \forall \mathbf{v} \in \mathbb{C}^n \tag{10.105}$$

for some $\gamma > 0$ independent of h. Such a bound will hold since matrix N corresponds to the discretization of a first order differential operator, while A corresponds to the discrete Laplacian.

The Oseen problem does not admit a saddle point interpretation due to non-self adjointness of N. As a result, saddle point iterative algorithms will not be applicable to solve (10.103). However, block matrix preconditioned algorithms accelerated by Krylov space methods can be employed, and we shall consider preconditioners having the block matrix form [EL8]:

$$M_1 \equiv \begin{bmatrix} \nu\, A + N & 0 \\ 0 & \frac{1}{\nu} C \end{bmatrix} \quad \text{and} \quad M_2 \equiv \begin{bmatrix} \nu\, A + N & B^T \\ 0 & -\frac{1}{\nu} C \end{bmatrix}, \tag{10.106}$$

where C is any symmetric positive definite preconditioner for $\left(BA^{-1}B^T\right)$. The following result expresses the eigenvalues of such preconditioned systems in terms of the eigenvalues of $(1/\nu)\, C^{-1}S$, where $S = B\,(\nu\, A + N)^{-1}\, B^T$. In practice, matrix $\nu\, A + N$ can be replaced by a suitable preconditioner.

Lemma 10.64. *Suppose the following conditions hold.*

1. *Let L denote the following nonsymmetric matrix:*

$$L \equiv \begin{bmatrix} \nu\, A + N & B^T \\ B & 0 \end{bmatrix}.$$

2. *Let M_1 and M_2 be the block matrix preconditioners defined in (10.106).*
3. *Let $\mu_i \in \mathbb{C}$ denote eigenvalues of $\nu\, C^{-1}S$ for $i = 1, \ldots, m$, where the Schur complement matrix $S = B\,(\nu\, A + N)^{-1}\, B^T$ is nonsymmetric.*

Then, the following results will hold:

1. *If λ is an eigenvalue of $M_1^{-1}L$, it will satisfy:*

$$\lambda \in \{1\} \cup \left(\cup_{i=1}^{m} \left\{ \frac{1 \pm \sqrt{1 + 4\mu_i}}{2} \right\} \right).$$

2. *If λ is an eigenvalue of $M_2^{-1}L$, it will satisfy:*

$$\lambda \in \{1\} \cup \left(\cup_{i=1}^{m} \{\mu_i\} \right).$$

Proof. See [EL8]. Identical to the proof given in Chap. 10.5. □

The next result applies Bendixson's lemma 8.3 to estimate the eigenvalues of $\nu\, C^{-1} S$ when $S = B\,(\nu\, A + N)^{-1}\, B^T$ and $C = \left(BA^{-1}B^T\right)$. We let F^H denote the complex conjugate transpose of F, i.e., $F^H = \overline{F}^T$.

Lemma 10.65. *Suppose the following conditions hold.*

1. *Let $S = B\,(\nu\, A + N)^{-1}\, B^T$ where $A = A^H > 0$ is real and $N = -N^T$ is real skew-symmetric, with $\tilde{N} \equiv A^{-1/2} N A^{-1/2}$ satisfying:*

$$\|\tilde{N}\| \leq \gamma,$$

   for some $\gamma > 0$ independent of h.
2. *Let $C = \left(BA^{-1}B^T\right)$ be symmetric positive definite*

Then, the eigenvalues of $\nu\, C^{-1} S$ will lie in a rectangular subregion of the complex plane whose size is independent of h.

Proof. We follow the proof in [EL8] and estimate the real parameters $\gamma_1 < \gamma_2$ and $\delta_1 < \delta_2$, see (8.14), employed in Bendixson's lemma 8.3. Consider first the Hermitian part D of $S = BK^{-1}B^T$, where $K \equiv \nu\, A + N$:

$$\begin{aligned} D = \tfrac{1}{2}\left(S + S^H\right) &= \tfrac{1}{2}B\left(K^{-1} + K^{-H}\right)B^H \\ &= \tfrac{1}{2}BK^{-1}\left(K + K^H\right)K^{-H}B^H \\ &= BK^{-1}\,(\nu A)\,K^{-H}B^H. \end{aligned}$$

Here we have used that $A^H = A$ and that $N^H + N = 0$. Note that:

$$K^{-1} = (\nu A + N)^{-1} = A^{-1/2}\left(\nu\, I + \tilde{N}\right)^{-1} A^{-1/2},$$

where $\tilde{N} \equiv A^{-1/2} N A^{-1/2}$. Substituting this into the expression for D yields:

$$\begin{aligned} D &= \nu\, BA^{-1/2}\left(\nu\, I + \tilde{N}\right)^{-1}\left(\nu\, I + \tilde{N}\right)^{-T} A^{-1/2}B^T \\ &= \nu\, BA^{-1/2}\left(\nu^2\, I + \tilde{N}^T\tilde{N}\right)^{-1} A^{-1/2}B^T \\ &= BA^{-1/2}\left(\nu\, I + \nu^{-1}\,\tilde{N}^T\tilde{N}\right)^{-1} A^{-1/2}B^T. \end{aligned}$$

Consider the Rayleigh quotient $\nu\left(\mathbf{z}^H D\mathbf{z}/\mathbf{z}^H C\mathbf{z}\right)$ and substitute the preceding:

$$\begin{aligned} \nu\,\frac{\mathbf{z}^H D\mathbf{z}}{\mathbf{z}^H C\mathbf{z}} &= \nu\,\frac{\mathbf{z}^H BA^{-1/2}\left(\nu\, I + \nu^{-1}\,\tilde{N}^T\tilde{N}\right)^{-1} A^{-1/2}B^T\mathbf{z}}{\mathbf{z}^H\left(BA^{-1}B^T\right)\mathbf{z}} \\ &= \frac{\mathbf{z}^H BA^{-1/2}\left(I + \nu^{-2}\,\tilde{N}^T\tilde{N}\right)^{-1} A^{-1/2}B^T\mathbf{z}}{\mathbf{z}^H BA^{-1}B^T\mathbf{z}}. \end{aligned}$$

Substituting $\mathbf{w} = A^{-1/2} B^T \mathbf{z}$ and employing $\tilde{N} = -\tilde{N}^H$ with $\|\tilde{N}\| \le \gamma$ yields:

$$1 \le \frac{\mathbf{w}^H \left(I + \nu^{-2} \, \tilde{N}^T \tilde{N} \right) \mathbf{w}}{\mathbf{w}^H \mathbf{w}} \le 1 + (\gamma^2/\nu^2), \qquad \text{for} \quad \mathbf{w} \in \mathbb{C}^m \setminus \mathbf{0}.$$

This yields $\gamma_1 = 1$ and $\gamma_2 = (\nu^2 + \gamma^2)/\nu^2$:

$$1 \, \le \, \nu \, \frac{\mathbf{z}^H D \mathbf{z}}{\mathbf{z}^H C \mathbf{z}} \, \le \, \frac{\nu^2 + \gamma^2}{\nu^2}, \text{ for } \mathbf{z} \in \mathbb{C}^m \setminus \mathbf{0}.$$

We next consider the skew-Hermitian part E and obtain:

$$\begin{aligned} E &= \frac{1}{2} B \left(K^{-1} - K^{-H} \right) B^H \\ &= \frac{1}{2} B K^{-1} \left(K^H - K \right) K^{-H} B^H \\ &= -B K^{-1} N K^{-H} B^{-H} \\ &= -B \left(\nu \, A + N \right)^{-1} N \left(\nu \, A + N \right)^{-H} B^{-H} \\ &= - \, B A^{-1/2} \left(\nu \, I + \tilde{N} \right)^{-1} \tilde{N} \left(\nu \, I + \tilde{N} \right)^{-H} A^{-1/2} B^{-H}. \end{aligned}$$

Thus, the Rayleigh quotient:

$$\begin{aligned} \left(\frac{\nu}{i} \right) \left(\frac{\mathbf{z}^H E \mathbf{z}}{\mathbf{z}^H C \mathbf{z}} \right) &= \nu \, i \left(\frac{\mathbf{z}^H B A^{-1/2} \left(\nu \, I + \tilde{N} \right)^{-1} \tilde{N} \left(\nu \, I + \tilde{N} \right)^{-H} A^{-1/2} B^{-H} \mathbf{z}}{\mathbf{z}^H \left(B A^{-1} B^H \right) \mathbf{z}} \right) \\ &= \nu \, i \left(\frac{\mathbf{w}^H \tilde{N} \mathbf{w}}{\mathbf{w}^H \left(\nu \, I + \tilde{N} \right) \left(\nu \, I + \tilde{N} \right)^H \mathbf{w}} \right) \\ &= \nu \, i \left(\frac{\mathbf{w}^H \tilde{N} \mathbf{w}}{\mathbf{w}^H \left(\nu^2 \, I + \tilde{N} \tilde{N}^H \right) \mathbf{w}} \right), \end{aligned}$$

where $\mathbf{w} \equiv \left(\nu \, I + \tilde{N} \right)^{-H} A^{-1/2} B^{-H} \mathbf{z}$. Since $\tilde{N}$ is skew-Hermitian, it will be unitarily diagonalizable with $\tilde{N} = i \, U \Lambda U^H$ for a real diagonal matrix Λ and a unitary matrix U. Substituting this in the preceding yields:

$$\begin{aligned} \left(\frac{\nu}{i} \right) \left(\frac{\mathbf{z}^H E \mathbf{z}}{\mathbf{z}^H C \mathbf{z}} \right) &= \nu \, i \left(\frac{\mathbf{w}^H \tilde{N} \mathbf{w}}{\mathbf{w}^H \left(\nu^2 \, I + \tilde{N} \tilde{N}^H \right) \mathbf{w}} \right) \\ &= -\nu \left(\frac{\mathbf{w}^H U \Lambda U^H \mathbf{w}}{\mathbf{w}^H \left(\nu^2 \, I + \, U \Lambda^2 U^H \right) \mathbf{w}} \right) = -\nu \left(\frac{\mathbf{v}^H \Lambda \mathbf{v}}{\mathbf{v}^H \left(\nu^2 \, I + \Lambda^2 \right) \mathbf{v}} \right), \end{aligned}$$

where $\mathbf{v} = U^H \mathbf{w}$. Since $\|\tilde{N}\| = \|\Lambda\| \le \gamma$, we obtain the bound:

$$\left| \frac{\mathbf{z}^H E \mathbf{z}}{\mathbf{z}^H C \mathbf{z}} \right| = \nu \left| \frac{\mathbf{v}^H \Lambda \mathbf{v}}{\mathbf{v}^H \left(\nu^2 \, I + \Lambda^2 \right) \mathbf{v}} \right| \le \max_{\lambda \in [0, \gamma]} \frac{\nu \, \lambda}{\nu^2 + \lambda^2} \le \frac{1}{2}.$$

This yields the bound $-\delta_1 = \delta_2 = \frac{1}{2}$. □

The next result considers the eigenvalues of $L\,M_i^{-1}$ when matrix $C = h^d I$.

Lemma 10.66. *Suppose the assumptions of the preceding lemma holds. Then, for the choice $C = h^d I$, the eigenvalues of $\nu\, C^{-1} S$ will lie in a rectangular subregion of the complex plane whose size is independent of h.*

Proof. Follows immediately from the observation that:

$$\nu \, \frac{\mathbf{z}^H S \mathbf{z}}{\mathbf{z}^H h^d I \mathbf{z}} = \nu \left(\frac{\mathbf{z}^H S \mathbf{z}}{\mathbf{z}^H B A^{-1} B^T \mathbf{z}} \right) \left(\frac{\mathbf{z}^H B A^{-1} B^T \mathbf{z}}{\mathbf{z}^H h^d I \mathbf{z}} \right),$$

and that $BA^{-1}B^T$ is spectrally equivalent to $h^d I$ independent of h. □

10.6.4 Time Dependent Stokes and Navier-Stokes Equation

We shall now describe algorithms to solve an implicit discretization of the following linearization of the Navier-Stokes equations:

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \nu\, \Delta \mathbf{u} + (\mathbf{w} \cdot \nabla)\, \mathbf{u} + (\mathbf{u} \cdot \nabla)\, \mathbf{w} + \nabla p = \mathbf{f}, \;\text{ in } \Omega \times (0,t) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \nabla \cdot \mathbf{u} = 0, \;\text{ in } \Omega \times (0,t), \end{cases} \tag{10.107}$$

with initial data $\mathbf{u}(x,0) = \mathbf{u}_0(x)$ and $\mathbf{u}(\cdot,t) = \mathbf{0}$ on $\partial\Omega$. Here $\mathbf{w}(x,t)$ denotes a known velocity field about which the Navier-Stokes equations was linearized, with a modified forcing term $\mathbf{f}(.)$, initial data $\mathbf{u}_0(x)$ and boundary data $\mathbf{u}(\cdot,t)$.

To simplify our discussion, we consider a *backward Euler* discretization in time and a finite element or finite difference spatial discretization of (10.107). Let $\tau > 0$ denote a time step and let $\mathbf{u}_h^{(k)}$, $\mathbf{p}_h^{(k)}$ and $\mathbf{f}_h^{(k)}$ denote the discrete velocity, pressure and forcing term at time $k\,\tau$. Then, the backward Euler discretization in time with a finite element or finite difference spatial discretization will yield the following linear system at each time $k\tau$:

$$\begin{bmatrix} M + \tau\,(\nu\, A + N) & \tau\, B^T \\ \tau\, B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_h^{(k)} \\ \mathbf{p}_h^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_h^{(k-1)} + \tau \mathbf{f}_h^{(k)} \\ \mathbf{0} \end{bmatrix}, \tag{10.108}$$

where $A\mathbf{u}_h$ denotes a discretization of $-\Delta \mathbf{u}$ with $N\mathbf{u}_h$ a discretization of $(\mathbf{w} \cdot \nabla)\, \mathbf{u}$,. Here M denotes the mass (Gram) matrix, which will be block diagonal with d diagonal blocks, $B^T \mathbf{p}_h$ corresponds to a discretization of ∇p and $B\mathbf{u}_h$ to a discretization of $-\nabla \cdot \mathbf{u}$. If a finite difference spatial discretization is employed, then $M = I$. For convenience, we have multiplied the discrete divergence constraint $B\mathbf{u}_h^{(k)} = \mathbf{0}$ by the time step τ in (10.108).

Time Dependent Stokes. The time dependent Stokes equation arises when $\mathbf{w} = \mathbf{0}$ in (10.107). In this case matrix $N = 0$ in (10.108):

$$\begin{bmatrix} M + \tau\, \nu\, A & \tau\, B^T \\ \tau\, B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_h^{(k)} \\ \mathbf{p}_h^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_h^{(k-1)} + \tau \mathbf{f}_h^{(k)} \\ \mathbf{0} \end{bmatrix}, \tag{10.109}$$

and system (10.108) will be symmetric indefinite. Saddle point algorithms from Chap. 10.2 to Chap. 10.5 can be employed, however, the preconditioners employed for $(M + \tau\, \nu\, A)$ and the Schur complement $\tau^2\, B(M + \tau\, \nu\, A)^{-1}B^T$ must be modified to ensure a rate of convergence independent of h and τ. Matrix $(M + \tau\, \nu\, A)$ in (10.108) is better conditioned than matrix $\nu\, A$ in the stationary case, and various effective preconditioners can be formulated for it. However, the Schur complement matrix associated with (10.109) for the pressure variables will have the form $S(\tau) = \tau^2\, B\, (M + \tau\, \nu\, A)^{-1}\, B^T$ and unlike the stationary case, this matrix can be *ill-conditioned* for $\tau \ll 1$.

The convergence rate of *Uzawa* and *block matrix* algorithms depend on the choice preconditioners for $(M + \tau\, \nu\, A)$ and for $S(\tau)$, and we shall now indicate preconditioners for them. Matrix $(M + \tau\, \nu\, A)$ will be block diagonal:

$$(M + \tau\, \nu\, A) \,=\, \text{blockdiag}\left(M^{(1)} + \tau\, \nu\, A^{(1)}, \ldots, M^{(d)} + \tau\, \nu\, A^{(d)}\right),$$

where each $(M^{(l)} + \tau\, \nu\, A^{(l)})$ corresponds to a discretization of the operator $(I - \tau\, \nu\, \Delta)$ with Dirichlet boundary conditions. Effective domain decomposition preconditioners have been described for such matrices in chapter 9. Importantly, if $\tau \,\le\, c\, h_0^2$ (where h_0 denotes the size of subdomains), then coarse space correction can typically be omitted in domain decomposition preconditioners, without deterioration in the rate of convergence.

As $\tau \to 0^+$ and as $\tau \to \infty$ we will obtain: $(S(\tau)/\tau^2)$ will satisfy:

$$B\, (M + \tau\, \nu\, A)^{-1}\, B^T \to \begin{cases} \left(BM^{-1}B^T\right), & \text{as } \tau \to 0 \\ \tau^{-1}\, \nu^{-1}\, \left(BA^{-1}B^T\right), & \text{as } \tau \to \infty. \end{cases}$$

Here, $\text{cond}\left(BA^{-1}B^T\right) \le c$, while $\text{cond}\left(BM^{-1}B^T\right) \le c\, h^{-2}$ where $c > 0$ is independent of h, see (10.97). As a result, care must be exercised in preconditioning $S(\tau)$. Below, we describe a heuristic preconditioner $S_0(\tau)$ of [BR9] for $S(\tau)$, motivated by the following observation:

$$\left(B\, (M + \tau\, \nu\, A)^{-1}\, B^T\right)^{-1} \to \begin{cases} \left(BM^{-1}B^T\right)^{-1}, & \text{as } \tau \to 0 \\ \tau^{-1}\, \nu^{-1}\, \left(BA^{-1}B^T\right)^{-1}, & \text{as } \tau \to \infty. \end{cases}$$

Motivated by this, the inverse $S_0(\tau)^{-1}$ of the preconditioner is defined as:

$$S_0(\tau)^{-1} = \left(BM^{-1}B^T\right)^{-1} + \tau\, \nu\, \left(BA^{-1}B^T\right)^{-1}.$$

It is shown in [BR9] that $\text{cond}\,(S_0(\tau), S(\tau)) \,\le\, c$ for $\tau \in [h^2, 1]$, where $c > 0$ is independent of h. Since $B\mathbf{u}_h$ corresponds to a discretization of $-\nabla \cdot \mathbf{u}$, and $B^T\mathbf{p}_h$ to a discretization of ∇p, matrix $\left(BM^{-1}B^T\right)\mathbf{p}_h$ formally corresponds to a mixed finite element discretization of $-\Delta\, p$ on Ω, with Neumann boundary conditions on $\partial\Omega$, see Chap. 10.7. Such discretizations are obtained using mixed formulations of elliptic equations.

Remark 10.67. Since $\left(BA^{-1}B^T\right)$ is spectrally equivalent to the mass matrix $M \asymp c\, h^d I$, we may approximate $\left(BA^{-1}B^T\right)^{-1}$ by $ch^{-d}I$ for some $c > 0$.

Similarly, we may approximate $\left(BM^{-1}B^T\right) \asymp h^{-d}\left(BB^T\right)$. As indicated in Chap. 10.7, matrix $\left(BM^{-1}B^T\right)$ will correspond to a discretization of $-\Delta$ with Neumann boundary conditions for the pressure, using mixed finite element methods. Since B is sparse, matrix BB^T may be assembled explicitly, and sparse direct, multigrid or domain decomposition solvers can be employed.

Chorin's Projection Method. Chorin's projection method [CH27, CH28], is a *noniterative* algorithm for obtaining an *approximate* solution to (10.109), by *decoupling* the computation of the velocity and pressure unknowns $\mathbf{u}_h^{(k)}$ and $\mathbf{p}_h^{(k)}$ at each time step. The algorithm can be motivated by considering a *finite difference* discretization of the time dependent Stokes equation, with $M = I$ in (10.109). From a matrix viewpoint, Chorin's projection method solves (10.109) approximately, by *modifying* the linear system by replacing the term $\tau B^T \mathbf{p}_h^{(k)}$ in the first block row by the term $\tau(I + \tau\,\nu\,A)B^T\mathbf{p}_h^{(k)}$. This modification formally introduces an additional term $\tau^2\,\nu\,AB^T\mathbf{p}_h^{(k)}$ in the discretization error. To distinguish between the original and modified schemes, we shall denote the solution to the modified scheme as $\tilde{\mathbf{u}}_h^{(k)}$ and $\tilde{\mathbf{p}}_h^{(k)}$:

$$\begin{bmatrix} (I + \tau\,\nu\,A) & \tau\,(I + \tau\,\nu\,A)\,B^T \\ \tau\,B & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}_h^{(k)} \\ \tilde{\mathbf{p}}_h^{(k)} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{u}}_h^{(k-1)} + \tau\mathbf{f}_h^{(k)} \\ \mathbf{0} \end{bmatrix}. \tag{10.110}$$

The above modified system enables decoupling the velocity and pressures. Indeed, since $(I + \tau\,\nu\,A)$ multiplies both terms in the first block equation of (10.110), multiplying this block row of (10.110) by $(I + \tau\,\nu\,A)^{-1}$ yields:

$$\tilde{\mathbf{u}}_h^{(k)} + \tau B^T\tilde{\mathbf{p}}_h^{(k)} = (I + \tau\,\nu\,A)^{-1}\left(\tilde{\mathbf{u}}_h^{(k-1)} + \tau\mathbf{f}_h^{(k)}\right). \tag{10.111}$$

Applying matrix B to the above equation, and using that $B\tilde{\mathbf{u}}_h^{(k)} = \mathbf{0}$ yields:

$$\tau\,\left(BB^T\right)\tilde{\mathbf{p}}_h^{(k)} = B\,(I + \tau\,\nu\,A)^{-1}\left(\tilde{\mathbf{u}}_h^{(k-1)} + \tau\mathbf{f}_h^{(k)}\right).$$

Solving this system yields the pressure $\tilde{\mathbf{p}}_h^{(k)}$. Here, as in preceding discussions, matrix $\left(BB^T\right)\tilde{\mathbf{p}}_h$ corresponds to a mixed formulation discretization of $-\Delta\tilde{p}$ with Neumann boundary conditions.

To compute $\tilde{\mathbf{u}}_h^{(k)}$ note that $\tilde{\mathbf{u}}_h^{(k)} \in \mathcal{K}_0^h = \text{Kernel}(B)$ the space of discrete divergence free velocities. The fundamental theorem of linear algebra yields $\text{Range}(B^T)^\perp = \text{Kernel}(B)$, so applying the Euclidean orthogonal projection $P_{\mathcal{K}_0^h} = I - B^T(BB^T)^{-1}B$ (onto the subspace $\mathcal{K}_0^h$) to (10.111) we obtain:

$$\tilde{\mathbf{u}}_h^{(k)} = P_{\mathcal{K}_0^h}\left(\tilde{\mathbf{u}}_h^{(k)} + \tau B^T\tilde{\mathbf{p}}_h^{(k)}\right) = P_{\mathcal{K}_0^h}\,(I + \tau\,\nu\,A)^{-1}\left(\tilde{\mathbf{u}}_h^{(k-1)} + \tau\mathbf{f}_h^{(k)}\right).$$

This computation also requires the solution of a linear system with coefficient matrix (BB^T), corresponding to a discretization of $-\Delta$ with Neumann boundary conditions. The pressure and velocity $\tilde{\mathbf{p}}_h^{(k)}$ and $\tilde{\mathbf{u}}_h^{(k)}$ can thus be updated in parallel. We summarize the resulting algorithm.

Algorithm 10.6.3 *(Chorin's Projection Method to Solve (10.109))*
Given $\tilde{\mathbf{u}}_h^{(k-1)}$, $\tilde{\mathbf{p}}_h^{(k-1)}$*:*

1. *Update the velocity:*

$$\tilde{\mathbf{w}}_h^{(k)} = (I + \tau\,\nu\,A)^{-1}\left(\tilde{\mathbf{u}}_h^{(k-1)} + \tau \mathbf{f}_h^{(k)}\right)$$

2. *In parallel compute:*

$$\begin{aligned} \tilde{\mathbf{p}}_h^{(k)} &= \tau^{-1}\left(BB^T\right)^{-1} B\tilde{\mathbf{w}}_h^{(k)} \\ \tilde{\mathbf{u}}_h^{(k)} &= P_{\mathcal{K}_0^h}\tilde{\mathbf{w}}_h^{(k)} \end{aligned}$$

Output: $\tilde{\mathbf{u}}_h^{(k)}$, $\tilde{\mathbf{p}}_h^{(k)}$

Remark 10.68. Chorin's scheme is stable in the Euclidean norm $\|\cdot\|$:

$$\begin{aligned} \|\tilde{\mathbf{u}}_h^{(k)}\| = \|P_{\mathcal{K}_0^h}\tilde{\mathbf{w}}_h^{(k)}\| \le \|\tilde{\mathbf{w}}_h^{(k)}\| &= \|\,(I + \tau\nu A)^{-1}\left(\tilde{\mathbf{u}}_h^{(k-1)} + \tau\,\mathbf{f}_h^{(k)}\right)\| \\ &\le \|\tilde{\mathbf{u}}_h^{(k-1)} + \tau\,\mathbf{f}_h^{(k)}\|, \end{aligned}$$

since $\|P_{\mathcal{K}_0^h}\| \le 1$ and since $\|(I + \tau\nu A)^{-1}\| \le 1$ when $A = A^T > 0$. Furthermore, since Chorin's projection scheme introduces the additional term $\tau^2 AB^T\tilde{\mathbf{p}}_h^{(k)}$ into the discretization, the local truncation error will formally be $O(\tau^2)$ if the true solution is sufficiently smooth. As a result, Chorin's projection scheme will be globally 1st order accurate in τ.

Remark 10.69. Chorin's projection method can also be motivated by the *Hodge decomposition* theorem [TE]. A special case of this result states that a sufficiently smooth velocity field $\mathbf{u}(x)$ on Ω with *zero* flux $\mathbf{n}\cdot\mathbf{u}$ on $\partial\Omega$, can be orthogonally decomposed within $\left(L^2(\Omega)\right)^d$ as sum of a *divergence free* velocity field $\mathbf{v}(x)$ and the gradient of scalar potential:

$$\mathbf{u}(x) = \mathbf{v}(x) + \nabla p, \qquad \text{where } \nabla\cdot\mathbf{v}(x) = 0.$$

This decomposition is seen to be orthogonal in the $\left(L^2(\Omega)\right)^d$ inner product since $\int_\Omega \mathbf{v}(x)\cdot\nabla p(x)\,dx = -\int_\Omega p(x)\,(\nabla\cdot\mathbf{v}(x))\,dx = 0$ using integration by parts when $\mathbf{n}\cdot\mathbf{v} = 0$ on $\partial\Omega$. This orthogonal decomposition enables elimination of the pressure term in the Navier-Stokes equation [TE] by an application of the $\left(L^2(\Omega)\right)^d$-orthogonal projection map $P_{\mathcal{K}_0}$ onto the space $\mathcal{K}_0$ of divergence free functions:

$$\begin{cases} \dfrac{\partial\mathbf{u}}{\partial t} = P_{\mathcal{K}_0}\left(\mathbf{f} + \nu\Delta\mathbf{u} - (\mathbf{u}\cdot\nabla)\mathbf{u}\right) \\ \mathbf{u}(0, x) = \mathbf{u}_0(x). \end{cases}$$

Chorin's projection scheme computes an update $\mathbf{w}_h^{(k)}$ at time $k\,\tau$ omitting the projection $P_{\mathcal{K}_0}$ in the evolution equation on $[(k-1)\tau, k\,\tau]$. The velocity update is then defined by applying the projection $\mathbf{u}_h^{(k)} \equiv P_{\mathcal{K}_0}\mathbf{w}_h^{(k)}$. The pressure $\mathbf{p}_h^{(k)}$ update is computed by orthogonality of the decomposition [TE].

Schwarz Projection Methods. Apart from Uzawa and block matrix methods, and Chorin's method, the *Schwarz projection algorithms* can also be employed to solve (10.108), see Chap. 10.4. Such Schwarz algorithms will employ $(M+\tau\,\nu\,A)$-orthogonal local projections, and the rate of convergence is heuristically expected to be independent of h and τ, and if a coarse space projection is employed, independent of h_0, where h_0 is the diameter of subdomains. If $\tau \leq c\,h_0^2$, in some cases, the coarse space term may be omitted. These algorithms are easily parallelized [FO, LI6, PA2, CA34, PA12, CO5].

Time Dependent Linearized Navier-Stokes Equation. Next consider the solution of system (10.108) arising from the implicit discretization of the time dependent *linearized* Navier-Stokes equation (10.107). When $\mathbf{w} \neq \mathbf{0}$, matrix N will not be zero, and for simplicity, we shall assume that N is skew-symmetric, i.e., $N^T = -N$. Since (10.108) is nonsymmetric, this linear system does not admit a saddle point interpretation (however, some saddle point algorithms may still converge due to positive definiteness of the symmetric part $(M + \tau\,\nu\,A)$ of $M + \tau\,(\nu A + N)$).

Block matrix preconditioning algorithms from Chap. 10.5 can be employed, with Krylov space acceleration. Here block diagonal and block tridiagonal preconditioners may be employed. Symmetric positive definite preconditioners of the form $M + \tau\nu A$ can be employed for $M + \tau\,(\nu A + N)$ and symmetric positive definite preconditioners $S_0(\tau)$ of the form:

$$S_0(\tau)^{-1} = \left(BM^{-1}B^T\right)^{-1} + \tau\,\nu\,\left(BA^{-1}B^T\right)^{-1},$$

can be employed for $S(\tau) = B\,(M + \tau\nu A + \tau N)^{-1}\,B^T$, the nonsymmetric Schur complement matrix.

Chorin's projection scheme is also easily extended to the linearized Navier-Stokes equations, as indicated below.

Algorithm 10.6.4 *(Chorin's Projection Method to Solve (10.108))*
Given $\tilde{\mathbf{u}}_h^{(k-1)}$, $\tilde{\mathbf{p}}_h^{(k-1)}$:

1. *Update the velocity:*

$$\tilde{\mathbf{w}}_h^{(k)} = (I + \tau\,\nu\,A + \tau\,N)^{-1}\left(\tilde{\mathbf{u}}_h^{(k-1)} + \tau\mathbf{f}_h^{(k)}\right)$$

2. *In parallel compute:*

$$\begin{aligned}\tilde{\mathbf{p}}_h^{(k)} &= \tau^{-1}\,\left(BB^T\right)^{-1}B\tilde{\mathbf{w}}_h^{(k)}\\ \tilde{\mathbf{u}}_h^{(k)} &= P_{\mathcal{K}_0^h}\tilde{\mathbf{w}}_h^{(k)}\end{aligned}$$

Output: $\tilde{\mathbf{u}}_h^{(k)}$, $\tilde{\mathbf{p}}_h^{(k)}$

Remark 10.70. As before, Chorin's scheme will be stable in the Euclidean norm $\|\cdot\|$ provided $\|\,(I + \tau\,\nu\,A + \tau\,N)^{-1}\,\| \leq 1 + c\tau$ for some c independent of h and τ. For higher order schemes and further analysis see [BE3, RA, E, WE2].

10.7 Applications to Mixed Formulations of Elliptic Equations

In this section, we describe saddle point algorithms to solve discretizations of mixed formulations of elliptic equations. Such formulations reduce a 2nd order self-adjoint coercive elliptic equation to an equivalent system of first order partial differential equations having a saddle point structure. Discretization of this first order system yields a saddle point linear system which *simultaneously approximates* the solution of the original elliptic equation, as well as a linear combination of the gradient of the solution. Applications involve fluid flow in a porous medium, where the flow velocity satisfies Darcy's law [DA, EW2]. Our discussion is organized as follows.

- Background on mixed formulations of elliptic equations.
- Properties of matrices A and B.
- Uzawa, penalty and block matrix algorithms.
- Projection algorithms.
- Alternative algorithms.

10.7.1 Background

In mathematical models of fluid flow in a porous medium [DA, EW2], it is assumed that the *velocity* of the fluid interspersed between porous rocks is *proportional* to the negative of the *pressure gradient*. This law, referred to as Darcy's law, together with the conservation of mass yields a first order system of partial differential equations:

$$\begin{cases} \mathbf{u} = -\,a(x)\nabla p, & \text{in } \Omega \quad \text{Darcy's law} \\ \nabla\cdot\mathbf{u} = \;\; f(x), & \text{in } \Omega \quad \text{Conservation of mass} \\ \mathbf{n}\cdot\mathbf{u} = -\,g(x), & \text{in } \partial\Omega \;\text{Flux boundary condition,} \end{cases} \tag{10.112}$$

where $p(\cdot)$ and $\mathbf{u}(\cdot)$ denote the pressure and Darcy velocity, respectively, of the fluid in $\Omega \subset \mathbb{R}^d$, while $a(\cdot)$ denotes a tensor (matrix) coefficient of size d, referred to as the *permeability*, representing the proportionality between the Darcy velocity and the pressure gradient. Here $f(\cdot)$ denotes the rate of injection of fluid (through wells) into the medium, $g(\cdot)$ denotes the inward flux of fluid on $\partial\Omega$, and $\mathbf{n}(\cdot)$ denotes the unit outward normal on $\partial\Omega$. Substituting the Darcy velocity $\mathbf{u}(x)$ into the conservation of mass equation and applying the flux boundary condition yields the following elliptic equation:

$$\begin{cases} -\nabla\cdot(a(x)\nabla p) = f(x), & \text{in} \quad \Omega \\ \mathbf{n}\cdot(a(x)\nabla p) = g(x), & \text{on} \quad \partial\Omega. \end{cases} \tag{10.113}$$

An application of divergence theorem yields the following requirement:

$$\int_\Omega f(x)\,dx + \int_{\partial\Omega} g(x)\,ds_x = 0. \tag{10.114}$$

Neumann problem (10.113) will be solvable when the preceding consistency requirement holds, and the solution $p(\cdot)$ will be unique up to a constant. System (10.112) is generally coupled to a larger system of partial differential equations modeling the dynamics of the flow within the porous medium, and the permeability $a(x)$ of the medium is typically assumed to be a piecewise smooth, symmetric positive definite tensor (matrix valued function) with possibly large jump discontinuities [DA, EW2]. We shall assume the bounds:

$$a_0 \, \|\xi\|^2 \, \leq \, \xi^T a(x) \xi \, \leq \, a_1 \, \|\xi\|^2, \qquad \forall \xi \in \mathbb{R}^d,$$

for some $0 < a_0 < a_1$. In applications, the Darcy velocity $\mathbf{u} = -a(x,y)\nabla p$ is expected to be *smooth*, even when $a(x)$ is discontinuous. This property motivates discretizing the first order system (10.112) or its mixed formulation (10.115) to numerically approximate $\mathbf{u}$, instead of discretizing (10.113) and subsequently differentiating the discrete pressure [RA4, EW4, BR33].

Mixed Formulation of (10.113) and its Discretization. A mixed formulation of the first order system (10.112) is obtained by multiplying Darcy's law by $a(x)^{-1}$ on the left, and retaining the other equations:

$$\begin{cases} a(x)^{-1} \, \mathbf{u} + \nabla p = \mathbf{0}, & \text{in } \Omega \\ \nabla \cdot \mathbf{u} = f(x), & \text{in } \Omega \\ \mathbf{n} \cdot \mathbf{u} = g(x), & \text{in } \partial\Omega. \end{cases} \tag{10.115}$$

The mixed formulation is a saddle point problem similar to Stokes equation, with $a(x)^{-1}\mathbf{u}$ replacing the term $-\nu\Delta\mathbf{u}$ in Stokes equation. For convenience, henceforth we shall assume that $g(\cdot) = 0$ on $\partial\Omega$.

A *weak formulation* of (10.115) can be obtained as follows. Multiply the first vector equation in (10.115) by a test function $\mathbf{v}(\cdot)$ satisfying $\mathbf{n}\cdot\mathbf{v} = 0$ on $\partial\Omega$ and integrate over Ω. The term $\int_\Omega \nabla p \cdot \mathbf{v}(x)\, dx$ can be integrated by parts to yield $-\int_\Omega p(x)\, \nabla \cdot \mathbf{v}(x)\, dx$, since $\mathbf{n}\cdot\mathbf{v} = 0$ on $\partial\Omega$. The conservation of mass equation in (10.115) can similarly be multiplied by a test function $-q(\cdot)$ and integrated over Ω. Choosing appropriate function spaces so that the resulting integrals are well defined [RA4], we obtain a weak formulation which seeks $\mathbf{u}(\cdot) \in H_0(div, \Omega)$ and $p(\cdot) \in L^2(\Omega)$ satisfying:

$$\begin{cases} \int_\Omega \mathbf{v}^T a(x)^{-1} \mathbf{u}\, dx - \int_\Omega p(\nabla\cdot\mathbf{v}) dx = 0, & \forall \mathbf{v} \in H_0(div, \Omega) \\ \qquad\qquad - \int_\Omega q(\nabla \cdot \mathbf{u}) dx = - \int_\Omega f(x) q(x)\, dx, & \forall q \in L^2(\Omega), \end{cases} \tag{10.116}$$

where $H_0(div, \Omega)$ is defined by:

$$\begin{cases} H(div, \Omega) & = \{\mathbf{v}(x) \in (L^2(\Omega))^d \, : \, \nabla\cdot\mathbf{v} \in L^2(\Omega)\}, \\ \|\mathbf{v}\|^2_{H(div,\Omega)} & = \|\mathbf{v}\|^2_{L^2} + \|\nabla\cdot\mathbf{v}\|^2_{L^2}, \text{ and} \\ H_0(div, \Omega) & = \{\mathbf{v} \in H(div, \Omega) : \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\}. \end{cases}$$

The flux $\mathbf{n}\cdot\mathbf{v}$ can be shown to be well defined for $\mathbf{v} \in H(div, \Omega)$, due to a trace theorem [RA4], so that $H_0(div, \Omega)$ is well defined.

Remark 10.71. To obtain a unique pressure, the pressure space $L^2(\Omega)$ should be replaced by the quotient space $L^2(\Omega)/\mathbb{R}$.

A mixed finite element discretization of (10.115) can be obtained by Galerkin approximation of (10.116), see [RA4, BR33]. Let $\tau_h(\Omega)$ denote a quasiuniform triangulation of Ω having grid size h. Using this triangulation, let $\mathbf{V}_h \subset H_0(div, \Omega)$ and $Q_h \subset L^2(\Omega)$ denote finite element spaces of dimension n and m, respectively, for the velocity and pressure. Then, a mixed finite element discretization of (10.116) will seek $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in Q_h$ such that:

$$\begin{cases} \mathcal{A}(\mathbf{u}_h, \mathbf{v}_h) + \mathcal{B}(\mathbf{v}_h, p_h) = 0, & \forall \mathbf{v}_h \in \mathbf{V}_h \\ \qquad\qquad\qquad \mathcal{B}(\mathbf{u}_h, q_h) = -(f, q_h), & \forall q_h \in Q_h, \end{cases} \tag{10.117}$$

where $\mathcal{A}(.,.) : \mathbf{V}_h \times \mathbf{V}_h \rightarrow \mathbb{R}$ and $\mathcal{B}(.,.) : \mathbf{V}_h \times Q_h \rightarrow \mathbb{R}$ are bilinear forms:

$$\begin{cases} \mathcal{A}(\mathbf{u}, \mathbf{v}) \equiv & \int_\Omega \left(\mathbf{v}(x)^T a(x)^{-1} \mathbf{u}(x) \right) dx \\ \mathcal{B}(\mathbf{v}, q) \equiv - & \int_\Omega q(x) \; (\nabla \cdot \mathbf{v}(x)) \, dx \\ (f, q) \equiv & \int_\Omega f(x) \, q(x) \, dx, \end{cases} \tag{10.118}$$

for $\mathbf{u}(\cdot), \mathbf{v}(\cdot) \in \mathbf{V}_h$ and $p(\cdot)$, $q(\cdot) \in Q_h$. Integration by parts yields:

$$\mathcal{B}(\mathbf{v}, p) = - \int_\Omega p(x) \; (\nabla \cdot \mathbf{v}(x)) \, dx = \int_\Omega \nabla p(x) \cdot \mathbf{v}(x) \, dx,$$

since $\mathbf{n} \cdot \mathbf{v} = 0$ on $\partial\Omega$. To ensure that the discretization (10.117) of (10.116) is *stable* and to ensure that system (10.119) is solvable, the finite element spaces $\mathbf{V}_h$ and Q_h will be assumed to satisfy the *coercivity* and *uniform inf-sup* conditions described in Lemma 10.53.

Unlike finite element spaces which are subspaces of Sobolev spaces, the spaces $\mathbf{V}_h \subset H_0(div, \Omega)$ and $Q_h \subset L^2(\Omega)$ are *discontinuous* across elements. The velocity flux $\mathbf{n} \cdot \mathbf{v}_h$, however, will be required to be continuous across inter-element faces (when $\Omega \subset \mathbb{R}^3$) or edges (when $\Omega \subset \mathbb{R}^2$). For the lowest order finite element spaces [RA4, BR33], the discrete velocities within each element are determined uniquely by the fluxes on the faces (when $\Omega \subset \mathbb{R}^3$) or edges (when $\Omega \subset \mathbb{R}^2$) of the elements.

A linear system corresponding to (10.117) can be constructed as follows. Let $\{\boldsymbol{\psi}_1(x), \ldots, \boldsymbol{\psi}_n(x)\}$ denote a basis for $\mathbf{V}_h$ and $\{q_1(x), \ldots, q_m(x)\}$ a basis for $Q_h \subset L^2(\Omega)$. Expand $\mathbf{u}_h(x)$ and $p_h(x)$ using this basis:

$$\begin{aligned} \mathbf{u}_h(x) &= \textstyle\sum_{i=1}^n (\mathbf{u}_h)_i \, \boldsymbol{\psi}_i(x) \\ p_h(x) &= \textstyle\sum_{i=1}^m (\mathbf{p}_h)_i \, q_i(x), \end{aligned}$$

where with some abuse of notation, we have used $\mathbf{u}_h(x)$ to denote a finite element function and $\mathbf{u}_h$ to denote its vector representation relative to the given basis. Substituting $\mathbf{v}_h(\cdot) = \boldsymbol{\psi}_i(\cdot)$ for $i = 1, \ldots, n$ and $q_h(\cdot) = q_i(\cdot)$ for

$i = 1, \ldots, m$ into the above yields the following *saddle point* linear system:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_h \\ \mathbf{p}_h \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_h \end{bmatrix}, \tag{10.119}$$

where we define the matrices A and B, and the vector $\mathbf{f}_h$ as:

$$(A)_{ij} = \mathcal{A}(\boldsymbol{\psi}_i, \boldsymbol{\psi}_j), \quad (B)_{ij} = \mathcal{B}(\boldsymbol{\psi}_j, q_i) \quad (\mathbf{f}_h)_i = -(f, q_i).$$

By construction, matrix A will be *symmetric* and sparse. It will reduce to a *mass* (Gram) matrix G when $a(x) = I$, where $\mathbf{v}_h^T G \mathbf{v}_h = \|\mathbf{v}_h\|_{0,\Omega}^2$. More generally, using upper and lower bounds for $a(x)$, we obtain that A satisfies:

$$\frac{1}{a_1} \left(\mathbf{v}_h^T G \mathbf{v}_h \right) \leq \left(\mathbf{v}_h^T A \mathbf{v}_h \right) \leq \frac{1}{a_0} \left(\mathbf{v}_h^T G \mathbf{v}_h \right). \tag{10.120}$$

By construction, matrix B^T will be sparse and correspond to a discretization of the gradient operator, while B will correspond to a discretization of the negative of the divergence operator.

The finite element spaces $\mathbf{V}_h \subset H_0(div, \Omega)$ and $Q_h \subset L^2(\Omega)$ will be *assumed* to satisfy the property:

$$\mathcal{B}(\mathbf{v}_h, q_h) = 0, \quad \forall q_h \in Q_h \Longrightarrow \nabla \cdot \mathbf{v}_h = 0, \quad \text{in } \Omega. \tag{10.121}$$

As a result, the *discrete divergence free* velocity space $\mathcal{K}_0^h$ will be a subspace of the *divergence free* velocity space $\mathcal{K}_0$:

$$\begin{aligned} \mathcal{K}_0^h &= \{ \mathbf{v}_h \in \mathbf{V}_h \, : \, \mathcal{B}(\mathbf{v}_h, q_h) = 0, \ \forall \, q_h \in Q_h \} \\ &\subset \mathcal{K}_0 = \{ \mathbf{v} \in H_0(div, \Omega) \, : \, \nabla \cdot \mathbf{v} = 0 \}. \end{aligned} \tag{10.122}$$

For a description of finite element spaces satisfying the above properties, see [RA4, BR33]. When (10.122) holds, *coercivity* condition (10.91) will hold:

$$\begin{cases} \|\mathbf{v}_h\|_{H(div,\Omega)}^2 = \|\mathbf{v}_h\|_{0,\Omega}^2 + \|\nabla \cdot \mathbf{v}_h\|_{0,\Omega}^2, \\ \qquad\qquad\quad = \|\mathbf{v}_h\|_{0,\Omega}^2, \\ \qquad\qquad\quad \leq a_1 \, \mathcal{A}(\mathbf{v}_h, \mathbf{v}_h), \\ \qquad\qquad\quad = a_1 \, \mathbf{v}_h^T A \mathbf{v}_h, \end{cases} \quad \text{when } \mathbf{v}_h \in \mathcal{K}_0^h, \tag{10.123}$$

since the $H(div, \Omega)$ norm reduces to the $\left(L^2(\Omega) \right)^d$ norm within the class of *divergence free* functions. Thus, the matrix A will satisfy the *coercivity condition* (10.91) within the subspace $\mathcal{K}_0^h$.

We shall *assume* that the spaces $\mathbf{V}_h$ and Q_h satisfy the *uniform inf-sup* condition (10.93) with $\beta > 0$ independent of h such that:

$$\sup_{\mathbf{v}_h \in \mathbf{V}_h \setminus \{\mathbf{0}\}} \frac{\mathcal{B}(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{H(div,\Omega)}} \geq \beta \, \|q_h\|_{0,\Omega}, \quad \forall q_h \in Q_h / \mathbb{R}. \tag{10.124}$$

Remark 10.72. Mixed finite element methods can also be formulated for more general boundary conditions [RA4, BR33]. The *Dirichlet* problem:

$$\begin{cases} -\nabla \cdot (a(x)\nabla p) = f(x), & \text{in } \Omega \\ \qquad\qquad\quad p = g_D(x), & \text{on } \partial\Omega, \end{cases} \tag{10.125}$$

will seek $\mathbf{u} \in H(div, \Omega)$ and $p \in L^2(\Omega)$ satisfying (10.126) and (10.127) below. The first block row of (10.116) will be replaced by:

$$\int_\Omega \mathbf{v}^T a(x)^{-1} \mathbf{u}\, dx - \int_\Omega p(\nabla \cdot \mathbf{v}) dx = -\int_{\partial\Omega} g_D\, \mathbf{n} \cdot \mathbf{v}\, ds_x, \tag{10.126}$$

for each $\mathbf{v} \in H(div, \Omega)$, since integration by parts yields:

$$\int_\Omega \mathbf{v} \cdot \nabla p\, dx = -\int_\Omega p\, (\nabla \cdot \mathbf{v})\, dx + \int_{\partial\Omega} p\, \mathbf{n} \cdot \mathbf{v}\, ds_x,$$

and since $p = g_D$ on $\partial\Omega$. The second row of (10.116) will remain unchanged:

$$-\int_\Omega q(\nabla \cdot \mathbf{u}) dx = -\int_\Omega f(x) q(x)\, dx, \ \ \forall q \in L^2(\Omega). \tag{10.127}$$

We omit further details.

10.7.2 Properties of A and B

We now summarize properties of matrix A in system (10.119). Matrix A will be sparse and symmetric positive definite of size n corresponding to a mass (Gram) matrix when $a(x) = I$. When $a(x) \neq I$, we will obtain that:

$$\frac{1}{a_1} \le \frac{\mathbf{v}_h^T A \mathbf{v}_h}{\mathbf{v}_h^T G \mathbf{v}_h} \le \frac{1}{a_0}, \quad \mathbf{v}_h \in \mathbf{V}_h \setminus \{\mathbf{0}\},$$

where G denotes the mass matrix. Thus, it will hold that:

$$\gamma_0\, a_1^{-1}\, h^d \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le \gamma_1\, a_0^{-1}\, h^d,$$

for $\gamma_0 < \gamma_1$ independent of h, since the mass matrix G is spectrally equivalent to $ch^d I$. As a result, $\text{cond}(A) \le c \left(\frac{a_1}{a_0}\right)$ for some $c > 0$ independent of h.

In addition to the above, the following property will hold for matrix B.

Lemma 10.73. *Matrix B in discretization (10.117) of (10.116) will satisfy:*

$$\left|\mathbf{q}_h^T B \mathbf{v}_h\right| \le c_*\, h^{-1}\, \|\mathbf{v}_h\|_A\, \|\mathbf{q}_h\|_M, \tag{10.128}$$

for some $c_* > 0$ independent of h, where M denotes the mass matrix for the pressure $\mathbf{q}_h^T M \mathbf{q}_h = \|q_h\|_{0,\Omega}^2$.

Proof. Let $\mathbf{v}_h(\cdot)$ and $q_h(\cdot)$ denote finite element functions with associated nodal vectors $\mathbf{v}_h \in \mathbb{R}^n$ and $\mathbf{q}_h \in \mathbb{R}^m$, respectively. Then, bound (10.128) will hold for some c_* independent of h since:

$$\begin{aligned} \left|\mathbf{q}_h^T B \mathbf{v}_h\right| = |\mathcal{B}(\mathbf{v}_h, q_h)| &= \left|\int_\Omega q_h(x)\ (\nabla \cdot \mathbf{v}_h(x))\ dx\right| \\ &\le \|\nabla \cdot \mathbf{v}_h\|_{0,\Omega}\ \|q_h\|_{0,\Omega} \\ &\le c_1\, h^{-1}\ \|\mathbf{v}_h\|_{0,\Omega}\ \|q_h\|_{0,\Omega} \\ &\le c_*\, h^{-1}\ \|\mathbf{v}_h\|_A\ \|q_h\|_{0,\Omega}, \end{aligned}$$

where an *inverse inequality* $\|\nabla \cdot \mathbf{v}_h\|_{0,\Omega} \le c_1\, h^{-1}\ \|\mathbf{v}_h\|_{0,\Omega}$ was employed. □

The Schur complement $S = (BA^{-1}B^T)$ corresponds to a discretization of the elliptic operator L where $L\,p \equiv -\nabla \cdot (a(x)\,\nabla p)$ with Neumann boundary conditions. We heuristically expect the following bounds to hold for S.

Lemma 10.74. *Let the uniform inf-sup condition (10.95) and bound (10.95) hold. Then, there will exist* $0 < \alpha_0 < \alpha_1$ *independent of* h*, such that:*

$$\beta\, a_0 \left(\mathbf{q}_h^T M \mathbf{q}_h\right)\ \le\ \mathbf{q}_h^T \left(BA^{-1}B^T\right)\mathbf{q}_h\ \le\ c_*\, a_1\, h^{-2}\ \left(\mathbf{q}_h^T M \mathbf{q}_h\right), \qquad (10.129)$$

where M *denotes the mass matrix associated with the pressure finite element space* Q_h *(which we assume to satisfy* $Q_h \subset L^2(\Omega)/\mathbb{R}$*).*

Proof. Follows by a modification of lemma 10.11 and bound (10.128). Indeed, it will hold that:

$$\begin{aligned} \left(\mathbf{q}_h^T BA^{-1}B^T \mathbf{q}_h\right) &= \left(\mathbf{q}_h^T BA^{-1}AA^{-1}B^T \mathbf{q}_h\right) \\ &= \|A^{-1}B^T \mathbf{q}_h\|_A^2 \\ &= \sup_{\mathbf{v}_h \neq \mathbf{0}} \left(\frac{\mathbf{v}_h^T AA^{-1}B^T \mathbf{q}_h}{\|\mathbf{v}_h\|_A}\right)^2 \\ &= \sup_{\mathbf{v}_h \neq \mathbf{0}} \left(\frac{\mathbf{q}_h^T B \mathbf{v}_h}{\|\mathbf{v}_h\|_A}\right)^2. \end{aligned} \qquad (10.130)$$

Now, using that $\|\mathbf{v}_h\|_{H(div,\Omega)} \ge a_0 \|\mathbf{v}_h\|_A$ and substituting the uniform inf-sup condition (10.124) into (10.130) yields:

$$\begin{aligned} \left(\mathbf{q}_h^T BA^{-1}B^T \mathbf{q}_h\right) &= \sup_{\mathbf{v}_h \neq \mathbf{0}} \left(\frac{\mathbf{q}_h^T B \mathbf{v}_h}{\|\mathbf{v}_h\|_A}\right)^2 \\ &\ge a_0\ \sup_{\mathbf{v}_h \neq \mathbf{0}} \left(\frac{\mathbf{q}_h^T B \mathbf{v}_h}{\|\mathbf{v}_h\|_{H(div,\Omega)}}\right)^2 \\ &\ge a_0\, \beta^2 \mathbf{q}_h^T M \mathbf{q}_h, \end{aligned}$$

which is a lower bound for $S = BA^{-1}B^T$. To obtain an upper bound, substitute (10.128) into (10.130) to obtain:

$$\begin{aligned} \left(\mathbf{q}_h^T BA^{-1}B^T \mathbf{q}_h\right) &= \sup_{\mathbf{v}_h \neq \mathbf{0}} \left(\frac{\mathbf{q}_h^T B \mathbf{v}_h}{\|\mathbf{v}_h\|_A}\right)^2 \\ &\le c_*^2\, h^{-2} a_1\, \mathbf{q}_h^T M \mathbf{q}_h. \end{aligned}$$

This yields the desired bound. □

Remark 10.75. If $q_h(x) = 1 \in Q_h$ for a *Neumann* boundary value problem, with nodal vector $\mathbf{1} = (1, \ldots, 1)^T$ associated with $q_h(x) = 1$, then matrix B^T will satisfy $B^T \mathbf{1} = \mathbf{0}$. This is because $\nabla q_h(x) = 0$, since $\mathcal{B}(\mathbf{v}_h, q_h) = 0$ for all $\mathbf{v}_h \in \mathbf{V}_h$, when $q_h(x) = 1$. In this case, matrix B will not have full rank and the saddle point matrix in (10.119) will be *singular*. However, system (10.119) will be consistent provided $\mathbf{1}^T \mathbf{f_h} = 0$. A full rank matrix B can be constructed, if desired, by choosing the pressure space Q_h to consist only of functions having *mean value zero*, i.e., $Q_h \subset L^2(\Omega)/\mathbb{R}$. This, however, will be cumbersome and we shall instead work with a singular coefficient matrix, and appropriately modify algorithms from the preceding sections.

Remark 10.76. Since each *singular value* of matrix B^T corresponds to the square root of an eigenvalue of BB^T, their range may be estimated as follows. When $a(x) = I$, bounds from Lemma 10.74 yield:

$$\beta \left(\mathbf{q}_h^T M \mathbf{q}_h\right) \leq \mathbf{q}_h^T \left(BA^{-1}B^T\right) \mathbf{q}_h \leq c_* \, h^{-2} \left(\mathbf{q}_h^T M \mathbf{q}_h\right). \tag{10.131}$$

Since M is spectrally equivalent to a matrix $h^d I$ of size m and since A is spectrally equivalent to a matrix $h^d I$ of size n, it follows that:

$$\gamma_0 \, \beta \left(\mathbf{q}_h^T \mathbf{q}_h\right) \leq \mathbf{q}_h^T \left(BB^T\right) \mathbf{q}_h \leq \gamma_1 \, h^{-2} \left(\mathbf{q}_h^T \mathbf{q}_h\right), \tag{10.132}$$

for $0 < \gamma_0 < \gamma_1$ independent of h. Thus, the singular values of B^T will range from $(\gamma_0 \, \beta)^{1/2}$ to $(\gamma_1)^{1/2} \, h^{-1}$ when $Q_h \subset L^2(\Omega)/\mathbb{R}$. If $q_h(x) = 1 \in Q_h$, then $B^T \mathbf{1} = \mathbf{0}$ and $\sigma_1(B) = 0 < \sigma_2(B) \leq \cdots \leq \sigma_m(B) \leq \gamma_1^{1/2} \, h^{-1}$.

Remark 10.77. When $a(x) = I$, mixed finite element discretization (10.119) of *Neumann* problem (10.113) shows that matrix $\left(BM^{-1}B^T\right)$ from Chap. 10.6 and matrix $S = \left(BA^{-1}B^T\right)$ from this section, correspond to a discretization of Neumann problem (10.113). Lemma 10.74 yields $\text{cond}(BA^{-1}B^T) = O(h^{-2})$.

10.7.3 Uzawa and Block Matrix Algorithms

Saddle point system (10.119) can be solved using algorithms from Chap. 10.2 to Chap. 10.5. Below, we indicate different choices of preconditioners A_0 for A and S_0 for S for use in Uzawa and block matrix algorithms.

Preconditioners for A. Since matrix A has a condition number independent of h, employ its *diagonal* A_0 as a preconditioner for A:

$$A_0 = \text{diag}(A).$$

In this case $\text{cond}(A_0, A) \leq c$, independent of h. However, heuristically such a diagonal preconditioner may also help eliminate dependence on the variations (a_1/a_0) in the coefficient $a(x)$, particularly when (a_1/a_0) is large.

Preconditioners for S. The Schur complement matrix $S = \left(BA^{-1}B^T\right)$ corresponds to a discretization of the *elliptic* operator L associated with the pressure. Since A^{-1} is generally *dense*, matrix S will also be dense. However, cell centered *finite difference* or *finite element* discretizations, can be employed to construct a *sparse* approximation S_0 of S, see [WH, WE, AL2, AR2]. Sparse direct solvers, or iterative solvers such as multigrid and domain decomposition solvers may then be employed.

An alternative sparse *algebraic* preconditioner $S_0 \equiv BA_0^{-1}B^T$ can be constructed using the diagonal matrix A_0. Since B and B^T are *sparse*, we may *assemble* S_0. By construction, the condition number will satisfy [AL2]:

$$\mathrm{cond}(S_0, S) \leq \mathrm{cond}(A_0, A),$$

since the associated Rayleigh quotients satisfy:

$$\frac{\mathbf{q}_h^T S \mathbf{q}_h}{\mathbf{q}_h^T S_0 \mathbf{q}_h} = \frac{\mathbf{q}_h^T B A^{-1} B^T \mathbf{q}_h}{\mathbf{q}_h^T B A_0^{-1} B^T \mathbf{q}_h} = \frac{\mathbf{w}_h^T A^{-1} \mathbf{w}_h}{\mathbf{w}_h^T A_0^{-1} \mathbf{w}_h}, \qquad \text{for} \quad \mathbf{w}_h = B^T \mathbf{q}_h.$$

We may solve linear systems of the form $S_0 \mathbf{p}_h = \mathbf{r}_h$ using either sparse direct solvers or iterative multigrid and domain decomposition solvers.

Remark 10.78. System (10.119) will be *singular* for the Neumann problem, with $B^T\mathbf{1} = \mathbf{0}$. In this case, we require $\mathbf{1}^T\mathbf{f}_h = 0$ for solvability of (10.119). The Schur complement matrix will also be singular with $S\,\mathbf{1} = \mathbf{0}$, and the discrete pressure $\mathbf{p}_h$ will be unique up to a constant. In Uzawa algorithms, the initial pressure iterate $\mathbf{p}_h^{(0)}$ must be chosen to satisfy $\mathbf{1}^T\mathbf{p}_h^{(0)} = 0$. Subsequent iterates will automatically satisfy $\mathbf{1}^T\mathbf{p}_h^{(k)} = 0$ provided $\mathbf{1}^T S_0^{-1} B = \mathbf{0}$. Since the Schur complement S will be singular, computation of the action $S_0^{-1}\mathbf{r}_h$ should ideally be implemented in the form $(I - P_0)S_0^{-1}(I - P_0)\mathbf{r}_h$ where:

$$P_0 \mathbf{w}_h \equiv \left(\frac{\mathbf{1}^T\mathbf{w}_h}{\mathbf{1}^T\mathbf{1}}\right)\mathbf{1},$$

denotes the Euclidean orthogonal projection onto $\mathrm{span}(\mathbf{1})$. When $\mathbf{1}^T\mathbf{r}_h = 0$, the pre-projection step can be omitted. If inexact Uzawa or block triangular preconditioners are employed, it may be necessary to *scale* the preconditioners A_0 and S_0 as described in Chap. 10.2 and Chap. 10.5.

Remark 10.79. We shall not consider penalty or regularization algorithms, except to note that since $\mathrm{cond}(BB^T) = O(h^{-2})$, the matrix $\left(A + \frac{1}{\epsilon}BB^T\right)$ will be ill-conditioned for $0 < \epsilon < \infty$. For an application of penalty methods to mixed formulations of elliptic equations, see [CA22]. See also [LA11] for an algorithm based on $A_0 = A + \epsilon^{-1}B^TB$ and $S_0 = I$.

Remark 10.80. Saddle point preconditioners [AL2] as described in Chap. 10.5 can be employed for mixed formulations of elliptic equations, using a diagonal matrix A_0, as efficient solvers can be obtained for $S_0 = BA_0^{-1}B^T$.

10.7.4 Projection Algorithms

Projected Gradient. Since matrix A has a condition number independent of h, both the projected gradient descent algorithm with a fixed step $\tau > 0$ and the projected CG algorithm to solve (10.119) will converge at a rate independent of h, but dependent on the magnitude of the jumps in $a(x)$. When these jumps have a large magnitude, the projected CG algorithm can be employed with the diagonal preconditioner A_0. In both cases, the projection will require the action of $(BB^T)^{-1}$ each iteration, requiring the solution of a sparse system.

Schwarz Projection Algorithms. Schwarz projection algorithms, which are highly parallelizable, can also be employed to solve system (10.119). They require the solution of smaller saddle point problems [EW8, MA31, MA32]. We shall employ non-overlapping and overlapping subdomains. Let $\Omega_1, \ldots, \Omega_l$ denote a *nonoverlapping* decomposition of Ω with subdomains of size h_0. We let $\Omega_1^*, \ldots, \Omega_l^*$ denote an *overlapping* decomposition of Ω with the extended subdomains $\Omega_i^* \equiv \{x \in \Omega \, : \, \mathrm{dist}(x, \Omega_i) < \beta\, h_0\}$, having overlap $\beta\, h_0$. We shall assume that the elements of $\tau_h(\Omega)$ align with Ω_i and Ω_i^*. Given finite element spaces $\mathbf{V}_h \subset H_0(div, \Omega)$ and $Q_h \subset L^2(\Omega)$ we define *subdomain* velocity and pressure spaces $\mathbf{V}_i$ and $\mathcal{Q}_i$ on Ω_i, and $\mathbf{V}_i^*$ and $\mathcal{Q}_i^*$ on Ω_i^*, as follows:

$$\begin{cases} \mathbf{V}_i = \mathbf{V}_h \cap H_0(div, \Omega_i) \text{ and } \mathbf{V}_i^* = \mathbf{V}_h \cap H_0(div, \Omega_i^*), & \text{for } 1 \le i \le l \\ \mathcal{Q}_i = Q_h \cap L^2(\Omega_i) \qquad\quad \text{and } \mathcal{Q}_i^* = Q_h \cap L^2(\Omega_i^*), & \text{for } 1 \le i \le l. \end{cases}$$

For $1 \le i \le l$ the spaces $\mathbf{V}_i$ and $\mathbf{V}_i^*$ will be extended outside Ω_i and Ω_i^*, respectively, by *zero* extension due to the zero flux requirement on $\partial\Omega_i$ and $\partial\Omega_i^*$. We shall assume that the *discrete divergence free* space $\mathcal{K}_0^h = \mathrm{Kernel}(B)$ is a subspace of the *divergence free* space $\mathcal{K}_0$ within $H_0(div, \Omega)$. This property will hold for most discretizations of mixed formulations of elliptic equations [RA4, BR33]. If the subdomains $\Omega_1, \ldots, \Omega_l$ form the elements of a *coarse triangulation* $\tau_{h_0}(\Omega)$ of Ω, we shall define *coarse* velocity and pressure spaces $\mathbf{V}_0$ and $\mathcal{Q}_0$ using finite elements defined on $\tau_{h_0}(\Omega)$:

$$\begin{cases} \mathbf{V}_0 = \mathbf{V}_{h_0} \cap H_0(div, \Omega) \\ \mathcal{Q}_0 = Q_{h_0} \cap L^2(\Omega). \end{cases}$$

Let n_i, n_i^*, m_i and m_i^* denote the dimension of $\mathbf{V}_i$, $\mathbf{V}_i^*$, $\mathcal{Q}_i$ and $\mathcal{Q}_i^*$, respectively. We shall let U_i and U_i^* denote matrices of size $n \times n_i$ and $n \times n_i^*$, whose columns span the space of nodal vectors associated with $\mathbf{V}_i$ and $\mathbf{V}_i^*$, respectively, Similarly, we let Q_i and Q_i^* denote matrices of size $m \times m_i$ and $m \times m_i^*$ whose columns span the space of nodal vectors associated with $\mathcal{Q}_i$ and Q_i^*, respectively. We define the matrices A_i, A_i^*, B_i and B_i^* as:

$$A_i = U_i^T A U_i, \;\; A_i^* = U_i^{*T} A U_i^*, \;\; B_i = Q_i^T B U_i \;\text{ and }\; B_i^* = Q_i^{*T} B U_i^*, \tag{10.133}$$

of size $n_i \times n_i$, $n_i^* \times n_i^*$, $m_i \times n_i$ and $m_i^* \times n_i^*$ respectively. We shall let $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^m$, $\mathbf{1}_i = (1, \ldots, 1)^T \in \mathbb{R}^{m_i}$, $\mathbf{1}_i^* = (1, \ldots, 1)^T \in \mathbb{R}^{m_i^*}$ span the subspaces $\mathrm{Kernel}(B^T)$, $\mathrm{Kernel}(B_i^T)$ and $\mathrm{Kernel}(B_i^{*T})$, respectively.

Remark 10.81. By choice of $\mathbf{V}_h$ and $\mathcal{Q}_h$, it will hold that:

$$\begin{cases} B_i \mathbf{w}_i = \mathbf{0} \implies U_i \mathbf{w}_i \subset \mathcal{K}_0^h \subset \mathcal{K}_0, & \text{for } 0 \le i \le l \\ B_i^* \mathbf{w}_i^* = \mathbf{0} \implies U_i^* \mathbf{w}_i^* \subset \mathcal{K}_0^h \subset \mathcal{K}_0, & \text{for } 1 \le i \le l. \end{cases} \tag{10.134}$$

Thus, the local divergence free updates will be globally divergence free.

Schwarz projection algorithms to solve (10.119) involve three steps. In the *first step*, a discrete velocity $\mathbf{u}_*$ is computed such that $B\mathbf{u}_* = \mathbf{f}_h$. In the *second step* the divergence free component $\mathbf{w} \equiv \mathbf{u}_h - \mathbf{u}_*$ is determined by a Schwarz projection algorithm. In the *third step*, the discrete pressure $\mathbf{p}_h$ is determined. Below, we describe the first step, to compute $\mathbf{u}_*$.

Algorithm 10.7.1 *(Algorithm to Compute $\mathbf{u}_*$ Satisfying $B\mathbf{u}_* = \mathbf{f}_h$)*

1. *Solve the following coarse grid problem:*

$$\begin{bmatrix} \mathbf{v}_0 \\ \mathbf{q}_0 \end{bmatrix} = \begin{bmatrix} U_0 & 0 \\ 0 & Q_0 \end{bmatrix} \begin{bmatrix} A_0 & B_0^T \\ B_0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_0^T \mathbf{0} \\ Q_0^T \mathbf{f}_h \end{bmatrix} \tag{10.135}$$

2. *For $1 \le i \le l$ solve these local problems in parallel:*

$$\begin{bmatrix} \mathbf{v}_i \\ \mathbf{q}_i \end{bmatrix} = \begin{bmatrix} U_i & 0 \\ 0 & Q_i \end{bmatrix} \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T (\mathbf{0} - A\,\mathbf{v}_0) \\ Q_i^T (\mathbf{f}_h - B\mathbf{v}_0) \end{bmatrix} \tag{10.136}$$

Output: $\mathbf{u}_* = \mathbf{v}_0 + \mathbf{v}_1 + \cdots + \mathbf{v}_l$

Remark 10.82. Each of the local problems in (10.136) will be *consistent* by construction of $\mathbf{v}_0$. The reader may verify $B\,\mathbf{u}_* = \mathbf{f}_h$ as follows. The coarse velocity $\mathbf{v}_0$ computed in (10.135) will not satisfy the constraint $B\mathbf{v}_0 \ne \mathbf{f}_h$ on the fine grid. However, the coarse solution $\mathbf{v}_0$ will provide flux boundary conditions on the boundary of each subdomain Ω_i compatible with $f(x)$. Indeed, by construction, $(\mathbf{f}_h - B\mathbf{v}_0)$ will have *mean value zero* on each Ω_i since:

$$\mathbf{1}_i^T Q_i^T (\mathbf{f}_h - B\mathbf{v}_0) = \int_\Omega \chi_i(x) \left(f(x) - \nabla \cdot \mathbf{v}_0(x) \right) dx = 0, \qquad \forall \chi_i(x) \in \mathcal{Q}_0,$$

where $\chi_i(x)$ denotes the characteristic function of subdomain Ω_i with nodal vector $Q_i \mathbf{1}_i \in \mathbb{R}^m$. Substituting for $\chi_i(x)$ in the above yields:

$$\int_{\Omega_i} \nabla \cdot \mathbf{v}_0(x)\, dx = \int_{\partial\Omega_i} \mathbf{n}(x) \cdot \mathbf{v}_0(x)\, ds_x = -\int_{\Omega_i} f(x)\, dx.$$

Since $(\mathbf{f}_h - B\,\mathbf{v}_0)$ has mean value zero within each subdomain Ω_i, it follows that *zero* flux boundary conditions $\mathbf{n} \cdot \mathbf{v}_i = 0$ posed on $\partial\Omega_i$ will be compatible with the mixed finite element discretization of $\nabla \cdot \mathbf{v}_i = (-f(x) - \nabla \cdot \mathbf{v}_0(x))$ within Ω_i. Thus, subproblems in (10.136) will be *well posed* for $1 \le i \le l$.

Remark 10.83. The component $\mathbf{u}_*$ satisfying $B\mathbf{u}_* = \mathbf{f}_h$, may alternatively be obtained in the form $\mathbf{u}_* = B^T \boldsymbol{\gamma}_*$ for some $\boldsymbol{\gamma}_* \in \mathbb{R}^m$ satisfying:

$$\left(BB^T\right)\boldsymbol{\gamma}_* = \mathbf{f}_h.$$

This formally yields $\mathbf{u}_* = B^T \left(BB^T\right)^\dagger \mathbf{f}_h$, where $\left(BB^T\right)^\dagger$ denotes the Moore-Penrose pseudoinverse of BB^T since B^T will be singular.

We next describe the computation of the divergence free velocity $\mathbf{w} = (\mathbf{u}_h - \mathbf{u}_*)$ and also the pressure $\mathbf{p}_h$. An additive or multiplicative Schwarz projection algorithm can be employed within $\mathcal{K}_0^h$, as in Chap. 10.4. Below, we list the multiplicative Schwarz algorithm to determine $\mathbf{w}$ and $\mathbf{p}_h$, see [MA31].

Algorithm 10.7.2 *(Multiplicative Schwarz Algorithm)*
Let $\mathbf{v}_h^{(0)} = \mathbf{0}$ *and* $\mathbf{p}_h^{(0)} = \mathbf{0}$ *denote starting iterates*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $i = 0, 1, \ldots, l$ do:
           $\begin{bmatrix} \mathbf{w}_i^{(k)} \\ \boldsymbol{\mu}_i^{(k)} \end{bmatrix} = \begin{bmatrix} U_i & 0 \\ 0 & Q_i \end{bmatrix} \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} U_i^T(-A\mathbf{v}_h^{(k+\frac{i}{l+1})} - B^T\mathbf{p}_h^{(k+\frac{i}{l+1})}) \\ \mathbf{0} \end{bmatrix}$
       Define $\gamma_i = \left(\mathbf{1}^T(\mathbf{p}_h^{(k+\frac{i}{l+1})} + \boldsymbol{\mu}_i^{(k)})/\mathbf{1}^T Q_i \mathbf{1}_i\right)$ and update:
           $\begin{bmatrix} \mathbf{v}_h^{(k+\frac{i+1}{l+1})} \\ \mathbf{p}_h^{(k+\frac{i+1}{l+1})} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_h^{(k+\frac{i}{l+1})} \\ \mathbf{p_h}^{(k+\frac{i}{l+1})} \end{bmatrix} + \begin{bmatrix} \mathbf{w}_i^{(k)} \\ \boldsymbol{\mu}_i^{(k)} - \gamma_i Q_i \mathbf{1}_i \end{bmatrix}$
3.     Endfor
4. Endfor
```

Output: $\mathbf{v}_h^{(k)} \approx \mathbf{w}$ *and* $\mathbf{p}_h^{(k)} \approx \mathbf{p}_h$

Once $\mathbf{w}$ has been computed, an approximate solution to (10.119) can be obtained as $\mathbf{u}_h \approx \mathbf{u}_* + \mathbf{v}_h^{(k)}$ and $\mathbf{p}_h^{(k)}$. The multiplicative Schwarz algorithm is *sequential*. However, it can be parallelized by *multicoloring* of the subdomains. Alternatively, the additive Schwarz algorithm may be employed.

Remark 10.84. By construction, all the velocity iterates in the multiplicative Schwarz algorithm will be discrete divergence free, i.e., $\mathbf{v}^{(k+\frac{i}{l+1})} \in \mathcal{K}_0^h$, while the pressure iterates $\mathbf{p}_h^{(k+\frac{i+1}{l+1})}$ will have mean value zero. For sufficiently large overlap β, the iterates $\mathbf{v}_h^{(l)}$ above will converge geometrically to the discrete solution $\mathbf{u}_h$ at a rate independent h, see [EW8, MA32]. The pressure $\mathbf{p}_h^{(k)}$ will converge similarly to $\mathbf{p}_h$. If a *coarse space* is included, then this rate of convergence will also be independent of h_0.

10.7.5 Alternative Algorithms for Mixed Formulations

We mention here several alternative approaches for solving system (10.119).

Algorithm of [GL14]. The algorithm of [GL14] is a *non-overlapping* domain decomposition algorithm for mixed formulations of elliptic equations. This algorithm was a precursor to the development of a variety of domain decomposition algorithms and introduced a *natural coarse space* to enforce certain constraints [BO7, WI6, MA31, FA14, FA15, CO10, MA17, FA14]. Below, we outline the key steps in this algorithm for solving a mixed finite element discretization (10.119) of an elliptic *Neumann* problem.

Let $\Omega_1, \ldots, \Omega_l$ form a *non-overlapping* decomposition of Ω with subdomains of size h_0 and a common interface $\Gamma = \left(\cup_{i=1}^{l} \partial \Omega_i \right) \cap \Omega$ (we use Γ to denote the subdomain interface, instead of B, since in this section, the latter denotes the discretization of the divergence operator). Based on this decomposition, the algorithm decomposes a discrete velocity $\mathbf{w}_h$ as follows:

$$\mathbf{w}_h = \sum_{i=1}^{l} U_i \, \mathbf{w}_i + G \, \boldsymbol{\alpha} + H \, \boldsymbol{\beta}, \tag{10.137}$$

where U_i is a matrix of size $n \times n_i$, G of size $n \times l$, H of size $n \times n_\beta$, $\mathbf{w}_i \in \mathbb{R}^{n_i}$, $\boldsymbol{\alpha} \in \mathbb{R}^l$ and $\boldsymbol{\beta} \in \mathbb{R}^{n_\beta}$. In the above, we let the i'th column of matrix G be the nodal vector corresponding to a discrete velocity $\boldsymbol{\chi}_i(x)$ satisfying:

$$\int_{\partial \Omega_i} \left(\mathbf{n}_i(x) \cdot \boldsymbol{\chi}_i(x) \right) \, ds_x \neq 0,$$

where $\mathbf{n}_i(x)$ is the unit exterior normal to $\partial \Omega_i$. For instance $\mathbf{n}_i(x) \cdot \boldsymbol{\chi}_i(x) = 1$ on $\partial \Omega_i \cap \Omega$ with zero flux on $\Gamma \setminus \partial \Omega_i$, so that $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_l$ are linearly independent. The columns of matrix U_i are nodal velocity vectors corresponding to a basis for $\mathbf{V}_h \cap H_0(div, \Omega_i)$ zero outside Ω_i. The columns of H form a nodal vector basis for the velocities $\mathbf{v}_h$ with support on Γ, satisfying:

$$\int_{\partial \Omega_i} \left(\mathbf{n}(x) \cdot \mathbf{v}_h(x) \right) \, ds_x = 0, \quad \text{for } i = 1, \ldots, l,$$

with zero flux on all interior edges or faces in $\cup_{i=1}^{l} \Omega_i$. By construction, the columns of G, H and $U_1, \ldots, U_l$ together form a basis for $\mathbf{V}_h \cap H_0(div, \Omega)$, and we may decompose $\mathbf{w}_h = \mathbf{w}_I + \mathbf{w}_\Gamma$ with $\mathbf{w}_I = \sum_{i=1}^{l} U_i \mathbf{w}_i$ corresponding to subdomain interiors, and $\mathbf{w}_\Gamma = G \boldsymbol{\alpha} + H \boldsymbol{\beta}$ corresponding to interface Γ.

Similarly, the pressure can be decomposed as:

$$\mathbf{p}_h = \sum_{i=1}^{l} Q_i \, \mathbf{p}_i + C \, \boldsymbol{\mu}, \tag{10.138}$$

where Q_i is of size $m \times (m_i - 1)$, C is $m \times l$, $\mathbf{p}_i \in \mathbb{R}^{(m_i - 1)}$ and $\boldsymbol{\mu} \in \mathbb{R}^l$. The columns of Q_i forms a basis for the discrete pressures in $Q_h \cap \left(L^2(\Omega_i) / \mathbb{R} \right)$ with mean value zero, while the i'th column of C denotes a nodal vector for the pressure corresponding to the characteristic function of Ω_i. The columns of Q_i and C thus together form a basis for $Q_h \cap L^2(\Omega)$.

The algorithm of [GL14] determines the solution to (10.119) in two steps. In the *first step*, a discrete velocity $\mathbf{u}_*$ is computed satisfying $B\mathbf{u}_* = \mathbf{f}_h$. This can be computed as in Algorithm 10.7.1.

In the *second* step, the components $\mathbf{w}_h = \mathbf{u}_h - \mathbf{u}_*$ and $\mathbf{p}_h$ are sought. By construction of $\mathbf{u}_*$ in the first step, $\mathbf{w}_h$ will satisfy $B\mathbf{w}_h = \mathbf{0}$, so that the finite element function $\mathbf{w}_h(x)$ associated with $\mathbf{w}_h$ will satisfy $\nabla \cdot \mathbf{w}_h(x) = 0$ in Ω, yielding that $\int_{\partial\Omega_i} \mathbf{n} \cdot \mathbf{w}_h \, ds_x = 0$ for $i = 1, \ldots, l$. As a result, the term $G\boldsymbol{\alpha}$ can be *omitted* in the expansion (10.137) of $\mathbf{w}_h$. The block unknowns $\mathbf{w}_h$ and $\mathbf{p}_h$ will solve the residual equation:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_h \\ \mathbf{p}_h \end{bmatrix} = \begin{bmatrix} -A\mathbf{u}_* \\ \mathbf{0} \end{bmatrix}. \tag{10.139}$$

Since $B\mathbf{w}_h = \mathbf{0}$, the problem to determine $\mathbf{w}_h \in \mathcal{K}_0^h$ will be *coercive* satisfying:

$$\mathbf{v}_h^T A \mathbf{w}_h = -\mathbf{v}_h^T A \mathbf{u}_*, \quad \forall \mathbf{v}_h \in \mathcal{K}_0^h. \tag{10.140}$$

The algorithm of [GL14] expands $\mathbf{w}_h$ using (10.137) omitting $G\,\boldsymbol{\alpha}$, and expands $\mathbf{p}_h$ using (10.138), and eliminates $\mathbf{w}_1, \ldots, \mathbf{w}_l$, $\mathbf{p}_1, \ldots, \mathbf{p}_l$ and $\boldsymbol{\mu}$ in system (10.139) to solve a positive definite symmetric Schur complement system for $\boldsymbol{\beta}$. To eliminate $\mathbf{w}_1, \ldots, \mathbf{w}_l$, $\mathbf{p}_1, \ldots, \mathbf{p}_l$ and $\boldsymbol{\mu}$ in system (10.139), substitute expression (10.137) for $\mathbf{w}_h$ omitting $G\,\boldsymbol{\alpha}$, and (10.138) for $\mathbf{p}_h$ into the residual equation (10.139), move the term involving $H\boldsymbol{\beta}$ to the right hand side:

$$\begin{cases} A\left(\sum_{i=1}^l U_i \mathbf{w}_i\right) + B^T\left(\sum_{i=1}^l Q_i \mathbf{p}_i + C\boldsymbol{\mu}\right) = -A\mathbf{u}_* - AH\boldsymbol{\beta} \\ \qquad\qquad\qquad\qquad B\left(\sum_{i=1}^l U_i \mathbf{w}_i\right) = -BH\boldsymbol{\beta}, \end{cases} \tag{10.141}$$

and solve for the block unknowns $\mathbf{w}_1, \ldots, \mathbf{w}_l$, $\mathbf{p}_1, \ldots, \mathbf{p}_l, \boldsymbol{\mu}$ as outlined below. *Step 1.* Multiply both sides of (10.141) by blockdiag(U_i^T, Q_i^T) to obtain:

$$\begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \mathbf{p}_i \end{bmatrix} = \begin{bmatrix} -U_i^T A\mathbf{u}_* - U_i^T AH\boldsymbol{\beta} \\ -Q_i^T BH\boldsymbol{\beta} \end{bmatrix},$$

where $A_i = U_i^T\, A\, U_i$ and $B_i = Q_i^T\, B\, U_i$. The other terms are zero because of disjoint subdomains $U_i^T\, A\, U_j = 0$ and $U_i^T\, B\, Q_j = 0$ when $i \neq j$, and since $U_i^T\, B^T\, C\, \boldsymbol{\mu} = 0$ because $C\,\boldsymbol{\mu}$ is constant within subdomains.
Step 2. To determine $\boldsymbol{\mu}$, multiply both sides of (10.141) by blockdiag$(G^T, 0)$ and move the terms involving $\mathbf{w}_i$ and $\mathbf{p}_i$ to the right hand side to obtain:

$$\left(G^T\, B^T\, C\right)\, \boldsymbol{\mu} = -G^T A \left(\mathbf{u}_* + H\,\boldsymbol{\beta} + \sum_{i=1}^l U_i \mathbf{w}_i\right) - G^T B^T \left(\sum_{i=1}^l Q_i \mathbf{p}_i\right), \tag{10.142}$$

where $\mathbf{w}_i$ and $\mathbf{p}_i$ are as obtained in step 1. Solve for $\boldsymbol{\mu}$.

Remark 10.85. The linear systems (10.140) can be solved in parallel for $i = 1, \ldots, l$. The matrix $\left(G^T \, B^T \, C\right)$ in (10.142) will be of size l. It will be singular for the Neumann problem since $B^T \mathbf{1} = \mathbf{0}$ and $C\,\mathbf{1} = \mathbf{1}$. However, it can easily be seen to be consistent, so any solution with mean value zero may be chosen.

To determine $\boldsymbol{\beta}$, denote by $E\,\boldsymbol{\beta}$ the *discrete divergence free* velocity extension obtained in step 1 when $\mathbf{u}_* = \mathbf{0}$:

$$E\,\boldsymbol{\beta} \equiv H\boldsymbol{\beta} + \sum_{i=1}^{l} \begin{bmatrix} U_i & 0 \end{bmatrix} \begin{bmatrix} A_i & B_i^T \\ B_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} -U_i^T \, A\, H\, \boldsymbol{\beta} \\ -Q_i^T \, B\, H\, \boldsymbol{\beta} \end{bmatrix}. \qquad (10.143)$$

By construction, $E\,\boldsymbol{\beta} \in \mathcal{K}_0^h$. A reduced Schur complement system to determine $\boldsymbol{\beta}$ can be obtained by taking inner products of the first block row of (10.141) with $E\,\boldsymbol{\omega}$ for each $\boldsymbol{\omega} \in \mathbb{R}^{n_\beta}$ and substituting for $\mathbf{w}_1, \ldots, \mathbf{w}_l$ in terms of $\boldsymbol{\beta}$. Since $E\,\boldsymbol{\omega} \in \mathcal{K}_0^h$ this yields a linear system of the following form:

$$\mathcal{S}\boldsymbol{\beta} = \mathbf{r}, \text{ where } \begin{cases} \mathcal{S} = \quad E^T \, A\, E, \\ \mathbf{r} \; = - \; E^T \, A\, \mathbf{u}_*. \end{cases} \qquad (10.144)$$

Matrix $\mathcal{S}$ will be of size n_β and is easily seen to be symmetric *positive definite*. It should not be assembled explicitly, instead its action can be computed algorithmically using steps 1 and 2 described previously, using $\mathbf{u}_* = \mathbf{0}$.

A preconditioned CG method can be employed to solve (10.144). Once $\boldsymbol{\beta}$ has been determined, the components $\mathbf{w}_i$, $\mathbf{p}_i$ and $\boldsymbol{\mu}$ can be determined as in steps 1 and 2. The original algorithm of [GL14] employed weighted sums of mass matrices on Γ to precondition $\mathcal{S}$, where the weights were chosen based on the coefficient of $a(\cdot)$ on subdomains adjacent to a face (when $\Omega \subset \mathbb{R}^3$) or edge (when $\Omega \subset \mathbb{R}^2$). The resulting rate of convergence depends mildly on h, but is independent of h_0, due to inclusion of the natural coarse space.

Algorithm of [CO10]. We next outline an algorithm of [CO10] for solving system (10.119). It is based on the solution of a larger saddle point system *equivalent* to (10.119), but obtained by *decoupling* the velocity flux in different elements, and by introducing new *Lagrange multiplier* variables to enforce the matching of such decoupled velocity variables. Below, we outline the construction of such an extended system.

Let $\mathbf{v}_h \in \mathbb{R}^n$ denote the nodal vector associated with a finite element velocity function $\mathbf{v}_h(x) \in \mathbf{V}_h \cap H_0(div, \Omega)$. By construction, the *flux* of $\mathbf{v}_h(x)$ will be *continuous* across elements. If κ_l and κ_j are adjacent elements, the *flux* of the velocity $\mathbf{v}_h(x)$ on a face $\partial\kappa_l \cap \partial\kappa_j$ (when $\Omega \subset \mathbb{R}^3$, or an edge $\partial\kappa_l \cap \partial\kappa_j$ when $\Omega \subset \mathbb{R}^2$) must be the same when computed from element κ_l or κ_j. The algorithm of [CO10] introduces extended finite element velocity functions with *arbitrary* nodal values for its flux within each element. These extended velocity functions $\tilde{\mathbf{v}}_h(x)$ will have *discontinuous* flux across elements.

Let $\tilde{\mathbf{v}}_h \in \mathbb{R}^{\tilde{n}}$ denote the nodal vector associated with such an extended velocity function $\tilde{\mathbf{v}}_h(x)$. Here $\tilde{n} > n$ due to multiple flux values on each face or edge. For $i = 1, 2$, let $\tilde{B}_i$ denote a matrix of size $r_i \times \tilde{n}$ chosen as follows. Matrix $\tilde{B}_2$ should be chosen with entries in $\{-1, 0, 1\}$ so that $\tilde{B}_2 \tilde{\mathbf{v}}_h = \mathbf{0}$, requires the flux of $\tilde{\mathbf{v}}_h$ to match on each face (or edge) $\partial\kappa_l \cap \partial\kappa_j$. In this case, the extended velocity vectors in Kernel$(\tilde{B}_2)$ can be represented in the form:

$$\text{Kernel}(\tilde{B}_2) = \{E\,\mathbf{v}_h \,:\, \mathbf{v}_h \in \mathbb{R}^n\},$$

where E is a matrix of size $\tilde{n} \times n$ with entries in $\{0, 1\}$ such that the extended velocity $\tilde{\mathbf{v}}_h = E\,\mathbf{v}_h$ has continuous flux across elements, i.e., $\tilde{B}_2\,E\mathbf{v}_h = \mathbf{0}$. Matrix $\tilde{B}_1$ of size $r_1 \times \tilde{n}$ should be chosen so that $\tilde{B}_1\,E\,\mathbf{v}_h = B\,\mathbf{v}_h = \mathbf{f}_h$, corresponds to the original divergence constraints in (10.119), with $r_1 = m$.

Associated with each element $\kappa \in \Omega_h$, we shall let $\tilde{A}_\kappa$ denote the element velocity stiffness matrix corresponding to $\mathcal{A}(.,.)$ integrated on κ. Define a block diagonal matrix $\tilde{A}$ of size $\tilde{n}$ as $\tilde{A} = \text{blockdiag}\left(\tilde{A}_{\kappa_1}, \ldots, \tilde{A}_{\kappa_{n_e}}\right)$, where $\kappa_1, \ldots, \kappa_{n_e}$ is an enumeration of the elements in the triangulation Ω_h. By construction, it will hold that:

$$\mathbf{v}_h^T \left(E^T \tilde{A} E\right) \mathbf{v}_h = \mathbf{v}_h A \mathbf{v}_h,$$

where A is as in (10.119).

Introducing new Lagrange multipliers $\boldsymbol{\mu}_h \in \mathbb{R}^{r_2}$ to enforce the constraints $\tilde{B}_2\,\tilde{\mathbf{u}}_h = \mathbf{0}$ yields the following saddle point problem equivalent to (10.119):

$$\begin{bmatrix} \tilde{A} & \tilde{B}_1^T & \tilde{B}_2^T \\ \tilde{B}_1 & 0 & 0 \\ \tilde{B}_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}_h \\ \mathbf{p}_h \\ \boldsymbol{\mu}_h \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix}, \tag{10.145}$$

where $\mathbf{g}_1 = \mathbf{f}_h$ and $\mathbf{g}_2 = \mathbf{0}$. The Lagrange multipliers $\boldsymbol{\mu}_h$ will approximate the *pressure* on the faces (or edges) across elements. Eliminating $\tilde{\mathbf{u}}_h$ in system (10.145) yields the following Schur complement system for $\left(\boldsymbol{\mu}_h^T, \mathbf{p}_h^T\right)^T$:

$$\begin{bmatrix} \tilde{B}_1\tilde{A}^{-1}\tilde{B}_1^T & \tilde{B}_1\tilde{A}^{-1}\tilde{B}_2^T \\ \tilde{B}_2\tilde{A}^{-1}\tilde{B}_1^T & \tilde{B}_2\tilde{A}^{-1}\tilde{B}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{p}_h \\ \boldsymbol{\mu}_h \end{bmatrix} = \begin{bmatrix} -\mathbf{f}_h \\ \mathbf{0} \end{bmatrix}.$$

This system corresponds to a *nonconforming* finite element discretization of the original elliptic equation [AR4]. It can be solved using a CG method with domain decomposition preconditioner for nonconforming discretizations of elliptic equations, see [SA7, CO8, CO10, BR23]. The rate of convergence will be poly-logarithmic in h and robust with respect to $a(.)$. Once this system has been solved for $\mathbf{p}_h$ and $\boldsymbol{\mu}_h$, the extended velocity can be obtained as $\tilde{\mathbf{u}}_h = -\tilde{A}^{-1}\left(\tilde{B}_1^T\mathbf{p}_h + \tilde{B}_2^T\boldsymbol{\mu}_h\right)$ and $\mathbf{u}_h = \left(E^T E\right)^{-1} E^T\tilde{\mathbf{u}}_h$, see [CO10].

Algorithm of [EW8]. The algorithm of [EW8] employs the *Helmholtz* decomposition in two dimensions, which represents the space of divergence free functions in *two* dimensions in terms of the *curl* of a stream function:

$$\begin{aligned} \mathcal{K}_0 &= \{\mathbf{v}(x) \,:\, \nabla \cdot \mathbf{v}(x) = 0 \text{ in } \Omega \text{ and } \mathbf{n} \cdot \mathbf{v} = 0 \text{ on } \partial\Omega\} \\ &= \{\text{curl}\,(\psi) \,:\, \psi(x) = 0 \text{ on } \partial\Omega\}\,, \end{aligned} \tag{10.146}$$

where $\text{curl}(\psi)(x) \equiv (\psi_{x_2}(x), -\psi_{x_1}(x))$. To solve (10.119), the algorithm of [EW8] employs three steps. In the *first* step a discrete velocity $\mathbf{u}_*$ is determined such that $B\mathbf{u}_* = \mathbf{f}_h$, as in Algorithm 10.7.1. In the *second* step the unknown *divergence free* component $\mathbf{w}_h = \mathbf{u}_h - \mathbf{u}_*$ is sought using the curl representation (10.146). It is shown in [EW8] that the discrete divergence free subspace of various mixed finite element spaces in two dimensions can be represented as the *curl* of stream functions from standard conforming finite element spaces $V_h \subset H_0^1(\Omega)$. Suppose $\psi_1(x), \ldots, \psi_{n-m}(x)$ denotes a basis for $V_h \subset H_0^1(\Omega)$ such that:

$$\mathcal{K}_0^h = \text{Kernel}(B) = \left\{ curl(\psi) \,:\, \psi \in V_h \cap H_0^1(\Omega) \right\}.$$

Then, we may let C be a matrix of size $n \times (m - n)$ whose columns satisfy:

$$\text{curl}(\psi_i)(x) = \sum_{j=1}^{n} C_{ij}\, \phi_j(x).$$

The positive definite problem (10.140) to determine $\mathbf{w}_h$ can be expressed:

$$C^T A C\, \mathbf{w} = -C^T A \mathbf{u}_*,$$

where $\mathbf{w}_h = C\mathbf{w}$. In [EW8], a Schwarz method is employed to solve the resulting conforming finite element discretization. In the *third* step, the pressure is determined as in Algorithm 10.7.2. Importantly, representation (10.146) enables analysis of the convergence of Schwarz algorithms for mixed formulations of elliptic equations in two dimensions, see [EW8, MA32].

Other Algorithms. Other algorithms for mixed formulations include the $H(div)$ and $H(curl)$ algorithms of [AR6], Schwarz algorithms of [ME2] and a block matrix method of [LA11], based on $A_0 = A + \delta^{-1} B^T B$ and $S_0 = I$.

10.8 Applications to Optimal Control Problems

Optimal control problems [LI2] involving partial differential equations arise in various engineering applications, and yield large saddle point problems. The problem we shall consider involves an *output* variable $y(\cdot)$ which solves an elliptic or a parabolic equation with an *input* boundary data or forcing term $u(\cdot)$, referred to as the *control*. The problem of interest is to choose the control $u(\cdot)$ so that $y(\cdot)$ closely *matches* a given *target* output $y_*(\cdot)$. As an example, the output variable $y(\cdot)$ may represent the stationary temperature in an object Ω, the input $u(\cdot)$ may represent a heat source on the boundary or interior of Ω, and the target $y_*(\cdot)$ may represent a desired temperature distribution within Ω. Here, $y(.)$ will solve an elliptic equation with Neumann data u, and the control problem will seek $u(\cdot)$ such that $y(\cdot)$ closely matches the target $y_*(\cdot)$. In optimal control theory [LI2], the "matching" between the output $y(\cdot)$ and the target $y_*(\cdot)$ is measured by a *performance functional* $J_0(\cdot)$:

$$J_0(y) = \frac{1}{2} \|y - y_*\|^2,$$

where $\|\cdot\|$ denotes some chosen norm. To obtain an optimal choice of control $u(.)$, we may seek to *minimize* $J_0(y)$ subject to the *constraint* that $y(\cdot)$ solves the partial differential equation with input data $u(\cdot)$. However, this typically results in an *ill-posed* problem, requiring the inversion of a compact operator to determine u, with eigenvalues clustered around zero. Such ill-posedness can be handled by *Tikhonov regularization*, which replaces $J_0(y)$ by a modified functional $J(y,u) = J_0(y) + \frac{\alpha}{2}\|u\|^2$, with small $\alpha > 0$ to limit the magnitude of the regularization term. We may then seek the minimum of $J(y,u)$, subject to the original constraints involving $y(\cdot)$ and $u(\cdot)$, and Lagrange multipliers can be introduced to reformulate it as a saddle point problem.

Our discussion in this section will focus on saddle point methods to solve a discretized optimal control problem [LI2, BI4, BI5, HE4, HE5, PR3, MA36]. Our focus will be on block matrix methods based on the solution of a reduced Schur complement system, referred to as a "Hessian" system, for the control variable u, see [HA, BI4, BI5, MA36]. We shall omit discussion of duality based algorithms, since in some applications, the leading diagonal block in the resulting saddle point system can be *singular*, without augmentation of the Lagrangian. In Chap. 10.8.1, we consider an elliptic optimal control problem with Neumann controls, and describe preconditioned Hessian algorithms that converge uniformly with respect to the mesh size h and the regularization parameter α. In Chap. 10.8.2, we describe a parabolic optimal control problem with control in the forcing term, and describe preconditioned Hessian algorithms that converge uniformly with respect to the mesh size h and time step τ. For alternative algorithms, see [BI4, BI5, HE4, HE5, PR3, MA36, GO5].

10.8.1 Elliptic Optimal Control Problems

In an elliptic control problem, the variable $y(\cdot)$ will solve an elliptic equation. More specifically, we shall assume that $y(\cdot)$ solves an elliptic equation on $\Omega \subset \mathbb{R}^d$ with Neumann control data $u(\cdot) \in L^2(\Gamma)$ on a segment $\Gamma \subset \partial\Omega$:

$$\begin{cases} -\Delta\, y(x) + \sigma\, y(x) = f(x), \text{ in } \Omega \\ \dfrac{\partial y(x)}{\partial n} = u(x), \text{ on } \Gamma \\ y(x) = 0, \qquad \text{on } \partial\Omega \setminus \Gamma, \end{cases} \tag{10.147}$$

where $\sigma > 0$ is a given parameter. The control data $u(\cdot)$ parameterizes $y(\cdot)$ in (10.147). So, given a target $y_*(\cdot) \in L^2(\Omega)$, we may seek to optimally match $y(\cdot)$ with $y_*(\cdot)$ by choosing $u(\cdot)$ which minimizes the functional:

$$J_0(y) = \frac{1}{2} \|y - y_*\|^2_{L^2(\Omega)}. \tag{10.148}$$

However, as discussed later, minimization of $J_0(\cdot)$ amongst solutions to (10.147) will be *ill-posed.* Instead, minimization of the regularized functional:

$$J(y,u) \equiv \frac{1}{2} \left(\|y - y_*\|^2_{L^2(\Omega_0)} + \alpha_1 \|u\|^2_{L^2(\Gamma)} + \alpha_2 \|u\|^2_{(H^{1/2}_{00}(\Gamma))'} \right), \tag{10.149}$$

for α_1, $\alpha_2 \geq 0$ will yield a well posed problem, provided either $\alpha_i > 0$. In the following sub-sections, we formulate the optimal control problem, its finite element discretization, and resulting saddle point system. We derive the reduced symmetric positive definite Hessian system for determining the discretized control $u(\cdot)$ and describe a preconditioner yielding a rate of convergence uniform with respect to the mesh and regularization parameters. Our discussion will closely follow [MA36, GO5]. For a discussion of alternative iterative solvers, the reader is referred to [BI4, BI5, HE4, HE5, PR3].

The Constrained Minimization Problem. Given $f(.) \in L^2(\Omega)$, define the *constraint set* $\mathcal{V}_f$ of solutions to elliptic problem (10.147) as:

$$\mathcal{V}_f \equiv \{(y,u) : \text{equation (10.147) holds}\}. \tag{10.150}$$

Given a target function $y_* \in L^2(\Omega)$, the *regularized* optimal control problem will seek to minimize the functional $J(\cdot,\cdot)$ within $\mathcal{V}_f$:

$$J(y,u) = \min_{(\tilde{y},\tilde{u}) \in \mathcal{V}_f} J(\tilde{y},\tilde{u}). \tag{10.151}$$

The constraint set $\mathcal{V}_f$ will be *closed* in $H^1_\Gamma(\Omega) \times (H^{1/2}_{00}(\Gamma))'$, where:

$$H^1_\Gamma(\Omega) \equiv \{w \in H^1(\Omega) : w = 0 \text{ on } \partial\Omega \setminus \Gamma\}, \tag{10.152}$$

and $(H^{1/2}_{00}(\Gamma))'$ is the dual space of $H^{1/2}_{00}(\Gamma) = [L^2(\Gamma), H^1_0(\Gamma)]_{1/2}$. As mentioned before, the minimization of $J(\cdot)$ within $\mathcal{V}_f$ will be *ill-posed* if $\alpha_1 = \alpha_2 = 0$, and the reason for this will be indicated later.

Remark 10.86. Recall that the dual Sobolev norm $\|u\|_{(H^{1/2}_{00}(\Gamma))'}$ is defined as:

$$\|u\|_{(H^{1/2}_{00}(\Gamma))'} \equiv \sup_{v \in H^{1/2}_{00}(\Gamma)} \frac{\int_\Gamma u\, v\, ds_x}{\|v\|_{H^{1/2}_{00}(\Gamma)}},$$

where $H^{1/2}_{00}(\Gamma) = \left[L^2(\partial\Omega), H^1_0(\Gamma)\right]_{1/2}$.

Remark 10.87. The regularization term in $J(.,.)$ alters the original functional $J_0(.)$ and the intended matching between $y(\cdot)$ and $y_*(\cdot)$ will also be altered. As a result, it will be important that α_1, $\alpha_2 \geq 0$ be *small* parameters in applications, to ensure that the minimization determines $y(\cdot)$ close to $y_*(\cdot)$. The traditional choice is $\alpha_1 > 0$ and $\alpha_2 = 0$. However, the choice $\alpha_1 = 0$ and $\alpha_2 > 0$ yields a weaker regularization term.

Weak Formulation and Discretization. Given $f(\cdot) \in L^2(\Omega)$, the constraint set $\mathcal{V}_f \subset \mathcal{V} \equiv H^1_\Gamma(\Omega) \times (H^{1/2}_{00}(\Gamma))'$ can be described weakly as:

$$\mathcal{V}_f \equiv \left\{(y,u) \in \mathcal{V} \,:\, \mathcal{A}(y,w) - < u, w > = (f,w), \;\; \forall w \in H^1_\Gamma(\Omega)\right\}, \quad (10.153)$$

where the forms are defined by:

$$\begin{cases} \mathcal{A}(u,w) \equiv \int_\Omega (\nabla u \cdot \nabla w + \sigma\, u\, w)\, dx, & \text{for } u, w \in H^1_\Gamma(\Omega) \\ (f,w) \equiv \int_\Omega f(x)\, w(x)\, dx, & \text{for } w \in H^1_\Gamma(\Omega) \\ < u, w > \equiv \int_\Gamma u(x)\, w(x)\, ds_x, & \text{for } u \in (H^{1/2}_{00}(\Gamma))',\, w \in H^{1/2}_{00}(\Gamma). \end{cases} \quad (10.154)$$

To ensure well posedness of (10.151), saddle point theory [GI3] requires an *inf-sup* condition to hold (in appropriately chosen norms), and requires $J(.,.)$ to be *coercive* within $\mathcal{V}_0$. The *inf-sup* condition can be shown to hold for this problem, however, $J(.,.) = J_0(\cdot)$ for $\alpha_1 = \alpha_2 = 0$ will not be *coercive* within $\mathcal{V}_0$, since the $L^2(\Omega)$ norm in $J_0(\cdot)$ is weaker than the $H^1(\Omega)$ norm for $y(\cdot)$. However, if $\alpha_2 > 0$ and $\alpha_1 = 0$, then for $y(\cdot) \in \mathcal{V}_0$, elliptic regularity theory for harmonic functions will yield *coercivity* of $J(.,.)$ within $\mathcal{V}_0$. Note that if $\alpha_2 = 0$ and $\alpha_1 > 0$, the term $\|u\|_{L^2(\Gamma)}$ is not strictly defined for $u \in (H^{1/2}_{00}(\Gamma))'$, however, it will be defined for finite element discretizations.

The solution (y,u) of (10.151) can be obtained from the *saddle point* (y,u,p) of the Lagrangian $\mathcal{L}(\cdot,\cdot,\cdot)$ with Lagrange multiplier $p \in H^1_\Gamma(\Omega)$:

$$\mathcal{L}(y,u,p) \equiv J(y,u) + (\mathcal{A}(y,p) - < u, p > - (f,p)), \quad (10.155)$$

The appropriate function space is $(y,u,p) \in H^1_\Gamma(\Omega) \times (H^{1/2}_{00}(\Gamma))' \times H^1_\Gamma(\Omega)$. A finite element discretization of the saddle point problem for $\mathcal{L}(.,.,.)$ in (10.155) can be obtained by discretizing its weak formulation [GI3], using finite element subspaces of $H^1_\Gamma(\Omega) \times (H^{1/2}_{00}(\Gamma))' \times H^1_\Gamma(\Omega)$, as outlined next. A discretization of $\alpha_2 \|u\|^2_{(H^{1/2}_{00}(\Gamma))'}$ will be based on an equivalent expression.

Given a triangulation $\tau_h(\Omega)$ of Ω, let $V_h(\Omega) \subset H^1_\Gamma(\Omega)$ denote a finite element space with restriction $V_h(\Gamma) \subset L^2(\Gamma)$ to Γ. A discretization of the saddle point problem for $\mathcal{L}(.,.,.)$ will seek an approximation $(y_h,\, u_h,\, p_h)$ of $(y,\, u, p)$ within $V_h(\Omega) \times V_h(\Gamma) \times V_h(\Omega)$. Given the standard finite element nodal basis $\{\phi_1(x), \ldots, \phi_n(x)\}$ and $\{\psi_1(x), \ldots, \psi_m(x)\}$ for $V_h(\Omega)$ and $V_h(\Gamma)$, respectively, let $\mathbf{y}$, $\mathbf{u}$, $\mathbf{p}$ be the nodal vectors associated with y_h, u_h, p_h:

$$y_h(x) = \sum_{i=1}^{n} \mathbf{y}_i\, \phi_i(x), \quad u_h(x) = \sum_{j=1}^{m} \mathbf{u}_j\, \psi_j(x), \quad p_h(x) = \sum_{l=1}^{n} \mathbf{p}_l\, \phi_l(x). \quad (10.156)$$

The following matrices M, A and Q shall be employed:

$$\begin{cases} M_{ij} \equiv \int_\Omega \phi_i(x)\, \phi_j(x)\, dx, & \text{for } 1 \le\, i\,, j\, \le n \\ A_{ij} \equiv \int_\Omega (\nabla\phi_i(x) \cdot \nabla\phi_j(x) + \sigma\, \phi_i(x)\, \phi_j(x))\, dx, & \text{for } 1 \le\, i\,, j\, \le n \\ Q_{ij} \equiv \int_{\partial\Omega} \psi_i(x)\, \psi_j(x)\, ds_x, & \text{for } 1 \le\, i\,, j\, \le m, \end{cases} \quad (10.157)$$

where M denotes the mass matrix of size n on Ω, and A denotes the stiffness matrix of size n with Neumann boundary conditions on Γ, and Q denotes the mass matrix of size m on Γ. It will hold that $M = M^T > 0$, $Q = Q^T > 0$, and when $\sigma > 0$, then $A = A^T > 0$. We shall order the nodal unknowns in $\mathbf{y}$ and $\mathbf{p}$ so that the nodes in the *interior* of Ω are ordered prior to the nodes on Γ. This will yield a partition $\mathbf{y} = \left(\mathbf{y}_I^T, \mathbf{y}_\Gamma^T\right)^T$ and $\mathbf{p} = \left(\mathbf{p}_I^T, \mathbf{p}_\Gamma^T\right)^T$. Based on this, we block partition A and define matrix B of size $n \times m$ as follows:

$$A = \begin{bmatrix} A_{II} & A_{I\Gamma} \\ A_{I\Gamma}^T & A_{\Gamma\Gamma} \end{bmatrix} \quad \text{and} \quad B \equiv \begin{bmatrix} 0 \\ -Q \end{bmatrix}, \quad B^T \equiv \begin{bmatrix} 0 & -Q^T \end{bmatrix}. \tag{10.158}$$

The following discrete forcing vectors will also be employed:

$$\begin{cases} (\mathbf{f}_1)_i = \int_\Omega y_*(x)\, \phi_i(x)\, dx, & \text{for} \quad 1 \le i \le n \\ (\mathbf{f}_2)_i = 0, & \text{for} \quad 1 \le i \le m \\ (\mathbf{f}_3)_i = \int_\Omega f(x)\, \phi_i(x)\, dx, & \text{for} \quad 1 \le i \le n. \end{cases} \tag{10.159}$$

Then, using the above, the *discrete* performance functional $J_h(\mathbf{y}, \mathbf{u})$ and the *discrete* constraint set $\mathcal{V}_f^h$ can be expressed as:

$$\begin{cases} J_h(\mathbf{y}, \mathbf{u}) = \frac{1}{2}\left((\mathbf{y} - \mathbf{y}_*)^T M (\mathbf{y} - \mathbf{y}_*) + \alpha_1\, \mathbf{u}^T Q \mathbf{u} + \alpha_2\, \mathbf{u}^T B^T A^{-1} B \mathbf{u}\right) \\ \qquad \mathcal{V}_f^h = \{(\mathbf{y}, \mathbf{u}) \,:\, A\,\mathbf{y} + B\,\mathbf{u} \,=\, \mathbf{f}_3\}, \end{cases}$$

where $\mathbf{y}_* = M^{-1}\mathbf{f}_1$, and the saddle point discretization of (10.151) will be:

$$\begin{bmatrix} M & 0 & A^T \\ 0 & G & B^T \\ A & B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \end{bmatrix}, \tag{10.160}$$

where matrix $G \equiv \alpha_1\, Q \,+\, \alpha_2\, (B^T A^{-1} B)$.

Given a finite element function $u_h(\cdot)$ on Γ, with associated nodal vector $\mathbf{u}$, it can be verified that $A^{-1}B\mathbf{u}$ is the nodal vector corresponding to the discrete harmonic extension of Neumann data u_h into Ω. Thus, $\mathbf{u}^T B^T A^{-1} B \mathbf{u}$ will denote the A-energy of the discrete harmonic extension of $\mathbf{u}$. The block matrix structure of A and B in (10.158), and the block structure of A^{-1}, will yield $B^T A^{-1} B = Q^T S^{-1} Q$, where $S = (A_{\Gamma\Gamma} - A_{I\Gamma}^T A_{II}^{-1} A_{I\Gamma})$, see Lemma 10.88. The discrete extension theorem and elliptic regularity theory will yield the following spectral equivalence (denoted $\asymp$), see Remark 10.90 and [MA36]:

$$\mathbf{u}^T \left(B^T A^{-1} B\right) \mathbf{u} \,=\, \mathbf{u}^T \left(Q^T S^{-1} Q\right) \mathbf{u} \asymp \|u_h\|^2_{(H_{00}^{1/2}(\partial\Omega))'}$$

As a consequence of the above, the regularization term can be expressed as:

$$\alpha_1\, \|u_h\|^2_{L^2(\Gamma)} + \alpha_2\, \|u_h\|^2_{(H_{00}^{1/2}(\Gamma))'} \asymp \mathbf{u}^T \left(\alpha_1\, Q + \alpha_2\, (B^T A^{-1} B)\right) \mathbf{u},$$

and this is what motivated the definition of $G \equiv \alpha_1 Q + \alpha_2 \left(B^T A^{-1} B\right)$. Solvability of (10.160), with norm bounds for $\|\mathbf{y}\|_A$, $\|\mathbf{u}\|_{QS^{-1}Q}$ and $\|\mathbf{p}\|_A$, can be analyzed using saddle point theory [GI3]. It requires the *coercivity* of $\left(\mathbf{y}^T M\mathbf{y} + \mathbf{u}^T G\mathbf{u}\right)$ within the constraint set $\mathcal{V}_0^h$, and that the bilinear form $\mathbf{p}^T\left(A\mathbf{y} + B\mathbf{u}\right)$ satisfy an *inf-sup* condition, both in the appropriate norms. When $\alpha_1 > 0$ or $\alpha_2 > 0$, $\left(\mathbf{y}^T M\mathbf{y} + \mathbf{u}^T G\mathbf{u}\right)$ will be coercive within $\mathcal{V}_0^h$, by discrete elliptic regularity theory, and the *inf-sup* condition will hold trivially for $\mathbf{p}^T\left(A\mathbf{y} + B\mathbf{u}\right)$, by discrete elliptic regularity theory.

Hessian System for u. The algorithm we describe for solving (10.160) will be based on the solution of a reduced Schur complement system for the control variable $\mathbf{u}$. When A is non-singular, solving the third block row in (10.160) formally yields $\mathbf{y} = A^{-1}\left(\mathbf{f}_3 - B\,\mathbf{u}\right)$. Solving the first block row for $\mathbf{p}$ yields $\mathbf{p} = A^{-T}\left(\mathbf{f}_1 - MA^{-1}\mathbf{f}_3 + MA^{-1}B\mathbf{u}\right)$. Substituting these into the second block row yields the following reduced system, referred to as the Hessian system, for the control variable $\mathbf{u}$:

$$\begin{cases} C\,\mathbf{u} = \tilde{\mathbf{f}}_2, \quad \text{where} \\ \;C \equiv \left(G + B^T A^{-T} M A^{-1} B\right) \\ \;\tilde{\mathbf{f}}_2 \equiv \left(\mathbf{f}_2 - B^T A^{-T}\mathbf{f}_1 + B^T A^{-T} M A^{-1}\mathbf{f}_3\right). \end{cases} \tag{10.161}$$

The Hessian matrix $C = C^T > 0$ will be symmetric and positive definite of size m, and system (10.161) can be solved using a PCG algorithm. Each matrix vector product with C requires the action of A^{-T} and A^{-1} (two applications of A^{-1} since $A = A^T$). If the action of A^{-1} is computed iteratively, this will result in *double iteration*, with inner and outer iterations. Once $\mathbf{u}$ has been determined by solving (10.161), we can determine $\mathbf{y}$ and $\mathbf{p}$ by solving:

$$A\,\mathbf{y} = \left(\mathbf{f}_3 - B\mathbf{u}\right) \quad \text{and} \quad A^T\,\mathbf{p} = \left(\mathbf{f}_1 - MA^{-1}\mathbf{f}_3 + MA^{-1}B\mathbf{u}\right). \tag{10.162}$$

In the following, we describe the spectral properties of the Hessian matrix C, and formulate a preconditioner C_0 which yields a rate of convergence independent of the mesh size h and α_1, α_2.

Since the Hessian C is a weighted sum of three matrices:

$$C = \alpha_1\,Q + \alpha_2\left(B^T A^{-1} B\right) + \left(B^T A^{-T} M A^{-1} B\right), \tag{10.163}$$

its properties will depend on α_1, α_2. The block structures of A and B yield:

$$A^{-1} = \begin{bmatrix} A_{II}^{-1} + A_{II}^{-1} A_{I\Gamma} S^{-1} A_{I\Gamma}^T A_{II}^{-1} & -A_{II}^{-1} A_{I\Gamma} S^{-1} \\ -S^{-1} A_{I\Gamma}^T A_{II}^{-1} & S^{-1} \end{bmatrix}, \quad B^T = \left[0 \;\; -Q^T\right], \tag{10.164}$$

where $S = \left(A_{\Gamma\Gamma} - A_{I\Gamma}^T A_{II}^{-1} A_{I\Gamma}\right)$. This yields $A^{-1}B = -E\,S^{-1}\,Q$ where:

$$E = \begin{bmatrix} -A_{II}^{-1} A_{I\Gamma} \\ I \end{bmatrix} \tag{10.165}$$

denotes the discrete A-harmonic extension of Dirichlet data on Γ into Ω. Also:

$$\begin{aligned} B^T A^{-1} B &= Q^T S^{-1} Q \\ B^T A^{-T} M A^{-1} B &= Q^T S^{-T} E^T M E S^{-1} Q. \end{aligned} \tag{10.166}$$

The next result describes a spectral equivalence for $(B^T A^{-T} M A^{-1} B)$.

Lemma 10.88. *Let $\Omega \subset R^d$ be a convex polygonal domain and let $\Gamma \subset \partial\Omega$ be smooth. The following equivalence will hold independent of h, α_1 and α_2:*

$$B^T A^{-T} M A^{-1} B \asymp Q^T S^{-T} Q^T S^{-1} Q S^{-1} Q, \tag{10.167}$$

where $S = (A_{\Gamma\Gamma} - A_{I\Gamma}^T A_{II}^{-1} A_{I\Gamma})$.

Proof. We shall employ the property that if matrices X and X_0 of size n satisfy $c_1 X_0 \le X \le c_2 X_0$ in the sense of quadratic forms, then:

$$c_1 (Y^T X_0 Y) \le (Y^T X Y) \le c_2 (Y^T X_0 Y),$$

will hold for any matrix Y of size $n \times p$. Thus, if $E^T M E \asymp Q^T S^{-1} Q$, we may substitute $X = E^T M E$, $X_0 = Q S^{-1} Q$, and $Y = S^{-1} Q$ into (10.166) to obtain (10.167). To verify $E^T M E \asymp Q^T S^{-1} Q$, given a finite element Dirichlet data v_h on Γ, let $E v_h$ denote its discrete harmonic extension into Ω. As a result of $H^2(\Omega)$ regularity for the Dirichlet problem, the equivalence:

$$\|E v_h\|_{L^2(\Omega)}^2 \asymp \|v_h\|_{(H_{00}^{1/2}(\Gamma))'}^2$$

will hold [PE2], where the dual norm is defined in Remark 10.86. If $\mathbf{v}_\Gamma$ and $E\mathbf{v}_\Gamma$ denote the nodal vectors associated with v_h and Ev_h, the equivalence of [PE2] can be expressed in matrix terms, see Remark 10.90, as:

$$\mathbf{v}_\Gamma^T \left(E^T M E\right) \mathbf{v}_\Gamma \asymp \mathbf{v}_\Gamma^T \left(Q^T S^{-1} Q\right) \mathbf{v}_\Gamma$$

Thus $E^T M E \asymp Q S^{-1} Q$ and the desired equivalence follows. □

Remark 10.89. As a corollary, we obtain (denoting $G = \alpha_1 Q + \alpha_2 (Q^T S^{-1} Q)$):

$$C = G + Q^T S^{-T} E^T M E S^{-1} Q \asymp G + (Q^T S^{-1} Q^T S^{-1} Q S^{-1} Q). \tag{10.168}$$

For α_1, α_2 "large", we will obtain $\lambda_{\min}(G) \ge \lambda_{\max}(Q^T S^{-T} E^T M E S^{-1} Q)$ and $\mathrm{cond}(G, C) \le 2$. For α_1, α_2 "small", $\lambda_{\max}(G) \le \lambda_{\min}(Q^T S^{-T} E^T M E S^{-1} Q)$ and $\mathrm{cond}(Q^T S^{-T} E^T M E S^{-1} Q, C) \le 2$. In the latter case, matrix C will be ill-conditioned with a condition number of $O(h^{-3})$. We shall later consider preconditioners for C which are uniformly effective for h, α_1, α_2.

Remark 10.90. Let $v_h, w_h \in V_h(\Gamma)$ with associated nodal vectors $\mathbf{v}$, $\mathbf{w}$ with the notation $\|\mathbf{w}\|_X = \left(\mathbf{w}^T X \mathbf{w}\right)^{1/2}$ for $X = X^T > 0$. Since S will be spectrally equivalent to matrix generating the fractional Sobolev inner product, there should be $c_1 > 0$ independent of h such that $\|w_h\|_{H_{00}^{1/2}(\Gamma)} \leq c_1 \|\mathbf{w}\|_S$. Then, the following bound will hold:

$$\begin{aligned}
\|\mathbf{v}\|_{Q^T S^{-1} Q} = \sup_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T Q \mathbf{v}}{\|\mathbf{w}\|_S} &\leq c_1 \sup_{w_h \in V_h(\Gamma)\backslash\{0\}} \frac{\int_\Gamma v_h\, w_h\, ds_x}{\|w_h\|_{H_{00}^{1/2}(\Gamma)}} \\
&\leq c_1 \sup_{w \in H_{00}^{1/2}(\Gamma)\backslash 0} \frac{\int_\Gamma v_h\, w\, ds_x}{\|w\|_{H_{00}^{1/2}(\Gamma)}} \\
&= c_1 \|v\|_{(H_{00}^{1/2}(\Gamma))'}
\end{aligned}$$

If $P_h w$ denotes the $L^2(\Gamma)$-orthogonal projection of w onto $V_h(\Gamma)$, then this projection will satisfy $\|P_h w\|_{H_{00}^{1/2}(\Gamma)} \leq c_2 \|w\|_{H_{00}^{1/2}(\Gamma)}$ for some $c_2 > 0$ independent of h (by interpolation and stability of the $L^2(\Gamma)$-projection P_h in the $H_0^1(\Gamma)$ and $L^2(\Gamma)$ norms [BR21]). As a result, the reverse bound should hold:

$$\begin{aligned}
\|v_h\|_{(H_{00}^{1/2}(\Gamma))'} &= \sup_{w \in H_{00}^{1/2}(\Gamma)\backslash\{0\}} \frac{\int_\Gamma v_h\, w\, ds_x}{\|w\|_{H_{00}^{1/2}(\Gamma)}} \\
&= \sup_{w \in H_{00}^{1/2}(\Gamma)\backslash\{0\}} \frac{\int_\Gamma v_h\, P_h w\, ds_x}{\|w\|_{H_{00}^{1/2}(\Gamma)}} \\
&\leq c_2 \sup_{w \in H_{00}^{1/2}(\Gamma)\backslash\{0\}} \frac{\int_\Gamma v_h\, P_h w\, ds_x}{\|P_h w\|_{H_{00}^{1/2}(\Gamma)}}.
\end{aligned}$$

Combining both bounds yields $\|\mathbf{v}\|_{QS^{-1}Q} \asymp \|v_h\|_{(H_{00}^{1/2}(\Gamma))'}$.

Remark 10.91. When $\alpha_1 = \alpha_2 = 0$, the *ill-posedness* of system (10.151) can be understood *heuristically* by studying the Hessian system (10.161). Under $H^2(\Omega)$ regularity assumptions for the Dirichlet problem on Ω, the bilinear form corresponding to $C = B^T A^{-T} M A^{-1} B$ will be a compact operator with eigenvalues clustered around zero, coercive for $u_h \in H^{-3/2}(\Gamma)$. Formally, its inverse C^{-1} will be a bounded map for a forcing term with $H^{3/2}(\Gamma)$ regularity. However, $\tilde{\mathbf{f}}_2 = \left(\mathbf{f}_2 - B^T A^{-T} \mathbf{f}_1 + B^T A^{-T} M A^{-1} \mathbf{f}_3\right)$ can be verified to represent a term with only $H^{1/2}(\Gamma)$ regularity, resulting in a formal solution $C^{-1}\tilde{\mathbf{f}}_2$ that is *unbounded* as $h \to 0$. However, when $\alpha_1 > 0$, the bilinear form corresponding to C will be bounded in $L^2(\Gamma)$, and when $\alpha_2 > 0$, it will be bounded for forcing terms with $H^{1/2}(\Gamma)$ regularity, yielding a bounded solution.

Remark 10.92. Inclusion of a regularization term alters the original ill-posed minimization problem, yielding a bounded solution, instead of an unbounded solution. Its effects can be understood *heuristically* by studying the minimization of the following least squares functional:

$$F(\mathbf{x}) = \frac{1}{2}\|H\mathbf{x} - \mathbf{b}\|^2$$

where H is a rectangular or singular matrix of dimension $m \times n$ with singular value decomposition $H = U\Sigma V^T$. A minimum of the functional $F(\cdot)$ will occur at $\mathbf{x}_* = H^\dagger \mathbf{b} = V\Sigma^\dagger U^T \mathbf{b}$. However, when H has a non-trivial null space, there will be an affine space of minima. For instance, if N is a matrix of dimension $n \times k$ whose columns span the null space of H, with $\text{Range}(N) = \text{Kernel}(H)$, then, the general minimum of $F(\cdot)$ will be $\mathbf{x}_* + N\boldsymbol{\beta}$ for any vector $\boldsymbol{\beta} \in \mathbb{R}^k$. We shall consider two alternative regularization terms:

$$F_1(\mathbf{x}) = F(\mathbf{x}) + \frac{\alpha}{2}\|P_N \mathbf{x}\|^2 \quad \text{and} \quad F_2(\mathbf{x}) = F(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}\|^2,$$

where P_N denotes the Euclidean orthogonal projection onto the null space N. The minimum of $F_1(\cdot)$ will occur at $\mathbf{x}_* = H^\dagger \mathbf{b}$ and it will be unique for $\alpha > 0$. The minimum of $F_2(\cdot)$ will solve $(H^T H + \alpha\, I)\mathbf{x} = H^T \mathbf{b}$. Using the singular value decomposition of H, we will obtain $\mathbf{x} = V\left(\Sigma^T \Sigma + \alpha\, I\right)^{-1} \Sigma^T U^T \mathbf{b}$ as the unique solution to the regularized functional $F_2(\cdot)$. The ith diagonal entry of $\left(\Sigma^T \Sigma + \alpha\, I\right)^{-1} \Sigma^T$ will be $\sigma_i/(\sigma_i^2 + \alpha)$, with $\sigma_i/(\sigma_i^2+\alpha) \to 1/\sigma_i$ as $\alpha \to 0^+$ when $\sigma_i > 0$, while if $\sigma_i = 0$, then $\sigma_i/(\sigma_i^2 + \alpha) = 0$. Thus, $\mathbf{x} \to \mathbf{x}_* = H^\dagger \mathbf{b}$ as $\alpha \to 0^+$. In our applications, $H^T H$ will correspond to $(B^T A^{-T} M A^{-1} B)$, while $F(\mathbf{x})$ will correspond to $J(\mathbf{y}(\mathbf{u}), \mathbf{u})$ with $\mathbf{x}$ corresponding to $\mathbf{u}$.

Remark 10.93. The regularization parameter $\alpha > 0$ must be appropriately small. When matrix H arises from the discretization of an *ill posed* problem, its singular values will *cluster* around 0, and the choice of parameter $\alpha > 0$ must balance the accuracy of the modes associated with the larger singular values of H and dampen the modes associated with the smaller singular values of H (resulting in a bounded solution in our applications).

Solution of the Hessian System $C\mathbf{u} = \tilde{\mathbf{f}}_2$. Two alternative approaches can be employed to *iteratively* solve the Hessian system. In the first approach, we solve $C\mathbf{u} = \tilde{\mathbf{f}}_2$ using a PCG algorithm with preconditioner $C_0 = C_0^T > 0$. In the second approach, the control problem for Neumann data $\mathbf{u}$ is transformed into an equivalent control problem for associated Dirichlet data $\mathbf{v} = S^{-1} Q\mathbf{u}$. Since $C = Q^T S^{-T}\, D\, S^{-1} Q$, the Dirichlet control $\mathbf{v}$ can be shown to solve:

$$D\,\mathbf{v} = \mathbf{g}, \quad \text{where } D = \alpha_1\,(S^T Q^{-1} S) + \alpha_2\, S + E^T M E, \quad \mathbf{g} = S^T Q^{-T}\, \tilde{\mathbf{f}}_2. \tag{10.169}$$

Here $D = D^T > 0$. To obtain $\mathbf{u}$, solve $D\,\mathbf{v} = \mathbf{g}$ using a PCG algorithm with preconditioner $D_0 = D_0^T > 0$, and compute $\mathbf{u} = Q^{-1} S\mathbf{v}$. The preconditioners we describe for D require weaker regularity assumptions than those for C.

In both approaches, the matrices C and D are the sums of products of matrices, and caution must be exercised when formulating a preconditioner for the products of matrices. Consider matrices $X = X^T > 0$ and $Y = Y^T > 0$ of size n, with effective preconditioners X_0 and Y_0 for X and Y, respectively.

Then, $Y_0^T X_0 Y_0$ need not be an effective preconditioner for $Y^T X Y$. Indeed, as an example, for $a > 0$ choose $X = X_0 = I$ with:

$$\begin{aligned} Y &= \begin{bmatrix} 2 & a^{\frac{1}{2}} \\ a^{\frac{1}{2}} & 2a \end{bmatrix}, Y_0 = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix}, Y^T Y = \begin{bmatrix} 4+a & 2(a^{\frac{1}{2}} + a^{\frac{3}{2}}) \\ 2(a^{\frac{1}{2}} + a^{\frac{3}{2}}) & a + 4a^2 \end{bmatrix}, \\ Y_0^T Y_0 &= \begin{bmatrix} 1 & 0 \\ 0 & a^2 \end{bmatrix}. \end{aligned}$$

Then, $(Y^T X Y) = Y^T Y$, $(Y_0^T X_0 Y_0) = Y_0^T Y_0$, with $\operatorname{cond}(X, X_0) = 1$ and $\operatorname{cond}(Y, Y_0) \le 3$, yet $\operatorname{cond}(Y^T Y, Y_0^T Y_0)$ depends on a and can be arbitrarily large. As another example, choose $X_0 = X$ and $Y_0 = I$ with:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix}, \quad Y = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, Y^T X Y = \begin{bmatrix} 4+a & 2+2a \\ 2+2a & 1+4a \end{bmatrix}, \quad Y_0^T X_0 Y_0 = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix},$$

Then, $\operatorname{cond}(X, X_0) = 1$ and $\operatorname{cond}(Y, Y_0) \le 3$, yet $\operatorname{cond}(Y^T X Y, Y_0^T X_0 Y_0)$ is dependent on a and can be arbitrarily large. Despite the above examples, the matrix $(Y^T X Y)$ will be spectrally equivalent to $(Y_0^T X_0 Y_0)$ under additional assumptions on X, Y, X_0 and Y_0, as noted below.

- Replacing the "inner" matrix X by a preconditioner X_0 will yield the following bounds trivially, if $Y^T X_0 Y$ is used as a preconditioner for $Y^T X Y$:
$$c_1 \le \frac{\mathbf{v}^T X \mathbf{v}}{\mathbf{v}^T X_0 \mathbf{v}} \le c_2 \implies c_1 \le \frac{\mathbf{v}^T (Y^T X Y) \mathbf{v}}{\mathbf{v}^T (Y^T X_0 Y) \mathbf{v}} \le c_2.$$
However, computing the action of $(Y^T X_0 Y)^{-1}$ requires the action of Y^{-1}.
- When the matrices X, Y, X_0 and Y_0 *commute*, the following bounds will hold trivially if $Y_0^T X_0 Y_0$ is used as a preconditioner for $Y^T X Y$:
$$\begin{cases} c_1 \le \dfrac{\mathbf{v}^T X \mathbf{v}}{\mathbf{v}^T X_0 \mathbf{v}} \le c_2 \\ d_1 \le \dfrac{\mathbf{v}^T Y \mathbf{v}}{\mathbf{v}^T Y_0 \mathbf{v}} \le d_2 \end{cases} \implies c_1 \, d_1^2 \le \frac{\mathbf{v}^T (Y^T X Y) \mathbf{v}}{\mathbf{v}^T (Y_0^T X_0 Y_0) \mathbf{v}} \le c_2 \, d_2^2.$$
However, such commutativity assumptions hold only rarely in practice.
- If matrix $X \asymp K_0^{\beta}$ and $(Y^T K_0^{\beta} Y) \asymp K_0^{\beta+2\alpha}$, then we may employ $K_0^{\beta+2\alpha}$ as a preconditioner for $(Y^T X Y)$. Indeed, the following bounds will hold:
$$\begin{cases} c_1 \le \dfrac{\mathbf{v}^T X \mathbf{v}}{\mathbf{v}^T K_0^{\beta} \mathbf{v}} \le c_2 \\ d_1^2 \le \dfrac{\mathbf{v}^T (Y^T K_0^{\beta} Y) \mathbf{v}}{\mathbf{v}^T (K_0^{\beta+2\alpha}) \mathbf{v}} \le d_2^2 \end{cases} \implies c_1 \, d_1^2 \le \frac{\mathbf{v}^T (Y^T X Y) \mathbf{v}}{\mathbf{v}^T (K_0^{\beta+2\alpha}) \mathbf{v}} \le c_2 \, d_2^2.$$
Such properties may hold under regularity assumptions. Importantly, there must be efficient ways to compute the action of the inverse of $K_0^{\beta+2\alpha}$.

Motivated by these properties, we shall describe preconditioners Q_0 and S_0 for Q and S, respectively, and formulate preconditioners for C and D.

FST Based Preconditioners. When $\Omega \subset \mathbb{R}^2$, we may precondition the matrices Q and S by *tridiagonal* matrices Q_0 and S_0, which are simultaneously diagonalized by the fast sine transform (FST). Let Q_0 denote the mass matrix associated with a uniform mesh of size h on Γ and let L_0 denote the discrete Laplace-Beltrami matrix associated with $L_B = -\frac{d^2}{ds_x^2}$ on a uniform mesh of size h on Γ (with zero Dirichlet conditions on $\partial\Gamma$). If Q and S are of size m, then Q_0 and L_0 will be tridiagonal matrices of size m given by:

$$Q_0 = \frac{h}{6}\,\text{tridiag}(1,4,1) \;\text{ and }\; L_0 = \frac{1}{h}\,\text{tridiag}(-1,2,-1).$$

Our preconditioner S_0 for S will be the interpolation matrix:

$$S \asymp S_0 \equiv [Q_0, L_0]_{1/2} = Q_0^{1/2}\left(Q_0^{-1/2} L_0 Q_0^{-1/2}\right)^{1/2} Q_0^{1/2}.$$

Matrices Q_0 and L_0 (and hence S_0) will be simultaneously diagonalized by the discrete sine transform matrix F of size m:

$$(F)_{ij} = \sqrt{\frac{2}{m+1}}\,\sin\left(\frac{i\,j\,\pi}{m+1}\right), \quad \text{for} \quad 1 \le i,\, j \le m.$$

The eigenvalues $(\Lambda_{Q_0})_{ii}$ in the spectral decomposition $Q_0 = F\Lambda_{Q_0}F^T$ are:

$$(\Lambda_{Q_0})_{ii} = \frac{1}{3\,(m+1)}\left(3 - 2\,\sin^2\left(\frac{i\,\pi}{2\,(m+1)}\right)\right), \quad \text{for} \quad 1 \le i \le m.$$

The eigenvalues $(\Lambda_{L_0})_{ii}$ in the spectral decomposition $L_0 = F\Lambda_{L_0}F^T$ are:

$$(\Lambda_{L_0})_{ii} = 4\,(m+1)\,\sin^2\left(\frac{i\,\pi}{2\,(m+1)}\right), \quad \text{for} \quad 1 \le i \le m.$$

Since Q_0 and L_0 are diagonalized by F, we obtain:

$$S_0 = F\Lambda_{S_0}F^T = F\left(\Lambda_{Q_0}^{1/4}\Lambda_{L_0}^{1/2}\Lambda_{Q_0}^{1/4}\right)F^T.$$

Replacing Q by Q_0 and S by S_0 in the expressions for C and D, we obtain the following preconditioners C_0 and D_0:

$$\begin{aligned} C_0 &= F\,\Lambda_{C_0}\,F^T = F\left(\alpha_1\,\Lambda_{Q_0} + \alpha_2\,\Lambda_{Q_0}^2\,\Lambda_{S_0}^{-1} + \Lambda_{Q_0}^4\,\Lambda_{S_0}^{-3}\right)F^T, \\ D_0 &= F\,\Lambda_{D_0}\,F^T = F\left(\alpha_1\,\Lambda_{S_0}\,\Lambda_{Q_0}^{-1}\,\Lambda_{S_0} + \alpha_2\,\Lambda_{S_0} + \Lambda_{Q_0}\,\Lambda_{S_0}^{-1}\,\Lambda_{Q_0}\right)F^T. \end{aligned} \tag{10.170}$$

Under regularity assumptions on Q and S, the above preconditioners will yield a rate of convergence independent of h and α_i (weaker assumptions for D). The action of C_0^{-1} and D_0^{-1} can be computed at a cost proportional to two FST's, once Λ_{Q_0} and Λ_{S_0} are computed analytically. The FST preconditioner can be generalized for $\Omega \subset R^3$, provided the grid on Γ can be mapped onto an uniform rectangular grid. It requires using two dimensional FST matrices to diagonalize the Laplace-Beltrami matrix L_0 and the mass matrix Q_0.

Multilevel Preconditioners. When the grid $\tau_h(\Gamma)$ restricted to Γ has a *hierarchical* structure, obtained by successive refinement of some coarse grid $\tau_{h_0}(\Gamma)$ on Γ, we can formulate multilevel preconditioners for C and D using hierarchical projections [BR17, BR20, OS, OS2]. We shall denote the refined grid sizes as $h = h_p < h_{p-1} < \cdots < h_1 < h_0$, where $h_i = (h_{i-1}/2)$ for $1 \leq i \leq p$. Let $V_h(\Gamma) = V_{h_p}(\Gamma)$ denote the restriction of the fine grid finite element space onto Γ, and let $V_{h_j}(\Gamma)$ denote the finite element space based on the triangulation $\tau_{h_j}(\Gamma)$. By construction, these spaces will be nested:

$$V_{h_0}(\Gamma) \subset V_{h_1}(\Gamma) \subset \cdots \subset V_{h_{p-1}}(\Gamma) \subset V_{h_p}(\Gamma).$$

Let $\mathcal{P}_j$ denote the $L^2(\Gamma)$-orthogonal projection onto $V_{h_j}(\Gamma)$. The projections $(\mathcal{P}_j - \mathcal{P}_{j-1})$ can be employed to provide a spectrally equivalent approximation of the discrete Laplace-Beltrami operator $\mathcal{L}_0$ on $V_h(\Gamma)$ endowed with the $L^2(\Gamma)$ inner product [BR20, BR17]:

$$\begin{array}{ll} I & = \mathcal{P}_0 + \sum_{j=1}^{p} (\mathcal{P}_j - \mathcal{P}_{j-1}) \\ \mathcal{L}_0 & \asymp h_0^{-2}\,\mathcal{P}_0 + \sum_{j=1}^{p} h_j^{-2}\,(\mathcal{P}_j - \mathcal{P}_{j-1}). \end{array}$$

The theory of Hilbert scales yields the following equivalences [BR17]:

$$\mathcal{S}_0 \asymp h_0^{-1}\mathcal{P}_0 + \sum_{j=1}^{p} h_j^{-1}(\mathcal{P}_j - \mathcal{P}_{j-1}),\; \mathcal{S}_0^{-1} \asymp h_0^{1}\mathcal{P}_0 + \sum_{j=1}^{p} h_j^{1}(\mathcal{P}_j - \mathcal{P}_{j-1}),$$

where $(\mathcal{S}_0\cdot,\cdot)_{L^2(\Gamma)}$ generates the $H^{1/2}_{00}(\Gamma)$ space inner product in $V_h(\Gamma)$. Under regularity assumptions on S, Q and S_0, and using $E^T M E \asymp Q^T S^{-1} Q$ and $Q \asymp h^{d-1} I$, we will obtain the equivalences:

$$\begin{array}{l} C \asymp \alpha_1\, h^{d-1}\, I + \alpha_2\, h^{2d-2}\, S_0^{-1} + h^{4d-4}\, S_0^{-3} \\ D \asymp \alpha_1\, h^{-d+1}\, S_0^2 + \alpha_2\, S_0 + h^{2d-2}\, S_0^{-1}. \end{array}$$

To obtain the hierarchical preconditioners for C and D, we shall substitute $S_0 \asymp Q_0^{1/2}\mathcal{S}_0 Q_0^{1/2} \asymp h^{d-1}\left(h_0^{-1}\,\mathcal{P}_0 + \sum_{j=1}^{p} h_j^{-1}\,(\mathcal{P}_j - \mathcal{P}_{j-1})\right)$, which is the relation between $\mathcal{S}_0$ and its matrix representation S_0. This will yield:

$$\begin{array}{l} C \asymp h^{d-1}\,\left(\left(\alpha_1 + \alpha_2\, h_0 + h_0^3\right) P_0 + \sum_{i=1}^{p}\left(\alpha_1 + \alpha_2\, h_i + h_i^3\right)(P_i - P_{i-1})\right) \\ D \equiv h^{d-1}\,\left(\left(\alpha_1\, h_0^{-2} + \alpha_2\, h_0^{-1} + h_0^1\right) P_0 \right.\\ \qquad\qquad \left. + \sum_{i=1}^{p}\left(\alpha_1\, h_i^{-2} + \alpha_2\, h_i^{-1} + h_i^1\right)(P_i - P_{i-1})\right), \end{array}$$

where P_i denotes the matrix representation of $\mathcal{P}_i$. In practice, to reduce the computational cost of applying each P_i, we may use an approximation $\tilde{P}_i \approx P_i$. The resulting action of the inverse of the preconditioners for C and D will be:

$$\begin{array}{l} C_0^{-1} = h^{-d+1}\,\left(\alpha_1 + \alpha_2\, h_0 + h_0^3\right)^{-1} \tilde{P}_0^T \tilde{P}_0 \\ \qquad + h^{-d+1} \sum_{i=1}^{p}\left(\alpha_1 + \alpha_2\, h_i + h_i^3\right)^{-1} (\tilde{P}_i - \tilde{P}_{i-1})^T (\tilde{P}_i - \tilde{P}_{i-1}) \\ D_0^{-1} = h^{d-1}\,\left(\alpha_1\, h_0^{-2} + \alpha_2\, h_0^{-1} + h_0^1\right)^{-1} \tilde{P}_0^T \tilde{P}_0 \\ \qquad + h^{d+-1} \sum_{i=1}^{p}\left(\alpha_1\, h_i^{-2} + \alpha_2\, h_i^{-1} + h_i^1\right)^{-1} (\tilde{P}_i - \tilde{P}_{i-1})^T (\tilde{P}_i - \tilde{P}_{i-1}). \end{array} \tag{10.171}$$

Under appropriate assumptions on $\tilde{P}_i \approx P_i$, see [BR17], the resulting preconditioners will be as effective as for exact implementation of P_i.

Augmented Lagrangian Algorithms. The disadvantage of solving (10.160) based on the Hessian system (10.161) is that the PCG algorithm requires the action of A^{-1} twice each iteration, since the Hessian $C{=}(G{+}B^TA^{-T}MA^{-1}B)$. If an efficient sparse direct solver is available for A (such as when $\Omega \subset \mathbb{R}^2$), this approach can be tractable. However, if the action of A^{-1} is computed *iteratively* within an inner loop, this will result in *double iteration*, and such double iteration can be avoided if a preconditioned MINRES algorithm is applied to iteratively solve (10.160), see [RU5, BR9, EL2, KL2, ZU].

We consider an augmented Lagrangian reformulation of system (10.160), see [GL7], since it will be easier to precondition the augmented system. Choose a weight matrix $W = W^T \geq 0$ of size n, and multiply the third block row of (10.160) by A^TW and add it to the first block row, and similarly multiply the third row of (10.160) by B^TW and add it to the second block row. This will yield the following augmented Lagrangian system:

$$\begin{bmatrix} M + A^TWA & A^TWB & A^T \\ B^TWA & G + B^TWB & B^T \\ A & B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 + A^TW\mathbf{f}_3 \\ \mathbf{f}_2 + B^TW\mathbf{f}_3 \\ \mathbf{f}_3 \end{bmatrix}, \qquad (10.172)$$

where by construction system (10.160) and (10.172) have the same solution. For convenience, we shall employ the notation:

$$K = \begin{bmatrix} M + A^TWA & A^TWB \\ B^TWA & G + B^TWB \end{bmatrix}, \quad K_0 = \begin{bmatrix} M_0 & 0 \\ 0 & C_0 \end{bmatrix}, \quad N^T = \begin{bmatrix} A^T \\ B^T \end{bmatrix}. \qquad (10.173)$$

The next result describes a choice of matrix W and preconditioner for (10.172) yielding preconditioned eigenvalues independent of h and α_i, see [GO6].

Lemma 10.94. *Suppose the following assumptions hold.*

1. *Let A_0 and M_0 be chosen such that $A_0M_0^{-1}A_0 \asymp AM^{-1}A$.*
2. *Let $C_0 \asymp (G + B^TA^{-T}MA^{-1}B)$ and let $W \equiv A_0^{-1}M_0A_0^{-1}$.*

Then, the following properties will hold for K, K_0 and N as defined in (10.173):

1. *The equivalences $K \asymp K_0$ and $(NK^{-1}N^T) \asymp (A_0M_0^{-1}A_0)$ will hold.*
2. *The generalized eigenvalues λ of:*

$$\begin{bmatrix} K & N^T \\ N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{p} \end{bmatrix} = \lambda \begin{bmatrix} K_0 & 0 \\ 0 & A_0M_0^{-1}A_0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{p} \end{bmatrix} \qquad (10.174)$$

will lie in $[-b, -a] \cup [a, b]$ where $0 < a < b$ are independent of h and α_i.

Proof. See [GO6] and [RU5, BR9, EL2, KL2, ZU]. □

Thus, solving system (10.172) using the MINRES algorithm [SA2] with the preconditioner blockdiag $\left(K_0, A_0 M_0^{-1} A_0\right)$ will yield robust convergence. This approach requires the action of A_0^{-1} twice each iteration (but not that of A^{-1}).

10.8.2 Parabolic Optimal Control Problems

In a parabolic control problem, variable $y(\cdot)$ will solve a parabolic equation. The computational time and memory required for solving a parabolic optimal control problem can be significantly larger than for an elliptic optimal control problem. Traditionally, the solution to temporal optimal control problems are sought by an application of the Pontryagin maximum principle and by solving a Riccati equation [LI2, LU4]. However, when the number n of output variables is large, the Riccati approach can become prohibitively expensive, since it requires computing and storing a dense matrix of size n at each time step. Thus, although we shall outline the Pontryagin maximum principle and Riccati equation approach, our focus will be on an alternative approach which solves an "all-at-once" spatial and temporal discretization of the parabolic optimal control problem, using a preconditioned MINRES algorithm.

The Control Problem. We consider an output variable $y(.,.)$ which solves a parabolic equation on $\Omega \times (0, \mathcal{T})$ with an input *control* $u(.,.)$ as the forcing:

$$\begin{cases} \dfrac{\partial y}{\partial t} + L\, y = u, & \text{in } \Omega \times (0, \mathcal{T}) \\ \qquad\quad y = 0, & \text{on } \partial\Omega \times (0, \mathcal{T}) \\ \quad y(0,.) = y_0(.), \text{ in } \Omega, & \end{cases} \tag{10.175}$$

where $L\, y \;\equiv\; -\Delta y$. Given a target output $y_*(.,.)$, we define a regularized performance functional $J(y,u)$ measuring the difference between y and y_*:

$$\begin{aligned} J(y,u) = \frac{1}{2} \int_0^{\mathcal{T}} &\left(\|y(t,.) - y_*(t,.)\|^2_{L^2(\Omega)} \,+\, \alpha\, \|u(t,.)\|^2_{L^2(\Omega)} \right) dt \\ &+ \frac{\beta}{2}\, \|y(\mathcal{T},.) - y_*(\mathcal{T},.)\|^2_{L^2(\Omega)}. \end{aligned} \tag{10.176}$$

Here $\alpha > 0$ is a regularization parameter and $\beta \ge 0$ is a weight. The control problem will seek to minimize $J(.,.)$ within a constraint set $\mathcal{V}_{y_0}$:

$$J(y,u) \;=\; \inf_{(\tilde{y},\tilde{u}) \in \mathcal{V}_{y_0}} J(\tilde{y},\tilde{u}), \tag{10.177}$$

where $\mathcal{V}_{y_0}$ consists of solutions to (10.175):

$$\mathcal{V}_{y_0} \;\equiv\; \{(y,u) \;:\; \text{equation (10.175) holds}\,\}. \tag{10.178}$$

To obtain a saddle point formulation of (10.177), let $p(.,.)$ denote a Lagrange multiplier to enforce the constraint (10.175). Define:

$$\mathcal{L}(y,u,p) = J(y,u) + \int_0^{\mathcal{T}} \int_{\Omega} p(t,x) \left(\frac{\partial y}{\partial t} + L\, y - u \right) dx\, dt,$$

as the Lagrangian functional. Formally, if (y, u, p) is the saddle point of $\mathcal{L}(., ., .)$, then (y, u) will solve the constrained minimization problem (10.177). An application of the Pontryagin maximum principle, and elimination of $u(., .)$, will yield a Hamiltonian system for $y(., .)$ and $p(., .)$. We shall describe the Hamiltonian system when we consider the Pontryagin maximum principle.

Semi-Discretized Control Problem. Let $\tau_h(\Omega)$ denote a triangulation of Ω and let $V_h(\Omega) \subset H_0^1(\Omega)$ denote a finite element space defined on $\tau_h(\Omega)$, with basis $\{\phi_1(.), \ldots, \phi_n(.)\}$. Let $U_h(\Omega)$ denote a finite element space for the controls on Ω with basis $\{\psi_1(.), \ldots, \psi_m(.)\}$. A semi-discretization of the saddle point formulation of (10.177), will seek approximations $y_h(t, .)$, $p_h(t, .) \in V_h(\Omega)$ and $u_h(t, .) \in U_h(\Omega)$ of $y(t, .)$, $p(t, .)$ and $u(t, .)$, respectively.

Let A and M denote the stiffness and mass matrices of size n, as defined in (10.157). Let B and Q be matrices of size $n \times m$ and $m \times m$:

$$B_{ij} \equiv \int_\Omega \phi_i(x)\, \psi_j(x)\, dx \quad \text{and} \quad Q_{ij} \equiv \int_\Omega \psi_i(x)\, \psi_j(x)\, dx.$$

Let $\mathbf{y}(t)$, $\mathbf{p}(t) \in \mathbb{R}^n$ and $\mathbf{u}(t) \in \mathbb{R}^m$ denote nodal vectors associated with $y_h(t, .)$, $p_h(t, .)$ and $u_h(t, .)$. A semi-discretization of the constraint $\mathcal{V}_{y_0}$ yields:

$$\mathcal{V}_{\mathbf{y}_0} = \left\{ (\mathbf{y}, \mathbf{u}) : \mathbf{y}' + M^{-1} A\, \mathbf{y} = M^{-1} B\, \mathbf{u}, \text{ for } 0 < t < \mathcal{T},\ \mathbf{y}(0) = \mathbf{y}_0 \right\}$$

where $\mathbf{y}'(t) = \frac{d\mathbf{y}}{dt}(t)$. The semi-discrete performance functional will be:

$$\begin{aligned} J_h(\mathbf{y}, \mathbf{u}) = \frac{1}{2} \int_0^{\mathcal{T}} &\left(\|\mathbf{y}(t) - \mathbf{y}_*(t)\|_M^2 + \alpha\, \|\mathbf{u}(t)\|_Q^2 \right) dt \\ &+ \frac{\beta}{2}\, \|\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\|_M^2, \end{aligned} \tag{10.179}$$

where $\mathbf{y}_*(t)$ is the discretized target function. Denote the Lagrange multiplier function for enforcing the constraints in $\mathcal{V}_{\mathbf{y}_0}$ as $\mathbf{p}(\cdot) \in \mathbb{R}^n$. We then define the semi-discrete *Lagrangian* $\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p})$ associated with $J_h(\mathbf{y}, \mathbf{u})$ in $\mathcal{V}_{\mathbf{y}_0}$ as:

$$\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p}) \equiv J_h(\mathbf{y}, \mathbf{u}) + \int_0^{\mathcal{T}} \mathbf{p}(t)^T \left(\mathbf{y}' + M^{-1} A \mathbf{y} - M^{-1} B \mathbf{u} \right) dt.$$

Formally, if $(\mathbf{y}(.), \mathbf{u}(.), \mathbf{p}(.))$ is the saddle point of $\mathcal{L}_h(., ., .)$, we expect $(\mathbf{y}, \mathbf{u})$ to minimize $J_h(\tilde{\mathbf{y}}, \tilde{\mathbf{u}})$ within $\mathcal{V}_{\mathbf{y}_0}$. Furthermore, we expect the finite element functions $(y_h(.), u_h(.), p_h(.))$ associated with $(\mathbf{y}(.), \mathbf{u}(.), \mathbf{p}(.))$ to approximate $(y(.), u(.), p(.))$. At the saddle point of $\mathcal{L}_h(., ., .)$, the paths $\mathbf{y}(t)$, $\mathbf{u}(t)$ and $\mathbf{p}(t)$ will satisfy a system of differential equations and an algebraic inequality, and these conditions are stated in the Pontryagin maximum principle.

The Pontryagin Maximum Principle. Our discussion of the Pontryagin maximum principle will be heuristic, and will only consider the case where $\mathcal{V}_{\mathbf{y}_0}$ involves initial conditions (i.e., no terminal constraints). The maximum principle derives a system of ordinary differential equations for $\mathbf{y}(.)$ and $\mathbf{p}(.)$

and an algebraic inequality characterizing $\mathbf{u}(.)$ at the saddle point of the Lagrangian $\mathcal{L}_h(.,.,.)$, see [LU4]. The following functional $H(\mathbf{p}, \mathbf{y}, \mathbf{u})$, referred to as the *Hamiltonian*, is associated with the Lagrangian functional $\mathcal{L}_h(.,.,.)$:

$$H(\mathbf{p}, \mathbf{y}, \mathbf{u}) \;=\; \frac{1}{2}\left(\|\mathbf{y}-\mathbf{y}_*\|_M^2 + \alpha\,\|\mathbf{u}\|_Q^2\right) + \mathbf{p}^T\left(M^{-1}A\mathbf{y} - M^{-1}B\mathbf{u}\right). \tag{10.180}$$

The Lagrangian can be expressed using the Hamiltonian as:

$$\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p}) = \int_0^{\mathcal{T}} \left(\mathbf{p}^T\mathbf{y}' + H(\mathbf{y}, \mathbf{u}, \mathbf{p})\right)\, dt + \frac{\beta}{2}\,\|\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\|_M^2.$$

We now state the minimum principle of Pontryagin.

Theorem 10.95 (Pontryagin). *Let $\mathbf{y}(.) \in \mathbb{R}^n$ and $\mathbf{u}(.) \in \mathbb{R}^m$ denote the output variable and control functions which minimize $J_h(.,.)$ within $\mathcal{V}_{\mathbf{y}_0}$. Then, $\mathbf{y}(.)$ and $\mathbf{u}(.)$, along with the Lagrange multiplier $\mathbf{p}(.) \in \mathbb{R}^n$ will satisfy:*

$$\begin{cases} \mathbf{y}' = -M^{-1}A\mathbf{y} + M^{-1}B\mathbf{u} & \text{with} \;\; \mathbf{y}(0) \;=\; \mathbf{y}_0 \\ \mathbf{p}' = A^T M^{-T}\mathbf{p} + M\left(\mathbf{y} - \mathbf{y}_*\right) & \text{with} \;\; \mathbf{p}(\mathcal{T}) \;=\; -\beta\, M\left(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\right) \end{cases}$$

together with the minimization requirement:

$$H(\mathbf{y}, \mathbf{u}, \mathbf{p}) \;\leq\; H(\mathbf{y}, \mathbf{v}, \mathbf{p})$$

for any other path $\mathbf{v}(.)$ for the control.

Proof. We shall outline the proof heuristically, see [LU4]. Requiring the Gateaux derivative of $\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p})$ with respect to $\mathbf{p}(.)$ to be zero will yield the constraints $\mathbf{y}' = -M^{-1}A\mathbf{y} + M^{-1}B\mathbf{u}$, since:

$$\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p} + \delta\mathbf{p}) - \mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p}) = \int_0^{\mathcal{T}} \delta\mathbf{p}(t)^T\left(\mathbf{y}' + M^{-1}A\mathbf{y} - M^{-1}B\mathbf{u}\right) dt.$$

Next, given a control $\mathbf{v}(.) = \mathbf{u}(.) + \delta\mathbf{u}(.)$, let $\mathbf{y}(.) + \delta\mathbf{y}(.)$ denotes the output variable such that $(\mathbf{y} + \delta\mathbf{y}, \mathbf{u} + \delta\mathbf{u}) \in \mathcal{V}_{\mathbf{y}_0}$ where $\delta\mathbf{y}(0) = \mathbf{0}$. We decompose

$$\delta\mathcal{L}_h \;\equiv\; \left(\mathcal{L}_h(\mathbf{y} + \delta\mathbf{y}, \mathbf{u} + \delta\mathbf{u}, \mathbf{p}) - \mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p})\right) \;=\; \delta\mathcal{L}_h^{(1)} + \delta\mathcal{L}_h^{(2)}$$

where each of the terms are defined as:

$$\begin{cases} \delta\mathcal{L}_h^{(1)} \equiv \left(\mathcal{L}_h(\mathbf{y} + \delta\mathbf{y}, \mathbf{u} + \delta\mathbf{u}, \mathbf{p}) - \mathcal{L}_h(\mathbf{y}, \mathbf{u} + \delta\mathbf{u}, \mathbf{p})\right) \\ \delta\mathcal{L}_h^{(2)} \equiv \left(\mathcal{L}_h(\mathbf{y}, \mathbf{u} + \delta\mathbf{u}, \mathbf{p}) - \mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p})\right). \end{cases}$$

Since $(\mathbf{y} + \delta\mathbf{y}, \mathbf{u} + \delta\mathbf{u}) \in \mathcal{V}_{\mathbf{y}_0}$ and $(\mathbf{y}, \mathbf{u}) \in \mathcal{V}_{\mathbf{y}_0}$, the constraints are satisfied, and it must hold that $\delta\mathcal{L}_h = \delta J_h = \left(J_h(\mathbf{y} + \delta\mathbf{y}, \mathbf{u} + \delta\mathbf{u}) - J_h(\mathbf{y}, \mathbf{u})\right) \geq 0$.

To evaluate $\delta\mathcal{L}_h^{(1)}$ and $\delta\mathcal{L}_h^{(2)}$, we shall employ the following alternative expression for $\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p})$ obtained using integration by parts:

$$\begin{aligned} \mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p}) = \left(\mathbf{p}(\mathcal{T})^T\mathbf{y}(\mathcal{T}) - \mathbf{p}(0)^T\mathbf{y}(0)\right) + \frac{\beta}{2}\,\|\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\|_M^2 \\ + \int_0^{\mathcal{T}} \left(-\mathbf{y}^T\mathbf{p}' + H(\mathbf{y}, \mathbf{u}, \mathbf{p})\right)\, dt. \end{aligned}$$

Employing this expression, we obtain:

$$\begin{aligned}\delta\mathcal{L}_h^{(1)} &= \delta\mathbf{y}(\mathcal{T})^T\,(\mathbf{p}(\mathcal{T}) + \beta\,M\,\left(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\right)) \\ &\quad + \int_0^{\mathcal{T}} \delta\mathbf{y}^T\,\left(-\mathbf{p}' + H_{\mathbf{y}}(\mathbf{y},\mathbf{u},\mathbf{p})\right)\,dt + O(\|\delta\|^2),\end{aligned}$$

where $O(\|\delta\|^2) = O(\|\delta\mathbf{y}\|^2 + \|\delta\mathbf{y}\|\,\|\delta\mathbf{u}\| + \|\delta\mathbf{u}\|^2)$. Thus, if we require:

$$\mathbf{p}' = H_{\mathbf{y}}(\mathbf{y},\mathbf{u}+\delta\mathbf{u},\mathbf{p}) \;\text{ and }\; \mathbf{p}(\mathcal{T}) = -\beta\,M\,\left(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\right)$$

we obtain $\mathbf{p}' = M(\mathbf{y}-\mathbf{y}_*) + A^TM^{-T}\mathbf{p}$ and $\delta\mathcal{L}_h^{(1)} = O(\|\delta\|^2)$. Evaluating $\delta\mathcal{L}_h^{(2)}$ employing the alternative expression for $\mathcal{L}_h(\mathbf{y},\mathbf{u},\mathbf{p})$ yields:

$$\delta\mathcal{L}_h^{(2)} = \int_0^{\mathcal{T}} \left(H(\mathbf{y},\mathbf{u}+\delta\mathbf{u},\mathbf{p}) - H(\mathbf{y},\mathbf{u},\mathbf{p})\right)\,dt.$$

Since it must hold that $\delta\mathcal{L}_h = \delta J_h \geq 0$ for arbitrary $\delta\mathbf{u}$, it must also hold that $H(\mathbf{y},\mathbf{u}+\delta\mathbf{u},\mathbf{p}) - H(\mathbf{y},\mathbf{u},\mathbf{p}) \geq 0$. □

Remark 10.96. The minimization requirement $H(\mathbf{y},\mathbf{u},\mathbf{p}) \leq H(\mathbf{y},\mathbf{v},\mathbf{p})$ can be reduced to $H_{\mathbf{u}}(\mathbf{y},\mathbf{u},\mathbf{p}) = 0$, since there are no inequality constraints:

$$H_{\mathbf{u}}(\mathbf{y},\mathbf{u},\mathbf{p}) = \alpha\,Q\mathbf{u} - B^TM^{-1}\mathbf{p} = 0 \implies \mathbf{u} = \frac{1}{\alpha}\,Q^{-1}B^TM^{-1}\mathbf{p}.$$

Substituting $\mathbf{u}(t) = \frac{1}{\alpha}\,Q^{-1}B^TM^{-1}\mathbf{p}$ into the equations yields:

$$\begin{cases}\mathbf{y}' = -M^{-1}A\mathbf{y} + \dfrac{1}{\alpha}\,M^{-1}\,B\,Q^{-1}\,B^T\,M^{-1}\mathbf{p} \quad\text{with}\quad \mathbf{y}(0) = \mathbf{y}_0 \\ \mathbf{p}' = A^TM^{-T}\mathbf{p} + M\,(\mathbf{y}-\mathbf{y}_*) \quad\text{with}\quad \mathbf{p}(\mathcal{T}) = -\beta\,M\,(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T}))\end{cases} \tag{10.181}$$

This system will have a *Hamiltonian* structure for the reduced Hamiltonian:

$$\begin{aligned}\mathcal{H}(\mathbf{y},\mathbf{p}) &\equiv H(\mathbf{y},\frac{1}{\alpha}\,Q^{-1}B^TM^{-1}\mathbf{p},\mathbf{p}) \\ &= \frac{1}{2}\|\mathbf{y}-\mathbf{y}_*\|_M^2 - \frac{1}{2\,\alpha}\mathbf{p}^TM^{-T}BQ^{-1}B^TM^{-1}\mathbf{p} + \mathbf{p}^TM^{-1}A\mathbf{y}.\end{aligned}$$

After elimination of $\mathbf{u}(t)$, the reduced system for $\mathbf{y}(t)$ and $\mathbf{p}(t)$ will be:

$$\begin{cases}\mathbf{y}' = -\mathcal{H}_{\mathbf{p}}(\mathbf{y},\mathbf{p}) \text{ with } \mathbf{y}(0) = \mathbf{y}_0 \\ \mathbf{p}' = \quad \mathcal{H}_{\mathbf{y}}(\mathbf{y},\mathbf{p}) \text{ with } \mathbf{p}'(\mathcal{T}) = -\beta\,M(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})).\end{cases}$$

The reduced Hamiltonian will be *constant* along the solution, since:

$$\begin{aligned}\frac{d\,\mathcal{H}(\mathbf{y}(t),\mathbf{p}(t))}{dt} &= \mathcal{H}_{\mathbf{y}}(\mathbf{y},\mathbf{p})\cdot\mathbf{y}' + \mathcal{H}_{\mathbf{p}}(\mathbf{y},\mathbf{p})\cdot\mathbf{p}' \\ &= -\mathcal{H}_{\mathbf{y}}(\mathbf{y},\mathbf{p})\cdot\mathcal{H}_{\mathbf{p}}(\mathbf{y},\mathbf{p}) + \mathcal{H}_{\mathbf{p}}(\mathbf{y},\mathbf{p})\cdot\mathcal{H}_{\mathbf{y}}(\mathbf{y},\mathbf{p}) = 0.\end{aligned}$$

Thus, $\mathcal{H}(\mathbf{y}(t),\mathbf{p}(t)) = \mathcal{H}(\mathbf{y}(0),\mathbf{p}(0))$ for $\forall t$.

Riccati Method. The Hamiltonian system (10.181) for $\mathbf{y}(t)$ and $\mathbf{p}(t)$ does not have a complete set of initial data. The requirement $\mathbf{y}(0) = \mathbf{y}_0$ yields n data constraints at time $t = 0$, while $\mathbf{p}(\mathcal{T}) = -\beta\, M\,(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T}))$ yields n constraints at time $t = \mathcal{T}$. In principle, we may seek the unknown initial data $\mathbf{p}(0) = \mathbf{p}_0$ by applying a *shooting* method. This will require solving system (10.181) with initial data $\mathbf{y}(0) = \mathbf{y}_0$ and $\mathbf{p}(0) = \mathbf{p}_0$ on the time interval $(0, \mathcal{T})$ and seeking $\mathbf{p}_0$ such that $\mathbf{p}(\mathcal{T}) = -\beta\, M\,(\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T}))$. Unfortunately, the shooting method can be unstable in our applications, due to positive and negative eigenvalues of a large magnitude for the coefficient matrix of the linear Hamiltonian system involving $\mathbf{y}(t)$ and $\mathbf{p}(t)$.

Instead, traditionally, the solution to system (10.181) is obtained when $\mathbf{y}_*(.) = \mathbf{0}$ by solving an associated matrix Riccati equation [LU4]. It seeks an $n \times n$ matrix function $W(t)$ so that the relation $\mathbf{p}(t) = W(t)\,\mathbf{y}(t)$ holds. Substituting this *ansatz* into the Hamiltonian system (10.181) yields:

$$\begin{cases} \mathbf{y}' = \left(-M^{-1}A + \dfrac{1}{\alpha}M^{-1}BQ^{-1}B^TM^{-1}W\right)\mathbf{y} \\ W'\,\mathbf{y} + W\,\mathbf{y}' = M\,\mathbf{y} + A^TM^{-1}W\mathbf{y}. \end{cases}$$

Multiplying the first row by $-W(t)$ adding it to the second row yields:

$$\left(W' - WM^{-1}A - A^TM^{-1}W + \frac{1}{\alpha}WM^{-1}BQ^{-1}B^TM^{-1}W - M\right)\mathbf{y} \;=\; \mathbf{0}.$$

Requiring the above equations to hold for arbitrary $\mathbf{y}(.)$ yields a first order matrix differential equation for $W(t)$. Imposing $\mathbf{p}(\mathcal{T}) = -\beta\, M\mathbf{y}(\mathcal{T})$ yields terminal conditions for $W(\mathcal{T})$. We obtain:

$$\begin{cases} W' = \left(WM^{-1}A + A^TM^{-1}W\right) - \dfrac{1}{\alpha}\left(WM^{-1}BQ^{-1}B^TM^{-1}W\right) + M \\ W(\mathcal{T}) = -\beta\, M \end{cases} \tag{10.182}$$

This first order differential equation for the $n \times n$ matrix $W(t)$ on $(0, \mathcal{T})$ is referred to as the *Riccati equation.* The Riccati differential equation for $W(t)$ has quadratic non-linearity, and can be solved *numerically* backwards in time on $(0, \mathcal{T})$ using the terminal data $W(\mathcal{T}) = -\beta\, M$. Since $W'(t)$ and $W(\mathcal{T})$ are symmetric, matrix $W(t)$ will also be symmetric. Furthermore, if $\beta = 0$, matrix $W(t)$ can also be shown to be positive semi-definite.

The solution $W(t)$ to the Riccati equation can be computed and stored *offline*. Given an observation $\mathbf{y}(t)$, the control $\mathbf{u}(t) = \frac{1}{\alpha}Q^{-1}B^TM^{-1}W\mathbf{y}(t)$ can be computed *instantaneously*, and thus, the Riccati based solution is useful in real time applications. However, the cost of computing and storing the $n \times n$ matrix function $W(t)$ at each discrete time can be prohibitive when n is large. In some applications, it may be sufficient to compute a stationary solution W_* of the Riccati equation using time marching. Using W_* can reduce computational costs.

Dynamic Programming. Dynamic programming provides an alternative approach for solving time dependent control problems [LU4]. Although the methodology is prohibitively expensive in applications to the control of parabolic equations, we shall outline it for its intrinsic interest.

The dynamic programming method determines the control $\mathbf{u}(.)$ based on an *optimal* value function $V(t, \mathbf{y})$ defined for $0 \leq t \leq \mathcal{T}$ and $\mathbf{y} \in \mathbb{R}^n$. The value function $V(t, \mathbf{y})$ represents the minimum value of the functional $J_h(.,.)$ when restricted to the time interval $(t, \mathcal{T})$ for sub-trajectories in $\mathcal{V}_{\mathbf{y}_0}$ that pass through $\mathbf{y}$ at time t. Given $0 \leq t_0 < \mathcal{T}$ and $\mathbf{w} \in \mathbb{R}^n$, define $\mathcal{V}_{(t_0,\mathbf{w})}$ as:

$$\mathcal{V}_{(t_0,\mathbf{w})} = \{ (\mathbf{y}(.), \mathbf{u}(.)) : \mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t), \mathbf{u}(t)), \text{ for } t_0 < t < \mathcal{T}, \ \mathbf{y}(t_0) = \mathbf{w} \},$$

where $\mathbf{f}(\mathbf{y}, \mathbf{u}) = -M^{-1}A\mathbf{y} + M^{-1}B\,\mathbf{u}$. Define the restricted functional:

$$J_{(t_0,\mathcal{T})}(\mathbf{y}, \mathbf{u}) \equiv \int_{t_0}^{\mathcal{T}} l(\mathbf{y}(t), \mathbf{u}(t))\, dt + \frac{\beta}{2} \|\mathbf{y}(\mathcal{T}) - \mathbf{y}_*(\mathcal{T})\|_M^2, \qquad (10.183)$$

where $l(\mathbf{y}, \mathbf{u}) = \frac{1}{2}\|\mathbf{y} - \mathbf{y}_*\|_M^2 + \frac{\alpha}{2}\|\mathbf{u}\|_Q^2$ and $J_{(t_0,\mathcal{T})}(\mathbf{y}, \mathbf{u})$ is the contribution to $J_h(.,.)$ along $(t_0, \mathcal{T})$. The value function $V(t_0, \mathbf{w})$ is then defined as:

$$V(t_0, \mathbf{w}) \equiv \inf_{(\tilde{y},\tilde{u}) \in \mathcal{V}_{(t_0,\mathbf{w})}} J_{(t_0,\mathcal{T})}(\tilde{\mathbf{y}}, \tilde{\mathbf{u}}).$$

In the following, we shall heuristically derive a partial differential equation for $V(.,.)$ employing the *principle of optimality* [LU4]. This principle says that if $t_0 < t_1$ and $(\mathbf{y}(.), \mathbf{u}(.))$ optimizes $J_{(t_0,\mathcal{T})}(.,.)$ then its restriction to $(t_1, \mathcal{T})$ will optimize $J_{(t_1,\mathcal{T})}(.,.)$. As a result, the optimal value $V(t_0, \cdot)$ will depend on the optimal value $V(t_1, \cdot)$, and the control $\mathbf{u}(\cdot)$ can be determined on (t_0, t_1) based on $V(.,.)$. To derive an equation for $V(.,.)$ let $\delta t > 0$ be an infinitesimally small time step and let $\mathbf{y}(t_0) = \mathbf{w}$ with $\mathbf{u}(t_0) = \mathbf{u}_0$. Then, using $J_{(t_0,\mathcal{T})}(\mathbf{y}, \mathbf{u}) = l(\mathbf{w}, \mathbf{u}_0)\,\delta t + J_{(t_0+\delta t,\mathcal{T})}(\mathbf{y}, \mathbf{u}) + O(\delta t^2)$ yields:

$$V(t_0, \mathbf{w}) = \lim_{\delta t \to 0^+} \inf_{\mathbf{u}_0} \{ l(\mathbf{w}, \mathbf{u}_0)\,\delta t + V(t_0 + \delta t, \mathbf{w} + \delta t\, \mathbf{f}(\mathbf{w}, \mathbf{u}_0)) \}. \qquad (10.184)$$

We substitute the following first order Taylor series expansion for $V(.,.)$:

$$\begin{aligned} V(t_0 + \delta t, \mathbf{w} + \delta t\, \mathbf{f}(\mathbf{w}, \mathbf{u}_0)) = V(t_0, \mathbf{w})) + \delta t\, V_t(t_0, \mathbf{w}) \\ + V_{\mathbf{y}}^T(t_0, \mathbf{w})\, \delta t\, \mathbf{f}(\mathbf{w}, \mathbf{u}_0) + O(\delta t^2), \end{aligned}$$

into (10.184). Since $V(t_0, \mathbf{w})$ does not depend on $\mathbf{u}_0$, canceling terms yields:

$$0 = \lim_{\delta t \to 0^+} \inf_{\mathbf{u}_0} \left\{ l(\mathbf{w}, \mathbf{u}_0)\,\delta t + \delta t\, V_t(t_0, \mathbf{w}) + V_{\mathbf{y}}^T(t_0, \mathbf{w})\, \delta t\, \mathbf{f}(\mathbf{w}, \mathbf{u}_0) \right\}.$$

Replacing t_0, $\mathbf{w}$, $\mathbf{u}_0$ by t, $\mathbf{y}$, $\mathbf{u}$, respectively, and simplifying yields:

$$V_t(t, \mathbf{y}) + \inf_{\mathbf{u}} \left\{ l(\mathbf{y}, \mathbf{u}) + V_{\mathbf{y}}^T(t, \mathbf{y})\, \mathbf{f}(\mathbf{y}, \mathbf{u}) \right\} = 0. \qquad (10.185)$$

This is referred to as the *Hamilton-Jacobi-Bellman* equation. By construction, $V(.,.)$ will satisfy the terminal condition $V(\mathcal{T}, \mathbf{y}) = \frac{\beta}{2}\|\mathbf{y} - \mathbf{y}_*\|_M^2$ at $t = \mathcal{T}$.

The Hamilton-Jacobi-Bellman equation is *hyperbolic* in nature. It will be *non-linear* when $\inf_{\mathbf{u}} \left\{ l(\mathbf{y},\mathbf{u}) + V_{\mathbf{y}}^T(t,\mathbf{y})\,\mathbf{f}(\mathbf{y},\mathbf{u}) \right\}$ is non-linear. Indeed, for $l(\mathbf{y},\mathbf{u}) = \frac{1}{2}\left((\mathbf{y}-\mathbf{y}_*)^T M (\mathbf{y}-\mathbf{y}_*) + \alpha\,\mathbf{u}^T Q \mathbf{u} \right)$ and $\mathbf{f}(\mathbf{y},\mathbf{u}) {=} M^{-1}(-A\mathbf{y} + B\mathbf{u})$, the first derivative test yields:

$$\alpha\, Q\mathbf{u} + B^T M^{-1} V_{\mathbf{y}} = \mathbf{0} \implies \mathbf{u} = -\frac{1}{\alpha}\, Q^{-1} B^T M^{-1} V_{\mathbf{y}}.$$

Substituting for $\mathbf{u}$ in the Hamilton-Jacobi-Bellman equation yields:

$$\begin{cases} V_t + \dfrac{1}{2}\left(\mathbf{y}^T M \mathbf{y} - \dfrac{1}{\alpha}\, V_{\mathbf{y}}^T M^{-1} B Q^{-1} B^T M^{-1} V_{\mathbf{y}} \right) - \mathbf{y}^T A^T M^{-1} V_{\mathbf{y}} \;=\; 0 \\ V(\mathcal{T},\mathbf{y}) \;=\; \dfrac{\beta}{2}\, \|\mathbf{y}-\mathbf{y}_*\|_M^2, \end{cases}$$

when $\mathbf{y}_* = \mathbf{0}$. If we formally seek a solution using the *ansatz*:

$$V(t,\mathbf{y}) \equiv -\frac{1}{2}\,\mathbf{y}^T W(t) \mathbf{y},$$

where $W(t)$ is an $n \times n$ symmetric matrix function. Substituting this ansatz into the Hamilton-Jacobi-Bellman yields the following equations:

$$\begin{cases} \mathbf{y}^T \left(-\dfrac{1}{2}\, W' + \dfrac{1}{2}\, M - \dfrac{1}{2\,\alpha}\, W^T M^{-1} B Q^{-1} B^T M^{-1} W + A^T M^{-1} W \right) \mathbf{y} \;=\; 0 \\ W(\mathcal{T}) \;=\; -\beta\, M. \end{cases}$$

Eliminating $\mathbf{y}$ yields the following matrix Riccati equations for $W(.)$:

$$\begin{cases} \quad W' = M - \dfrac{1}{\alpha}\, W^T M^{-1} B Q^{-1} B^T M^{-1} W + (A^T M^{-1} W + W^T M^{-1} A) \\ W(\mathcal{T}) = -\beta\, M. \end{cases}$$

This is identical to the Ricatti equation (10.182) described earlier for the reduced Hamiltonian system. The solution $W(t)$ can be determined by time marching. However, as mentioned before, since $\mathbf{y} \in \mathbb{R}^n$ and n is large in our applications, the computation and storage of $W(.)$ can be prohibitively expensive. Once $V(.,.)$ has been determined, the control $\mathbf{u}(.)$ will satisfy:

$$\mathbf{u}(t) \;=\; -\frac{1}{\alpha}\, Q^{-1} B^T M^{-1} V_{\mathbf{y}}(t,\mathbf{y}(t)) \;=\; \frac{1}{\alpha}\, Q^{-1} B^T M^{-1} W(t) \mathbf{y}(t),$$

by local optimization of (10.185).

Hessian System. Since the dynamic programming and Riccati equation based solution of control problem (10.177) are prohibitively expensive for large n (as it requires computing a matrix $W(t)$ of size n for $0 < t < \mathcal{T}$), we outline an alternative *iterative* approach for determining the optimal control $\mathbf{u}(.)$, based on the solution of a reduced Hessian system. This iterative approach will not yield a *feedback* solution, as in the Riccati method. However, we shall outline *heuristic* modifications of this iterative approach, which will yield approximations of the optimal control $\mathbf{u}(.)$ in real time applications.

Our discussion will consider a *full discretization* of the parabolic optimal control problem with time step $\tau = (\mathcal{T}/l)$ and spatial mesh size h. We denote the discrete times as $t_i = i\,\tau$ for $1 \le i \le l$. For simplicity, we shall assume that the discrete control $\mathbf{u}(t)$ corresponding to the finite element function $u_h(t,x)$ is constant on each time interval (t_i, t_{i+1}). The discrete output variable $\mathbf{y}(t)$ corresponding to the finite element function $y_h(t,x)$ will be assumed to continuous and piecewise linear, i.e., $\mathbf{y}(t)$ will be linear on each (t_i, t_{i+1}). We shall let $\mathbf{y}_i = \mathbf{y}(t_i) \in \mathbb{R}^n$ denote the nodal vector solution at time t_i. Similarly, $\mathbf{u}_i \in \mathbb{R}^m$ will denote the discrete control for $t \in [t_{i-1}, t_i)$ for $1 \le i \ne l$. Denote the nodal vectors associated with $\mathbf{y}(.)$, $\mathbf{p}(.)$ and $\mathbf{u}(.)$ at the discrete times as:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_l \end{bmatrix} \in \mathbb{R}^{n\,l}, \quad \mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_l \end{bmatrix} \in \mathbb{R}^{n\,l}, \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{u}_l \end{bmatrix} \in \mathbb{R}^{m\,l}.$$

For simplicity, we shall assume that $y_*(0,\cdot) = y_0(\cdot)$, so that $\mathbf{y}(0) = \mathbf{y}_*(0)$. We define a block matrix G of size $m\,l$ and K of size $n\,l$ as:

$$G = \alpha\,\tau \begin{bmatrix} Q & & & 0 \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \\ 0 & & & Q \end{bmatrix} \quad \text{and} \quad K = \frac{\tau}{6} \begin{bmatrix} 4M & M & & & 0 \\ M & 4M & M & & \\ & \ddots & \ddots & \ddots & \\ & & M & 4M & M \\ 0 & & & M & \gamma M \end{bmatrix},$$

where $\gamma = 2 + (6\beta/\tau)$. Then, the discretized functional $J_h(.,.)$ will satisfy:

$$J_{h,\tau}(\mathbf{y},\mathbf{u}) = \frac{1}{2}(\mathbf{Y} - \mathbf{Y}_*)^T K (\mathbf{Y} - \mathbf{Y}_*) + \frac{1}{2}\mathbf{U}G\mathbf{U},$$

where $\mathbf{Y}_* = \left(\mathbf{y}_*(t_1)^T, \ldots, \mathbf{y}_*(t_l)^T\right)^T \in \mathbb{R}^{n\,l}$ denotes a discretization of the target output. To obtain a spatio-temporal discretization of the constraint set $\mathcal{V}_{\mathbf{y}_0}$, we apply a θ-scheme to discretize $M\,\mathbf{y}'(t) + A\,\mathbf{y}(t) = B\,\mathbf{u}(t)$ in time. This will yield a large system of linear equations for the evolution problem:

$$\mathcal{V}_{h,\tau} = \{(\mathbf{Y},\mathbf{U}) : E\,\mathbf{Y} + N\,\mathbf{U} = \mathbf{F}\}$$

where matrices E and N will be described later for the backward Euler scheme. The initial data $\mathbf{y}(0) = \mathbf{y}_0$ is included in $\mathbf{F}$.

For the backward Euler scheme, $E \in \mathbb{R}^{nl \times nl}$ and $N \in \mathbb{R}^{nl \times ml}$ satisfy:

$$E = \begin{bmatrix} (M + \tau A) & & & \\ -M & \ddots & & \\ & \ddots & \ddots & \\ & & -M & (M + \tau A) \end{bmatrix} \quad \text{and} \quad N = -\tau \begin{bmatrix} B & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & B \end{bmatrix},$$

while the forcing term $\mathbf{F} \in \mathbb{R}^{n\,l}$ is defined as $\mathbf{F} = \begin{bmatrix} M\mathbf{y}_0 & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}^T$. The following discrete Lagrangian functional $\mathcal{L}_{h,\tau}(\mathbf{Y}, \mathbf{U}, \mathbf{P})$ will be associated with the minimization of $J_{h,\tau}(\mathbf{Y}, \mathbf{U})$ within the constraint set $\mathcal{V}_{h,\tau}$:

$$\mathcal{L}_{h,\tau}(\mathbf{Y}, \mathbf{U}, \mathbf{P}) = J_{h,\tau}(\mathbf{Y}, \mathbf{U}) + \mathbf{P}^T \left(E\,\mathbf{Y} + N\,\mathbf{U} - \mathbf{F} \right).$$

The system for determining the saddle point $(\mathbf{Y}, \mathbf{U}, \mathbf{P})$ of $\mathcal{L}_{h,\tau}(.,.,.)$ is:

$$\begin{bmatrix} K & 0 & E^T \\ 0 & G & N^T \\ E & N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ \mathbf{U} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} K\mathbf{Y}_* \\ \mathbf{0} \\ \mathbf{F} \end{bmatrix}. \tag{10.186}$$

The properties of matrices K, G, E and N are as follows. Matrix $K = K^T > 0$ is spectrally equivalent to $\tau\, h^d\, I$ of size nl. Matrix $G = G^T > 0$ will be spectrally equivalent to $\alpha\, \tau\, h^d\, I$ of size ml, provided Q is well conditioned. Matrix E is block lower bi-diagonal, and its diagonal blocks are ill-conditioned. The properties of the rectangular matrix N of size $nl \times ml$, depend on the choice of the control basis. In the special case that $m = n$ and $B = M$, matrix N will be spectrally equivalent to $h^d I$.

The Hessian system for $\mathbf{U}$ can be obtained by eliminating $\mathbf{Y} = E^{-1}$ $(\mathbf{F} - N\mathbf{u})$ using the third block row and eliminating $\mathbf{P} = E^{-T}K(\mathbf{Y}_* - \mathbf{Y})$ using the first block row, and substituting these into the second block row, yielding:

$$\begin{cases} C\,\mathbf{U} = \mathbf{g} \;\; \text{where} \\ \quad C \equiv (G + N^T E^{-T} K E^{-1} N) \\ \quad \mathbf{g} \equiv (N^T E^{-T} K E^{-1} \mathbf{F} - N^T E^{-T} K \mathbf{Y}_*). \end{cases} \tag{10.187}$$

The Hessian matrix C is symmetric positive definite, and system (10.187) can be solved by a PCG algorithm. Matrix $N^T E^{-T} K E^{-1} N$ is *ill-conditioned*, and corresponds to the discretization of a *compact* operator whose eigenvalues are bounded and cluster around zero. Thus, the addition of the α dependent regularization term G to $N^T E^{-T} K E^{-1} N$ shifts the eigenvalues of C away from zero. If $\alpha = O(1)$, matrix C will be well conditioned, while if $\alpha \to 0^+$, matrix C will be ill-conditioned, and require a preconditioner C_0.

Solving the Hessian System. Our discussion will be restricted to a few *special cases*. The *first case* arises when $\alpha = O(1)$. In this case, the Hessian C will be well conditioned, and system (10.187) can be solved using a PCG algorithm with preconditioner I, see [SC2]. The *second case* arises when the number m of control basis and the number l of time steps are both “small” with $m\,l \ll n$. In this case, the dense matrix $N^T E^{-T} K E^{-1} N$ of size $m\,l$ can be assembled and stored and a direct solver can be used to solve (10.187).

The *third case* arises when $m = n$ and matrices $B = M = Q$. In this case $G = \alpha\, N$ and a preconditioner C_0 can be formulated based on the identity:

$$\begin{cases} C = \alpha\, N + N^T E^{-T} K E^{-1} N \\ \quad = N^T E^{-T} \left(\alpha\, E^T N^{-1} E + K \right) E^{-1} N. \end{cases} \tag{10.188}$$

To obtain a preconditioner C_0, we shall formally replace each block submatrix M, A and $(M + \tau\, A)$ in the matrices N, K and E by spectrally equivalent approximations $M_0 \asymp M$, $A_0 \asymp A$ and $(M_0 + \tau A_0) \asymp (M + \tau\, A)$, where we require that $M_0 = M_0^T$ and $A_0 = A_0^T$ are *simultaneously diagonalizable*. One such choice is $M_0 = h^d I$, $A_0 = A$ and $(M_0 + \tau A_0) = (h^d I + \tau A)$, however, in applications we shall choose $M_0 \asymp M$ and $A_0 \asymp A$ that are simultaneously diagonalized by the fast sine transform (FST) or hierarchical projections.

Let N_0, K_0 and E_0 denote the matrices obtained when we replace the block submatrices M, A and $(M + \tau A)$ of N, K and E by M_0, A_0 and $(M_0 + \tau A_0)$. We formally define the preconditioner C_0 as:

$$\begin{cases} C_0 \;\;\, \equiv N_0^T E_0^{-T} \left(\alpha\, E_0^T N_0^{-1} E_0 + K_0 \right) E_0^{-1} N_0 \\ C_0^{-1} = N_0^{-1} E_0 \left(\alpha\, E_0^T N_0^{-1} E_0 + K_0 \right)^{-1} E_0^T N_0^{-T}. \end{cases} \tag{10.189}$$

Let V_0 denote the unitary matrix which diagonalizes $M_0 = V_0^T \Lambda_{M_0} V_0$ and $A_0 = V_0^T \Lambda_{A_0} V_0$, simultaneously. Then, the block submatrices in N_0, E_0 and K_0, and hence in C_0 and C_0^{-1}, will also be diagonalized by V_0. This property will yield a fast and efficient algorithm for computing the action of C_0^{-1}.

Let $V \equiv \text{blockdiag}(V_0, \ldots, V_0)$ denote the block diagonal matrix of size $n\,l \times n\,l$ whose diagonal blocks are V_0. Then $V C_0^{-1} V^T$ will have the form:

$$V C_0^{-1} V^T \;=\; \tilde{N}_0^{-1} \tilde{E}_0 \left(\alpha\, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 + \tilde{K}_0 \right)^{-1} \tilde{E}_0^T \tilde{N}_0^{-T}. \tag{10.190}$$

where $\tilde{K}_0$, $\tilde{E}_0$ and $\tilde{N}_0$ are obtained by replacing M_0, A_0 and $(M_0 + \tau A_0)$ by the *diagonal matrices* Λ_{M_0}, Λ_{A_0} and $(\Lambda_{M_0} + \tau \Lambda_{A_0})$, respectively:

$$\tilde{K}_0 = \frac{\tau}{6} \begin{bmatrix} 4\Lambda_{M_0} & \Lambda_{M_0} & & & 0 \\ \Lambda_{M_0} & 4\Lambda_{M_0} & \Lambda_{M_0} & & \\ & \ddots & \ddots & \ddots & \\ & & \Lambda_{M_0} & 4\Lambda_{M_0} & \Lambda_{M_0} \\ 0 & & & \Lambda_{M_0} & \gamma\Lambda_{M_0} \end{bmatrix} \quad \text{and} \quad \tilde{N}_0 = -\tau \begin{bmatrix} \Lambda_{M_0} & & & & 0 \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ 0 & & & & \Lambda_{M_0} \end{bmatrix},$$

and:

$$\tilde{E}_0 = \begin{bmatrix} (\Lambda_{M_0} + \tau \Lambda_{A_0}) & & & \\ -\Lambda_{M_0} & \ddots & & \\ & \ddots & \ddots & \\ & & -\Lambda_{M_0} & (\Lambda_{M_0} + \tau \Lambda_{A_0}) \end{bmatrix}.$$

Importantly, matrix $\left(\alpha \, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 \, + \, \tilde{K}_0\right)$ will be block tridiagonal:

$$(\alpha \, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 \, + \, \tilde{K}_0) = \begin{bmatrix} Z_1 & Z_3 & & & \\ Z_3 & Z_1 & Z_3 & & \\ & \ddots & \ddots & \ddots & \\ & & Z_3 & Z_1 & Z_3 \\ & & & Z_3 & Z_2 \end{bmatrix}$$

where Z_1, Z_2 and Z_3 are the following diagonal matrices:

$$\begin{cases} Z_1 = \dfrac{4\,\tau}{6} \, \Lambda_{M_0} + \dfrac{\alpha}{\tau} \, (\Lambda_{M_0} + \tau \, \Lambda_{A_0})^2 \Lambda_{M_0}^{-1} + \dfrac{\alpha}{\tau} \, \Lambda_{M_0} \\ Z_2 = \dfrac{\gamma\tau}{6} \, \Lambda_{M_0} + \dfrac{\alpha}{\tau} \, (\Lambda_{M_0} + \tau \, \Lambda_{A_0})^2 \Lambda_{M_0}^{-1} \\ Z_3 = \dfrac{\tau}{6} \, \Lambda_{M_0} - \dfrac{\alpha}{\tau} \, (\Lambda_{M_0} + \tau \, \Lambda_{A_0}). \end{cases} \tag{10.191}$$

The action of C_0^{-1} can now be computed using the expression:

$$C_0^{-1} = V^T \tilde{N}_0^{-1} \tilde{E}_0 \left(\alpha \, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 \, + \, \tilde{K}_0\right)^{-1} \tilde{E}_0^T \tilde{N}_0^{-T} V. \tag{10.192}$$

As C_0^{-1} is a product of several matrices, this involves consecutive products with V, $\tilde{N}_0^{-T}$, $\tilde{E}_0^T$, $\left(\alpha \, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 \, + \, \tilde{K}_0\right)^{-1}$, $\tilde{E}_0$, $\tilde{N}_0^{-1}$ and V^T. Computing the action of $\left(\alpha \, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 \, + \, \tilde{K}_0\right)^{-1}$ corresponds to solving a block tridiagonal linear system with the coefficient matrix $\left(\alpha \, \tilde{E}_0^T \tilde{N}_0^{-1} \tilde{E}_0 \, + \, \tilde{K}_0\right)$ in (10.191). However, since Z_1, Z_2 and Z_3 are diagonal matrices, permuting the rows and columns of (10.191) so that indices modulo n belong to the same block will yield a block diagonal matrix whose diagonal blocks are tridiagonal. The permuted system can be solved using a direct solver at a cost of $O(n\,l)$ and the total computational cost will be proportional to $2\,l$ applications of V_0, and solving l tridiagonal linear systems system of size n. We omit the details.

Remark 10.97. If M and A are diagonalized by the FST, then we can choose $M_0 = M$ and $A_0 = A$, yielding $C_0 = C$. In this case, the preceding method will yield a fast direct solver, with multiplication by V costing $O(l\,n\,\log(n))$. Generally, however, C_0 will be a formal preconditioner for C.

Remarks on Real Time Applications. In real time applications, at each time t_i, the control $\mathbf{u}_{i+1}$ for $t \in (t_i, t_{i+1})$ must be computed in real time. If the output state $\mathbf{y}(.)$ is observed at each time t_i, this will provide additional information for determining the control. One of the advantages of the Riccati approach is that once matrix $W(.)$ has been computed and stored, the control $\mathbf{u}_{i+1}$ can be expressed as an instantaneous "feedback" function of the observed output state $\mathbf{y}_i$. However, when n is large, storage and multiplication with matrix $W(.)$ can be prohibitively expensive. The alternative approach based on the solution of the reduced Hessian (10.187) may also not be viable in real time applications, without additional modifications.

To reduce the computational costs in real time applications, the optimal control problem on $[0, \mathcal{T}]$ may be replaced by a local optimal control problem on a smaller time interval, yielding an *approximate* solution. For instance, if the output state $\mathbf{y}(t_i)$ is observed at time t_i, we may seek an approximate control $\hat{\mathbf{u}}_{i+1}$ on (t_i, t_{i+1}) by solving an optimal control problem on $[t_i, t_i + l_0\,\tau]$ choosing $1 \le l_0 \ll l$. To obtain a local tracking function, the global tracking function $\mathbf{y}_*(t)$ on $[0, \mathcal{T}]$ can be restricted to $(t_i, t_i + l_0\tau)$, or the current state $\mathbf{y}(t_i)$ and the terminal target $\mathbf{y}_*(\mathcal{T})$ may be interpolated. Then, the Hessian system (10.187) can be replaced by the following smaller local Hessian system, resulting from the control problem on $(t_i, t_i + l_0\tau)$:

$$\begin{cases} \hat{C}\,\hat{\mathbf{U}}_i = \hat{\mathbf{g}}_i, \quad \text{where} \\ \quad \hat{C} = (\hat{G} + \hat{N}^T \hat{E}^{-T} \hat{K} \hat{E}^{-1} \hat{N}) \;\text{ and} \\ \hat{\mathbf{U}}_i = \left(\hat{\mathbf{u}}_{i+1}^T, \ldots, \hat{\mathbf{u}}_{i+l_0}^T\right)^T, \end{cases} \tag{10.193}$$

where $\tilde{\mathbf{g}}_i$ is computed based on $\mathbf{y}(t_i)$ and the local tracking function (omitting the $\mathbf{y}_*(\mathcal{T})$ term). Here, $\hat{G}$, $\hat{N}$, $\hat{E}$ and $\hat{K}$ have the same block structure as G, N, E and K, respectively, with l_0 blocks, instead of l blocks. The local Hessian $\hat{C}$ will be of size $m\,l_0$, and if $m \ll n$ and $l_0 \ll l$, a direct solver can be used.

In the limiting case $l_0 = 1$, we obtain $\hat{G} = \alpha\,\tau Q$, $\hat{N} = -\tau\,B$, $\hat{E} = (M + \tau A)$, $\hat{K} = \tau\,M$ and $\mathbf{F} = M\mathbf{y}_i$ and $K\,\mathbf{Y}_* = \tau\,M\,\mathbf{y}_*(t_{i+1})$. This will yield:

$$\begin{cases} \hat{C}\,\hat{\mathbf{u}}_i = \hat{\mathbf{g}}_i, \quad \text{where} \\ \quad \hat{C} = \left(\alpha\,\tau Q + \theta_i\,\tau^2\,B^T(M + \tau A)^{-1} M (M + \tau A)^{-1} B\right) \\ \quad \hat{\mathbf{g}}_i = \theta_i\,\tau^2\,B^T(M + \tau A)^{-1} M\, \left((M + \tau A)^{-1} M\mathbf{y}_i - \mathbf{y}_*(t_{i+1})\right), \end{cases}$$

where $\theta_i = \frac{2\tau}{6}$ for $i < l$ and $\theta_i = \frac{\gamma\tau}{6}$ for $i = l$. This linear system will be of size m, and may be solved using a direct solver. Other heuristic choices of $K\,\mathbf{Y}_*$ and $\hat{\mathbf{g}}_i$ may also be used.

Remark 10.98. Alternative approximate control problems may be obtained. For instance, instead of the local Hessian system, we may solve the global Hessian system with a larger time step $\tau = (\mathcal{T}/l_0)$ for small l_0. This will yield an approximate control $\hat{\mathbf{U}} = \left(\hat{\mathbf{u}}_1^T, \ldots, \hat{\mathbf{u}}_{l_0}^T\right)^T$ to accuracy $O(\mathcal{T}/l_0)$.

11

Non-Matching Grid Discretizations

A *non-matching grid* is a collection of *overlapping* or *non-overlapping* grids, with associated subdomains that cover a domain, where the grids are obtained by the independent triangulation of the subdomains, without requirement to match with the grids adjacent to it, see Fig. 11.1. In this chapter, we describe several methods for the *global discretization* of a self adjoint and coercive *elliptic equation* on a non-matching grid:

- *Mortar element* discretization of an elliptic equation.
- *Chimera* (composite grid or Schwarz) discretization of an elliptic equation.
- Alternative non-matching grid discretizations of an elliptic equation.

Each non-matching grid discretization is based on a *hybrid formulation* of the underlying elliptic equation on its associated subdomain decomposition. The mortar element method, for instance, is formulated for a *non-overlapping* non-matching grid, and employs a Lagrange multiplier hybrid formulation of the elliptic equation, which enforces *weak matching* of the solution across adjacent subdomains [MA4, BE18, BE23, BE6, BE4, WO, WO4, WO5, KI], while the Chimera discretization is a finite difference discretization on an *overlapping* grid that enforces *strong matching* of the solution across adjacent grids, using a Schwarz formulation [ST, ST6, GR16, HE9, HE10, GO7, CA17].

Chap. 11.1 describes the hybrid formulations used in the mortar and Chimera discretizations of an elliptic equation. Chap. 11.2 and Chap. 11.3 describe the saddle point and the non-conforming versions of the mortar element discretization. The former yields a saddle point system, while the latter yields a positive definite system. Chap. 11.4 describes the Chimera discretization of an elliptic equation. It yields a non-symmetric linear system. Chap. 11.5 heuristically outlines the *Steklov-Poincare*, *least squares-control* and *partition of unity* discretizations. Chap. 11.6 outlines heuristic discretizations of a parabolic equation on a non-matching space-time grid. Alternative approaches are described in [DO4, PH, TH, KU7, CA7, AC5, HU3].

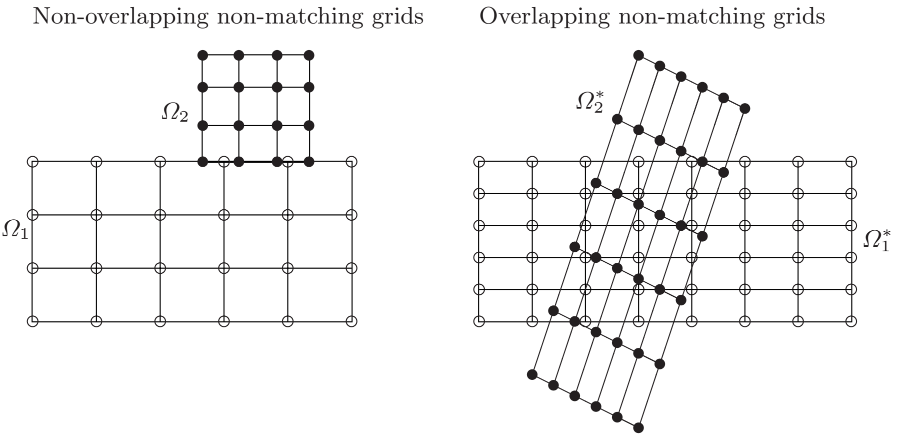

Non-overlapping non-matching grids
Overlapping non-matching grids
Ω_2
Ω_1
Ω_2^*
Ω_1^*

Fig. 11.1. Two subdomain non-matching grids

11.1 Multi-Subdomain Hybrid Formulations

In this section, we describe the multi-subdomain hybrid formulations used to construct the *mortar element* and *Chimera* discretizations of a self adjoint and coercive elliptic equation. The mortar element discretization employs the Lagrange multiplier formulation which enforces *weak continuity* of the local solutions across non-overlapping subdomains, while the Chimera discretization employs the Schwarz hybrid formulation which enforces *strong continuity* of the solution across overlapping subdomains.

We consider the following self adjoint and coercive elliptic equation:

$$\begin{cases} L\,u \equiv -\nabla \cdot (a(x)\nabla u) + c(x)u = f(x), & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega \end{cases} \tag{11.1}$$

where $f(x) \in L^2(\Omega)$ and $a(x) = a(x)^T > 0$ is a matrix function satisfying $\lambda_{min}\,(a(x)) \geq a_0 > 0$ and the scalar function $c(x)$ satisfies $c(x) \geq 0$. A weak formulation of elliptic equation (11.1) will seek $u \in H_0^1(\Omega)$ satisfying:

$$\begin{cases} \mathcal{A}(u,v) = (f,v)\,, \qquad \forall v \in H_0^1(\Omega), \qquad\qquad \text{where} \\ \mathcal{A}(u,v) = \int_\Omega \left(a(x)\,\nabla u \cdot \nabla v + c(x)\,u\,v\right)\,dx \\ \quad (f,v) = \int_\Omega f(x)\,v(x)\,dx. \end{cases} \tag{11.2}$$

An equivalent minimization formulation of (11.1) will seek $u \in H_0^1(\Omega)$:

$$J(u) = \min_{v\,\in\,H_0^1(\Omega)} J(v), \tag{11.3}$$

where $J(v) \equiv \frac{1}{2}\mathcal{A}(v,v) - (f,v)$. The Lagrange multiplier formulation will be derived using (11.3), and the Schwarz formulation using (11.1).

11.1.1 Lagrange Multiplier Hybrid Formulation

The multi-subdomain Lagrange multiplier hybrid formulation provides the framework for constructing the mortar element discretization of (11.1). This hybrid formulation employs a *non-overlapping* decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, and seeks the minimum to (11.3) by solving an equivalent constrained minimization problem, which replaces v by $(v_1, \ldots, v_p)$, where v_l is defined on Ω_l. Given v_l on Ω_l, we let $J_l(v_l) \equiv \frac{1}{2}\mathcal{A}_{\Omega_l}(v_l, v_l) - (f, v_l)_{\Omega_l}$ denote the local energy, see (11.12), and the total energy as $J_*(v_1, \ldots, v_p) \equiv \sum_{l=1}^{p} J_l(v_l)$. This will be defined even if v_l does not match v_j on $\partial\Omega_l \cap \partial\Omega_j$. However, by construction, if $v_l = v$ on each subdomain Ω_l, then $J_*(v_1, \ldots, v_p) = J(v)$.

The minimum of (11.3) may now be sought by minimizing $J_*(v_1, \ldots, v_p)$ subject to the *constraint* that v_l *weakly matches* v_j for all $\partial\Omega_l \cap \partial\Omega_j \neq \emptyset$. Weak matching between v_l and v_j requires $\int_{\partial\Omega_l \cap \partial\Omega_j} (v_l - v_j)\mu_{lj}\, ds_x = 0$ for $\mu_{lj} \in Y_{lj}$ where Y_{lj} is an appropriately chosen *multiplier space*. When $\Omega \subset \mathbb{R}^d$, weak matching is applied only for $\partial\Omega_l \cap \partial\Omega_j$ of dimension $(d-1)$. Finite element approximation of this hybrid formulation will yield the mortar element discretization, where finite element spaces must be chosen for approximating each v_l, and *discrete multiplier spaces* for approximating the spaces Y_{lj} on each $(d-1)$ dimensional interface $\partial\Omega_l \cap \partial\Omega_l$. Discrete multiplier spaces are constructed using finite element basis functions defined on the triangulation of $\partial\Omega_l \cap \partial\Omega_j$ obtained by restricting the triangulation from either Ω_l or Ω_j (the side chosen is called the non-mortar side). In this section, we describe the hybrid formulation. Multiplier spaces are described in Chaps. 11.2 and 11.3.

We elaborate the details. Let $\Omega_1, \ldots, \Omega_p$ be non-overlapping subdomains:

$$\overline{\Omega} = \overline{\Omega}_1 \cup \cdots \cup \overline{\Omega}_p \quad \text{with} \quad \Omega_i \cap \Omega_j = \emptyset, \quad \text{for } i \neq j.$$

Define $B^{(l)} = (\partial\Omega_l \backslash \partial\Omega)$ and $B_{[l]} = (\partial\Omega_l \cap \partial\Omega)$ as the *interior* and *exterior* boundary segments of Ω_l. Let $\mathcal{O}(l)$ denote the subdomains *adjacent* to Ω_l:

$$\mathcal{O}(l) \equiv \{j : \partial\Omega_l \cap \partial\Omega_j \neq \emptyset\}. \tag{11.4}$$

By construction $j \in \mathcal{O}(l) \Leftrightarrow l \in \mathcal{O}(j)$. We let $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$ denote the common interface between Ω_l and Ω_j. Using this notation, we can express:

$$\overline{B^{(l)}} = \cup_{\{j \in \mathcal{O}(l)\}} B_{lj}. \tag{11.5}$$

For each subdomain Ω_l and for each of its boundary segments B_{lj}, the user must *assign* a "side" $j \in \mathcal{I}(l)$ as the "non-mortar" side of B_{lj}. We require:

$$\mathcal{I}(l) \subset \mathcal{O}(l), \tag{11.6}$$

such that if $B_{lj} \neq \emptyset$ then either $j \in \mathcal{I}(l)$ or $l \in \mathcal{I}(j)$, *but not both.* When $\Omega \subset \mathbb{R}^d$, we define $\mathcal{I}_*(l) \subset \mathcal{I}(l)$ as the indices of segments B_{lj} of dimension $(d-1)$. Only interfaces B_{lj} of dimension $(d-1)$ will be used for matching.

When $B_{lj} \neq \emptyset$, the "side" of B_{lj} approached from Ω_j shall be referred to as the *nonmortar side* if $j \in \mathcal{I}_*(l)$, while the "side" of B_{lj} approached from Ω_l shall be referred to as the *mortar side*. Using the index set $\mathcal{I}(l)$ or $\mathcal{I}_*(l)$, we may decompose the interface $B = \cup_{l=1}^{p} \overline{B^{(l)}}$ as follows:

$$B \;=\; \cup_{l=1}^{p} \left(\cup_{\{j \in \mathcal{I}(l)\}} B_{lj} \right) \;=\; \cup_{l=1}^{p} \left(\cup_{\{j \in \mathcal{I}_*(l)\}} B_{lj} \right). \tag{11.7}$$

To derive a multi-subdomain Lagrange multiplier hybrid formulation of (11.3), let $u \in H_0^1(\Omega)$ denote the desired solution. On each subdomain, let u_l denote the restriction of u to $\overline{\Omega}_l$:

$$u_l \equiv \left(u|_{\overline{\Omega}_l} \right) \quad \text{for } \; 1 \leq l \leq p.$$

By construction, it will hold that $u_l \in H^1_{0,B_{[l]}}(\Omega_l)$, where:

$$H^1_{0,B_{[l]}}(\Omega_l) = \left\{ v_l \in H^1(\Omega_l) \; : \; v_l = 0 \; \text{ on } \; B_{[l]} \right\}, \tag{11.8}$$

with the boundary data of u_j having the following regularity [GR8]:

$$\left(u_l|_{B_{lj}} \right) \; \in \; H^{1/2}(B_{lj}), \quad \forall j \in \mathcal{O}(l).$$

Since $u \in H^1(\Omega)$, the following *strong matching* conditions will hold between u_l and u_j on each intersubdomain interface B_{lj}:

$$[u]_{lj} = 0 \qquad \text{on} \quad B_{lj}, \;\; \forall j \in \mathcal{I}_*(l), \;\; \forall l, \tag{11.9}$$

where $[u]_{lj} \equiv u_l - u_j$ on B_{lj}. Such strong matching conditions can be replaced by equivalent *weak matching* conditions as follows:

$$\int_{B_{lj}} [u]_{lj} \, \psi_{lj}(x) \, ds_x = 0, \quad \forall \psi_{lj}(x) \in H^{-1/2}(B_{lj}), \;\; \forall j \in \mathcal{I}_*(l), \;\; \forall l, \tag{11.10}$$

where $H^{-1/2}(B_{lj})$ denotes the *dual* space of $H_{00}^{1/2}(B_{lj})$.

The hybrid formulation of (11.3) based on $\Omega_1, \ldots, \Omega_p$ will minimize an *extended* energy functional $J_*(\cdot)$ subject to *constraints*. Given $w_l \in H^1_{0,B_{[l]}}(\Omega_l)$ for $1 \leq l \leq p$ where each w_l may not match across subdomains, define:

$$J_*(w) = \frac{1}{2} \mathcal{A}_*(w, w) - (f, w)_* \quad \text{for } \; w = (w_1, \ldots, w_p) \tag{11.11}$$

where $\mathcal{A}_*(w, w) = \sum_{l=1}^{p} \mathcal{A}_{\Omega_l}(w_l, w_l)$ and $(f, w)_* = \sum_{l=1}^{p} (f, w_l)_{\Omega_l}$ with:

$$\begin{cases} \mathcal{A}_{\Omega_l}(v_l, w_l) = \int_{\Omega_l} \left(a(x) \nabla v_l \cdot \nabla w_l + \, c(x) \, v_l \, w_l \right) \, dx \\ \quad (f, w_l)_{\Omega_l} = \int_{\Omega_l} f(x) \, w_l \, dx. \end{cases} \tag{11.12}$$

By construction, if $w_l(x) = v(x)$ on each Ω_l for some $v \in H^1(\Omega)$, then:

$$J(v) = J_*(w) \quad \text{for } \ w = (w_1, \ldots, w_p). \tag{11.13}$$

Define $X = \pi_{l=1}^p \left(H^1_{0,B_{[l]}}(\Omega_l) \right)$ with norm $\|w\|_X = \left(\sum_{l=1}^p \|w_l\|^2_{1,\Omega_l} \right)^{1/2}$ as the function space for $J_*(w)$. Then, if $w \in X$ and constraint (11.9) or (11.10) holds, there will exist $v \in H^1(\Omega)$ such that $w_l = v$ on each Ω_l, yielding (11.13). Motivated by this, we define a constraint set $\mathcal{K}_0$:

$$\mathcal{K}_0 \equiv \left\{ w \in X \, : \, \int_{B_{lj}} [w]_{lj} \, \psi_{lj} \, ds_x = 0, \forall \psi_{lj} \in H^{-1/2}(B_{lj}), \forall j \in \mathcal{I}_*(l), \forall \, l \right\}, \tag{11.14}$$

where $w = (w_1, \ldots, w_p)$ and $[w]_{lj} = w_l - w_j$ on each B_{lj}. The preceding observations suggest that when u is the solution to (11.3) and $u_l = u$ on each Ω_l, then $(u_1, \ldots, u_p)$ will satisfy:

$$J_*(u_1, \ldots, u_p) = \min_{(w_1, \ldots, w_p) \in \mathcal{K}_0} J_*(w_1, \ldots, w_p), \tag{11.15}$$

yielding a *constrained* minimization problem equivalent to (11.3).

As in Chap. 10, problem (11.15) can be reformulated as a saddle point problem by introducing Lagrange multipliers to enforce the constraints. We define a Lagrange *multiplier* space Y and its associated norm as follows:

$$Y \equiv \Pi_{l=1}^p \left(\Pi_{j \in \mathcal{I}_*(l)} H^{-1/2}(B_{lj}) \right), \ \ \|\psi\|_Y = (\sum_{l=1}^p \sum_{j \in \mathcal{I}_*(l)} \|\psi_{lj}\|^2_{-1/2,B_{lj}})^{1/2}, \tag{11.16}$$

for $\psi = \left((\psi_{lj})_{j \in \mathcal{I}_*(l)} \right)_{l=1}^p \in Y$. Define a bilinear form $\mathcal{M}_*(.,.) : X \times Y \to \mathbb{R}$:

$$\mathcal{M}_*(w, \psi) = \sum_{l=1}^p \sum_{j \in \mathcal{I}_*(l)} \int_{B_{lj}} [w]_{lj} \, \psi_{lj} \, ds_x, \tag{11.17}$$

for $w = (w_1, \ldots, w_p) \in X$ and $[w]_{lj} = w_l - w_j$. Then $\mathcal{K}_0$ can be expressed:

$$\mathcal{K}_0 = \{ v \in X \, : \, \mathcal{M}_*(v, \psi) = 0, \ \ \forall \, \psi \in Y \} \, . \tag{11.18}$$

Let $\mathcal{L}(w, \phi) = (\frac{1}{2}\mathcal{A}_*(w, w) - (f, w)_*) + \mathcal{M}_*(w, \phi)$ denote a Lagrangian functional associated with (11.15), with Lagrange multiplier $\phi \in Y$, where $\mathcal{A}_*(.,.) : X \times X \to \mathbb{R}$ and $(f, \cdot)_* : X \to \mathbb{R}$ are as defined in (11.12). Then, the *saddle point* $(u, \psi) \in X \times Y$ of $\mathcal{L}(\cdot, \cdot)$ will satisfy:

$$\begin{cases} \mathcal{A}_*(u, v) + \mathcal{M}_*(v, \psi) = (f, v)_*, & \forall v \in X \\ \qquad\qquad\qquad \mathcal{M}_*(u, \phi) = 0, & \forall \phi \in Y, \end{cases} \tag{11.19}$$

where $u \in X$ denotes the solution of (11.15). It is shown in (11.21) that each component $\psi_{lj}(\cdot)$ on B_{lj} of $\psi \in Y$ corresponds to Neumann *flux* data of $u(\cdot)$

on B_{lj} for $j \in \mathcal{I}_*(l)$. The Lagrange multiplier ψ can be eliminated from (11.19) as indicated in the following. The second row of (11.19) shows that $u \in \mathcal{K}_0$. Substituting $v(\cdot) \in \mathcal{K}_0 \subset X$ into the first row of (11.19) yields:

$$\mathcal{A}_*(u, v) = (f, v)_*, \quad \forall v \in \mathcal{K}_0, \tag{11.20}$$

since $\mathcal{M}_*(v, \phi) = 0$ for $v \in \mathcal{K}_0$. Thus, we may seek $u \in \mathcal{K}_0$ by solving (11.20), which will be a coercive problem, provided $\mathcal{A}_*(.,.)$ is coercive in $\mathcal{K}_0$.

Lemma 11.1. *Let $\mathcal{K}_0$ be as in (11.18).*

1. *Equip the space X with the following norm:*

$$\|w\|_X^2 = \sum_{l=1}^{p} \|w_l\|_{1,\Omega_l}^2.$$

2. *Equip the space Y with the following norm:*

$$\|\psi\|_Y^2 = \sum_{l=1}^{p} \sum_{j \in \mathcal{I}_*(l)} \|\psi_{lj}\|_{-1/2,B_{lj}}^2.$$

Then, the following will hold for some $c > 0$ and $\beta > 0$.

1. *The bilinear form $\mathcal{A}_*(.,.)$ will be X-coercive within the subspace $\mathcal{K}_0$:*

$$\mathcal{A}_*(w, w) \geq c\,\|w\|_X^2, \qquad \forall w \in \mathcal{K}_0.$$

2. *The following inf-sup condition will hold:*

$$\sup_{v \in X \setminus \{0\}} \frac{\mathcal{M}_*(v, \psi)}{\|v\|_X} \geq \beta\,\|\psi\|_Y, \quad \forall \psi \in Y.$$

3. *Saddle point problem (11.19) will be uniquely solvable and u will solve (11.20).*

Proof. See [BE18, BE4, WO5]. □

Remark 11.2. Given $\psi \in Y$, let $u = (u_1, \ldots, u_p) \in X$ solve the first row $\mathcal{A}_*(u, v) = (f, v)_* - \mathcal{M}_*(v, \psi)$ of (11.19). Then u_l solves the Neumann problem:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u_l) + c(x)\,u_l = f(x), & \text{in } \Omega_l \\ \mathbf{n} \cdot (a\nabla u_l) = -\psi_{lj}(x), & \text{on } B_{lj}, \ \forall j \in \mathcal{I}_*(l) \\ \mathbf{n} \cdot (a\nabla u_l) = \psi_{jl}(x), & \text{on } B_{lj}, \ \forall l \in \mathcal{I}_*(j) \\ u_l = 0, & \text{on } B_{[l]}, \end{cases} \tag{11.21}$$

with $\pm\psi_{lj}$ corresponding to the Neumann *flux* data of u_l on B_{lj}.

The mortar element method discretizes (11.1) on a nonmatching grid by Galerkin approximation of saddle point formulation (11.19), or nonconforming approximation of (11.20). In the saddle point approach, spaces X and Y are replaced by finite element or spectral subspaces $X_h \subset X$ and $Y_h \subset Y$, where subspace Y_h is referred to as the *multiplier* space. The discretization seeks an approximation (u_h, ψ_h) of (u, ψ) whose associated linear system is a saddle point system. If an approximation ψ_h of the subdomain fluxes ψ is not desired, then the approximation u_h of u can be sought directly by discretizing (11.20) using a nonconforming approximation $\mathcal{K}_0^h$ of $\mathcal{K}_0$. This will yield a symmetric positive definite linear system, as described in Chap. 11.3.

11.1.2 Schwarz Hybrid Formulation

The Schwarz hybrid formulation of (11.1) is constructed on a *non-overlapping* subdomain decomposition of Ω. It replaces (11.1) by a coupled system of partial differential equations for the restrictions of u to the subdomains. Let $\Omega_1^*, \ldots, \Omega_p^*$ form an overlapping decomposition of Ω with boundary segments:

$$B^{(l)} = (\partial\Omega_l^* \backslash \partial\Omega) \qquad \text{and} \qquad B_{[l]} = (\partial\Omega_l^* \cap \partial\Omega).$$

Ideally, if $\{\Omega_l\}_{l=1}^p$ is a non-overlapping decomposition of Ω, with subdomains of size h_0, define $\Omega_l^{\beta h_0} = \{x \in \Omega \,:\, \text{dist}(x, \Omega_l) \leq \beta\, h_0\}$ and choose $\Omega_l^* = \Omega_l^{\beta\, h_0}$ for some $0 < \beta < 1$. Let $\chi_1(x), \ldots, \chi_p(x)$ denote a smooth *partition of unity* subordinate to $\Omega_1^{\epsilon h_0}, \ldots, \Omega_p^{\epsilon\, h_0}$ for some $0 < \epsilon \ll \beta$ (so that $\Omega_l^{\epsilon h_0} \subset \Omega_l^*$):

$$\begin{cases} \chi_l(x) \geq 0, & \text{in } \Omega, \qquad\quad \text{for } 1 \leq l \leq p \\ \chi_l(x) \leq 1, & \text{in } \Omega, \qquad\quad \text{for } 1 \leq l \leq p \\ \chi_l(x) = 0, & \text{in } \Omega \backslash \overline{\Omega_l^{\epsilon\, h_0}}, \quad \text{for } 1 \leq l \leq p \\ \chi_1(x) + \cdots + \chi_p(x) = 1, & \text{in } \Omega. \end{cases}$$

Then, since $\chi_l(x) = 0$ on $B^{(l)}$, it will hold that $\sum_{j \neq l} \chi_j(x) = 1$ on each $B^{(l)}$. If we define $u_l \equiv u$ on Ω_l^* as the restriction of the solution u of (11.1) to Ω_l^*, then by construction the following equations will be satisfied by $u_l(\cdot)$:

$$\begin{cases} L\, u_l = f(x), & \text{in } \Omega_l^* \\ u_l = \sum_{j \neq l} \chi_j\, u_j, & \text{on } B^{(l)} \quad \text{for } 1 \leq l \leq p. \\ u_l = 0, & \text{on } B_{[l]} \end{cases} \tag{11.22}$$

The above corresponds to a *coupled system* of partial differential equations for the unknowns $(u_1(x), \ldots, u_p(x))$. If $c(x) \geq c_0 > 0$ and suitable regularity conditions hold for $f(x)$ and the subdomains, and the overlap is sufficiently large, this coupled system is *heuristically* expected to be well posed in the maximum norm, and to satisfy a *contraction property*, see Chap. 15.2 and [ST, CA17]. Given non-matching overset grids on the overlapping subdomains, a discretization of (11.1) can be obtained by discretizing each local equation in (11.22) using a finite difference scheme, and discretizing the interface matching conditions on each $B^{(l)}$, using weighted *interpolation*. This is described in Chap. 11.4, and yields a *non-symmetric* linear system, even though (11.1) is self adjoint. Although the discretization is simple to formulate, its stability and accuracy is sensitive to the amount of *overlap* between the subdomains.

Remark 11.3. An alternative *overlapping* subdomains based hybrid formulation was introduced in [CA7] for two subdomains, using a *partition of unity*, and extended to the multi-subdomain case in [AC5]. These hybrid formulations enable the construction of alternative non-matching grid discretizations.

Remark 11.4. Given the overlapping subdomains $\Omega_1^*, \ldots, \Omega_p^*$, the Lagrange multiplier formulation (11.19) can be extended to the case of *overlapping* subdomains to yield an alternative hybrid formulation of (11.1). However, the stability of the resulting formulation may again depend on the amount of overlap between the subdomains. To derive such a hybrid formulation, let $u_l = u$ on Ω_l^* denote the restriction of the solution u of (11.1) to Ω_l^*. Then, u_l will solve a Neumann problem on Ω_l^* for *unknown* data $-\psi_l$ on $B^{(l)}$:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u_l) + c(x)\, u_l = f(x), & \text{in } \Omega_l^* \\ \mathbf{n}_l \cdot (a\nabla u_l) = -\psi_l(x), & \text{on } B^{(l)} \\ u_l = 0, & \text{on } B_{[l]}. \end{cases} \tag{11.23}$$

Here $\mathbf{n}_l(x)$ denotes the exterior unit normal on $B^{(l)}$. These local fluxes $(\psi_1, \ldots, \psi_p)$ must be chosen so that $u_l = \sum_{j \neq l} \chi_j\, u_j$ on $B^{(l)}$, given a partition of unity $\chi_1(x), \ldots, \chi_p(x)$. To obtain a saddle point problem for determining $(u_1, \ldots, u_p)$ and $(\psi_1, \ldots, \psi_p)$, we define X and a multiplier space Y as:

$$X = \Pi_{l=1}^p \left(H^1_{0,B_{[l]}}(\Omega_l^*) \right) \quad \text{and} \quad Y = \Pi_{l=1}^p \left(H^{1/2}(B^{(l)}) \right)'.$$

Define $\mathcal{A}_*(.,.) : X \times X \to \mathbb{R}$ and linear functional $(f, \cdot)_* : X \to \mathbb{R}$ as:

$$\begin{cases} \mathcal{A}_*(v, w) = \sum_{l=1}^p \int_{\Omega_l^*} (a(x)\nabla v_l \cdot \nabla w_l + \, c(x)\, v_l\, w_l)\, dx \\ (f, w)_* = \sum_{l=1}^p \int_{\Omega_l^*} (f(x)\, w_l)\, dx. \end{cases} \tag{11.24}$$

Let $\mathcal{M}_*(.,.) : X \times Y \to \mathbb{R}$ denote the following bilinear form:

$$\mathcal{M}_*(w, \psi) = \sum_{l=1}^p \int_{B^{(l)}} \left(w_l(x) - \sum_{j \neq l} \chi_j(x) w_j(x) \right) \psi_l(x)\, ds_x. \tag{11.25}$$

Then, solve the saddle point problem which seeks $(u, \psi) \in X \times Y$ such that:

$$\begin{cases} \mathcal{A}_*(u, v) + \mathcal{M}_*(v, \psi) = (f, v)_*, & \forall v \in X \\ \mathcal{M}_*(u, \phi) = 0, & \forall \phi \in Y. \end{cases} \tag{11.26}$$

By construction, the constraints $u_l = \sum_{j \neq l} \chi_j u_j$ will hold on each $B^{(l)}$ for $u = (u_1, \ldots, u_p)$. We leave it to the reader to verify the stability and well posedness of Lagrange multiplier formulation (11.26). An overset grid discretization of (11.1) can be obtained by Galerkin approximation of (11.26), based on subspaces $X_h \subset X$ or $Y_h \subset Y$, yielding a saddle point system. A space $\Pi_{l=1}^{p} Y_h(B^{(l)})$ of multipliers may be chosen, with $Y_h(B^{(l)})$ based on the triangulation of Ω_l^*, as in mortar element discretizations.

Remark 11.5. The local solutions in a non-matching grid discretization may be mildly *discontinuous* across the subdomains. To obtain a *continuous* global solution, a conforming finite element solution can be constructed by modifying the non-matching grid [KU7], or a *partition of unity* method can be employed as in [HU3, BA6, BA7], or the solutions combined using a partition of unity.

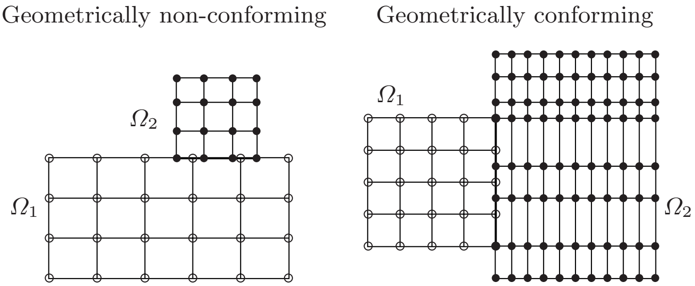
11.2 Mortar Element Discretization: Saddle Point Approach

In this section, we describe the saddle point version of a mortar element discretization of Dirichlet problem (11.1). Such a discretization is obtained by Galerkin approximation of (11.19) using finite element spaces $X_h \subset X$ and $Y_h \subset Y$. To obtain *stable* and *accurate* schemes, care must be exercised in the selection of these subspaces. The choice of X_h will be standard, however, the choice of the *discrete multiplier* space Y_h will be novel. Our discussion will first focus on *two* subdomain mortar discretizations before considering extensions to many subdomains. Two alternative choices of discrete multiplier spaces Y_h will be described, a *continuous* multiplier space and a *discontinuous* multiplier space, the latter equipped with an easily computed *biorthogonal basis* (dual basis). Iterative algorithms for solving the resulting systems will be outlined.

11.2.1 Notation

Let $\Omega_1, \ldots, \Omega_p$ be a nonoverlapping decomposition of Ω, with $B^{(l)} = \partial\Omega_l \cap \Omega$ and $B_{[l]} = \partial\Omega_l \cap \partial\Omega$. As before, $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$ will denote the common interface between Ω_l and Ω_j, with $\mathcal{O}(l)$ denoting the indices such that $B_{lj} \neq \emptyset$ when $j \in \mathcal{O}(l)$. It will thus hold that $j \in \mathcal{O}(l) \Leftrightarrow l \in \mathcal{O}(j)$.

Local Triangulations. We assume that each subdomain $\Omega_l \subset \mathbb{R}^d$ is triangulated by a *quasiuniform* grid $\mathcal{T}_{h_l}(\Omega_l)$ with elements of size h_l. On each boundary segment $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$ of dimension $(d-1)$, we let $\mathcal{T}_{h_l}(B_{lj})$ denote the restriction of triangulation $\mathcal{T}_{h_l}(\Omega_l)$ to B_{lj}. We shall assume, see assumption *(A.1)*, that the restriction of triangulation $\mathcal{T}_{h_l}(\Omega_l)$ to boundary segment B_{lj} and the restriction of triangulation $\mathcal{T}_{h_j}(\Omega_j)$ to boundary segment B_{lj}, triangulates B_{lj}. Mortar element discretizations select only *one* specific triangulation for each nonempty segment B_{lj}, either $\mathcal{T}_{h_l}(B_{lj})$ or $\mathcal{T}_{h_j}(B_{lj})$. We let $\mathcal{I}(l) \subset \mathcal{O}(l)$ denote indices such that $\mathcal{T}_{h_j}(B_{lj})$ is the chosen triangulation


Geometrically non-conforming

Geometrically conforming

Fig. 11.2. Two subdomain non-matching grids

of B_{lj} for each $j \in \mathcal{I}(l)$. When there are three or more subdomains, we only consider segments B_{lj} of dimension $(d-1)$ in typical mortar element discretizations. We denote the subset of indices j within $\mathcal{I}(l)$ for which B_{lj} is of dimension $(d-1)$ as $\mathcal{I}_*(l) \subset \mathcal{I}(l)$. The segments B_{lj} for $j \in \mathcal{I}_*(l)$ and $1 \le l \le p$ will partition $B = \cup_{l=1}^{p} B^{(l)}$.

Definition 11.6. *When $j \in \mathcal{I}_*(l)$, the "side" of B_{lj} approached from Ω_j is referred to as the "nonmortar" side, while the "side" of B_{lj} approached from Ω_l is referred to as the "mortar" side. Nodal unknowns on the mortar side of B_{lj} will be "master" variables, and on the nonmortar side "slave" variables.*

We shall henceforth focus primarily on nonmatching grids which satisfy certain *geometrical conformity* assumptions, see Fig. 11.2. Typically, condition *(A.1)* will always be assumed to hold, while the stronger condition *(A.2)* will be assumed to hold when improved theoretical bounds are desired.

Geometrical Conformity. We caution the reader that our terminology on geometrical conformity differs from that used in the literature.
(A.1) A nonmatching grid satisfies condition *(A.1)* if each local triangulation $\mathcal{T}_{h_l}(\Omega_l)$ restricted to B_{lj} *triangulates* this interface. Each restricted triangulation shall be denoted $\mathcal{T}_{h_l}(B_{lj})$ for $j \in \mathcal{I}(l)$. The local triangulations $\mathcal{T}_{h_l}(B_{lj})$ and $\mathcal{T}_{h_j}(B_{lj})$ of B_{lj} may not match on B_{lj}.

(A.2) A nonmatching grid satisfies condition *(A.2)*, if *(A.1)* holds and if each local triangulation $\mathcal{T}_{h_l}(B_{lj})$ *matches* $\mathcal{T}_{h_j}(B_{lj})$ on ∂B_{lj} for $j \in \mathcal{I}(l)$. This requires both local triangulations to share the same nodes on ∂B_{lj}.

Mortar element discretizations of (11.1) will be obtained by the Galerkin approximation of (11.19) using finite element spaces $X_h \subset X$ and $Y_h \subset Y$ defined on the nonmatching grid, where:

$$X = \Pi_{l=1}^{p} \left(H^1_{0,B_{[l]}}(\Omega_l) \right) \quad \text{and} \quad Y = \Pi_{l=1}^{p} \left(\Pi_{j \in \mathcal{I}_*(l)} H^{-1/2}(B_{lj}) \right). \tag{11.27}$$

Discrete solutions $u_h \in X_h \subset X$ and $\psi_h \in Y_h \subset Y$ will be sought satisfying:

$$\begin{cases} \mathcal{A}_*(u_h, v_h) + \mathcal{M}_*(v_h, \psi_h) = (f, v_h)_*, & \forall v_h \in X_h \\ \mathcal{M}_*(u_h, \phi_h) = 0, & \forall \phi_h \in Y_h, \end{cases} \tag{11.28}$$

where $\mathcal{A}_*(u,v) \equiv \sum_{l=1}^{p} \mathcal{A}_{\Omega_l}(u_l,v_l)$ and $(f,v)_* \equiv \sum_{l=1}^{p} (f,v_l)_{\Omega_l}$ with:

$$\begin{cases} \mathcal{A}_{\Omega_l}(u_l,v_l) \equiv \int_{\Omega_l} (a(x)\nabla u_l \cdot \nabla v_l + c(x)\, u_l\, v_l)\, dx \\ \quad (f,v_l)_{\Omega_l} \equiv \int_{\Omega_l} f(x)\, v_l(x)\, dx \\ \mathcal{M}_*(u,\phi) \equiv \sum_{l=1}^{p} \sum_{j \in \mathcal{I}_*(l)} \int_{B_{lj}} [u]_{lj}\, \phi_{lj}\, ds_x, \end{cases} \tag{11.29}$$

and $[u]_{lj} = u_l - u_j$ with $u = (u_1, \ldots, u_p)$, $v = (v_1, \ldots, v_p) \in X$. Expanding $u_h \in X_h$ and $\phi_h \in Y_h$ relative to bases for X_h and Y_h will yield a saddle point linear system, whose block structure shall be described later.

Local Finite Element Spaces. We shall let $X_{h_l} \subset H^1_{0,B_{[l]}}(\Omega_l)$ denote a conforming finite element subspace defined on triangulation $\mathcal{T}_{h_l}(\Omega_l)$. If $w_{h_l} \in X_{h_l}$ denotes a local finite element function, we shall associate a nodal vector $\mathbf{w}_{h_l} = \left(\mathbf{w}_I^{(l)^T}, \mathbf{w}_B^{(l)^T}\right)^T$ where $\mathbf{w}_I^{(l)}$ denotes a vector of nodal values corresponding to *interior nodes* of $\mathcal{T}_{h_l}(\Omega_l)$, while $\mathbf{w}_B^{(l)}$ denotes a vector of nodal values on $B^{(l)}$. We shall let $n_I^{(l)}$ and $n_B^{(l)}$ denote the number of nodes of $\mathcal{T}_{h_l}(\Omega_l)$ in the *interior* and on $B^{(l)}$, respectively, with n_l denoting $n_l = n_I^{(l)} + n_B^{(l)}$.

Local Stiffness Matrices and Load Vectors. We shall let $A^{(l)}$ and $\mathbf{f}^{(l)}$ denote the local stiffness matrix and load vector corresponding to:

$$\mathcal{A}_{\Omega_l}(v_{h_l}, w_{h_l}) = \begin{bmatrix} \mathbf{v}_I^{(l)} \\ \mathbf{v}_B^{(l)} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{IB}^{(l)^T} & A_{BB}^{(l)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(l)} \\ \mathbf{w}_B^{(l)} \end{bmatrix}, \; (f, v_{h_l})_{\Omega_l} = \begin{bmatrix} \mathbf{v}_I^{(l)} \\ \mathbf{v}_B^{(l)} \end{bmatrix}^T \begin{bmatrix} \mathbf{f}_I^{(l)} \\ \mathbf{f}_B^{(l)} \end{bmatrix},$$

where $A_{II}^{(l)}$, $A_{IB}^{(l)}$ and $A_{BB}^{(l)}$, are matrices of size $n_I^{(l)}$, $n_I^{(l)} \times n_I^{(l)}$ and $n_B^{(l)}$, with $\mathbf{f}_I^{(l)}$ and $\mathbf{f}_B^{(l)}$ of size $n_I^{(l)}$ and $n_B^{(l)}$, respectively.

Guidelines for Choosing $\boldsymbol{X_h}$ and $\boldsymbol{Y_h}$. Theoretical analysis [GI3, BR33] of saddle point problems suggests three requirements for X_h and Y_h.

- For the *stability* of discretization (11.28), given X_h, subspace Y_h must be compatible with X_h, satisfying *inf-sup* condition (11.31) uniformly in h.
- For *stability*, bilinear form $\mathcal{A}_*(.,.)$ must satisfy $\mathcal{A}_*(v_h, v_h) \geq \alpha \, \|v_h\|_X^2$ for $v_h \in \mathcal{K}_0^h$, i.e., be *coercive*, where:

  $$\mathcal{K}_0^h = \{v_h \in X_h : \mathcal{M}_*(v_h, \psi_h) = 0, \, \forall \psi_h \in Y_h\},$$

  and $\alpha > 0$ is independent of h.
- For the *accuracy* of discretization (11.28), subspaces X_h and Y_h must have *approximation properties* such as (11.30) and (11.32).

Motivated by the preceding, the subspaces X_h and Y_h are chosen as follows.

Subspace $\boldsymbol{X_h}$. On each subdomain Ω_l, let $X_{h_l}(\Omega_l) \subset H^1_{0,B_{[l]}}(\Omega_l)$ denote a finite element space of continuous piecewise polynomial functions of degree q_l with *zero* Dirichlet boundary conditions on $B_{[l]}$:

$$X_{h_l}(\Omega_l) \equiv \{u_{h_l} \in C(\overline{\Omega}_l) : u|_\tau \in P_{q_l}(\tau), \, \forall \tau \in \mathcal{T}_{h_l}(\Omega_l), \, u_{h_l} = 0 \text{ on } B_{[l]}\},$$

where $\tau \in \mathcal{T}_{h_l}(\Omega_l)$ denotes an element and $P_{q_l}(\tau)$ denotes polynomials of degree q_l or less on τ. Define $X_h \subset X$ as the product space:

$$X_h \equiv \left(\Pi_{l=1}^{p} X_{h_l}(\Omega_l)\right) \subset X,$$

equipped with the inherited norm. Such a choice of finite element space on each subdomain Ω_l will ensure an *approximation property* of the form:

$$\inf_{v_{h_l} \in X_{h_l}(\Omega_l)} \|u - v_{h_l}\|_{1,\Omega_l} \leq C\, h_l^{q_l} \|u\|_{q_l+1,\Omega_l}, \qquad (11.30)$$

for sufficiently smooth u. For simplicity, we shall henceforth assume that all triangulations involve triangular elements in two dimensions and tetrahedral elements in three dimensions, and that each finite element space $X_{h_l}(\Omega_l)$ consists of continuous *piecewise linear* functions, i.e., $q_l = 1$.

Subspace $\boldsymbol{Y_h}$. For mortar element discretizations, the multiplier space Y for the *flux*, and defined by (11.27), will admit *discontinuous* functions on each interface segment B_{lj}, since $L^2(B_{lj}) \subset H^{-1/2}(B_{lj})$. The uniform *inf-sup* condition [GI3, BR33] corresponds to the requirement that $M_h : X_h \rightarrow Y_h'$ (where Y_h' denotes the dual space of Y_h) induced by $\mathcal{M}_*(.,.) : X_h \times Y_h \rightarrow \mathbb{R}$ be *surjective*, satisfying:

$$\sup_{v_h \in X_h} \frac{\mathcal{M}_*(v_h, \psi_h)}{\|v_h\|_X} \geq \beta \|\psi_h\|_Y, \qquad \forall \psi_h \in Y_h, \qquad (11.31)$$

for $\beta > 0$ independent of h, i.e., $\|M_h^{\dagger}\| \leq (1/\beta)$.

The requirement that $\mathcal{A}_*(.,.)$ be *coercive* within $\mathcal{K}_0^h$ will be satisfied for arbitrary subspaces X_h and Y_h *provided* $c(x) \geq c_0 > 0$. This is because, $\mathcal{A}_*(v, v)$ will be equivalent to $\|v\|_X^2$ in X using (11.12) for $c(x) \geq c_0 > 0$. When $c(x) = 0$, coercivity will be lost in the interior subdomains (as in the FETI method of Chap. 4). If X_h satisfies (11.30), then ideally Y_h should be chosen with compatible approximation property. Let $Y_h(B_{lj})$ denote the discrete multiplier space on B_{lj}. If $\phi \in Y$ is sufficiently smooth, we require:

$$\inf_{\phi_h \in Y_h(B_{lj})} \|\phi - \phi_h\|_{Y_h(B_{lj})} \leq c \max\{h_l^{q_l}, h_j^{q_j}\} \|\phi\|_{-1/2+q_*,B_{lj}}. \qquad (11.32)$$

Generally, subspace Y_h will be pivotal for the *stability* and *accuracy* of the mortar element discretization. Once subspace X_h has been chosen, subspace Y_h must be selected so that a uniform *inf-sup* condition holds, and so that an approximation property compatible with the accuracy of X_h holds. The multiplier space Y defined by (11.27) will represent the *flux* on each interface segment B_{lj} in $\cup_{l=1}^{p} \left(\cup_{j \in \mathcal{I}_*(l)} B_{lj}\right)$ and must include the *constant* functions on B_{lj}. The multiplier space Y and the discrete multiplier space Y_h will be:

$$Y = \Pi_{l=1}^{p} \left(\Pi_{j \in \mathcal{I}_*(l)} H^{-1/2}(B_{lj})\right) \text{ and } Y_h = \Pi_{l=1}^{p} \left(\Pi_{j \in \mathcal{I}_*(l)} Y_h(B_{lj})\right),$$

where each space $Y_h(B_{lj})$ denotes the discrete multipliers on B_{lj}. The space $Y_h(B_{lj})$ is typically represented as *span*$\{\psi_1^{lj}(x), \ldots, \psi_{m_{lj}}^{lj}(x)\}$.

When $j \in \mathcal{I}(l)$ or $j \in \mathcal{I}_*(l)$, the space $Y_h(B_{lj})$ will be constructed using triangulation $\mathcal{T}_{h_j}(B_{lj})$ of B_{lj}, and required to satisfy the following properties:

- $Y_h(B_{lj})$ has the same *dimension* as $X_{h_j}(B_{lj}) \cap H_0^1(B_{lj})$.
- $Y_h(B_{lj})$ contains *constant* functions on B_{lj}.
- $Y_h(B_{lj})$ is at most of degree q_j in each element of $\mathcal{T}_{h_j}(B_{lj})$.

Since the *inf-sup* condition requires that the map $M_h : X_h \rightarrow Y_h'$ induced by $\mathcal{M}_*(.,.)$ be *surjective*, and since $\mathcal{M}_*(.,.)$ depends only on the degrees of freedom on B, this *restricts* the dimension of space Y_h not to exceed the degrees of freedom on B.

Remark 11.7. When there are three or more subdomain, there will be segments B_{lj} with dimension is $(d-2)$ or lower, such as when B_{lj} is a *cross point* in two dimensions, or an *edge* or *cross point* in three dimensions. In mortar element discretizations, the discrete multiplier spaces $Y_h(B_{lj})$ are defined only on segments B_{lj} of dimension $(d-1)$, when $\Omega \subset \mathbb{R}^d$. When B_{lj} is triangulated by $\mathcal{T}_{h_j}(B_{lj})$, the dimension of $Y_h(B_{lj})$ on segment B_{lj} will correspond to the number of *interior nodes* of triangulation $\mathcal{T}_{h_j}(B_{lj})$ on B_{lj}.

Next, we describe the multiplier space $Y_h(B_{lj})$ for two subdomains.

11.2.2 Two Subdomain Discretizations

Two subdomain mortar element discretizations have a simple structure, and in the multi-subdomain case, each multiplier space $Y_h(B_{lj})$ is defined based on the two subdomain case. Let Ω_1 and Ω_2 be polygonal subdomains forming a nonoverlapping decomposition of $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$. Let each subdomain Ω_l be triangulated by a quasiuniform triangulation $\mathcal{T}_{h_l}(\Omega_l)$ with grid size h_l. We shall assume that either conformity condition *(A.1)* or *(A.2)* holds, and let $B_{12} = \partial\Omega_1 \cap \partial\Omega_2$ denote the common interface between the two subdomains. Let j_* denote the *nonmortar* side with $\mathcal{T}_{h_{j_*}}(B_{12})$ as the triangulation of B_{12}.

In *two dimensions*, a two subdomain interface B_{12} will either be a loop (homeomorphic to a circle) or an arc (homeomorphic to the unit interval). When B_{12} is a loop, conformity condition *(A.1)* requires both local triangulations to triangulate B_{12}. If B_{12} is an arc, then conformity condition *(A.1)* requires the endpoints of B_{12} be nodes of both local triangulations. For two dimensional domains *(A.1)* and *(A.2)* are equivalent.

In *three dimensions*, geometrical conformity is more restrictive. A two subdomain interface B_{12} will be two dimensional, and homeomorphic either to the surface of a sphere (when one subdomain is floating) or to a rectangle. In the former case, condition *(A.1)* requires both triangulations to triangulate B_{12} (and condition *(A.2)* can be ignored), while in the latter case, ∂B_{12} will be a loop, and condition *(A.2)* will hold only if the nodes of both local triangulations match on ∂B_{12}.

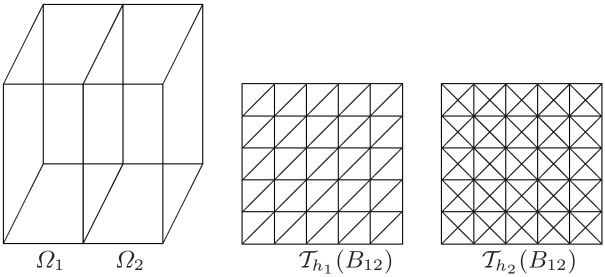


Fig. 11.3. A non-matching grid in three dimensions satisfying *(A.2)*

Remark 11.8. The nonmatching grid on the left side of Fig. 11.2 violates *(A.1)*, since the endpoints of B_{12} are not nodes of $\mathcal{T}_{h_1}(\Omega_1)$, by contrast the non-matching grid on the right side of Fig. 11.2 satisfies *(A.1)*. When conformity condition *(A.1)* is violated on Ω_l, zero boundary conditions may not be accurately imposed on $\partial\Omega_1 \cap \partial\Omega$. For the *three dimensional* domain in Fig. 11.3, triangulations $\mathcal{T}_{h_1}(B_{12})$ and $\mathcal{T}_{h_2}(B_{12})$ do not match on B_{12}. However, they do match on ∂B_{12}, yielding that condition *(A.2)* is satisfied.

In the two grid case, given subspaces $X_h \subset X$ and $Y_h \subset Y$, the mortar element discretization of (11.1) will seek $u_h \in X_h$ and $\psi_h \in Y_h$ satisfying:

$$\begin{cases} \mathcal{A}_*(u_h, v_h) + \mathcal{M}_*(v_h, \psi_h) = (f, v_h)_*, & \forall v_h \in X_h \\ \qquad\qquad\qquad \mathcal{M}_*(u_h, \phi_h) = 0, & \forall \phi_h \in Y_h, \end{cases} \tag{11.33}$$

where $\mathcal{A}_*(v, w) = \sum_{l=1}^{2} \mathcal{A}_{\Omega_l}(v_l, w_l)$ and $(f, w)_* = \sum_{l=1}^{2} (f, w_l)_{\Omega_l}$ with:

$$\begin{cases} \mathcal{A}_{\Omega_l}(v, w) = \int_{\Omega_l} \left(a(x)\nabla v_l \cdot \nabla w_l + c(x)\, v_l\, w_l\right) dx \\ \quad (f, w)_{\Omega_l} = \int_{\Omega_l} f(x)\, w_l(x)\, dx \\ \mathcal{M}_*(v, \psi) = \int_{B_{12}} \left(v_1(x) - v_2(x)\right)\, \psi(x)\, ds_x. \end{cases} \tag{11.34}$$

Given a basis for X_h and Y_h, let $\mathbf{u}^{(l)}$ and $\boldsymbol{\psi}$ denote the coefficient vectors of $u_{h_l}(x) \in X_{h_l}(\Omega_l)$ and $\psi_h(x) \in Y_h$, respectively, relative to each basis. If $X_{h_l}(\Omega_l)$ is a finite element space, let $\mathbf{u}_I^{(l)}$ and $\mathbf{u}_B^{(l)}$ denote vectors of nodal values in the interior of Ω_l and on B_{12}, of size $n_I^{(l)}$ and $n_B^{(l)}$, respectively. Also let m denote the dimension of Y_h. Then, discretization (11.33) will yield the following saddle point system:

$$\begin{bmatrix} A^{(1)} & 0 & M^{(1)^T} \\ 0 & A^{(2)} & -M^{(2)^T} \\ M^{(1)} & -M^{(2)} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(2)} \\ \mathbf{0} \end{bmatrix}, \tag{11.35}$$

where for $l = 1, 2$:

$$A^{(l)} = \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{IB}^{(l)^T} & A_{BB}^{(l)} \end{bmatrix}, \;\; M^{(l)} = \begin{bmatrix} 0 & M_B^{(l)} \end{bmatrix}, \;\; \mathbf{u}^{(l)} = \begin{bmatrix} \mathbf{u}_I^{(l)} \\ \mathbf{u}_B^{(l)} \end{bmatrix}, \;\; \mathbf{f}^{(l)} = \begin{bmatrix} \mathbf{f}_I^{(l)} \\ \mathbf{f}_B^{(l)} \end{bmatrix}, \tag{11.36}$$

with the matrices and vectors defined by:

$$\begin{cases} \mathcal{A}_{\Omega_l}(u_{h_l}, u_{h_l}) = \mathbf{u}^{(l)^T} A^{(l)} \mathbf{u}^{(l)}, \\ (f, u_{h_l})_{\Omega_l} = \mathbf{u}^{(l)^T} \mathbf{f}^{(l)}, \\ \mathcal{M}_* (u_h, \phi_h) = \boldsymbol{\phi}^T \left(M^{(1)} \mathbf{u}^{(1)} - M^{(2)} \mathbf{u}^{(2)} \right) \\ \qquad\qquad\quad = \boldsymbol{\phi}^T \left(M_B^{(1)} \mathbf{u}_B^{(1)} - M_B^{(2)} \mathbf{u}_B^{(2)} \right). \end{cases} \tag{11.37}$$

Here block matrices $A_{II}^{(l)}$, $A_{IB}^{(l)}$, $A_{BB}^{(l)}$ and $M_B^{(l)}$ will be of size $n_I^{(l)}$, $n_I^{(l)} \times n_I^{(l)}$, $n_B^{(l)}$ and $m \times n_B^{(l)}$, respectively, while the vectors $\mathbf{f}_I^{(l)}$, $\mathbf{f}_B^{(l)}$ and $\mathbf{f}^{(l)}$ will be of size $n_I^{(l)}$, $n_B^{(l)}$ and n_l, respectively. In the two subdomain case, we note that if $\boldsymbol{\phi}$ denotes the coefficient vector associated with $\phi_h(x) \in Y_h$, then:

$$\boldsymbol{\phi}^T M_B^{(l)} \mathbf{u}_B^{(l)} = \int_{B_{12}} u_{h_l}(x)\, \phi(x)\, ds_x. \tag{11.38}$$

Matrix $M_B^{(l)}$ will be *rectangular* of size $m \times n_B^{(l)}$. If Dirichlet conditions are imposed on $\partial\Omega$, then nodal unknowns on ∂B_{12} will be zero. If the multiplier space $Y_h(B_{12})$ can be chosen so that $M_B^{(1)}$ or $M_B^{(2)}$ is an invertible square matrix, it would enable us to solve for $\mathbf{u}_B^{(1)}$ or $\mathbf{u}_B^{(2)}$ in terms of the other:

$$\mathbf{u}_B^{(1)} = \left(M_B^{(1)} \right)^{-1} M_B^{(2)} \mathbf{u}_B^{(2)} \quad \text{or} \quad \mathbf{u}_B^{(2)} = \left(M_B^{(2)} \right)^{-1} M_B^{(1)} \mathbf{u}_B^{(1)}.$$

This will be particularly useful for explicitly deriving a symmetric positive definite linear system for $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$, by eliminating the multipliers $\boldsymbol{\psi}$.

Choice of Subspace X_h. Each subdomain space $X_{h_l}(\Omega_l)$ can be chosen to be any *conforming* finite element space defined on triangulation $\mathcal{T}_{h_l}(\Omega_l)$ and satisfying $X_{h_l}(\Omega_l) \subset H^1_{0,B_{[l]}}(\Omega_l)$. The *two subdomain* space X_h will be:

$$X_h = X_{h_1}(\Omega_1) \times X_{h_2}(\Omega_2).$$

Typically, each $X_{h_l}(\Omega_l)$ is chosen to consist of continuous functions which are polynomials of degree q_l on each element of $\mathcal{T}_{h_l}(\Omega_l)$. In this case, the approximation error will satisfy (11.30).

We shall describe two alternative choices of multiplier spaces Y_h for *two* subdomain decompositions, motivated by stability and approximation considerations [MA4, BE18, PH, BE23, BE6, BE4, BR2, WO, WO4, WO5, KI].

- A space Y_h of *continuous* piecewise polynomials defined on triangulation $\mathcal{T}_{h_{j_*}}(B_{12})$, and having the same dimension as $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$ (or $X_{h_{j_*}}(B_{12})$ when B_{12} is homeomorphic to a loop or a sphere). The space Y_h must contain the *constant* functions.

- A space Y_h of *discontinuous* piecewise polynomials defined on triangulation $\mathcal{T}_{h_{j_*}}(B_{12})$, generated by a basis *biorthogonal* (dual) to the standard nodal basis for $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$ (or $X_{h_{j_*}}(B_{12})$ when B_{12} is homeomorphic to a loop or a sphere), containing *constant* functions.

For simplicity, our description will focus on piecewise linear finite elements, and j_* will denote the index of the *nonmortar* (slave variable) side.

Continuous Multiplier Space $\boldsymbol{Y_h}$. Continuous multipliers were used in early mortar element methods [MA4, BE18, PH, BE23, BE6, BE4, BR2]. When condition *(A.1)* holds, let $X_{h_{j_*}}(B_{12})$ denote the restriction of finite element space $X_{h_{j_*}}(\Omega_{j_*})$ to triangulation $\mathcal{T}_{h_{j_*}}(B_{12})$, with degree q_{j_*} polynomials in each element. The space $Y_h = Y_h(B_{12})$ must satisfy the following.

- Y_h has the same *dimension* as $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$.
- Y_h contains *constant* functions on B_{12}.
- Y_h is at most of degree q_{j_*} in each element of $\mathcal{T}_{h_{j_*}}(B_{12})$.

We shall separately describe the construction of Y_h for $\Omega \subset \mathbb{R}^2$ and $\Omega \subset \mathbb{R}^3$.

Consider a two subdomain nonmatching grid in *two dimensions*, in which the interface B_{12} is an *arc*. If $X_{h_{j_*}}(\Omega_l)$ consists of polynomials of degree q_{j_*} in each element, define the *continuous* multiplier space Y_h to consist of polynomials of degree q_{j_*} in each element of $\mathcal{T}_{h_{j_*}}(B_{12})$ except those touching the boundary ∂B_{12}, in which case the degree should be $(q_{j_*} - 1)$:

$$Y_h = \left\{ \psi_h \in C(B_{12}) : \psi_h|_e \in P_\alpha(e) \text{ where } \begin{pmatrix} \alpha = q_{j_*} & \text{if } \overline{e} \cap \partial B_{12} = \emptyset \\ \alpha = q_{j_*} - 1 & \text{if } \overline{e} \cap \partial B_{12} \neq \emptyset \end{pmatrix} \right\}, \tag{11.39}$$

where e denotes an element of $\mathcal{T}_{h_{j_*}}(B_{12})$.

For *piecewise linear* finite elements (i.e., $q_{j_*} = 1$), functions in Y_h will be constant on elements adjacent to the boundary. Let $x_0, x_1, \ldots, x_m, x_{m+1}$ denote the nodes of $\mathcal{T}_{h_{j_*}}(B_{12})$, arranged so that consecutive nodes define elements of $\mathcal{T}_{h_{j_*}}(B_{12})$, with x_0, x_{m+1} corresponding to the endpoints of B_{12}. Let $S_{h_{j_*}}(B_{12}) \subset H^1(B_{12})$ denote the standard continuous piecewise linear finite element space on triangulation $\mathcal{T}_{h_{j_*}}(B_{12})$, with standard nodal basis functions $\phi_0, \phi_1, \ldots, \phi_m, \phi_{m+1}$ for $S_{h_{j_*}}(B_{12})$. Then, the multiplier space Y_h will be defined as *span* of the following basis functions $\{\psi_1, \ldots, \psi_m\}$, see Fig. 11.4:

$$\begin{aligned} Y_h &= \text{span}\{\psi_1, \psi_2, \ldots, \psi_{m-1}, \psi_m\} \\ &= \text{span}\{\phi_0 + \phi_1, \phi_2, \ldots, \phi_{m-1}, \phi_m + \phi_{m+1}\} \subset S_{h_{j_*}}(B_{12}), \end{aligned}$$

where $\psi_1 = \phi_0 + \phi_1$, $\psi_j = \phi_j$ for $j = 2, \ldots, (m-1)$ and $\psi_m = \phi_m + \phi_{m+1}$.

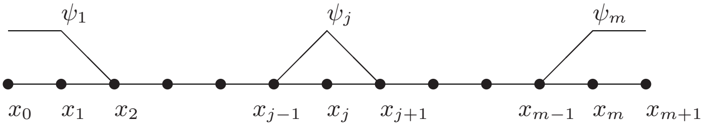


Fig. 11.4. Sample *continuous* basis functions for Y_h on an interface B_{12}

Remark 11.9. If one of the subdomains is *floating*, then B_{12} will be a *loop*. In this case, we can define $Y_h = X_{h_{j_*}}(B_{12}) = S_{h_{j_*}}(B_{12})$ as the space of continuous piecewise linear finite elements on B_{12}.

In *three dimensions*, we shall only consider the case where $q_{j_*} = 1$ and the two subdomain interface B_{12} is homeomorphic to a rectangle (since when B_{12} is homeomorphic to a sphere, we may define $Y_h = X_{h_{j_*}}(B_{12})$). When $\Omega \subset \mathbb{R}^3$, interface B_{12} will have a one dimensional boundary ∂B_{12}. To define a *continuous* multiplier space Y_h on B_{12}, we shall assume condition *(A.1)* when zero Dirichlet boundary conditions are imposed on ∂B_{12}, and condition *(A.2)* more generally. Unlike when B_{12} is one dimensional, it will not be possible to define a continuous piecewise linear multiplier space Y_h satisfying (11.39) with dimension equal to $\dim\left(X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})\right)$. To illustrate this, note that if $x_l \in \partial B_{12}$ is a boundary node adjacent to nodes x_i, x_j in the interior of B_{12}, then any continuous function that is linear in all the interior elements of $\mathcal{T}_{h_{j_*}}(B_{12})$ and constant on elements intersecting ∂B_{12} must have the same nodal values at x_i and x_j. This cannot hold if all interior nodes represent true degrees of freedom. To avoid such difficulties, the *continuous* piecewise linear multiplier space Y_h will permit linear functions on elements adjacent to ∂B_{12}, see [BE4, BE6, BR2], to ensure that constants are included and that $\dim(Y_h)$ equals the number m of interior nodes of triangulation $\mathcal{T}_{h_{j_*}}(B_{12})$.

The *continuous* piecewise linear multiplier space $Y_h(B_{12}) \subset L^2(B_{12})$ on the two dimensional interface B_{12} will be a *subspace* of $S_{h_{j_*}}(B_{12})$ of continuous piecewise linear functions on the triangulation $\mathcal{T}_{h_{j_*}}(B_{12})$. Let $x_1, \ldots, x_{m_*}$ denote the nodes of $\mathcal{T}_{h_{j_*}}(B_{12})$ with associated nodal basis $\phi_1, \ldots, \phi_{m_*}$:

$$S_{h_{j_*}}(B_{12}) = \text{span}\{\phi_1, \ldots, \phi_{m_*}\}.$$

Two nodes x_i and x_j on B_{12} will be said to be *adjacent* if they belong to a common triangular element $e \in \mathcal{T}_{h_{j_*}}(B_{12})$. We define:

$$\begin{aligned}
\mathcal{N}_I(B_{12}) \; &= \text{indices of nodes in the } \textit{interior} \text{ of } B_{12} \\
\mathcal{N}_{I_0}(B_{12}) &= \text{indices of } \textit{interior} \text{ nodes not adjacent to } \partial B_{12} \subset \mathcal{N}_I(B_{12}) \\
\mathcal{N}_B(B_{12}) &= \text{indices of nodes on the } \textit{boundary } \partial B_{12}.
\end{aligned}$$

We let m denote the number of indices in $\mathcal{N}_I(B_{12})$, i.e., the number of *interior* nodes, and $(m_* - m)$ the number of indices in $\mathcal{N}_B(B_{12})$.

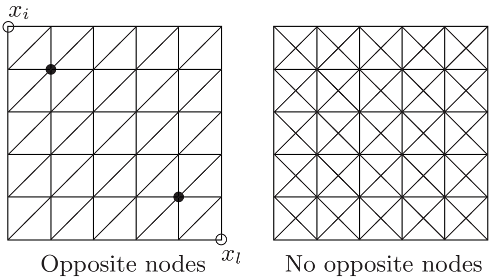
The *continuous* multiplier space $Y_h(B_{12}) \subset L^2(B_{12})$ will be defined as the subspace generated by m basis functions $\psi_1, \ldots, \psi_m$ from $S_{h_{j_*}}(B_{12})$:

$$Y_h = Y_h(B_{12}) = \text{span}\{\psi_1, \ldots, \psi_m\} \subset S_{h_{j_*}}(B_{12}).$$

We shall associate one basis function ψ_j with each *interior node* x_j and define

$$\mathcal{B}(x_j) = \{l \in \mathcal{N}_B(B_{12}) \,:\, x_l \text{ and } x_j \text{ are adjacent}\},$$

as the set of indices of nodes on ∂B_{12} adjacent to x_j. For each boundary node $x_i \in \partial B_{12}$, let $d\,(x_i)$ denote the number of *interior nodes* adjacent to x_i:



Opposite nodes No opposite nodes

Fig. 11.5. Triangulations with and without "opposite" nodes

$$d(x_i) = |\{l \in \mathcal{N}_I(B_{12}) \, : \, x_i \text{ and } \, x_l \, \text{are adjacent}\}| \, .$$

Some of the nodes x_i on ∂B_{12} may not have interior nodes adjacent to it, i.e., $d\,(x_i) = 0$. This can happen if the only triangular element containing x_i has all its three vertices lying on ∂B_{12}. Such nodes will be referred to as "opposite" nodes [BE4, BE6, BR2], see nodes x_i and x_l in Fig. 11.5. If x_i is an "opposite" node, then the element e to which it belongs will have all its nodes on ∂B_{12}, so that if constants must be included in $Y_h(B_{12})$ on B_{12}, then at least one of the basis functions ψ_j must have its support intersecting e. Accordingly, if x_i is an opposite node and x_j is an interior node *closest* to it, let $\mathcal{O}(x_j) = \{i\}$ denote the opposite node associated with x_j.

The basis functions generating the *continuous* multiplier space $Y_h(B_{12})$ is:

$$\psi_j(x) = \begin{cases} \phi_j(x), & \text{if } j \in \mathcal{N}_{I_0}(B_{12}) \\ \phi_j(x) + \sum_{l \in \mathcal{B}(j)} \frac{1}{d(x_l)} \, \phi_l(x), & \text{if } \mathcal{O}(x_j) = \emptyset \\ \phi_j(x) + \sum_{l \in \mathcal{B}(j)} \frac{1}{d(x_l)} \, \phi_l(x) + \phi_{\mathcal{O}(j)}(x), & \text{if } \mathcal{O}(x_j) \neq \emptyset, \end{cases}$$

for $1 \leq j \leq m$. By construction, ψ_j will be continuous and piecewise linear with $\psi_j = \phi_j$ when x_j is an interior node not adjacent to nodes on ∂B_{12}. Since $\psi_i(x_j) = \delta_{ij}$ for each interior node x_j, they will be linearly independent. Terms of the form $\sum_{l \in \mathcal{B}(j)} \frac{1}{d(x_l)} \, \phi_l(x)$ and $\phi_{\mathcal{O}(j)}(x)$ are included so that constants belong to Y_h, with $\psi_1 + \cdots + \psi_m = 1$ on B_{12}. The term $\phi_{\mathcal{O}(j)}(x)$ will be nonzero only if there is an "opposite node" associated with x_j.

Matrix Representation. The constraint $\mathcal{M}_*(u_h, \phi_h) = 0$ for all $\phi_h \in Y_h$ can be expressed in matrix form as:

$$\mathcal{M}_*(u_h, \phi_h) = \int_{B_{12}} (u_{h_1}(x) - u_{h_2}(x)) \; \phi(x) \, ds_x = 0 \Leftrightarrow \boldsymbol{\phi}^T M \mathbf{u}_h = 0, \quad (11.40)$$

where $\mathbf{u}_h \in \mathbb{R}^n$ and $\boldsymbol{\phi} \in \mathbb{R}^m$ denote the nodal vectors associated with the finite element functions $u_h(x) = (u_{h_1}(x), u_{h_2}(x)) \in X_h$ and $\phi_h(x) \in Y_h$, expanded relative to a basis $\{u_1, \ldots, u_n\}$ for X_h and $\{\psi_1, \ldots, \psi_m\}$ for Y_h. Matrix M will be of size $m \times n$ with entries $M_{ij} = \mathcal{M}_*(u_j, \psi_i)$.

We may block partition $\mathbf{u}_h$ and matrix M as follows. Order the basis for $X_h = X_{h_1}(\Omega_1) \times X_{h_2}(\Omega_2)$ so that the first n_1 basis functions form a basis for $X_{h_1}(\Omega_1)$ while the remaining n_2 form a basis for $X_{h_2}(\Omega_2)$. Further, order the basis for each $X_{h_l}(\Omega_l)$ with the $n_I^{(l)}$ basis functions for the *interior* nodes ordered before the $n_B^{(l)}$ basis functions for nodes on B_{12}. Let:

$$\mathbf{u}_h = \left(\mathbf{u}^{(1)^T}, \mathbf{u}^{(2)^T}\right)^T = \left(\mathbf{u}_I^{(1)^T}, \mathbf{u}_B^{(1)^T}, \mathbf{u}_I^{(2)^T}, \mathbf{u}_B^{(2)^T}\right)^T,$$

denote the resulting block partitioned vector associated with $u_{h_l}(x) \in X_{h_l}(\Omega_l)$ for $l = 1, 2$. Then constraint (11.40) can be equivalently expressed as:

$$M\,\mathbf{u}_h = \mathbf{0} \Leftrightarrow M_B^{(1)}\,\mathbf{u}_B^{(1)} - M_B^{(2)}\,\mathbf{u}_B^{(2)} = \mathbf{0} \tag{11.41}$$

where $M = \left[M_I^{(1)} \quad M_B^{(1)} \quad -M_I^{(2)} \quad -M_B^{(2)} \right]$ with $M_I^{(l)} = 0$ of size $m \times n_I^{(l)}$ and *sparse* matrix $M_B^{(l)}$ of size $m \times n_B^{(l)}$ whose entries are given by:

$$\left(M_B^{(l)}\right)_{ij} = \int_{B_{12}} \psi_i(x)\, u_j^{(l)}(x)\, ds_x, \quad \text{for } l = 1, 2. \tag{11.42}$$

Here $\{u_j^{(l)}(x)\}$ denotes a nodal basis for $X_{h_l}(B_{12}) \cap H_0^1(B_{12})$ on $\mathcal{T}_{h_l}(B_{12})$. If the unknowns on ∂B_{12} are zero, then $m = n_B^{(l_*)}$ and $M_B^{(j_*)}$ will be *square.*

Remark 11.10. Entries of $M_B^{(l)}$ in (11.42) can be computed by subassembly:

$$\left(M_B^{(l)}\right)_{ij} = \sum_{\sigma \in \mathcal{T}_{h_{j_*}}(B_{12})} \left(\int_\sigma \psi_i(x)\, u_j^{(l)}(x)\, ds_x \right), \tag{11.43}$$

based on the elements $\sigma \in \mathcal{T}_{h_{j_*}}(B_{12})$. When $l = j_*$, both $\psi_i(x)$ and $u_j^{(l)}(x)$ will be *polynomials* on each element $\sigma \in \mathcal{T}_{h_{j_*}}(B_{12})$. In this case matrix $M_B^{(l)}$ can be computed *exactly* using an exact quadrature rule for each elemental integral in (11.43). When $l \neq j_*$, the functions $\psi_i(x)$ and $u_j^{(l)}(x)$ will be defined on different triangulations of B_{12}, and the preceding will not apply. However, each integral on σ in (11.43) can be evaluated *approximately*, using a quadrature rule for piecewise smooth functions with accuracy $O(h_l^{q_l})$, where q_l denotes the degree of $\phi_j(x)$, see [CA38, MA5]. If $m = n_B^{(j_*)}$, then $M_B^{(j_*)}$ will be *square*, and solving constraint (11.41) will express $\mathbf{u}_B^{(j_*)}$ as *slave* variables:

$$\begin{cases} \mathbf{u}_B^{(1)} = \left(M_B^{(1)}\right)^{-1} M_B^{(2)}\,\mathbf{u}_B^{(2)}, & \text{if } j_* = 1 \\ \mathbf{u}_B^{(2)} = \left(M_B^{(2)}\right)^{-1} M_B^{(1)}\,\mathbf{u}_B^{(1)}, & \text{if } j_* = 2. \end{cases}$$

This expression is computationally expensive to evaluate when matrix $\left(M_B^{(j_*)}\right)^{-1}$ is *dense*, and motivates the construction of a *discontinuous* multiplier space $Y_h(B_{12}) \subset L^2(B_{12})$ which yields a *diagonal* matrix $M_B^{(j_*)}$.

Discontinuous Multiplier Space $\boldsymbol{Y_h}$. The motivation for employing a multiplier space $Y_h \subset L^2(B_{12})$ of *discontinuous* functions is that a basis $\{\psi_i\}_{i=1}^m$ can be constructed for Y_h at low computational cost, such that the mass matrix $M_B^{(j_*)}$ in (11.41) reduces to a *diagonal* matrix. This will yield an efficient *master-slave* expression for the nodal unknowns in the *interior* of the *nonmortar* side, in terms of the nodal unknowns on the *mortar* side. Discontinuous multiplier spaces were originally proposed in [WO4, WO5], and extended in [KI], and are based on a finite element technique for *interpolating* nonsmooth boundary data [SC6]. This applies even with unknowns on ∂B_{12}.

When condition *(A.1)* holds, discretizing the weak continuity condition on interface B_{12} using a multiplier space $Y_h = Y_h(B_{12})$ yields (11.40):

$$M_B^{(1)}\mathbf{u}_B^{(1)} - M_B^{(2)}\mathbf{u}_B^{(2)} = \mathbf{0},$$

where the local mass matrices are defined by:

$$\left(M_B^{(l)}\right)_{ij} = \int_{B_{12}} \psi_i(x)\, u_j^{(l)}(x)\, ds_x, \qquad \text{for } \ l = 1, 2, \tag{11.44}$$

for $\{\psi_i\}$ and $\{u_j^{(l)}\}$ denoting basis functions for Y_h and $X_{h_l}(B_{12}) \cap H_0^1(B_{12})$ respectively. Generally, each matrix $M_B^{(l)}$ will be rectangular, however, when *zero* Dirichlet boundary conditions are imposed on ∂B_{12} and $m = dim(Y_h)$ equals the dimension of $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$ (which equals the number of interior nodes in $\mathcal{T}_{h_{j_*}}(B_{12})$), then mass matrix $M_B^{(j_*)}$ will be square. To obtain a *diagonal* matrix $M_B^{(j_*)}$, the basis $\{\psi_i\}_{i=1}^m$ for $Y_h(B_{12}) \subset L^2(B_{12})$ must be *biorthogonal* to the nodal basis $\{u_j^{(j_*)}\}_{j=1}^m$ for $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$:

$$\int_{B_{12}} \psi_i(x)\, u_j^{(j_*)}(x)\, ds_x = \left(M_B^{(j_*)}\right)_{ij} = \gamma_i\, \delta_{ij}, \quad \text{for } \ 1 \le i,\, j \le m, \tag{11.45}$$

for δ_{ij} denoting the Kronecker delta and $\gamma_i > 0$ a scaling factor typically chosen so that $\int_{B_{12}} \psi_i(x)\, ds_x = 1$.

Remark 11.11. For the Dirichlet problem (11.1) (with no nodal unknowns on ∂B_{12}), a biorthogonal (dual) basis $\{\psi_j\}_{j=1}^m$ consisting of *continuous* finite element functions on $\mathcal{T}_{h_{j_*}}(B_{12})$ satisfying (11.45) can be constructed at high computational cost as follows. If $M_B^{(j_*)}$ is the mass matrix corresponding to the standard continuous multiplier space $Y_h = \text{span}\{\hat{\psi}_1, \ldots, \hat{\psi}_m\}$, then define a biorthogonal basis $\{\psi_j\}_{j=1}^m$ as follows:

$$\psi_j(x) = \sum_{l=1}^m \left(M_B^{(j_*)^{-T}}\right)_{jl} \hat{\psi}_l(x), \quad \text{for } \ j = 1, \ldots, m. \tag{11.46}$$

By construction, biorthogonality will hold. However, since m can be large, it will be *computationally expensive* to assemble the *dense* matrix $M_B^{(j_*)^{-T}}$.

For convenience, let $\{\phi_i\}_{i=1}^m$ denote the finite element nodal basis for $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$, where $\phi_i(x) = u_i^{(j_*)}(x)$ for $1 \le i \le m$. To construct a *discontinuous* basis $\{\psi_i\}_{i=1}^m \subset L^2(B_{12})$ *biorthogonal* to $\{\phi_j\}_{j=1}^m$, note that:

- Each integral on B_{12} can be decomposed using the subassembly identity:
$$\begin{aligned} \int_{B_{12}} \psi_i(x)\, \phi_j(x)\, ds_x &= \sum_{\sigma \in \mathcal{T}_{h_{j_*}}(B_{12})} \int_\sigma \psi_i(x)\, \phi_j(x)\, ds_x \\ &= \sum_{\sigma \in \mathcal{T}_{h_{j_*}}(B_{12})} \int_\sigma \psi_{i;\sigma}(x)\, \phi_{j;\sigma}(x)\, ds_x, \end{aligned} \tag{11.47}$$
where $\psi_{i;\sigma}(x)$, $\phi_{j;\sigma}(x)$ denote restrictions of $\psi_i(x)$, $\phi_j(x)$ to element σ.
- Since each $\phi_j(x)$ is a standard nodal basis for $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$, its restriction $\phi_{j;\sigma}(x)$ will correspond to one of the *local* nodal basis for $P_{q_{j_*}}(\sigma)$ on σ. Define an *element mass* matrix $M_\sigma^{(h_{j_*})}$ of size d_*:
$$\left(M_\sigma^{(h_{j_*})}\right)_{ij} = \int_\sigma \phi_i^{(\sigma)}(x)\, \phi_j^{(\sigma)}(x)\, ds_x, \qquad 1 \le i, \;\; j \le d_*,$$
where $\{\phi_l^{(\sigma)}(x)\}_{l=1}^{d_*}$ denotes the standard local nodal basis for $P_{q_{j_*}}(\sigma)$.
- Since d_* will be small, we may *explicitly* compute its inverse $M_\sigma^{(h_{j_*})^{-1}}$ and use it construct another basis $\{\psi_l^{(\sigma)}(x)\}_{l=1}^{d_*}$ for $P_{q_{j_*}}(\sigma)$:
$$\psi_j^{(\sigma)}(x) = \sum_{l=1}^{d_*} \left(M_\sigma^{(j_*)^{-1}}\right)_{jl} \phi_l^{(\sigma)}(x), \qquad 1 \le j \le d_*, \tag{11.48}$$
where $\{\psi_l^{(\sigma)}(x)\}_{l=1}^{d_*}$ are *biorthogonal* to $\{\phi_l^{(\sigma)}(x)\}_{l=1}^{d_*}$, satisfying:
$$\int_\sigma \phi_i^{(\sigma)}(x)\, \psi_j^{(\sigma)}(x)\, ds_x = \delta_{ij}, \qquad \text{for } 1 \le i, \;\; j \le d_*. \tag{11.49}$$
- If $x_1, \ldots, x_m$ denotes an ordering of all *interior* nodes in $\mathcal{T}_{h_{j_*}}(B_{12})$, for each $x_j \in \sigma$ define $1 \le \text{index}(x_j, \sigma) \le d_*$ as the index of node x_j in the *local ordering* of nodes on σ. Then, by construction, it will hold that:
$$\phi_{j;\,\sigma}(x) = \phi_{\tilde{j}}^{(\sigma)}(x), \quad \text{on } \sigma, \qquad \text{when } \tilde{j} = \text{index}(x_j, \sigma). \tag{11.50}$$
Substituting identity (11.50) into expression (11.47), it is easily seen that for each $x_i \in \sigma$ it is *sufficient* to define $\psi_{i;\,\sigma}(x)$ as:
$$\psi_i(x) = \gamma_{i,\sigma}\, \psi_{\tilde{i}}^{(\sigma)}(x) \quad \text{on } \sigma, \quad \text{for } \tilde{i} = \text{index}(x_i, \sigma), \tag{11.51}$$
where $\psi_{\tilde{i}}^{(\sigma)}(x)$ is given by (11.48) and $\gamma_{i,\sigma} = \int_\sigma \phi_{\tilde{i}}^{(\sigma)}(x)\, ds_x > 0$. If σ is adjacent to ∂B_{12}, however, multiple choices of $\psi_{i;\,\sigma}(x)$ will be available.

The above basis $\{\psi_i\}_{i=1}^m$ will satisfy (11.45) with $\gamma_i = \sum_\sigma \gamma_{i,\sigma}$ with each $\psi_i(x)$ having *support* in the elements containing x_i. To ensure that *constant* functions are in $Y_h = \text{span}\{\psi_1(x), \ldots, \psi_m(x)\}$ a careful choice of $\psi_{i;\sigma}(x)$ on elements σ adjacent to ∂B_{12} must be employed, without violating global biorthogonality.

We now apply the general guidelines (11.51) to construct *piecewise linear* dual basis functions, see [WO4, WO5, KI]. We first consider a uniform grid on a *one dimensional* interface B_{12} with nodes $x_0, x_1, \ldots, x_m, x_{m+1}$ and endpoints x_0, x_{m+1}. Let $\{\phi_l(x)\}_{l=1}^m$ denote the standard piecewise linear nodal basis for $X_{h_{j*}}(B_{12}) \cap H_0^1(B_{12})$ satisfying $\phi_i(x_j) = \delta_{ij}$. For $1 \leq i \leq m$, node x_i will belong to elements $\sigma_{i-1} = [x_{i-1}, x_i]$ and $\sigma_i = [x_i, x_{i+1}]$, while x_0 and x_{m+1} will only belong to $\sigma_0 = [x_0, x_1]$ and $\sigma_m = [x_m, x_{m+1}]$, respectively. On each element σ, the space $P_1(\sigma)$ will be of dimension $d_* = 2$, and spanned by the two local nodal basis $\{\phi_1^{(\sigma)}(x), \phi_2^{(\sigma)}(x)\}$. The *element mass* matrix and its inverse on an element σ will have the form:

$$M_\sigma^{(h_{j*})} = \frac{h_\sigma}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad M_\sigma^{(h_{j*})^{-1}} = \frac{2}{h_\sigma}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix},$$

where h_σ denotes the length of σ. For $\gamma_{i,\sigma} = \int_\sigma \phi_i^{(\sigma)}(x)\,dx = (h_\sigma/2)$, a local basis in $P_1(\sigma)$ biorthogonal to $\{\phi_1^{(\sigma)}(x), \phi_2^{(\sigma)}(x)\}$ will be:

$$\begin{cases} \psi_1^{(\sigma)}(x) = 2\,\phi_1^{(\sigma)}(x) - \phi_2^{(\sigma)}(x) \\ \psi_2^{(\sigma)}(x) = 2\,\phi_2^{(\sigma)}(x) - \phi_1^{(\sigma)}(x). \end{cases}$$

The global dual basis functions $\{\psi_i(x)\}_{i=1}^m$ will be constructed in terms of the elemental dual basis functions $\{\psi_j^{(\sigma)}(x)\}$ using (11.51). Each $\psi_i(x)$ will have *support* among σ with $x_i \in \sigma$. On the elements σ_0 and σ_m adjacent to $\partial B_{12} = \{x_0, x_{m+1}\}$, any linear combination of $\psi_1^{(\sigma)}(x)$ and $\psi_2^{(\sigma)}(x)$ can be used without violating global biorthogonality. To include *constants* in Y_h, we must require $\psi_1(x)$ and $\psi_m(x)$ to be constant on σ_0 and σ_m, respectively, since all other basis functions will be zero on these elements. Since $\psi_1^{(\sigma)}(x) + \psi_2^{(\sigma)}(x) = 1$ on σ, this constant can be chosen to be 1, yielding $\psi_1 + \cdots + \psi_m(x) = 1$ on B_{12}. The resulting basis dual to $\{\phi_j(x)\}_{j=1}^m$ will be:

$$\psi_l(x) = \begin{cases} 1 & \text{on } \sigma_0 & \text{for } l = 1 \\ -\phi_{l-1}(x) + 2\phi_l(x) & \text{on } \sigma_{l-1} & \text{for } 2 \leq l \leq m \\ 2\phi_l(x) - \phi_{l+1}(x), & \text{on } \sigma_l & \text{for } 1 \leq l \leq m-1 \\ 1 & \text{on } \sigma_m & \text{for } l = m \\ 0, & \text{elsewhere.} \end{cases} \qquad (11.52)$$

The above discontinuous dual basis functions are plotted in Fig. 11.6.

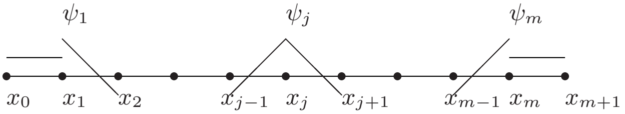


Fig. 11.6. Sample basis functions $\{\psi_j\} \subset Y_h$ biorthgonal to $\{\phi_i\}$ on B_{12}

When interface B_{12} is *two dimensional*, constructing a uniform grid dual basis for $\{\phi_l(x)\}_{l=1}^m$ using (11.51) will be complicated near the boundary [KI]. As before, let $\{x_l\}_{l=1}^m$ denote the interior nodes of B_{12} and $\{\phi_l(x)\}_{l=1}^m$ the standard piecewise linear nodal basis for $X_{h_{j*}}(B_{12}) \cap H_0^1(B_{12})$ satisfying $\phi_i(x_j) = \delta_{ij}$. On each element $\sigma \in \mathcal{T}_{h_{j*}}(B_{12})$ let $\{\phi_l^{(\sigma)}\}_{l=1}^3$ denote local nodal basis for $P_1(\sigma)$. The *element mass* matrix and its inverse will be:

$$M_\sigma^{(h_{j*})} = \frac{|\sigma|}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad M_\sigma^{(h_{j*})^{-1}} = \frac{3}{|\sigma|}\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix},$$

where $|\sigma|$ denotes the area of triangular element σ. A *local* basis $\{\psi_l^{(\sigma)}(x)\}_{l=1}^3$ dual to $\{\phi_l^{(\sigma)}(x)\}_{l=1}^3$ satisfying $\gamma_{j,\sigma} = \int_\sigma \phi_j^{(\sigma)}(x) dx_x$ will thus be:

$$\psi_j^{(\sigma)}(x) = 3\phi_j^{(\sigma)}(x) - \sum_{l \neq j} \phi_l^{(\sigma)}(x) \quad \text{on} \quad \sigma, \quad \text{for} \quad 1 \leq j \leq 3. \tag{11.53}$$

Each global dual basis function $\psi_j(x)$ will have support in the elements σ containing x_j. When defining $\psi_j(x)$ on elements σ adjacent to ∂B_{12}, containing more than one interior node on it, several choices of local dual basis will be possible which yields global biorthogonality. However, requiring that *constants* be included in Y_h can restrict the available choices, as noted below.

To simplify our discussion, we shall use the notation $\{y_{l,\sigma}\}_{l=1}^3$ to denote the three vertices of a triangular element $\sigma \in \mathcal{T}_{h_{j*}}(B_{12})$. The local nodal basis will be denoted $\{\phi_l^{(\sigma)}(x)\}_{l=1}^3$ satisfying $\phi_i^{(\sigma)}(y_{j,\sigma}) = \delta_{ij}$ for $1 \leq i, j \leq 3$, and its local dual basis will be denoted $\{\psi_l^{(\sigma)}(x)\}_{l=1}^3$ as in (11.53). Below, we describe $\psi_i(x)$ on each element σ containing x_i (except in cases D and E). The construction depends on the number of interior nodes in σ, see Fig. 11.7:

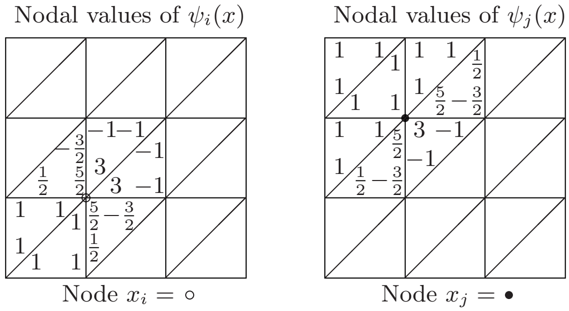
Case A. All *three* nodes $y_{1,\sigma}$, $y_{2,\sigma}$, $y_{3,\sigma}$ of σ are *interior* nodes in B_{12}. In this case, if $x_i = y_{1,\sigma}$, define:

$$\psi_i(x)|_\sigma = \psi_1^{(\sigma)}(x) = \left(3\phi_1^{(\sigma)}(x) - \phi_2^{(\sigma)}(x) - \phi_3^{(\sigma)}(x)\right), \quad \text{on } \sigma.$$

Case B. Only *two* nodes, $x_i = y_{1,\sigma}$ and $x_j = y_{2,\sigma}$ are *interior* nodes in B_{12}. In this case, $\psi_1^{(\sigma)}(x) + c_3\, \psi_3^{(\sigma)}(x)$ and $\psi_2^{(\sigma)}(x) + d_3\, \psi_3^{(\sigma)}(x)$ will both be locally biorthogonal to $\phi_1^{(\sigma)}(x)$ and $\phi_2^{(\sigma)}(x)$ for arbitrary c_3, $d_3 \in \mathbb{R}$. Choosing $c_3 = d_3 = \frac{1}{2}$ will ensure that $\gamma_{l,\sigma} = \int_\sigma \phi_{l,\sigma}(x)\, ds_x$ for $l = i, j$:

$$\psi_i(x)|_\sigma = \left(\psi_1^{(\sigma)}(x) + \frac{1}{2}\psi_3^{(\sigma)}(x)\right) = \tfrac{1}{2}\left(5\phi_1^{(\sigma)}(x) - 3\phi_2^{(\sigma)}(x) + \phi_3^{(\sigma)}(x)\right)$$

$$\psi_j(x)|_\sigma = \left(\psi_2^{(\sigma)}(x) + \frac{1}{2}\psi_3^{(\sigma)}(x)\right) = \tfrac{1}{2}\left(5\phi_2^{(\sigma)}(x) - 3\phi_1^{(\sigma)}(x) + \phi_3^{(\sigma)}(x)\right).$$


Nodal values of $\psi_i(x)$

Nodal values of $\psi_j(x)$

Node $x_i = \circ$

Node $x_j = \bullet$

Fig. 11.7. Nonzero nodal values of discontinuous dual basis functions

Case C. Only *one* node $y_{l,\sigma}$ is an *interior* node. Suppose that $x_i = y_{l,\sigma}$ is the interior node, then any linear combination of the local dual basis functions $\{\psi_j^{(\sigma)}(x)\}_{j=1}^3$ can be employed to define $\psi_i(x)$ without violating global biorthogonality. However, since *constants* must be included in Y_h, this restricts the choice to $c_1 = c_2 = c_3 = 1$, yielding:

$$\psi_i(x)\big|_\sigma = 1 \quad \text{on} \quad \sigma,$$

since $\psi_1^{(\sigma)}(x) + \psi_2^{(\sigma)}(x) + \psi_3^{(\sigma)}(x) = 1$ on σ.

Case D. None of the nodes $y_{1,\sigma}$, $y_{2,\sigma}$ and $y_{3,\sigma}$ are interior nodes of B_{12}. In this case, all three nodes must lie on ∂B_{12}. To ensure that *constants* are included in Y_h, let x_i denote the interior node closest to triangle σ and define:

$$\psi_i(x)\big|_\sigma = 1 \quad \text{on} \quad \sigma.$$

Case E. If $x_i \notin \sigma$ and it is not *case D*, then we define $\psi_i(x)\big|_\sigma = 0$ on σ.

Remark 11.12. Below, see [KI], we summarize the definition of each dual basis function $\psi_i(x)$ on an element σ containing *interior* node x_i (except in cases D and E), where we denote the vertices of σ as $x_i = y_{1,\sigma}$, $y_{2,\sigma}$ and $y_{3,\sigma}$:

$$\psi_i(x) \equiv \begin{cases} 3\phi_1^{(\sigma)}(x) - \phi_2^{(\sigma)}(x) - \phi_3^{(\sigma)}(x), & \text{Case A with } x_i = y_{1,\sigma} \\ \frac{1}{2}\left(5\phi_1^{(\sigma)}(x) - 3\phi_2^{(\sigma)}(x) + \phi_3^{(\sigma)}(x)\right), & \text{Case B with } x_i = y_{1,\sigma} \\ 1, & \text{Case C with } x_i = y_{1,\sigma} \\ 1, & \text{Case D with } x_i \notin \sigma \\ 0, & \text{Case E with } x_i \notin \sigma. \end{cases} \tag{11.54}$$

In *case D*, we assume that x_i is the *interior* node closest to triangle σ. For the above choice of dual basis functions, the following will hold:

$$\psi_1(x) + \cdots + \psi_m(x) = 1,$$

ensuring that *constants* are included in $Y_h = \text{span}\{\psi_1, \ldots, \psi_m\}$.

A mortar discretization of (11.1). using continuous or discontinuous spaces Y_h will yield a *stable* discretization [BE18, BE6, BE4, WO4, WO5, KI].

Remark 11.13. When *zero* Dirichlet conditions are imposed on ∂B_{12}, and if $m = dim(Y_h) = dim(X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12}))$, then matrix $M_B^{(j_*)}$ will be $m \times m$ for continuous and discontinuous spaces Y_h. Solving (11.41) will yield:

$$\begin{cases} \mathbf{u}_B^{(1)} = \left(M_B^{(1)}\right)^{-1} M_B^{(2)}\, \mathbf{u}_B^{(2)}, & \text{if } j_* = 1 \\ \mathbf{u}_B^{(2)} = \left(M_B^{(2)}\right)^{-1} M_B^{(1)}\, \mathbf{u}_B^{(1)}, & \text{if } j_* = 2, \end{cases}$$

The above mass matrices can be computed using subassembly and quadratures. For the discontinuous space Y_h, the *diagonal* matrix $M_B^{(j_*)}$ will satisfy:

$$\left(M_B^{(j_*)}\right)_{ii} = \int_{B_{12}} \psi_i(x)\, \phi_i(x)\, ds_x = \int_{B_{12}} \phi_i(x)\, ds_x, \qquad \text{for} \quad 1 \le i \le m.$$

When both grids *match* on B_{12} and *zero* boundary conditions are imposed on ∂B_{12}, then it will hold that $M_B^{(1)} = M_B^{(2)}$ and $\mathbf{u}_B^{(1)} = \mathbf{u}_B^{(2)}$ and the global discretization will reduce to the conforming finite element discretization.

Remark 11.14. The preceding construction of a basis biorthogonal to a nodal basis for $X_{h_{j_*}}(B_{12}) \cap H_0^1(B_{12})$ applies even when the nodal values on ∂B_{12} are *nonzero*. This property will be employed in multisubdomain discretizations. For instance, if *Neumann boundary conditions* are imposed on any segment of $\partial \Omega$ containing ∂B_{12}, then the nodal unknowns on ∂B_{12} will not be zero (as in the Dirichlet case). If $\{x_l\}_{l=1}^m$ denotes the *interior* nodes of $\mathcal{T}_{h_{j_*}}(B_{12})$ and $\{\phi_l(x)\}_{l=1}^m$ the standard piecewise linear nodal basis satisfying $\phi_i(x_j) = \delta_{ij}$, then a basis $\{\psi_l(x)\}_{l=1}^m$ biorthogonal (dual) to $\{\phi_l(x)\}_{l=1}^m$ can be constructed using the local dual basis $\{\psi_i^{(\sigma)}(x)\}$. When B_{12} is *one dimensional*:

$$\psi_i(x) \equiv \begin{cases} 2\phi_1^{(\sigma)}(x) - \phi_2^{(\sigma)}(x), & \text{if } x_i = y_{1,\sigma} \subset \sigma \\ 0, & \text{if } x_i \notin \sigma, \end{cases} \tag{11.55}$$

where element σ has vertices $y_{1,\sigma}$ and $y_{2,\sigma}$. When B_{12} is two dimensional:

$$\psi_i(x) \equiv \begin{cases} 3\phi_1^{(\sigma)}(x) - \phi_2^{(\sigma)}(x) - \phi_3^{(\sigma)}(x), & \text{if } x_i = y_{1,\sigma} \subset \sigma \\ 0, & \text{if } x_i \notin \sigma, \end{cases} \tag{11.56}$$

where σ is a triangular element with vertices $y_{1,\sigma}$, $y_{2,\sigma}$ and $y_{3,\sigma}$. The mass matrix $M_B^{(j_*)}$ will be rectangular of size $m \times n_B^{(j_*)}$. However, the nodes can be ordered so that its leading $m \times m$ sub-matrix is *diagonal*:

$$\left(M_B^{(j_*)}\right)_{ii} = \int_{B_{12}} \psi_i(x)\, \phi_i(x)\, ds_x = \int_{B_{12}} \phi_i(x)\, ds_x, \qquad \text{for} \quad 1 \le i \le m.$$

If $\{x_l\}_{l=1}^{m_*}$ denotes all nodes of $\mathcal{T}_{h_{j_*}}(B_{12})$ and $\{\phi_l(x)\}_{l=1}^{m_*} \subset X_{h_{j_*}}(B_{12})$ is the standard nodal basis satisfying $\phi_i(x_j) = \delta_{ij}$, then $\{\psi_l(x)\}_{l=1}^{m_*}$ biorthogonal to $\{\phi_l(x)\}_{l=1}^{m_*}$ can also be constructed. Matrix $M_B^{(j_*)}$ will be diagonal of size m_*.

Remark 11.15. When $q_{j_*} = 1$, each $\psi_l(x)$ will be linear on element σ and satisfy $\psi_l(x) \in D_{h_{j_*}}(B_{12})$, where:

$$D_{h_{j_*}}(B_{12}) \equiv \left\{ v(x) \in L^2(B_{12}) : v(x)|_\sigma \in P_{q_{j_*}}(\sigma),\, \forall \sigma \in \mathcal{T}_{h_{j_*}}(B_{12}) \right\}.$$

Since each $v(x) \in D_{h_{j_*}}(B_{12})$ is a polynomial of degree q_{j_*} on each element σ, it will hold that $\dim(D_{h_{j_*}}(B_{12})) = n_e\, d_*$, where n_e denotes the number of elements in $\mathcal{T}_{h_{j_*}}(B_{12})$ and $d_* = \dim(P_{q_{j_*}}(\sigma))$.

Saddle Point System. Let $\mathbf{u}^{(l)} = \left(\mathbf{u}_I^{(l)^T}, \mathbf{u}_B^{(l)^T}\right)^T \in \mathbb{R}^{n_l}$ denote the block partitioned nodal vector corresponding to $u_{h_l}(x) \in X_{h_l}(\Omega)$ for $l = 1, 2$ and let $\boldsymbol{\psi} \in \mathbb{R}^m$ denote the nodal vector associated with $\psi_h(x) \in Y_h$. Then, the saddle point linear system obtained by discretization of (11.19) on a *two subdomain* nonmatching grid, will have the block structure:

$$\begin{bmatrix} A^{(1)} & 0 & M^{(1)^T} \\ 0 & A^{(2)} & -M^{(2)^T} \\ M^{(1)} & -M^{(2)} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(2)} \\ \mathbf{0} \end{bmatrix}, \qquad (11.57)$$

where for $l = 1, 2$:

$$A^{(l)} = \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{IB}^{(l)^T} & A_{BB}^{(l)} \end{bmatrix}, \quad M^{(l)} = \begin{bmatrix} 0 & M_B^{(l)} \end{bmatrix}, \quad \mathbf{u}^{(l)} = \begin{bmatrix} \mathbf{u}_I^{(l)} \\ \mathbf{u}_B^{(l)} \end{bmatrix}, \quad \mathbf{f}^{(l)} = \begin{bmatrix} \mathbf{f}_I^{(l)} \\ \mathbf{f}_B^{(l)} \end{bmatrix}, \qquad (11.58)$$

with the entries of $A^{(l)}$ and $\mathbf{f}^{(l)}$ satisfying:

$$\begin{cases} \mathcal{A}_{\Omega_l}(u_{h_l}, u_{h_l}) = \mathbf{u}^{(l)^T} A^{(l)} \mathbf{u}^{(l)^T} \\ \quad (f, u_{h_l})_{\Omega_l} = \mathbf{f}^{(l)^T} \mathbf{u}^{(l)}, \end{cases} \qquad (11.59)$$

and with matrix $M^{(l)}$ satisfying:

$$\begin{aligned} \mathcal{M}_*((u_{h_1}, u_{h_2}), \psi_h) &= \boldsymbol{\psi}^T \big(M^{(1)}\mathbf{u}^{(1)} - M^{(2)}\mathbf{u}^{(2)}\big) \\ &= \boldsymbol{\psi}^T \big(M_B^{(1)}\mathbf{u}_B^{(1)} - M_B^{(2)}\mathbf{u}_B^{(2)}\big). \end{aligned} \qquad (11.60)$$

Let $n_I^{(l)}$, $n_B^{(l)}$ and $n_l = n_I^{(l)} + n_B^{(l)}$ denote the sizes of $\mathbf{u}_I^{(l)}$, $\mathbf{u}_B^{(l)}$ and $\mathbf{u}^{(l)}$, respectively, with $m = n_B^{(j_*)}$. Then matrices $A_{II}^{(l)}$, $A_{IB}^{(l)}$, $A_{BB}^{(l)}$ and $M_B^{(h_l)}$ will be of size $n_I^{(l)}$, $n_I^{(l)} \times n_I^{(l)}$, $n_B^{(l)}$ and $m \times n_B^{(l)}$, respectively. The vectors $\mathbf{f}_I^{(l)}$, $\mathbf{f}_B^{(l)}$ and $\mathbf{f}^{(l)}$ will be of size $n_I^{(l)}$, $n_B^{(l)}$ and n_l, respectively.

11.2.3 Multi-Subdomain Discretizations

We shall now describe mortar discretizations of (11.1) on a multi-subdomain non-matching grid in $\Omega \subset \mathbb{R}^d$. The methodology will be more complicated since the segments $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$ can have dimension $(d-1)$, $(d-2)$ or lower. However, it will often be sufficient to consider discretizations which enforce intersubdomain matching on interfaces B_{lj} of dimension $(d-1)$. The discrete solution, however, may be *nonconforming* even if the grids match.

We decompose Ω into p nonoverlapping subdomains $\Omega_1, \ldots, \Omega_p$, and let $\mathcal{O}(l)$ denote an index set such that $B_{lj} = \partial\Omega_l \cap \partial\Omega_j \neq \emptyset$ for $j \in \mathcal{O}(l)$. We let $\mathcal{T}_{h_l}(\Omega_l)$ denote a quasiuniform triangulation of Ω_l with mesh size h_l and assume for simplicity that either assumption *(A.1)* or assumption *(A.2)* holds for each non-empty B_{lj}. Thus, the nonmatching grid on the left side of Fig. 11.8 will be considered (since assumption *(A.1)* holds), while the nonmatching grid on the right side of Fig. 11.8 will not be considered (since Ω_1 and Ω_3, and Ω_2 and Ω_3 violate *(A.1)*). We let $\mathcal{I}(l) \subset \mathcal{O}(l)$ be an index set, such that if $j \in \mathcal{I}(l)$, then we select $\mathcal{T}_{h_j}(B_{lj})$ as the triangulation of B_{lj}, obtained by restricting $\mathcal{T}_{h_j}(\Omega_j)$ to B_{lj}. The "side" of B_{lj} approached from Ω_j for $j \in \mathcal{I}_*(l)$ is referred to as the *nonmortar* side, and the nodal unknowns in the interior of B_{lj} on the nonmortar side Ω_j will be *slave* variables, while the nodal unknowns on B_{lj} from the mortar side Ω_l will be *master* variables.

For multiple subdomains, two *alternative* saddle point formulations of (11.1) may be employed to construct the mortar element discretizations. In the first version, constraints are matched *weakly* on each segment B_{lj} for $j \in \mathcal{I}(l)$ and $1 \leq l \leq p$, while in the second version, the discretization is simplified by matching only on segments B_{lj} of dimension $(d-1)$ when $\Omega \subset \mathbb{R}^d$, see [MA4, BE18, PH, BE23, BE6, BE4, WO, WO4, WO5, KI], without significantly altering its accuracy. Thus, in the second version, matching along *cross point* segments B_{lj} in two dimensions (such as B_{13} and B_{24} on the left side of Fig. 11.8) and *edges* or *cross points* B_{lj} in three dimensions, are omitted.

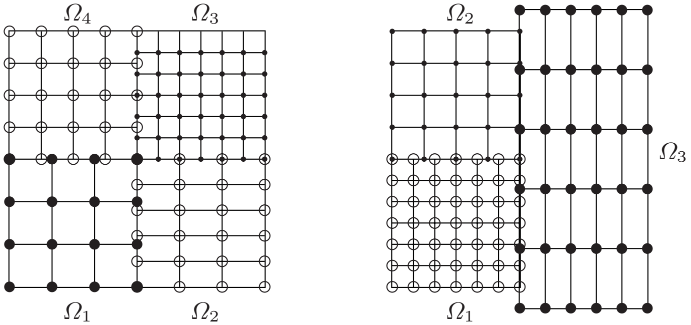


Fig. 11.8. Sample multi-subdomain non-matching grids

The function space $X = \Pi_{l=1}^{p}\left(H^1_{0,B_{[l]}}(\Omega_l)\right)$ will be the same in both versions of (11.19). The function space for the multipliers may be $\tilde{Y}$ or Y:

$$\tilde{Y} = \Pi_{l=1}^{p}\left(\Pi_{j\in\mathcal{I}(l)} H^{-1/2}(B_{lj})\right) \text{ and } Y = \Pi_{l=1}^{p}\left(\Pi_{j\in\mathcal{I}_*(l)} H^{-1/2}(B_{lj})\right). \tag{11.61}$$

Here $\tilde{Y}$ is only a *formal* expression, since $H^{-1/2}(B_{lj})$ will not be appropriate for B_{lj} of dimension $(d-2)$ or lower. However, Y is well defined. We define:

$$\begin{cases} \mathcal{M}_*(w,\psi) = \sum_{l=1}^{p}\sum_{j\in\mathcal{I}(l)}\int_{B_{lj}}[w]_{lj}\,\psi_{lj}\,ds_x, & \text{for } w\in X,\ \psi\in\tilde{Y} \\ \mathcal{M}_*(w,\psi) = \sum_{l=1}^{p}\sum_{j\in\mathcal{I}_*(l)}\int_{B_{lj}}[w]_{lj}\,\psi_{lj}\,ds_x, & \text{for } w\in X,\ \psi\in Y, \end{cases} \tag{11.62}$$

where $[w]_{lj} = w_l - w_j$ for $w = (w_1,\ldots,w_p)\in X$.

A mortar element discretization of (11.1) is now obtained by Galerkin approximation of (11.19) by enforcing weak matching on all nonempty B_{lj}, or by enforcing weak matching on all non-empty B_{lj} of dimension $(d-1)$. This will seek $u_h\in X_h$ and $\psi_h\in Y_h\subset\tilde{Y}$ or $\psi_h\in Y_h\subset Y$ satisfying:

$$\begin{cases} \mathcal{A}_*(u_h,v_h) + \mathcal{M}_*(v_h,\psi_h) = (f,v_h)_*, & \forall v_h\in X_h \\ \qquad\qquad\qquad \mathcal{M}_*(u_h,\phi_h) = 0, & \forall \phi_h\in Y_h, \end{cases} \tag{11.63}$$

for finite element spaces $X_h\subset X$ and $Y_h\subset\tilde{Y}$ or $Y_h\subset Y$ where:

$$\begin{cases} \mathcal{A}_*(w,w) = \sum_{l=1}^{p}\int_{\Omega_l}\left(a(x)\nabla v_l\cdot\nabla w_l + c(x)\,v_l\,w_l\right)dx \\ \quad (f,w)_* = \sum_{l=1}^{p}\int_{\Omega_l}\left(f(x)\,w_l\right)dx, \end{cases} \tag{11.64}$$

for $w=(w_1,\ldots,w_p)\in X$ and $\mathcal{M}_*(.,.)$ defined by (11.62), with $[w]_{lj} = w_l - w_j$ and $\psi\in Y_h$. Expanding $u_h\in X_h$ and $\psi_h\in Y_h$ relative to a basis for X_h and Y_h will yield a saddle point linear system, to be described later.

Choice of Subspace X_h. The choice of the subspace $X_h\subset X$ for the multisubdomain case will be analogous to that for the two subdomain case. On each subdomain Ω_l, let $X_{h_l}(\Omega_l)$ denote a *conforming* finite element space defined on triangulation $\mathcal{T}_{h_l}(\Omega_l)$ of degree q_l on each element satisfying:

$$X_{h_l}(\Omega_l)\subset H^1_{0,B_{[l]}}(\Omega_l).$$

Define $X_h = \Pi_{l=1}^{p} X_{h_l}(\Omega_l)$. Then, approximation property (11.30) will hold:

$$\inf_{v_{h_l}\in X_{h_l}(\Omega_l)} \|u - v_{h_l}\|_{1,\Omega_l} \le C\,h_l^{q_l}\,\|u\|_{q_l+1,\Omega_l}, \tag{11.65}$$

for sufficiently smooth u.

Choice of Subspace Y_h. The choice of the finite element space Y_h will depend on whether weak matching is enforced on all nonempty segments B_{lj} (in which case $Y_h\subset\tilde{Y}$) or whether weak matching is enforced only on nonempty segments B_{lj} of dimension $(d-1)$ (in which case $Y_h\subset Y$). The latter case is computationally simpler, and also yields optimal order accuracy.

If weak matching is enforced on all segments B_{lj} for $j \in \mathcal{I}(l)$, then:

$$Y_h = \Pi_{l=1}^{p} (\Pi_{j \in \mathcal{I}(l)} Y_{h_j}(B_{lj})).$$

Here, some of the segments B_{lj} may be of dimension $(d-2)$ or lower. When assumption *(A.1)* holds, on each segment B_{lj} for $j \in \mathcal{I}(l)$ and $1 \leq l \leq p$, we let $Y_{h_j}(B_{lj}) \subset L^2(B_{lj})$ denote a *continuous* or *discontinuous* multiplier space defined on the triangulation $\mathcal{T}_{h_j}(B_{lj})$ and satisfying:

- $Y_{h_j}(B_{lj})$ has dimension m_{lj} (number of interior nodes of $\mathcal{T}_{h_j}(B_{lj})$ on B_{lj}).
- $Y_{h_j}(B_{lj})$ consists of piecewise polynomials of degree q_j (typically $q_j = 1$).
- $Y_{h_j}(B_{lj})$ contains *constant* functions (if $q_j = 1$).

We let $z_1^{(l,j)}, \ldots, z_{m_{lj}}^{(l,j)}$ denote the *interior* nodes of triangulation $\mathcal{T}_{h_j}(B_{lj})$. Then, each local space $Y_{h_j}(B_{lj})$ can be defined as:

$$Y_{h_j}(B_{lj}) = \text{span}\{\psi_1^{(l,j)}(x), \ldots, \psi_{m_{lj}}^{(l,j)}(x)\} \subset L^2(B_{lj}),$$

where each continuous or discontinuous multiplier basis $\psi_i^{(l,j)}(x)$ is associated with the interior node $z_i^{(l,j)}$ of the triangulation $\mathcal{T}_{h_j}(B_{lj})$. The constraint $\mathcal{M}_*(u_h, \psi) = 0$ for each $\psi \in Y_h$ will yield m equations:

$$\int_{B_{lj}} \left(u_{h_l}(x) - u_{h_j}(x)\right) \psi(x)\, ds_x = 0, \quad \forall \psi(x) \in Y_{h_j}(B_{lj}),$$

for $j \in \mathcal{I}(l)$ and $1 \leq l \leq p$, where $m = \sum_{l=1}^{p} \sum_{j\, \mathcal{I}(l)} m_{lj}$, $n = (n_1 + \cdots + n_p)$.

Remark 11.16. If assumption *(A.2)* holds for each B_{lj}, we may *alternatively* apply *strong* matching between $u_{h_l}(x)$ and $u_{h_j}(x)$ on all *interior* nodes $z_1^{(l,j)}, \ldots, z_{m_{lj}}^{(l,j)}$ of $\mathcal{T}_{h_j}(B_{lj})$ for B_{lj} having dimension $(d-2)$ or lower:

$$u_{h_l}(z_r^{(l,j)}) = u_{h_j}(z_r^{(l,j)}), \qquad \text{for} \quad 1 \leq r \leq m_{lj}.$$

On the segments B_{lj} of dimension $(d-1)$, the standard two subdomain mortar based *weak* matching can be applied based on $Y_{h_j}(B_{lj})$. If all the grids match on B_{lj}, then the global solution will be conforming.

If matching is enforced only on segments B_{lj} of dimension $(d-1)$, then:

$$Y_h = \Pi_{l=1}^{p} \left(\Pi_{j \in \mathcal{I}_*(l)} Y_{h_j}(B_{lj})\right),$$

where the segments B_{lj} of dimension $(d-1)$ do not include *cross points* when $\Omega \subset \mathbb{R}^2$, or *cross points* and *edges* when $\Omega \subset \mathbb{R}^3$. Despite omission of the segments B_{lj} of dimension $(d-2)$ or lower, the segments of dimension $(d-1)$ will *cover* B. On each $(d-1)$ dimensional segment B_{lj} separating Ω_l and Ω_j, we shall let $Y_{h_j}(B_{lj}) \subset H^{-1/2}(B_{lj})$ denote a *continuous* or *discontinuous* multiplier space (as for a two subdomain interface B_{12}) satisfying:

- $Y_{h_j}(B_{lj})$ has dimension m_{lj} (number of interior nodes in $\mathcal{T}_{h_j}(B_{lj})$).
- $Y_{h_j}(B_{lj})$ consists of piecewise polynomials of degree q_j (typically $q_j = 1$).
- $Y_{h_j}(B_{lj})$ contains *constant* functions (if $q_j = 1$).

The constraint $\mathcal{M}_*(u_h, \psi) = 0$ for $\psi \in Y_h$ will yield a total of m equations:

$$\int_{B_{lj}} \left(u_{h_l}(x) - u_{h_j}(x)\right)\, \psi(x)\, ds_x = 0, \quad \forall \psi(x) \in Y_{h_j}(B_{lj}),$$

for $j \in \mathcal{I}_*(l)$, $1 \leq l \leq p$, $m = \sum_{l=1}^{p} \sum_{j\, \mathcal{I}_*(l)} m_{lj}$, and $n = (n_1 + \cdots + n_p)$. The multiplier space Y_h will contain constants on B, and the resulting mortar element discretization will have optimal order convergence (as when $Y_h \subset \tilde{Y}$).

Saddle Point System. Given $u_h(x) = (u_{h_1}(x), \ldots, u_{h_p}(x)) \in X_h$, we shall let $\mathbf{u} \in \mathbb{R}^n$ denote the vector of nodal values associated with $u_h(x)$, where each $\mathbf{u}$ is block partitioned as follows:

$$\mathbf{u} = \left(\mathbf{u}^{(1)^T}, \ldots, \mathbf{u}^{(p)^T}\right)^T \quad \text{with} \quad \mathbf{u}^{(l)} = \left(\mathbf{u}_I^{(l)^T}, \mathbf{u}_B^{(l)^T}\right)^T,$$

where $\mathbf{u}_I^{(l)} \in \mathbb{R}^{n_I^{(l)}}$ and $\mathbf{u}_B^{(l)} \in \mathbb{R}^{n_B^{(l)}}$ denote nodal vectors corresponding to values of $u_{h_l}(x) \in X_{h_l}(\Omega_l)$ on nodes of $\mathcal{T}_{h_l}(\Omega_l)$ in the *interior* of Ω_l and on $B^{(l)}$, respectively. Similarly, we shall let $\boldsymbol{\psi} \in \mathbb{R}^m$ denote the coefficient vector associated with $\psi(x) \in Y_h$.

A matrix representation of discretization (11.19) or (11.63) will involve:

$$A^{(l)} = \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{IB}^{(l)^T} & A_{BB}^{(l)} \end{bmatrix}, \quad M^{(l)} = \begin{bmatrix} 0 & M_B^{(l)} \end{bmatrix}, \quad \mathbf{u}^{(l)} = \begin{bmatrix} \mathbf{u}_I^{(l)} \\ \mathbf{u}_B^{(l)} \end{bmatrix}, \quad \mathbf{f}^{(l)} = \begin{bmatrix} \mathbf{f}_I^{(l)} \\ \mathbf{f}_B^{(l)} \end{bmatrix} \tag{11.66}$$

where $A^{(l)}$ and $\mathbf{f}^{(l)}$ satisfy:

$$\begin{cases} \mathcal{A}_{\Omega_l}(u_{h_l}, u_{h_l}) = \mathbf{u}^{(l)^T} A^{(l)} \mathbf{u}^{(l)} \\ \quad (f, u_{h_l})_{\Omega_l} = \mathbf{f}^{(l)^T} \mathbf{u}^{(l)}, \end{cases} \tag{11.67}$$

and for $Y_h \subset \tilde{Y}$, the matrices $M^{(l)}$ satisfy:

$$\begin{aligned} \mathcal{M}_*(u_h, \psi_h) &= \textstyle\sum_{l=1}^{p} \sum_{j \in \mathcal{I}(l)} \int_{B_{lj}} \left(u_{h_l}(x) - u_{h_j}(x)\right)\, \psi^{(l,j)}(x)\, ds_x \\ &= \boldsymbol{\psi}^T \left(M_B^{(1)} \mathbf{u}_B^{(1)} + \cdots + M_B^{(p)} \mathbf{u}_B^{(p)}\right) \\ &= \boldsymbol{\psi}^T \left(M^{(1)} \mathbf{u}^{(1)} + \cdots + M^{(p)} \mathbf{u}^{(p)}\right) \\ &= \boldsymbol{\psi}^T M \mathbf{u}. \end{aligned} \tag{11.68}$$

For $Y_h \subset Y$, the index set $\mathcal{I}(l)$ above must be replaced by $\mathcal{I}_*(l) \subset \mathcal{I}(l)$. The saddle point linear system corresponding to (11.19) or (11.63) will be:

$$\begin{bmatrix} A^{(1)} & & 0 & M^{(1)^T} \\ & \ddots & & \vdots \\ 0 & & A^{(p)} & M^{(2)^T} \\ M^{(1)} & \cdots & M^{(p)} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(1)} \\ \vdots \\ \mathbf{u}^{(p)} \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)} \\ \vdots \\ \mathbf{f}^{(p)} \\ \mathbf{0} \end{bmatrix}. \tag{11.69}$$

Here $n_I^{(l)}$, $n_B^{(l)}$ and $n_l = n_I^{(l)} + n_B^{(l)}$ denote the sizes of $\mathbf{u}_I^{(l)}$, $\mathbf{u}_B^{(l)}$ and $\mathbf{u}^{(l)}$, respectively, while m denotes the dimension of Y_h. Matrices $A_{II}^{(l)}$, $A_{IB}^{(l)}$, $A_{BB}^{(l)}$ and $M_B^{(l)}$ will be of size $n_I^{(l)}$, $n_I^{(l)} \times n_I^{(l)}$, $n_B^{(l)}$ and $m_l \times n_B^{(l)}$, respectively. Vectors $\mathbf{f}_I^{(l)}$, $\mathbf{f}_B^{(l)}$ and $\mathbf{f}^{(l)}$ will be of size $n_I^{(l)}$, $n_B^{(l)}$ and n_l, respectively.

A more compact representation of saddle point system (11.69) is:

$$\begin{bmatrix} A & M^T \\ M & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}, \tag{11.70}$$

involving the block partitioned matrices

$$A = \begin{bmatrix} A^{(1)} & & 0 \\ & \ddots & \\ 0 & & A^{(p)} \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} M^{(1)} & \cdots & M^{(p)} \end{bmatrix}, \tag{11.71}$$

with partitioned vectors $\mathbf{u} = \left(\mathbf{u}^{(1)^T}, \ldots, \mathbf{u}^{(p)^T}\right)^T$ and $\mathbf{f} = \left(\mathbf{f}^{(1)^T}, \ldots, \mathbf{f}^{(p)^T}\right)^T$. Iterative algorithms for solving the system (11.70) are considered next.

11.2.4 Saddle Point Iterative Solvers

Saddle point system (11.70) can be solved by modifying the FETI algorithm from Chap. 4, or the block matrix preconditioned algorithms from Chap. 10.5. We outline variants of the FETI algorithm, and other block preconditioners, see [AC2, KU7, LE8, KU8, HO5, AC6, KL8, FA11, FA10, LA, ST5, ST4].

Employing the block matrix structure of submatrices $A^{(l)}$ and $M^{(l)}$, and vectors $\mathbf{u}^{(l)}$, $\mathbf{f}^{(l)}$ in (11.69), we may reduce this system to a smaller, but equivalent, saddle point system. To obtain this, substitute the expressions from (11.63) into (11.69), and reorder the blocks as follows:

$$\mathbf{u}_I = \left(\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_I^{(p)^T}\right)^T \quad \text{and} \quad \mathbf{u}_B = \left(\mathbf{u}_B^{(1)^T}, \ldots, \mathbf{u}_B^{(p)^T}\right)^T.$$

This will yield the following reordered system:

$$\begin{bmatrix} A_{II} & A_{IB} & 0 \\ A_{IB}^T & A_{BB} & M_B^T \\ 0 & M_B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ \mathbf{f}_B \\ \mathbf{0} \end{bmatrix}, \tag{11.72}$$

where the block submatrices A_{II}, A_{IB} and A_{BB} are defined as follows:

$$A_{II} = \begin{bmatrix} A_{II}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{II}^{(p)} \end{bmatrix}, \; A_{IB} = \begin{bmatrix} A_{IB}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{IB}^{(p)} \end{bmatrix}, \; A_{BB} = \begin{bmatrix} A_{BB}^{(1)} & & 0 \\ & \ddots & \\ 0 & & A_{BB}^{(p)} \end{bmatrix},$$

with matrix $M_B = \begin{bmatrix} M_B^{(1)} & \cdots & M_B^{(p)} \end{bmatrix}$, while the load vectors satisfy:

$$\mathbf{f}_I = \left(\mathbf{f}_I^{(1)^T}, \ldots, \mathbf{f}_I^{(p)^T}\right)^T \text{ and } \mathbf{f}_B = \left(\mathbf{f}_B^{(1)^T}, \ldots, \mathbf{f}_B^{(p)^T}\right)^T.$$

Solving for $\mathbf{u}_I$ using the first block row of (11.72) yields $\mathbf{u}_I = A_{II}^{-1}(\mathbf{f}_I - A_{IB}\mathbf{u}_B)$. Substituting this expression into the second block row of (11.72) yields the following reduced saddle point system for $\mathbf{u}_B$ and $\boldsymbol{\psi}$:

$$\begin{bmatrix} S_{\mathcal{E}\mathcal{E}} & M_B^T \\ M_B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_B \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_B \\ \mathbf{0} \end{bmatrix}, \tag{11.73}$$

where $\tilde{\mathbf{f}}_B \equiv \left(\mathbf{f}_B - A_{IB}^T A_{II}^{-1}\mathbf{f}_I\right)$ and $S_{\mathcal{E}\mathcal{E}} = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB})$ satisfies:

$$S_{\mathcal{E}\mathcal{E}} = \text{blockdiag}\left((A_{BB}^{(1)} - A_{IB}^{(1)^T} A_{II}^{(1)^{-1}} A_{IB}^{(1)}), \ldots, (A_{BB}^{(p)} - A_{IB}^{(p)^T} A_{II}^{(p)^{-1}} A_{IB}^{(p)})\right).$$

The solution to (11.69) can be obtained by solving (11.73) for $\mathbf{u}_B$ and $\boldsymbol{\psi}$, and subsequently $\mathbf{u}_I = A_{II}^{-1}\left(\mathbf{f}_I - A_{IB}\mathbf{u}_B\right)$. Here M_B and $M_B^{(l)}$ will be of size $m \times n_B$ and $m \times n_B^{(l)}$ respectively, and $S_{\mathcal{E}\mathcal{E}}$ of size $n_B = (n_B^{(1)} + \cdots + n_B^{(p)})$.

We shall first describe the solution of (11.73) using variants of the FETI algorithm [FA11, FA10]. For convenience, we assume that coefficient $c(x) = 0$. In this case the subdomain stiffness matrices $A^{(l)}$ and Schur complement matrices $S^{(l)} = (A_{BB}^{(l)} - A_{IB}^{(l)^T} A_{II}^{(l)^{-1}} A_{IB}^{(l)})$ will be *singular* on floating subdomains. Let $Z^{(l)}$ be a matrix of size $n_B^{(l)} \times d_l$ such that $\text{Kernel}(S^{(l)}) = \text{Range}(Z^{(l)})$. We let $Z = \text{blockdiag}(Z^{(1)}, \ldots, Z^{(p)})$ of size $n_B \times d$ denote a matrix such that $\text{Kernel}(S_{\mathcal{E}\mathcal{E}}) = \text{Range}(Z)$, where $n_B = (n_B^{(1)} + \cdots + n_B^{(p)})$ and $d = (d_1 + \cdots + d_p)$. In this case, the first block equation of (11.73) will be solvable only if:

$$Z^T\left(\tilde{\mathbf{f}}_B - M_B^T\boldsymbol{\psi}\right) = \mathbf{0}. \tag{11.74}$$

When the above compatibility condition holds, $\mathbf{u}_B$ will satisfy:

$$\mathbf{u}_B = S_{\mathcal{E}\mathcal{E}}^{\dagger}\left(\tilde{\mathbf{f}}_B - M_B^T\boldsymbol{\psi}\right) + Z\boldsymbol{\alpha}, \tag{11.75}$$

where $S_{\mathcal{E}\mathcal{E}}^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $S_{\mathcal{E}\mathcal{E}}$ and $\alpha \in \mathbb{R}^d$. If this expression is substituted into the second block row of (11.73), we obtain:

$$M_B^T S_{\mathcal{E}\mathcal{E}}^{\dagger}\left(\tilde{\mathbf{f}}_B - M_B^T\boldsymbol{\psi}\right) + M_B Z\boldsymbol{\alpha} = \mathbf{0}. \tag{11.76}$$

Combining the above equation with the compatibility condition (11.74) will yield the following system of equations for $\boldsymbol{\psi} \in \mathbb{R}^m$ and $\boldsymbol{\alpha} \in \mathbb{R}^d$:

$$\begin{cases} K\,\boldsymbol{\psi} - G\,\boldsymbol{\alpha} = \mathbf{d} \\ \qquad\quad G^T\boldsymbol{\psi} = \mathbf{e}, \end{cases} \tag{11.77}$$

where matrices $K = M_B S_{\mathcal{E}\mathcal{E}}^{\dagger} M_B^T$ and $G = M_B Z$ are of sizes m and $m \times d$, with $\mathbf{d} = M_B S_{\mathcal{E}\mathcal{E}}^{\dagger} \tilde{\mathbf{f}}_B \in \mathbb{R}^m$ and $\mathbf{e} = Z^T \tilde{\mathbf{f}}_B \in \mathbb{R}^d$. A projection method will solve (11.77). Decompose $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \tilde{\boldsymbol{\psi}}$ using $\boldsymbol{\psi}_0 = G(G^T G)^{-1}\mathbf{e}$ (so that $G^T \boldsymbol{\psi}_0 = \mathbf{e}$). Then $\tilde{\boldsymbol{\psi}} \in \mathbb{R}^m$ and $\boldsymbol{\alpha} \in \mathbb{R}^d$ will solve the following linear system:

$$\begin{cases} K\,\tilde{\boldsymbol{\psi}} - G\,\boldsymbol{\alpha} = \tilde{\mathbf{d}} \\ \qquad\; G^T\,\tilde{\boldsymbol{\psi}} = \mathbf{0}, \end{cases} \quad \text{where} \quad \tilde{\mathbf{d}} = \mathbf{d} - K\,\boldsymbol{\psi}_0. \tag{11.78}$$

When $\text{Kernel}(M_B) \cap \text{Range}(Z) = \{\mathbf{0}\}$, matrix K will satisfy $K = K^T > 0$. We may *formally* eliminate $\boldsymbol{\alpha}$ by applying the (Euclidean) orthogonal projection $P_0 = I - G(G^T G)^{-1} G^T$ to the first block row above, and seek $\tilde{\boldsymbol{\psi}}$ within the *subspace* $\text{Kernel}(G^T)$, using that $\text{Range}(G)^{\perp} = \text{Kernel}(G^T)$. This will yield:

$$P_0\,K\,\tilde{\boldsymbol{\psi}} = P_0\,\tilde{\mathbf{d}}, \tag{11.79}$$

where $\boldsymbol{\alpha}$ can be determined as $\boldsymbol{\alpha} = (G^T G)^{-1}\,G^T(K\,\tilde{\boldsymbol{\psi}} - \tilde{\mathbf{d}})$. Since typically $d \le p$, an application of the projection P_0 will be computationally tractable.

To determine $\tilde{\boldsymbol{\psi}} \in \mathbb{R}^m$, system (11.79) can be solved using a *projected* CG algorithm, as described in Chap. 4 (see also [ST4]). Projection P_0 will provide *global transfer* of information. We list three *Dirichlet preconditioners* D_i for $P_0\,K$, see [LA, LA2, KL8, ST4], where the action of D_i^{-1} has the form:

$$\begin{cases} D_1^{-1} = P_0 \sum_{l=1}^{p} M_B^{(l)} S^{(l)} M_B^{(l)^T} \\ D_2^{-1} = P_0\,R^{-T} \left(\sum_{l=1}^{p} M_B^{(l)} S^{(l)} M_B^{(l)^T} \right) R^{-1} \\ D_3^{-1} = P_0 \left(M_B M_B^T \right)^{-T} \left(\sum_{l=1}^{p} M_B^{(l)} S^{(l)} M_B^{(l)^T} \right) \left(M_B M_B^T \right)^{-1}. \end{cases} \tag{11.80}$$

Here $R = \text{blockdiag}(M_B M_B^T)$ denotes the *block diagonal* matrix obtained from $M_B M_B^T$, where each block corresponds to Lagrange multipliers variables associated with interior nodes on each nonmortar interface B_{lj} for $j \in \mathcal{I}_*(l)$. Preconditioner D_1 is the original Dirichlet preconditioner [FA15, FA14], while preconditioners D_2 and D_3 were proposed in [LA] and [KL8], respectively. Theoretical results indicate that:

$$\text{cond}(D_i, P_0\,K) \;\le\; C\,(1 + \log(h_0/h))^3\,, \quad \text{for} \;\; i = 2, 3,$$

where h_0 denotes the diameter of the subdomains. The iterative algorithm converges faster for discontinuous mortar spaces with dual basis, see [ST4]. The performance of preconditioner D_1 deteriorates for nonmatching grids.

Remark 11.17. When coefficient $c(x) \ge c_0 > 0$, matrices A and $S_{\mathcal{E}\mathcal{E}}$ will be *nonsingular*, yielding $P_0 = I$. Then $\boldsymbol{\psi}$ will solve the following reduced problem:

$$K\,\boldsymbol{\psi} = \mathbf{d}, \;\; \text{where} \quad K = M_B\,S_{\mathcal{E}\mathcal{E}}^{-1}\,M_B^T \;\; \text{and} \;\; \mathbf{d} = M_B S_{\mathcal{E}\mathcal{E}}^{-1} \tilde{\mathbf{f}}_B. \tag{11.81}$$

In this case, the preconditioners in (11.80) will not provide *global* transfer of information. Such transfer may be included by employing the K-orthogonal projection Q described in Chap. 4. Let C denote an $m \times d$ matrix defined by $C = M_B \tilde{Z}$, where $\tilde{Z} = \text{blockdiag}(\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(p)})$ of size $n_B \times d$ with $\text{Range}(Z^{(i)}) = \text{Kernel}(\tilde{S}^{(i)})$ where $\tilde{S}^{(i)}$ denotes the subdomain Schur complement when $c(x) = 0$. Then, system (11.81) will be equivalent to:

$$K\boldsymbol{\psi} = \mathbf{d} \quad \text{and} \quad C^T K \boldsymbol{\psi} = C^T K \mathbf{d}, \tag{11.82}$$

since the second block equation is redundant. Decompose $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \tilde{\boldsymbol{\psi}}$ choosing $\boldsymbol{\psi}_0 = C(C^T K C)^{-1} C^T K \mathbf{d}$ so that $C^T K \tilde{\boldsymbol{\psi}} = \mathbf{0}$. Then $\tilde{\boldsymbol{\psi}}$ will solve:

$$K\tilde{\boldsymbol{\psi}} = \tilde{\mathbf{d}} \quad \text{with} \quad C^T K \tilde{\boldsymbol{\psi}} = \mathbf{0}, \tag{11.83}$$

where $\tilde{\mathbf{d}} = \mathbf{d} - K\boldsymbol{\psi}_0$. The solution to (11.83) can be sought by solving $K\tilde{\boldsymbol{\psi}} = \tilde{\mathbf{d}}$ within the subspace $\mathcal{G}_0 = \text{Kernel}(C^T K)$, using a *projected* conjugate gradient method. Dirichlet preconditioners can be obtained by replacing projection P_0 in (11.80) by the K-orthogonal projection $Q = I - C\left(C^T K C\right)^{-1} C^T K$.

Block matrix preconditioned algorithms may also be used to solve (11.70) using preconditioners of the form $\tilde{L}_i$ for the coefficient matrix L of the saddle point system [AC2, AC3, LE9, LE8, KU8, AC6]:

$$L = \begin{bmatrix} A & M^T \\ M & 0 \end{bmatrix}, \tilde{L}_1 = \begin{bmatrix} \tilde{A} & 0 \\ M & -\tilde{K} \end{bmatrix}, \tilde{L}_2 = \begin{bmatrix} \tilde{A} & 0 \\ 0 & -\tilde{K} \end{bmatrix}, \tilde{L}_3 = \begin{bmatrix} \tilde{A} & M^T \\ M & 0 \end{bmatrix},$$

where $\tilde{A}$ and $\tilde{K}$ denote preconditioners for matrices A and $K = MA^{\dagger}M^T$, respectively. Such iterative solvers were described in Chap. 10.5.

Since $A = \text{blockdiag}(A^{(1)}, \ldots, A^{(p)})$ where $A^{(l)}$ corresponds to the subdomain stiffness matrix on Ω_l with Neumann boundary conditions on $B^{(l)}$, standard preconditioners $\tilde{A}^{(l)}$ can be employed for each $A^{(l)}$. We may define $\tilde{A} = \text{blockdiag}(\tilde{A}^{(1)}, \ldots, \tilde{A}^{(p)})$ such that $\text{cond}\big(\tilde{A}, A\big) \leq C$ independent of h_l.

To construct a preconditioner for $K = MA^{\dagger}M^T$ in (11.70) note that:

$$MA^{\dagger}M^T = \sum_{l=1}^{p} M^{(l)} A^{(l)^{\dagger}} M^{(l)^T}, \tag{11.84}$$

due to the block structure of A and M. Employing the block structure of each subdomain stiffness matrix $A^{(l)}$ and $M^{(l)}$, we obtain:

$$\begin{aligned} M^{(l)} A^{(l)^{\dagger}} M^{(l)^T} &= \begin{bmatrix} 0 \\ M_B^{(l)^T} \end{bmatrix}^T \begin{bmatrix} A_{II}^{(l)} & A_{IB}^{(l)} \\ A_{IB}^{(l)^T} & A_{BB}^{(l)} \end{bmatrix}^{\dagger} \begin{bmatrix} 0 \\ M_B^{(l)^T} \end{bmatrix} \\ &= M_B^{(l)} S^{(l)^{\dagger}} M_B^{(l)^T}, \end{aligned} \tag{11.85}$$

where $S^{(l)} = (A_{BB}^{(l)} - A_{IB}^{(l)^T} A_{II}^{(l)^{-1}} A_{IB}^{(l)})$ denotes the Schur complement of $A^{(l)}$. This yields the equivalence:

$$M A^{\dagger} M^T = \sum_{l=1}^{p} M_B^{(l)} \, S^{(l)^{\dagger}} \, M_B^{(l)^T} = M_B S^{\dagger} M_B^T = K. \tag{11.86}$$

As a result, Dirichlet preconditioners described earlier can be employed. Thus, when $c(x) = 0$, the following preconditioner $\tilde{K}$ can be employed for K:

$$\tilde{K}^{-1} = T_0 + P_0 \, \left(M_B M_B^T\right)^{-T} \left(\sum_{l=1}^{p} M_B^{(l)} S^{(l)} M_B^{(l)^T} \right) \left(M_B M_B^T\right)^{-1} P_0,$$

where $P_0 = I - G\left(G^T G\right)^{-1} G^T$ and $T_0 = G\left(G^T K G\right)^{-1} G^T$. Alternative preconditioners D_i from (11.80) may also be suitably substituted. When the coefficient $c(x) \geq c_0 > 0$, a preconditioner $\tilde{K}$ for K can be defined as:

$$\tilde{K}^{-1} = \tilde{T}_0 + Q \, \left(M_B M_B^T\right)^{-T} \left(\sum_{l=1}^{p} M_B^{(l)} S^{(l)} M_B^{(l)^T} \right) \left(M_B M_B^T\right)^{-1} Q^T,$$

where $Q = I - C(C^T K C)^{-1} C^T K$ and $\tilde{T}_0 = C(C^T K C)^{-1} C^T$.

Remark 11.18. In the special case of a saddle point preconditioner $\tilde{L}$:

$$\tilde{L} = \begin{bmatrix} \tilde{A} & M_B^T \\ M_B & 0 \end{bmatrix},$$

the following choice was suggested in [AC2, AC3] for $\tilde{A}^{(l)}$ on $\Omega_l \subset \mathbb{R}^d$, where $\tilde{A} = \text{blockdiag}(\tilde{A}^{(1)}, \cdots, \tilde{A}^{(p)})$:

$$\tilde{A}^{(l)} = h_l^{d-2} \, \alpha_l \, (I^{(l)} - P^{(l)}) + \gamma_l \, h_l^{d-1} \, \beta_l P^{(l)}, \quad \forall 1 \leq l \leq p,$$

for $\gamma_l = \text{diam}(\Omega_l)$, $I^{(l)}$ the identity of size corresponding to the number of unknowns in $X_{h_l}(\Omega_l)$ and $P^{(l)}$ is the matrix that maps a function defined on $\partial\Omega_l$ onto its mean value, based on a preconditioner of [BR13]. A condition number $\text{cond}(\tilde{L}, L) \leq C \max_l \, (h_l/\gamma_l)$ is proved in [AC2, AC3] with C independent of h_l, γ_l (for aspect ratios bounded by a constant), and α_l and β_l. A linear system with coefficient matrix $\tilde{A}$ can be solved at a cost proportional to the number of unknowns in $\mathbf{u}$.

Remark 11.19. Alternative Neumann-Neumann, Schwarz and multigrid algorithms are described in [LE, DR6, GO11, BR3, DR7, DR8].

11.2.5 Accuracy of Mortar Element Discretizations

In the following, we state without proof, theoretical results on the *stability* and *accuracy* of mortar element saddle point discretizations. The reader is referred to [BE18, BE6, BE4, BE21, HO5, BR2, WO4, KI] for proofs and details. We shall focus only on the convergence of mortar element-saddle point discretizations based on *continuous* and *discontinuous* multiplier spaces Y_h.

Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of Ω in which each Ω_l is triangulated by a quasiuniform local grid $\mathcal{T}_{h_l}(\Omega_l)$ for $1 \le l \le p$. We shall assume that the local grids satisfy conformity condition *(A.2)*, and assume that the *nonmortar* segments B_{lj} have indices $j \in \mathcal{I}_*(l)$ for $1 \le l \le p$, where the dimension of each nonmortar B_{lj} is $(d-1)$ when $\Omega \subset \mathbb{R}^d$.

We employ the notation $\mathcal{A}_*(.,.)$ and $\mathcal{M}_*(.,.)$ to denote the bilinear forms:

$$\begin{cases} \mathcal{A}_*(u,v) = \sum_{l=1}^p \int_{\Omega_l} \left(a(x)\, \nabla u \cdot \nabla v + c(x)\, u\, v\right)\, dx \\ \mathcal{M}_*(u,\psi) = \sum_{l=1}^p \sum_{j \in \mathcal{I}_*(l)} \int_{B_{lj}} \left(u_l(x) - u_j(x)\right)\, \psi_{l,j}(x)\, ds_x, \end{cases} \tag{11.87}$$

where $u,\, v \in X$ and $\psi \in \tilde{Y}$. The following result concerns the stability and convergence of mortar element discretizations. Bounds for $u - u_h$ will reduce to standard finite element bounds when the grids *match* and when h_l are all of the same size, showing that mortar element discretizations are of optimal order in h_l. Here $\|\cdot\|_{1,\Omega_l}$ will denote a Sobolev norm while $\|\cdot\|_{Y_h(B_{lj})}$ would denote the norm $\left(H_{00}^{1/2}(B_{lj})\right)'$.

Lemma 11.20. *Suppose the following conditions hold.*

- *Let the coefficient $a(x) = a_l$ on each Ω_l in (11.1).*
- *Let the solution u to (11.1) satisfy $u \in H^{1+s}(\Omega_l)$ for $0 < s \le 2$ on Ω_l.*
- *Let the nonmortar sides be chosen so that $a_j \ge a_l$ and $h_j \ge h_l$ for each $j \in \mathcal{I}(l)$ and $1 \le l \le p$.*
- *Let $X_{h_l}(\Omega_l)$ denote piecewise linear finite element spaces with associated continuous or discontinuous multiplier space $Y_h \subset Y$.*

Then the following results will hold.

1. *The uniform inf-sup condition (11.31) will hold.*
2. *The coercivity condition:*

$$\mathcal{A}_*(v_h, v_h) \ge \alpha\, \|v_h\|_X^2,$$

will hold for $v_h \in \mathcal{K}_0^h$, where:

$$\mathcal{K}_0^h = \{v_h \in X_h : \mathcal{M}_*(v_h, \psi_h) = 0,\ \forall \psi_h \in Y_h\},$$

for $\alpha > 0$ independent of $h_1, \ldots, h_p$.

3. The following error bound will hold:

$$\begin{aligned}&\sum_{l=1}^{p} \|u - u_{h_l}\|^2_{1,\Omega_l} + \sum_{l=1}^{p} \sum_{j \in \mathcal{I}_*(l)} \|\psi - \psi_h\|^2_{Y_h(B_{lj})} \\ &\quad \le C \left(\sum_{l=1}^{p} h_l^{2\,s} \|u\|^2_{1+s,\Omega_l} + \sum_{l=1}^{p} h_l^{2\,s} \|u\|^2_{1/2+s,B_{lj}} \right),\end{aligned}$$

where $C > 0$ is independent of $h_1, \ldots, h_p$.

Proof. General results [GI3, BR33] on the stability and accuracy of discretizations of saddle point problems will yield result 3, provided results 1 and 2 hold, and provided suitable approximation properties hold for the spaces X_h and Y_h, see [BE18, BE6, BE4, BE21, HO5, BR2, WO4, KI]. The proof of the *inf-sup* condition will depend on each term $\int_{B_{lj}} (u_l - u_j)\, \psi_{l,j}\, ds_x$ in $\mathcal{M}_*(u, \psi)$. Given $\psi_{l,j}(x) \in Y_{h_j}(B_{lj})$ define $u_{j,l}(x) = 0$ on B_{lj} and construct $u_{l,j}(x) \in \left(X_{h_l}(B_{lj}) \cap H^{1/2}_{00}(B_{lj}) \right)$ satisfying:

$$\int_{B_{lj}} (u_{l,j} - u_{j,l})\, \psi_{l,j}\, ds_x \ge \|\psi_{l,j}\|^2_{Y_h(B_{lj})} \quad \text{with} \quad \|u_{l,j}\| \le \gamma_{l,j}\, \|\psi_{l,j}\|,$$

with $\gamma_{l,j}$ independent of h_l, see preceding references. Define $u_l = \sum_{j \in \mathcal{I}_*(l)} u_{l,j}$ and extend it discrete harmonically into $X_{h_l}(\Omega_l)$. It can then easily be verified that the inf-sup bound will depend only on $\gamma_{l,j}$ and the maximum number of nonmortars on each subdomain. Proof of the coercivity condition will be trivial when $c(x) \ge c_0 > 0$, see preceding references. Once the inf-sup and coercivity conditions have been proved, the discretization error will depend solely on Sobolev norm approximation errors within X_h and Y_h, which will follow by finite element techniques and Sobolev interpolation theory. □

11.3 Mortar Element Discretization: Nonconforming Approach

The nonconforming approach to the mortar element discretization of elliptic equation (11.1) solves a self adjoint and coercive problem for the solutions $u_{h_l}(x)$ on the subdomain Ω_l, by eliminating the flux unknowns $\psi_{lj}(x)$ on the interfaces B_{lj}. Recall that the mortar element-saddle point discretization of elliptic equation (11.1) seeks $u_h \in X_h$ and $\psi_h \in Y_h$ satisfying:

$$\begin{cases} \mathcal{A}_*(u_h, v_h) + \mathcal{M}_*(v_h, \psi_h) = (f, v_h)_*, & \forall v_h \in X_h \\ \mathcal{M}_*(u_h, \phi_h) = 0, & \forall \phi_h \in Y_h, \end{cases} \tag{11.88}$$

for subspaces $X_h \subset X$ and $Y_h \subset Y$ (or $Y_h \subset \tilde{Y}$) with bilinear forms:

$$\begin{cases} \mathcal{A}_*(w, w) = \sum_{l=1}^{p} \int_{\Omega_l} \left(a(x) \nabla v_l \cdot \nabla w_l + c(x)\, v_l\, w_l \right) dx \\ (f, w)_* = \sum_{l=1}^{p} \int_{\Omega_l} \left(f(x)\, w_l \right) dx \\ \mathcal{M}_*(w, \psi) = \sum_{l=1}^{p} \sum_{j \in \mathcal{I}_*(l)} \int_{B_{lj}} [w]_{lj}\, \psi_{lj}\, ds_x, \end{cases} \tag{11.89}$$

for $[w]_{lj} = w_l - w_j$ for $w = (w_1, \ldots, w_p) \in X$ and $\psi \in Y$. This not only yields approximations $u_{h_l}(x)$ to the solution $u(x)$ on each subdomain Ω_l, but also approximations $\psi_{lj}(x)$ of the flux $\psi(x)$ on each mortar segment B_{lj}. However, the flux unknowns $\psi_h(x)$ can be *eliminated* (as will be shown later), yielding a self adjoint *coercive* problem for determining the subdomain solutions $\mathbf{u}_{h_l}$ for $1 \leq l \leq p$. These subdomain solutions will be identical to that obtained in the saddle point approach, and since they do not match strongly across the interfaces B_{lj}, the global solution will generally be $H^1(\Omega)$-*nonconforming.*

The mortar element-saddle point discretization (11.88) can be reduced to a self adjoint coercive problem within the following subspace $\mathcal{K}_0^h$:

$$\mathcal{K}_0^h = \{v_h(x) \in X_h \, : \, \mathcal{M}_*(v_h, \phi_h) = 0, \; \forall \phi_h \in Y_h\}. \tag{11.90}$$

Substituting $v_h \in \mathcal{K}_0^h$ into the first row of (11.88) and using $\mathcal{M}_*(v_h, \phi_h) = 0$ for all $\phi_h \in Y_h$, yields a reduced problem for $u_h \in \mathcal{K}_0^h$ satisfying:

$$\mathcal{A}(u_h, v_h) = (f, v_h), \quad \forall v_h \in \mathcal{K}_0^h. \tag{11.91}$$

The bilinear form $\mathcal{A}(.,.)$ is self adjoint, but will also be *coercive* within $\mathcal{K}_0^h$, when the saddle point discretization (11.88) is stable. Thus problem (11.91) will be uniquely solvable.

Remark 11.21. Generally, $\mathcal{K}_0^h$ will *not* be $H^1(\Omega)$ conforming, and hence the name *nonconforming* approach. However, if the local grids *match*, and the multiplier space Y_h uses all segments B_{lj}, so that $Y_h \subset \tilde{Y}$, then $\mathcal{K}_0^h \subset H^1(\Omega)$ will be a *conforming* finite element space defined on the global grid.

In matrix terms, the saddle point linear system associated with the saddle point discretization (11.88) of (11.1) is:

$$\begin{bmatrix} A & M^T \\ M & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}, \tag{11.92}$$

involving the block partitioned matrices

$$A = \begin{bmatrix} A^{(1)} & & 0 \\ & \ddots & \\ 0 & & A^{(p)} \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} M^{(1)} & \cdots & M^{(p)} \end{bmatrix}, \tag{11.93}$$

with partitioned vectors $\mathbf{u} = \left(\mathbf{u}^{(1)^T}, \ldots, \mathbf{u}^{(p)^T}\right)^T$ and $\mathbf{f} = \left(\mathbf{f}^{(1)^T}, \ldots, \mathbf{f}^{(p)^T}\right)^T$. A reduced system for $\mathbf{u}$ can be obtained by eliminating the block vector $\boldsymbol{\psi}$ by taking inner products of the first block row in the above system using vectors $\mathbf{v} \in \text{Kernel}(M)$. The reduced problem will seek $\mathbf{u} \in \text{Kernel}(M)$ satisfying:

$$\mathbf{v}^T A \, \mathbf{u} = \mathbf{v}^T \mathbf{f}, \quad \forall \mathbf{v} \in \text{Kernel}(M), \tag{11.94}$$

since $\mathbf{v}^T M^T = \mathbf{0}$ when $\mathbf{v} \in \text{Kernel}(M)$. We shall now describe how to compute an explicit *parametric representation* of the subspace Kernel(M). We shall separately consider the *two subdomain* and *multisubdomain* cases.

11.3.1 Two Subdomain Case

In the two subdomain case, the constraint $M\,\mathbf{v} = \mathbf{0}$ has the form:

$$M_B^{(1)}\mathbf{v}_B^{(1)} - M_B^{(2)}\mathbf{v}_B^{(2)} = \mathbf{0}.$$

If *zero* boundary conditions are imposed on ∂B_{12} and j_* denotes *nonmortar* side, then $M_B^{(j_*)}$ will be a square matrix and yield the representation:

$$\begin{cases} \mathbf{v}_B^{(1)} = R_{12}\mathbf{v}_B^{(2)} = {M_B^{(1)}}^{-1} M_B^{(2)}\mathbf{v}_B^{(2)}, & \text{if } j_* = 1 \\ \mathbf{v}_B^{(2)} = R_{21}\mathbf{v}_B^{(1)} = {M_B^{(2)}}^{-1} M_B^{(1)}\mathbf{v}_B^{(1)}, & \text{if } j_* = 2. \end{cases}$$

In this case Kernel(M) will have the following parametric representation:

$$\text{Kernel}(M) = \begin{cases} \left({\mathbf{v}_I^{(1)}}^T, R_{12}{\mathbf{v}_B^{(2)}}^T, {\mathbf{v}_I^{(2)}}^T, {\mathbf{v}_B^{(2)}}^T \right)^T, & \text{if } j_* = 1 \\ \left({\mathbf{v}_I^{(1)}}^T, {\mathbf{v}_B^{(1)}}^T, {\mathbf{v}_I^{(2)}}^T, R_{21}{\mathbf{v}_B^{(1)}}^T \right)^T, & \text{if } j_* = 2, \end{cases}$$

for arbitrary $\mathbf{v}_I^{(l)} \in \mathbb{R}^{n_I^{(l)}}$ for $1 \le l \le 2$ and $\mathbf{v}_B^{(j_*)} \in \mathbb{R}^{n_B^{(j_*)}}$. Substituting this representation into (11.94) yields the following linear system when $j_* = 1$:

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)}R_{12} \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ R_{12}^T{A_{IB}^{(1)}}^T & {A_{IB}^{(2)}}^T & R_{12}^T A_{BB}^{(1)} R_{12} + A_{BB}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ R_{12}^T\mathbf{f}_B^{(1)} + \mathbf{f}_B^{(2)} \end{bmatrix}.$$

Similarly, when $j_* = 2$ the following linear system will be obtained:

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)} \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)}R_{21} \\ {A_{IB}^{(1)}}^T & R_{21}^T{A_{IB}^{(2)}}^T & A_{BB}^{(1)} + R_{21}^T A_{BB}^{(2)} R_{21} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B^{(1)} + R_{21}^T\mathbf{f}_B^{(2)} \end{bmatrix}.$$

The resulting reduced system will be symmetric and positive definite.

Remark 11.22. When a *discontinuous* multiplier space Y_h is employed, the matrices $R_{12} = {M_B^{(1)}}^{-1} M_B^{(2)}$ or $R_{21} = {M_B^{(2)}}^{-1} M_B^{(1)}$ will be sparse and computable at low computational cost, since matrix $M_B^{(j_*)}$ will be a *diagonal* matrix. For a *continuous* multiplier space Y_h, matrix ${M_B^{(j_*)}}^{-1}$ will be *dense.*

Remark 11.23. When the grids *match*, it will hold that $M_B^{(1)} = M_B^{(2)}$. Provided these matrices are *square*, then $R_{12} = R_{21} = I$, and the global system will be equivalent to a conforming finite element discretization.

11.3.2 Multisubdomain Case

In the multisubdomain case, constructing a parametric representation of Kernel(M) will be more complicated. Below, we outline this for the case of p subdomains of $\Omega \subset \mathbb{R}^d$, when the multiplier space Y_h is defined based only on segments B_{lj} of dimension $(d-1)$, with indices $j \in \mathcal{I}_*(l)$ for $1 \le l \le p$:

$$Y_h = \Pi_{l=1}^{p} \left(\Pi_{j \in \mathcal{I}_*(l)} Y_{h_j}(B_{lj}) \right),$$

with *discontinuous* multiplier spaces $Y_{h_j}(B_{lj})$. The following may be noted.

- On each subdomain boundary $\partial\Omega_j$, the nodal unknowns in the *interior* of *nonmortar* segments B_{lj} (i.e., if $j \in \mathcal{I}_*(l)$ for some l) will be dependent (*slave*) variables. All other unknowns associated with the nodes on $\partial\Omega_j$ will correspond to independent (*master*) variables.
- On each *nonmortar* side, a relation of the form $\mathbf{u}^{(j)}_{B_{lj}} = M^{(j)^{-1}}_{B_{lj}} M^{(l)}_{B_{lj}} \mathbf{u}^{(l)}_{\overline{B}_{lj}}$ will hold, where $\mathbf{u}^{(l)}_{\overline{B}_{lj}}$ denotes a vector of nodal unknowns associated with $u_{h_l}(x)$ on all nodes of B_{lj}, while $\mathbf{u}^{(j)}_{B_{lj}}$ will denote a nodal vector associated with $u_{h_j}(x)$ on the *interior* nodes of B_{lj}.

Collecting together the above master variables, we shall employ the notation:

$$\mathbf{u}^{(l)}_B = R_l \boldsymbol{\mu}_B, \tag{11.95}$$

where $\boldsymbol{\mu}_B$ denotes a nodal vector of all *master* unknowns on B. In this notation, the reduced symmetric positive definite linear system (11.94) will be:

$$\begin{bmatrix} A^{(1)}_{II} & & 0 & A^{(1)}_{IB} R_1 \\ & \ddots & & \vdots \\ 0 & & A^{(p)}_{II} & A^{(p)}_{IB} R_p \\ R_1^T A^{(1)^T}_{IB} & \cdots & R_p^T A^{(p)^T}_{IB} & R^T A_{BB} R \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(1)}_I \\ \vdots \\ \mathbf{u}^{(p)}_I \\ \boldsymbol{\mu}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)}_I \\ \vdots \\ \mathbf{f}^{(p)}_I \\ R^T \mathbf{f}_B \end{bmatrix}, \tag{11.96}$$

where

$$\begin{cases} R^T A_{BB} R = R_1^T A^{(1)}_{BB} R_1 + \cdots + R_p^T A^{(p)}_{BB} R_p \\ \quad R^T \mathbf{f}_B = R_1^T \mathbf{f}^{(1)}_B + \cdots + R_p^T \mathbf{f}^{(p)}_B. \end{cases} \tag{11.97}$$

Linear system (11.96) will be symmetric and positive definite by construction.

11.3.3 Iterative Solvers

System (11.96) can be solved using a preconditioned CG method, with Neumann-Neumann, Schwarz or multigrid methods [LE, DR6, GO11, BR3]. Below, we outline a heuristic variant of the Neumann-Neumann algorithm to solve (11.96), which we express as:

$$\begin{bmatrix} A_{II} & A_{IB}R \\ R^T A_{IB}^T & R^T A_{BB} R \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \boldsymbol{\mu}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ R^T \mathbf{f}_B \end{bmatrix}, \tag{11.98}$$

where $R^T A_{BB} R$ and $R^T \mathbf{f}_B$ are defined by (11.97), and R_l by (11.95).

A Schur complement method can be formally employed, by eliminating $\mathbf{u}_I = A_{II}^{-1}(\mathbf{f}_I - A_{IB} R \boldsymbol{\mu}_B)$. This will yield the system:

$$R^T \left(A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB} \right) R\, \boldsymbol{\mu}_B = R^T \left(\mathbf{f}_B - A_{IB}^T A_{II}^{-1} \mathbf{f}_I \right). \tag{11.99}$$

For discontinuous multiplier spaces, the parameterization $\mathbf{u}_B^{(l)} = R_l \boldsymbol{\mu}_B$ can be chosen with each R_l to be of full rank, provided the number of interior nodes on each mortar side exceeds the number of interior nodes on its nonmortar side. To construct a Neumann-Neumann preconditioner for the Schur complement system, note that when $c(x) = 0$ some of the matrices $S^{(l)}$ will be *singular*. In this case, let $R_0 = \text{span}\left\{ R_1^T Z^{(1)}, \ldots, R_p^T Z^{(p)} \right\}$ and $\text{Range}(R_0)$ as the *coarse space*, where $\text{Range}(Z^{(k)}) = \text{Kernel}(S^{(k)})$. If $c(x) \geq c_0 > 0$, each $S^{(l)}$ will be *nonsingular*, and so define $\text{Range}(Z^{(k)}) = \text{Kernel}(\tilde{S}^{(k)})$ where $\tilde{S}^{(k)}$ is the subdomain Schur complement resulting when $c(x) = 0$. The inverse of the formal Neumann-Neumann preconditioner $\tilde{S}$ will be:

$$\tilde{S}^{-1} = R_0^T S^{(0)^{-1}} R_0 + \sum_{l=1}^{p} R_l^T S^{(l)^{\dagger}} R_l,$$

where $S^{(0)} \equiv R_0 S R_0^T$. Rigorous convergence bounds are not known [LE].

11.4 Schwarz Discretizations on Overlapping Grids

In certain applications, it may be advantageous to employ overlapping non-matching grids, see Fig. 11.9. Different techniques may be employed to discretize an elliptic equation on such grids [ST, ST6, GR16, HE9, HE10, GO7], [CA17, AC5], and techniques, such as Chimera, were formulated prior to the development of domain decomposition methodology, employing finite difference or finite volume methods locally. Such discretizations are also related to composite grid schemes [BA5, BE15, MC4, GR9, MC3, FE3]. In this section, we describe the discretization of an elliptic equation on an overlapping non-matching grid, using the Schwarz hybrid formulation.

11.4.1 Schwarz Hybrid Formulation

We shall consider the elliptic equation:

$$\begin{cases} L\,u \equiv -\nabla \cdot (a(x)\,\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega, \end{cases} \tag{11.100}$$

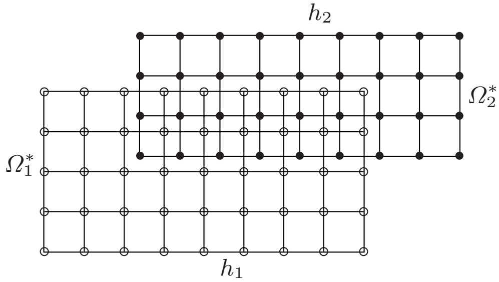


Fig. 11.9. Two overlapping non-matching grids

where $c(x) \geq c_0 > 0$. Let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping decomposition of Ω, obtained by extension of a nonoverlapping decomposition $\Omega_1, \ldots, \Omega_p$ of diameter h_0. Ideally, let $\Omega_l^{\beta h_0} = \{x \in \Omega : \operatorname{dist}(x, \Omega_l) \leq \beta\, h_0\}$ and choose $\Omega_l^* = \Omega_l^{\beta h_0}$ for $\beta < 1$. Let $\{\chi_l(x)\}_{l=1}^p$ denote a *partition of unity* subordinate to $\{\Omega_l^{\epsilon h_0}\}_{l=1}^p$ for some $0 < \epsilon \ll \beta$, where $\Omega_l^{\epsilon h_0} \subset \Omega_l^*$. By construction:

$$\sum_{j \neq l} \chi_j(x) = 1, \quad \text{for } x \in \partial\Omega_l \cap \Omega, \quad \text{for } 1 \leq l \leq p.$$

If $w_l(x)$ denotes the restriction of the solution $u(x)$ of (11.100) to Ω_l^*, then the following coupled system of equations will be satisfied by $w_l(x)$ for $1 \leq l \leq p$:

$$\begin{cases} L\, w_l = f, & \text{in } \Omega_l^* \\ \quad w_l = \sum_{j \neq l} \chi_j\, w_j, & \text{on } B^{(l)} = \partial\Omega_l \cap \Omega \\ \quad w_l = 0, & \text{on } B_{[l]} = \partial\Omega_l \cap \partial\Omega, \end{cases} \tag{11.101}$$

Under suitable regularity assumptions, this coupled system will be well posed.

11.4.2 Nonmatching Grid Discretization

Let each subdomain Ω_l^* be triangulated by a quasiuniform triangulation $\mathcal{T}_{h_l}(\Omega_l^*)$ with grid size h_l, where the local grids need not match on regions of overlap. Given such overlapping grids, a discretization of (11.100) can be obtained by discretization of its Schwarz hybrid formulation (11.101). Let $\mathbf{w}_I^{(l)}$ denote the vector of *interior* nodal values of the discrete solution on Ω_l^*, and $\mathbf{w}_B^{(l)}$ the vector of nodal values of the discrete solution on the boundary segment $B^{(l)} = \partial\Omega_l^* \cap \Omega$. Then, the matrix form of a finite difference or finite volume discretization of (11.101) will have the form:

$$\begin{cases} A_{II}^{(l)}\, \mathbf{w}_I^{(l)} + A_{IB}^{(l)}\, \mathbf{w}_B^{(l)} = \mathbf{f}_I^{(l)}, \\ \qquad\qquad\qquad\;\; \mathbf{w}_B^{(l)} = \sum_{j \neq l} C^{(l,j)}\, \mathbf{w}^{(j)}, \end{cases} \quad \text{for } 1 \leq l \leq p. \tag{11.102}$$

Here $A_{II}^{(l)}\, \mathbf{w}_I^{(l)} + A_{IB}^{(l)}\, \mathbf{w}_B^{(l)} = \mathbf{f}_I^{(l)}$ denotes a discretization of $L\, w_l = f$ in Ω_l^*, while $\mathbf{w}_B^{(l)} = \sum_{j \neq l} C^{(l,j)}\, \mathbf{w}^{(j)}$ denotes a discretization of $w_l = \sum_{j \neq l} \chi_j\, w_j$,

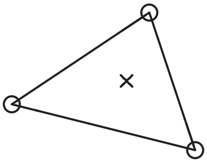


Fig. 11.10. Example of an interpolation stencil

where $\mathbf{w}^{(j)} = \left(\mathbf{w}_I^{(j)^T}, \mathbf{w}_B^{(j)^T}\right)^T$. The stencil for computing the boundary values of $\mathbf{w}_B^{(l)}$ on $B^{(l)}$ can be obtained by using suitably accurate finite difference or finite element *interpolation* to define the nodal values of $\mathbf{w}^{(l)}$ on $B^{(l)}$ using $\mathbf{w}^{(j)}$ for $j \neq l$. For instance, if x_r denotes a node from triangulation $\mathcal{T}_{h_l}(\Omega_l^*)$ lying on $B^{(l)}$ lying within a element of $\mathcal{T}_{h_j}(\Omega_j^*)$, then each term $\chi_j(x_r) w_j(x_r)$ can be approximated using the nodal values of $w_j(x)$ on the vertices of the element and summed up, see Fig. 11.10. Define:

$$\begin{cases} \mathbf{w}_I = \left(\mathbf{w}_I^{(1)^T}, \ldots, \mathbf{w}_I^{(p)^T}\right)^T \\ \mathbf{w}_B = \left(\mathbf{w}_B^{(1)^T}, \ldots, \mathbf{w}_B^{(p)^T}\right)^T \\ \mathbf{f}_I = \left(\mathbf{f}_I^{(1)^T}, \ldots, \mathbf{f}_I^{(p)^T}\right)^T. \end{cases}$$

Then, the above coupled system can be expressed in block matrix form as:

$$\begin{bmatrix} A_{II} & A_{IB} \\ -C & I \end{bmatrix} \begin{bmatrix} \mathbf{w}_I \\ \mathbf{w}_B \end{bmatrix} \begin{bmatrix} \mathbf{f}_I \\ \mathbf{0} \end{bmatrix}, \tag{11.103}$$

where $A_{II} = \text{blockdiag}(A_{II}^{(1)}, \ldots, A_{II}^{(p)})$, $A_{IB} = \text{blockdiag}(A_{IB}^{(1)}, \ldots, A_{IB}^{(p)})$, and C is a block submatrix with blocks $C^{(l,j)}$ for $l \neq j$ and zero diagonal blocks. Under suitable assumptions, system (11.103) will be *nonsingular* [CA17].

11.4.3 Accuracy of the Discretization

Theoretical results in [ST, CA17] and computational results in [ST6, GR16], [HE9, HE10, GO7] show that when the *interpolation* stencils are at least as accurate as the truncation errors of the local discretization schemes, when the overlap is sufficiently large, and when each local scheme satisfies a discrete maximum principle, then the global discretization error of a Schwarz discretization will be of *optimal* order, see Chap. 15.

Lemma 11.24. *Suppose the following conditions hold.*

1. *Let the local discretization error on Ω_l^* be $O(h_l^{q_l})$.*
2. *Let the local interpolation stencil $C^{(l)}$ on $B^{(l)}$ be $O(\mathcal{E}_l)$ accurate, with $\|C^{(l)}\|_\infty \leq 1$.*
3. *Let w_l and w_{h_l} denote the exact and discrete solution restricted to the grid points of $\Omega_l^* \cup B^{(l)}$.*
4. *Let a discrete maximum principle hold on each local grid.*

Then, the following bound will hold:

$$\sum_{l=1}^{p} \|w_l - w_{h_l}\|_\infty \leq \alpha \sum_{l=1}^{p} (O(h_l^{q_l}) + O(\mathcal{E}_l)),$$

where α will depend on higher order derivatives of the exact solution $u(x)$ and the amount of overlap, but will be independent of h_l.

Proof. See [ST, CA17] for finite difference schemes satisfying a discrete maximum principle, and see [AC5] for the finite element case. □

Remark 11.25. Generally, the stability bound improves as the overlap factor β increases with $\Omega_l^* = \Omega_l^{\beta h_0}$. It also improves as the parameter ϵ used in the partition of unity $\{\chi_l\}_{l=1}^p$ subordinate to $\{\Omega_l^{\epsilon h_0}\}_{l=1}^p$, decreases. In the discrete case, a discontinuous partition of unity obtained as $\epsilon \to 0^+$ would also be sufficient. In practice, the inter-subdomain interpolation stencil can be chosen to be a *convex* combination of nodal values from nodes on adjacent grids.

11.4.4 Iterative Solvers

We shall now describe sequential and parallel Schwarz iterative solvers for (11.102). Let $\mathbf{w}^{(l;k)}$ to denote the k'th iterate approximating:

$$\mathbf{w}^{(l)} = \left(\mathbf{w}_I^{(l)^T}, \mathbf{w}_B^{(l)^T}\right)^T.$$

Algorithm 11.4.1 *(Sequential Schwarz Algorithm)*

```
1. Let $\mathbf{v}^{(l;0)} = \mathbf{w}^{(l;0)}$ for $1 \leq l \leq p$
2. For $k = 0, 1, \ldots,$ until convergence do:
3.     For $l = 1, \ldots, p$ solve:
           $\begin{cases} A_{II}^{(l)} \mathbf{w}_I^{(l;k+\frac{l}{p})} + A_{IB}^{(l)} \mathbf{w}_B^{(l;k+\frac{l}{p})} = \mathbf{f}_I^{(l)}, \\ \mathbf{w}_B^{(l;k+\frac{l}{p})} = \sum_{j \neq l} C^{(l,j)} \mathbf{v}^{(j;k+\frac{(l-1)}{p})}, \end{cases}$
4.     Define: $\mathbf{v}^{(l;k+\frac{l}{p})} = \mathbf{w}^{(l;k+\frac{(l-1)}{p})}$
5.     Endfor
6. Endfor
```

A parallel version of the above algorithm is outlined below.

Algorithm 11.4.2 *(Parallel Schwarz Algorithm)*

```
1. For $k = 0, 1, \ldots,$ until convergence do:
2.     For $l = 1, \ldots, p$ in parallel solve:
           $$\begin{cases} A_{II}^{(l)}\, \mathbf{w}_I^{(l;k+1)} + A_{IB}^{(l)}\, \mathbf{w}_B^{(l;k+1)} = \mathbf{f}_I, \\ \mathbf{w}_B^{(l;k+1)} = \sum_{j \neq l} C^{(l,j)}\, \mathbf{w}^{(j;k)}, \end{cases}$$
3.     Endfor
4. Endfor
```

The above algorithms correspond to the block Gauss-Seidel and Jacobi algorithms to solve (11.103). Both of the above algorithms can be shown to converge geometrically at a rate independent of the mesh sizes $\{h_l\}$, given sufficiently *large overlap* amongst the subdomains, see [CA17]. Since a coarse space is not included, the rate of convergence can deteriorate with increasing number of subdomains. Using a *zero* starting guess, and employing *one* iteration of the preceding algorithms, a preconditioner can be defined to accelerate the solution of the nonsymmetric system (11.103).

11.5 Alternative Nonmatching Grid Discretization Methods

In this section, we outline alternative techniques for discretizing a self adjoint elliptic equation on a nonmatching grid. We first outline a heuristic discretization based on the Steklov-Poincaré formulation [AG2, AG, DO4, GA15] on a nonoverlapping nonmatching grid, see Fig. 11.11. We next outline a heuristic discretization based on the least square-control approach [AT, GL13, GU3]. It is applicable on overlapping or nonoverlapping grids. Next, the partition of unity approach is outlined [HU3]. It is applicable either on overlapping or nonoverlapping grids. In each case, our discussion is *heuristic* in nature, since such discretizations have not been analyzed (except for the partition of unity approach [HU3]). For alternative approaches, see [PH, DO4, CA7].

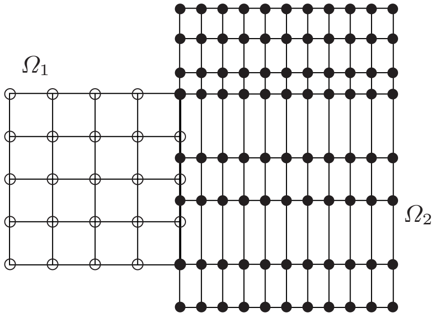


Fig. 11.11. Non-overlapping non-matching grid

11.5.1 Steklov-Poincaré Approach

A non-overlapping non-matching grid discretization of (11.100) can also be obtained using the Steklov-Poincaré approach [AG2, AG, DO4, GA15]. For simplicity, we only consider a *two* subdomain *nonoverlapping* decomposition Ω_1 and Ω_2 of Ω. An extension to multi-subdomain decompositions will be possible, but more involved. For $1 \le l \le 2$, let $\mathcal{T}_{h_l}(\Omega_l)$ denote a quasiuniform triangulation of Ω_l with grid size h_l, which does not necessarily match on $B_{12} = \partial\Omega_1 \cap \partial\Omega_2$. Then, a nonmatching grid discretization of (11.100) can be obtained by discretizing its Steklov-Poincaré formulation.

Accordingly, recall the Steklov-Poincaré formulation associated with (11.100) described in Chap. 1.3 is based on the *transmission* boundary conditions. Let $w_l(x) = u(x)$ on Ω_l. Then, we seek $w_l(x)$ on each subdomain Ω_l satisfying:

$$\begin{cases} Lw_1 = f, & \text{in } \Omega_1 \\ w_1 = w_2, & \text{on } B_{12} \\ w_1 = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} Lw_2 = f, & \text{in } \Omega_2 \\ \mathbf{n}\cdot(a\nabla w_2) = \mathbf{n}\cdot(a\nabla w_1), & \text{on } B_{12} \\ w_2 = 0, & \text{on } B_{[2]} \end{cases} \qquad (11.104)$$

where $B_{[l]} = \partial\Omega_l \cap \partial\Omega$ and $\mathbf{n}$ is the normal to Ω_2 on $B_{12} = \partial\Omega_1 \cap \partial\Omega_2$.

We shall denote the local discretization of $L\,w_1 = f$ on Ω_1 as:

$$A_{II}^{(1)}\mathbf{w}_I^{(1)} + A_{IB}^{(1)}\mathbf{w}_B^{(1)} = \mathbf{f}_I^{(1)},$$

where $\mathbf{w}_I^{(1)}$ and $\mathbf{w}_B^{(1)}$ denote vectors of nodal values in the *interior* of Ω_1 and on $B^{(1)} = B_{12} = \partial\Omega_1 \cap \Omega$, while $\mathbf{f}_I^{(1)}$ denotes the forcing vector in the *interior* of Ω_1. Similarly, we denote the discretization of $L\,w_2 = f$ on Ω_2 with flux boundary conditions $\mathbf{n}\cdot(a\nabla w_2) = g$ on B_{12} as:

$$\begin{cases} A_{II}^{(2)}\mathbf{w}_I^{(2)} + A_{IB}^{(2)}\mathbf{w}_B^{(2)} & = \mathbf{f}_I^{(2)} \\ A_{BI}^{(2)}\mathbf{w}_I^{(2)} + A_{BB}^{(2)}\mathbf{w}_B^{(2)} & = \mathbf{f}_B^{(2)} + \mathbf{g}_B, \end{cases}$$

where $\mathbf{f}_B^{(2)} + \mathbf{g}_B$ denotes the discrete flux on B_{12}.

Discretization of the *transmission* boundary conditions require care, to ensure that the global discretization is *stable*. In particular, the total number of equations must equal the total number unknowns. For finite element discretizations, we may discretize the matching condition $w_1 = w_2$ on B_{12} by applying a mortar element discretization of its *weak* form:

$$\int_{B_{12}} \psi_{12}(x)\ (w_1(x) - w_2(x))\ ds_x = 0, \quad \forall \psi_{12}(x) \in Y_h(B_{12}),$$

using a multiplier space $Y_h(B_{12}) \subset H^{-1/2}(B_{12})$, yielding:

$$M_B^{(1)}\mathbf{w}_B^{(1)} - M_B^{(2)}\mathbf{w}_B^{(2)} = \mathbf{0}.$$

If a discontinuous multiplier space $Y_h = Y_{h_{j_*}}(B_{12})$ is employed, where j_* denotes the index of the *nonmortar* side for $j_* = 1, 2$, then $M_B^{(j_*)}$ will be a diagonal matrix. This will yield a parameterization (master-slave relation):

$$\begin{cases} \mathbf{w}_B^{(1)} = R_1 \mathbf{w}_B^{(2)} = M_B^{(1)^{-1}} M_B^{(2)} \mathbf{w}_B^{(2)}, & \text{if } j_* = 1 \\ \mathbf{w}_B^{(2)} = R_2 \mathbf{w}_B^{(1)} = M_B^{(2)^{-1}} M_B^{(1)} \mathbf{w}_B^{(1)}, & \text{if } j_* = 2. \end{cases}$$

This will yield the following mappings and their adjoints:

$$\begin{cases} R_1 : X_{h_2}(B_{12}) \to X_{h_1}(B_{12}), & R_1^T : X_{h_1}(B_{12})' \to X_{h_2}(B_{12})', \text{ when } j_* = 1 \\ R_2 : X_{h_1}(B_{12}) \to X_{h_2}(B_{12}), & R_2^T : X_{h_2}(B_{12})' \to X_{h_1}(B_{12})', \text{ when } j_* = 2. \end{cases} \tag{11.105}$$

If $n_B^{(l)}$ denotes the number of interior nodes of $\mathcal{T}_{h_l}(\Omega_l)$ on $B^{(l)} = B_{12}$, this will denote the dimension of $X_{h_l}(B_{12})$ for $l = 1, 2$. In this case, matrix R_1 will be of size $n_B^{(1)} \times n_B^{(2)}$, while R_2 will be of size $n_B^{(2)} \times n_B^{(1)}$. To discretize:

$$\int_{B_{12}} \phi_{12}(x)\, \mathbf{n} \cdot (a\nabla w_1(x) - a\nabla w_2(x))\, ds_x = 0, \quad \forall \phi_{12}(x) \in Z_h(B_{12}), \tag{11.106}$$

i.e., flux matching, we may choose a subspace $Z_h(B_{12}) \subset H_{00}^{1/2}(B_{12})$.

To ensure that the total number of equations equals the number of unknowns, the sum of the dimension of $Y_h(B_{12})$ and $Z_h(B_{12})$ must equal $(n_B^{(1)} + n_B^{(2)})$. If a mortar element discretization was employed to weakly match $w_1 = w_2$ on B_{12}, for a multiplier space $Y_{h_{j_*}}(B_{12})$, dimension of $Y_h(B_{12})$ will be $n_B^{(j_*)}$. In this case, a simple choice for $Z_h(B_{12}) \subset H_{00}^{1/2}(B_{12})$ can be obtained using *adjoint* map $R_{j_*}^T$ as indicated below. To be specific, suppose $j_* = 1$. Then, $\mathbf{w}_B^{(1)} = R_1 \mathbf{w}_B^{(2)}$ will denote a parameterization (slave-master relation), and in this case the flux can be mapped using R_1^T, yielding:

$$A_{BI}^{(2)} \mathbf{w}_I^{(2)} + A_{BB}^{(2)} \mathbf{w}_B^{(2)} = -R_1^T \left(A_{BI}^{(1)} \mathbf{w}_I^{(1)} + A_{BB}^{(1)} \mathbf{w}_B^{(1)} \right) + R_1^T \mathbf{f}_B^{(1)} + \mathbf{f}_B^{(2)}.$$

Substituting that $\mathbf{w}_B^{(1)} = R_1 \mathbf{w}_B^{(2)}$, and combining the discretization of each equation in the Steklov-Poincaré system yields the following global system:

$$\begin{bmatrix} A_{II}^{(1)} & 0 & A_{IB}^{(1)} R_1 \\ 0 & A_{II}^{(2)} & A_{IB}^{(2)} \\ R_1^T A_{BI}^{(1)} & A_{BI}^{(2)} & R_1^T A_{BB}^{(1)} R_1 + A_{BB}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ R_1^T \mathbf{f}_B^{(1)} + \mathbf{f}_B^{(2)} \end{bmatrix}. \tag{11.107}$$

In the special case that $A_{BI}^{(1)} = A_{IB}^{(1)^T}$ and $A_{BI}^{(2)} = A_{IB}^{(2)^T}$, this linear system will be symmetric and positive definite. Additionally, when the grids match, it will hold that $R_1 = I$ and this system will reduce to the standard conforming discretization. General theoretical results are not known on the convergence and stability of such discretizations. This discretization, can in principle be extended to multi-subdomain non-overlapping nonmatching grids.

11.5.2 Least Squares-Control Approach

Least squares-control formulations, see [AT, GL13, GU3, GU2], have not been studied for constructing nonmatching grid discretizations. However, we outline how elliptic equation (11.100) can be discretized on a nonmatching grid using the least square-control hybrid formulation, as noted in Chap. 1.5.

Our discussion will focus on *overlapping* nonmatching grids, though it will be evident that the methodology immediately extends to *non-overlapping* nonmatching grids. Let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping decomposition of Ω with $B^{(l)} = \partial\Omega_l^* \cap \Omega$ and $B_{[l]} = \partial\Omega_l^* \cap \partial\Omega$. We shall assume that each Ω_l^* is obtained from a nonoverlapping subdomain Ω_l by extension. An equivalent least squares-control formulation of (11.100) can be obtained as follows. Let $w_l(x) = u(x)$ on Ω_l^*. Then, we seek $w_1(x), \ldots, w_p(x)$ defined on $\Omega_1^*, \ldots, \Omega_p^*$, respectively, satisfying the following constrained minimization problem:

$$J(w_1, \ldots, w_p) = \min_{(v_1,\ldots,v_p)\in\mathcal{K}} J(v_1, \ldots, v_p) \tag{11.108}$$

where we define $\mathcal{K}$ as a constraint set of local solutions to (11.100):

$$\mathcal{K} = \left\{ (v_1, \ldots, v_p) \, : \, \begin{array}{rl} L\, v_l = f, & \text{in } \Omega_l^* \\ v_l = 0, & \text{on } B_{[l]} \quad \text{for } 1 \le l \le p \\ \mathbf{n}\cdot(a\nabla v_l) = g_l, & \text{on } B^{(l)} \end{array} \right\},$$

for $g_l(x)$ denoting *unknown* fluxes parameterizing the subdomain solutions $v_l(x)$ for $1 \le l \le p$. Here, $J(v_1, \ldots, v_p)$ is a *nonnegative* functional that can be defined in many alternative ways, provided $J(\cdot)$ is minimized when the true solution is obtained on each subdomain. For instance, we may define:

$$J(v_1, \ldots, v_p) = \sum_{l=1}^{p} \sum_{j\in\mathcal{I}_*(l)} \|v_l - v_j\|^2_{L^2(B_{lj})},$$

where $B_{lj} = \partial\Omega_l \cap \partial\Omega_j$ and $\mathcal{I}_*(l)$ denotes indices of subdomains such that $B_{lj} \neq \emptyset$ has dimension $(d-1)$, as in preceding sections.

To obtain a nonmatching grid discretization of (11.108), let $\mathcal{T}_{h_l}(\Omega_l^*)$ denote a quasiuniform triangulation of Ω_l^* with grid size h_l. For simplicity, we shall assume that the restriction of each triangulation $\mathcal{T}_{h_l}(\Omega_l^*)$ to Ω_l yields a triangulation $\mathcal{T}_{h_l}(\Omega_l)$. Define a discrete functional $J_h(\cdot)$ as follows:

$$J_h(v_1, \ldots, v_p) = \sum_{l=1}^{p} \sum_{j\in\mathcal{I}_*(l)} \|v_j - P_{jl}\, v_l\|^2_{L^2(B_{lj})}, \tag{11.109}$$

where $P_{jl}v_l \in \left(X_{h_j}(B_{lj}) \cap H^{1/2}(B_{lj})\right)$ denotes an oblique projection of v_l:

$$\int_{B_{lj}} (P_{jl}v_l)\, \psi\, ds_x = \int_{B_{lj}} v_l\, \psi\, ds_x, \quad \forall \psi \in Y_{h_j}(B_{lj}),$$

where $Y_{h_j}(B_{lj})$ denotes a local nonmortar multiplier space.

The least squares-control discretization will seek $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(p)}$:

$$\mathbf{J}_h(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(p)}) = \min_{(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)}) \in \mathcal{K}_h} \mathbf{J}_h(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)}) \tag{11.110}$$

where

$$\mathcal{K}_h = \left\{ (\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)}) \, : \begin{array}{ll} A_{II}^{(l)} \mathbf{v}_I^{(l)} + A_{IB}^{(l)} \mathbf{v}_B^{(l)} & = \mathbf{f}_I^{(l)} \\ A_{IB}^{(l)^T} \mathbf{v}_I^{(l)} + A_{BB}^{(l)} \mathbf{v}_B^{(l)} & = \mathbf{g}_B^{(l)}. \end{array} \text{ for } 1 \leq l \leq p \right\},$$

with $\mathbf{g}_B^{(l)}$ denoting *unknown* fluxes parameterizing $\mathcal{K}_h$. Here:

$$\mathbf{J}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)}) = \sum_{l=1}^{p} \sum_{j \in \mathcal{I}_*(l)} \|\mathbf{v}^{(j)} - P_{jl} \mathbf{v}^{(l)}\|_{0, B_{lj}}^2,$$

denotes the differences between local solutions on common interfaces. The preceding constrained minimization problem can be reduced to a unconstrained minimization problem, using the parameterization of $\mathcal{K}_h$ in terms of the discrete fluxes $\mathbf{g}_B^{(1)}, \ldots, \mathbf{g}_B^{(p)}$. We omit further details.

11.5.3 Partition of Unity Approach

The partition of unity method [HU3] constructs an $H^1(\Omega)$ solution to an elliptic equation (11.100) using overlapping (or nonoverlapping) subdomains and a partition of unity [BA6, BA7]. To obtain a Galerkin approximation, a globally $H^1(\Omega)$-conforming finite dimensional space is constructed from local finite element spaces on the nonmatching grids, using the partition of unity. Apart from yielding a *conforming* solution, the accuracy of such discretizations do not deteriorate with decreasing overlap between the grids [HU3].

We describe the partition of unity method for overlapping subdomains. Let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping decomposition of Ω obtained by extending a nonoverlapping decomposition $\Omega_1, \ldots, \Omega_p$. We let $B_{[l]} = \partial \Omega_l^* \cap \partial \Omega$. On each subdomain Ω_l^*, let $\mathcal{T}_{h_l}(\Omega_l^*)$ denote a quasiuniform triangulation of grid size h_l and $X_{h_l}(\Omega_l^*) \subset H_{0, B_{[l]}}^1(\Omega_l^*)$ a finite element space on Ω_l^*. Let $\chi_1(x), \ldots, \chi_p(x)$ denote a *partition of unity* subordinate to the subdomains $\Omega_1^*, \ldots, \Omega_p^*$. Using the local finite element spaces and the partition of unity, define a global finite dimensional space $X_h(\Omega)$ as follows:

$$X_h(\Omega) = \left\{ \sum_{l=1}^{p} \chi_l(x) \, w_l(x) \, : \, w_l(x) \in X_{h_l}(\Omega_l^*) \right\} \subset H_0^1(\Omega). \tag{11.111}$$

A Galerkin approximation of (11.100) will seek $u_h(x) \in X_h(\Omega)$ satisfying:

$$\mathcal{A}(u_h, v_h) = (f, v_h), \quad \forall v_h \in X_h(\Omega), \tag{11.112}$$

where

$$\begin{cases} \mathcal{A}(u,v) = \int_{\Omega} \left(a(x)\nabla u \cdot \nabla v + c(x)\, u\, v\right)\, dx \\ \quad (f,v) = \int_{\Omega} f(x)\, v(x)\, dx. \end{cases}$$

Theoretical results in [HU3] show that this discretization is accurate of optimal order. Importantly, this accuracy does not deteriorate as the overlap between the subdomains decreases. However, a computational disadvantage of this approach is that explicit construction of a basis for $X_h(\Omega)$ can be difficult, since when the overlap is sufficiently large, smooth partitions of unity can be constructed for which the nodal basis on each subdomain do not form a global basis for $X_h(\Omega)$. In this case, the standard assembled stiffness matrix can be *singular*, see [HU3].

Remark 11.26. A similar approach can be employed for non-overlapping subdomains $\Omega_1, \ldots, \Omega_p$ of Ω. However, a partition of unity sub-ordinate to $\overline{\Omega}_1, \ldots, \overline{\Omega}_p$ will be *discontinuous* across $\partial\Omega_l$. Despite this, the preceding can be extended for non-overlapping subdomains [HU3].

11.6 Applications to Parabolic Equations

In time dependent problems, it can be computationally advantageous to use different *time steps* in different space-time regions [EW5, JA], with smaller time steps in regions of rapid change in the solution, and larger time steps in regions of slower change. Such choices in the time step can reduce the total computational cost, but will generally result in nonmatching space-time grids. In this section, we *heuristically* describe discretizations of a parabolic equation on *nonmatching* space-time grids.

Our discussion will focus on a parabolic equation of the form:

$$\begin{cases} u_t + Lu = f(x,t), & \text{in } \Omega \times (0,T) \\ \qquad\quad u = 0, & \text{on } \partial\Omega \times (0,T) \\ \qquad\quad u = u_0(x), & \text{in } \Omega \text{ when } t = 0, \end{cases} \tag{11.113}$$

where

$$Lu = -\nabla \cdot (a\nabla u) + cu. \tag{11.114}$$

Three alternative discretizations will be considered, one based on a Schwarz hybrid formulation (involving overlapping space-time grids), another based on a Steklov-Poincaré formulation (for nonoverlapping space-time grids) and one based on a least squares-control formulation (for overlapping or nonoverlapping space-time grids). The methods we shall consider are:

- *Schwarz hybrid formulation.* Here, the spatial domain Ω is partitioned into p overlapping subregions $\{\Omega_l^*\}_{l=1}^{p}$ with different time steps $\tau_l \equiv (T/n_l)$ and grid size h_l on each space-time subregion $\Omega_l^* \times [0,T]$. A global discretization of the parabolic equation is obtained by discretizing its Schwarz hybrid formulation on the space-time regions $\{\Omega_l^* \times (0,T)\}_{l=1}^p$.

- *Steklov-Poincaré formulation.* Here, the spatial domain is partitioned into p nonoverlapping domains $\{\Omega_l\}_{l=1}^p$ with different time steps $\tau_l = (T/n_l)$ and grid size h_l on each space-time subregion $\Omega_l \times [0,T]$. A global discretization is obtained by discretization of the Steklov-Poincaré hybrid formulation of the parabolic equation. We consider only the case $p = 2$.
- *Least squares-control formulation.* Here, the spatial domain is partitioned into overlapping or nonoverlapping subdomains, with different time steps $\tau_l = (T/n_l)$ and grid size h_l in each space-time region. A least squares-control hybrid formulation of the parabolic equation is then discretized.

11.6.1 Schwarz Formulation

Let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping decomposition of the spatial domain Ω, so that $\{\Omega_l^* \times [0,T]\}_{l=1}^p$ forms an overlapping cover of the space-time domain $\Omega \times [0,T]$. Let $\{\chi_l(x)\}_{l=1}^p$ form a spatial partition of unity on Ω subordinate to the subdomains $\{\Omega_l^*\}_{l=1}^p$. By construction, it will satisfy:

$$1 = \sum_{j \neq l} \chi_j(x) \quad \text{on } B^{(l)} = \partial\Omega_l^* \cap \Omega, \quad \text{for } 1 \leq l \leq p.$$

On each space-time region let $w_l(x,t)$ denote the restriction of the solution $u(x,t)$ to $\Omega_l^* \times [0,T]$. Then, the following coupled system of partial differential equations can be shown to be well posed and equivalent to the original parabolic equation, provided $c(x) \geq c_0 > 0$ and other assumptions hold:

$$\begin{cases} \frac{\partial w_l}{\partial t} + L\, w_l = f, & \text{on } \Omega_l^* \times (0,T), \\ \qquad\quad w_l = \sum_{j \neq l} \chi_j\, w_j, & \text{on } B^{(l)} \times (0,T), \\ \qquad\quad w_l = u_0(x), & \text{on } \Omega_l^* \text{ when } t = 0, \end{cases} \qquad \text{for } 1 \leq l \leq p. \quad (11.115)$$

We shall consider a quasiuniform spatial triangulation $\mathcal{T}_{h_l}(\Omega_l^*)$ of each spatial subdomain Ω_l^* with grid size h_l, and a time step $\tau_l = (T/n_l)$ on $\Omega_l^* \times [0,T]$. The resulting space-time grids may not match, see Fig. 11.12. A space-time discretization of the original parabolic equation (11.113) can be constructed on the nonmatching space-time grid, by discretizing the Schwarz hybrid formulation (11.115). On each local space-time region $\Omega_l^* \times (0,T)$, a stable explicit or implicit scheme can be employed to discretize $\frac{\partial w_l}{\partial t} + L\, w_l = f$. The intersubdomain matching conditions $w_l = \sum_{j \neq l} \chi_j\, w_j$ can be appropriately discretized using a local *interpolation* stencil (or a mortar element matching in the finite element case), with convex weights.

The resulting steps can be summarized as follows:

- Let $\mathbf{w}^{(l);k}$ denote the discrete solution on Ω_l^* at time $k\,\tau_l$ for $0 \leq k \leq n_l$. If a backward Euler Scheme is employed in time, and a finite difference

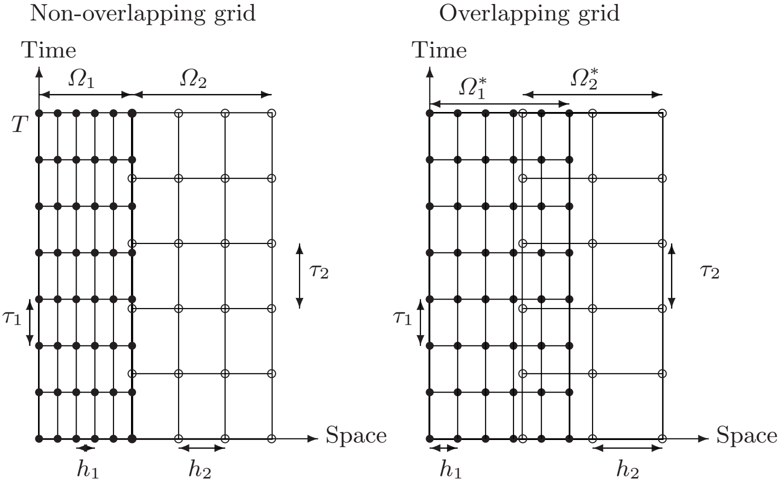


Fig. 11.12. Non-matching space-time grids

discretization in space on Ω_l^*, then the following equations will hold on each space-time region $\Omega_l^* \times (0,T)$ for $0 \le k \le (n_l - 1)$ and $1 \le l \le p$:

$$\begin{cases} (I + \tau_l A_{II}^{(l)})\mathbf{w}_I^{(l);k+1} + \tau_l\, A_{IB}^{(l)}\mathbf{w}_B^{(l);k+1} = \mathbf{w}_I^{(l);k} + \tau_l \mathbf{f}_I^{(l);k+1} \\ \qquad\qquad\qquad\qquad\qquad\quad \mathbf{w}_B^{(l);k+1} = \mathbf{g}_B^{(l);k+1}, \end{cases} \tag{11.116}$$

where $\mathbf{g}_B^{(l);k+1}$ denotes the boundary data obtained by interpolating the discrete solution on adjacent space-time grids (thus coupling the different local equations). An explicit scheme may alternatively be employed, provided the local time step τ_l satisfies stability restrictions.

- The boundary data $\mathbf{g}_B^{(l);k+1}$ on $B^{(l)} \times (0,T)$ is obtained by discretizing the matching condition $w_l(x,t) = \sum_{j \ne l} \chi_j(x)\, w_j(x,t)$ at the different discrete times $k\,\tau_l$, using *interpolation* stencils for each term $\chi_j(x) w_j(x,t)$ involving its nodal values on adjacent space-time grid points. Non-negative, convex weights are ideal. We shall denote the resulting stencil as:

$$\mathbf{w}_B^{(l);k+1} = \mathbf{g}_B^{(l);k+1} = \sum_{\tilde{k}} \sum_{j \ne l} \mathcal{I}_{\tilde{k},j}^{(l),k+1} \mathbf{w}_j^{(j);\tilde{k}}. \tag{11.117}$$

Substituting (11.117) into (11.116) couples the local discretizations.

The above coupled system of algebraic equations involving all the unknowns on each of the different space-time grids, can be expressed compactly. Let $\mathbf{W}_I^{(l)}$, $\mathbf{W}_B^{(l)}$, $\mathbf{W}_I$, $\mathbf{W}_B$ and $\mathbf{f}_I^{(l)}$ denote the following block vectors:

$$\begin{cases} \mathbf{W}_I^{(l)} &= \left(\mathbf{w}_I^{(l);1^T}, \ldots, \mathbf{w}_I^{(l);n_l^T}\right)^T \\ \mathbf{W}_B^{(l)} &= \left(\mathbf{w}_B^{(l);1^T}, \ldots, \mathbf{w}_B^{(l);n_l^T}\right)^T \\ \mathbf{W}_I &= \left(\mathbf{W}_I^{(1)^T}, \ldots, \mathbf{W}_I^{(p)^T}\right)^T \\ \mathbf{W}_B &= \left(\mathbf{W}_B^{(1)^T}, \ldots, \mathbf{W}_B^{(p)^T}\right)^T \\ \mathbf{f}_I^{(l)} &= \left(\mathbf{u}_I^{(l);0^T} + \tau_l \mathbf{f}_I^{(l);1^T}, \tau_l \mathbf{f}_I^{(l);2^T}, \ldots, \tau_l \mathbf{f}_I^{(l);n_l^T}\right)^T. \end{cases}$$

Define the following block matrices and vectors for $1 \le l \le p$:

$$G_{II}^{(l)} = \begin{bmatrix} (I + \tau_l\, A_{II}^{(l)}) & & \\ -I & \ddots & \\ & -I & (I + \tau_l\, A_{II}^{(l)}) \end{bmatrix}, \quad G_{IB}^{(l)} = \begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix},$$

and additionally define:

$$G_{II} = \begin{bmatrix} G_{II}^{(1)} & & \\ & \ddots & \\ & & G_{II}^{(p)} \end{bmatrix}, \quad G_{IB} = \begin{bmatrix} G_{IB}^{(1)} & & \\ & \ddots & \\ & & G_{IB}^{(p)} \end{bmatrix}, \quad \mathbf{f}_I = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \vdots \\ \mathbf{f}_I^{(p)} \end{bmatrix}.$$

Then, the coupled *global* discretization can be expressed *compactly* as:

$$\begin{bmatrix} G_{II} & G_{IB} \\ -\mathcal{I} & I \end{bmatrix} \begin{bmatrix} \mathbf{W}_I \\ \mathbf{W}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I \\ \mathbf{0} \end{bmatrix}, \tag{11.118}$$

where $\mathcal{I}$ denotes the matrix associated with *interpolation* stencil, which expresses the boundary value $\mathbf{W}_B = \mathcal{I}\,\mathbf{W}_I$. Importantly, the diagonal blocks of $\mathcal{I}$ will be zero (since we require that the interpolation stencil does not employ boundary nodal values from adjacent grids).

System (11.118) can be solved by a sequential or parallel Schwarz type iterative algorithm. For instance, each *iteration* of a parallel Schwarz algorithm will update the old solution as follows:

$$\begin{cases} \mathbf{W}_B^{\text{new}} = \mathcal{I}\,\mathbf{W}_I^{\text{old}} \\ \mathbf{W}_I^{\text{new}} = G_{II}^{-1}\left(\mathbf{f}_I - G_{IB}\mathbf{W}_B^{\text{new}}\right). \end{cases}$$

In practice, each local solution need only be stored on the grid points of the boundary $B^{(l)} \times \{0, \tau_l, \ldots, T\}$. The following result concerns the accuracy of the global discretization on $\Omega \times (0, T)$.

Lemma 11.27. *Suppose the following conditions hold.*

1. *Let each local scheme on $\Omega_l^* \times (0,T)$ be accurate of order $O\big(h_l^{q_l} + \tau_l^{r_l}\big)$.*
2. *Let the interpolation $\mathcal{I}^{(l)}$ stencil be accurate of order $O(\mathcal{E}_l)$.*
3. *Let $\|\mathcal{I}\|_\infty \le 1$.*
4. *Let a discrete maximum principle hold on each space-time grid.*

Then, if $U^{(l)} - W^{(l)}$ denotes the error at all grid points of the space-time grid $\mathcal{T}_{h_l}(\Omega_l^*) \times \{0, \tau_l, \ldots, T\}$, the following bound will hold:

$$\sum_{l=1}^{p} \|U^{(l)} - W^{(l)}\|_\infty \le C \left(\sum_{l=1}^{p} (h_l^{q_l} + \tau_l^{r_l}) + \sum_{l=1}^{p} \mathcal{E}_l \right),$$

where C will depend on higher order derivatives of the exact solution $u(x,t)$ and the amount of overlap, but will be independent of h_l and τ_l.

Proof. See [MA33]. □

11.6.2 Steklov-Poincaré Approach

We next outline a *heuristic* discretization of parabolic equation (11.113) on a non-overlapping non-matching space-time grid, based on a Steklov-Poincaré formulation, see [AG2, GA15, QU4], of (11.113). For simplicity, we consider only *two* nonoverlapping space-time grids. Accordingly, let $\Omega_1 \times (0,T)$ and $\Omega_2 \times (0,T)$ denote cylindrical nonoverlapping space-time regions where Ω_1 and Ω_2 form a nonoverlapping decomposition of Ω. We let $B^{(l)} = \partial\Omega_l \cap \Omega$ and $B_{[l]} = \partial\Omega_l \cap \partial\Omega$. On each spatial region Ω_l let $\mathcal{T}_{h_l}(\Omega_l)$ denote a quasiuniform grid, and let the time interval $[0,T]$ be partitioned as $\{0, \tau_l, 2\,\tau_l, \ldots, T\}$, where $\tau_l = (T/n_l)$ denotes the local time step. The two local space-time grids need not match on $B_{12} \times [0,T]$, where $B_{12} = \partial\Omega_1 \cap \partial\Omega_2$.

A discretization of the parabolic equation (11.113) can be obtained on the nonmatching space-time grid by discretizing its equivalent Steklov-Poincaré formulation. Let $w_l(x,t)$ denote the restriction of $u(x,t)$ to $\Omega_l \times [0,T]$. Then, the following coupled system will be solved by $w_1(x,t)$ and $w_2(x,t)$:

$$\left\{ \begin{array}{rll}
\frac{\partial w_1}{\partial t} + L\, w_1 = f, & \text{in} & \Omega_1 \times (0,T) \\
w_1 = w_2, & \text{on} & B_{12} \times (0,T) \\
w_1 = 0, & \text{on} & B_{[1]} \times (0,T) \\
w_1 = u_0(x), & \text{when} & t = 0 \\
\frac{\partial w_2}{\partial t} + L\, w_2 = f, & \text{in} & \Omega_2 \times (0,T) \\
\frac{\partial w_2}{\partial t} + \mathbf{n} \cdot (a \nabla w_2) = \frac{\partial w_1}{\partial t} + \mathbf{n} \cdot (a \nabla w_1), & \text{on} & B_{12} \times (0,T) \\
w_2 = 0, & \text{on} & B_{[1]} \times (0,T) \\
w_2 = u_0(x), & \text{when} & t = 0,
\end{array} \right. \qquad (11.119)$$

where $\mathbf{n}$ denotes the unit exterior normal to $B^{(2)}$.

Remark 11.28. When $w_1 = w_2$ on $B_{12} \times (0,T)$ it will also hold that $\frac{\partial w_1}{\partial t} = \frac{\partial w_2}{\partial t}$ on $B_{12} \times (0,T)$, so that the *flux* transmission condition can also be stated:

$$\mathbf{n} \cdot (a\nabla w_2) = \mathbf{n} \cdot (a\nabla w_1), \quad \text{on } B_{12} \times (0,T).$$

However, we shall retain the terms involving the time derivatives.

On each local space-time grid, the equation $\frac{\partial w_l}{\partial t} + L\,w_l = f$ can be discretized using an *explicit* or *implicit* scheme in time, and a finite element method in space. A backward Euler scheme in time and a finite element method in space will yield the following for $0 \le k \le (n_l - 1)$ and $1 \le l \le p$:

$$(M_{II}^{(l)} + \tau_l A_{II}^{(l)})\mathbf{w}_I^{(l);k+1} + (M_{IB}^{(l)} + \tau_l\, A_{IB}^{(l)})\mathbf{w}_B^{(l);k+1} = M^{(l)}\mathbf{w}^{(l);k} + \tau_l \mathbf{f}_I^{(l);k+1}, \tag{11.120}$$

where $\mathbf{w}_I^{(l);k+1}$ and $\mathbf{w}_B^{(l);k+1}$ denote the nodal unknowns on Ω_l and $B^{(l)}$ at time $(k+1)\tau_l$, respectively, and $M^{(l)} = [M_{II}^{(l)}\ M_{IB}^{(l)}]$, $\mathbf{w}^{(l);k} = [\mathbf{w}_I^{(l);k^T}\ \mathbf{w}_B^{(l);k^T}]^T$ and $\mathbf{f}_I^{(l);k+1}$ is the discrete forcing term on Ω_l at time $(k+1)\tau_l$.

The preceding local discretizations can be expressed more compactly as follows. Define block vectors for $1 \le l \le 2$:

$$\begin{cases} \mathbf{W}_I^{(l)} = \left(\mathbf{w}_I^{(l);1^T}, \ldots, \mathbf{w}_I^{(l);n_l^T}\right)^T \\ \mathbf{W}_B^{(l)} = \left(\mathbf{w}_B^{(l);1^T}, \ldots, \mathbf{w}_B^{(l);n_l^T}\right)^T \\ \mathbf{F}_I^{(l)} = \left(M^{(l)}\mathbf{w}_I^{(l);0^T} + \tau_l \mathbf{f}_I^{(l);1^T}, \tau_l \mathbf{f}_I^{(l);2^T}, \ldots, \tau_l \mathbf{f}_I^{(l);n_l^T}\right)^T. \end{cases}$$

Define the following $n_l \times n_l$ block matrices using $D_{II}^{(l)} = (M_{II}^{(l)} + \tau_l\, A_{II}^{(l)})$:

$$G_{II}^{(l)} = \begin{bmatrix} D_{II}^{(l)} & & \\ -M_{II}^{(l)} & \ddots & \\ & -M_{II}^{(l)} & D_{II}^{(l)} \end{bmatrix}, \quad G_{IB}^{(l)} = \begin{bmatrix} D_{IB}^{(l)} & & \\ -M_{IB}^{(l)} & \ddots & \\ & -M_{IB}^{(l)} & D_{IB}^{(l)} \end{bmatrix},$$

and $D_{IB}^{(l)} = (M_{IB}^{(l)} + \tau_l\, A_{IB}^{(l)})$. Then, each local discretization (11.120) is:

$$G_{II}^{(l)}\mathbf{W}_I^{(l)} + G_{IB}^{(l)}\mathbf{W}_B^{(l)} = \mathbf{F}_I^{(l)}, \quad \text{for} \quad 1 \le l \le 2, \tag{11.121}$$

where *transmission* conditions must additionally be imposed for $\mathbf{W}_B^{(l)}$. When both space-time grids match on $B_{12} \times (0,T)$, the space-time boundary data must match $\mathbf{W}_B^{(1)} = \mathbf{W}_B^{(2)}$. When both local space-time grids match, the flux transmission condition $\frac{\partial w_2}{\partial t} + \mathbf{n} \cdot (a\nabla w_2) = \frac{\partial w_1}{\partial t} + \mathbf{n} \cdot (a\nabla w_1)$ can be *heuristically* discretized on $B_{12} \times (0,T)$ using a backward Euler method as:

$$\begin{aligned} &\textstyle\sum_{l=1}^{2} \left((M_{BB}^{(l)} + \tau_l\, A_{BB}^{(l)})\mathbf{w}_B^{(l);k+1} + (M_{BI}^{(l)} + \tau_l A_{BI}^{(l)})\mathbf{w}_I^{(l);k+1}\right) \\ &\quad = \textstyle\sum_{l=1}^{p} \left(M_{BB}^{(l)}\mathbf{w}_B^{(1);k} + M_{BI}^{(l)}\mathbf{w}_I^{(1);k} + \tau_1 \mathbf{f}_B^{(1);k+1}\right), \end{aligned} \tag{11.122}$$

for $0 \leq k \leq (n_l - 1)$ with $n_1 = n_2$ and $\tau_1 = \tau_2$. This can be expressed as:

$$\sum_{l=1}^{2} \left(G_{BB}^{(l)} \mathbf{W}_B^{(l)} + G_{BI}^{(l)} \mathbf{W}_I^{(l)} \right) = \sum_{l=1}^{2} \mathbf{F}_B^{(l)}, \tag{11.123}$$

where $\mathbf{F}_B^{(l)} = \left(M_{BB}^{(l)} \mathbf{w}_B^{(l);0^T} + M_{BI}^{(l)} \mathbf{w}_I^{(l);0^T} + \tau_l \mathbf{f}_B^{(l);1^T}, \tau_l \mathbf{f}_B^{(l);2^T}, \ldots, \tau_l \mathbf{f}_B^{(l);n_l^T} \right)^T$, and the block entries of $G_{BB}^{(l)}$ and $G_{BI}^{(l)}$ are the same as for $G_{II}^{(l)}$ and $G_{IB}^{(l)}$ with B and I interchanged. By combining the local equations and the interface matching conditions yields the following global discretization:

$$\begin{bmatrix} G_{II}^{(1)} & 0 & G_{IB}^{(1)} \\ 0 & G_{II}^{(2)} & G_{IB}^{(2)} \\ G_{BI}^{(1)} & G_{BI}^{(2)} & G_{BB}^{(1)} + G_{BB}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{W}_I^{(1)} \\ \mathbf{W}_I^{(2)} \\ \mathbf{W}_B \end{bmatrix} = \begin{bmatrix} \mathbf{F}_I^{(1)} \\ \mathbf{F}_I^{(2)} \\ \mathbf{F}_B^{(1)} + \mathbf{F}_B^{(2)} \end{bmatrix},$$

in the case of *matching* space-time grids, where $\mathbf{W}_B = \mathbf{W}_B^{(1)} = \mathbf{W}_B^{(2)}$.

When the space-time grid is *nonmatching*, $\mathbf{W}_B^{(1)} \neq \mathbf{W}_B^{(2)}$. Indeed, these vectors may be of different sizes. To discretize the transmission boundary conditions, we shall assume for simplicity that the discrete solution is *piecewise linear* in time on each local space-time grid. The matching conditions $w_1 = w_2$ on $B_{12} \times (0, T)$ can be discretized using the weak form:

$$\int_0^T \int_{B_{12}} (w_1(x,t) - w_2(x,t))\psi(x,t)\, ds_x\, dt = 0, \ \forall \psi(x,t) \in Y_{h,t}(B_{12} \times [0,T]),$$

where $Y_{h,t}\,(B_{12} \times [0,T])$ denotes a *nonmortar* multiplier space defined on the triangulation $\mathcal{T}_{h_{j_*}}(B_{12}) \times \{0, \tau_{j_*}, \ldots, T\}$, where $j_* = 1, 2$ denotes the *nonmortar* side. This will yield a master-slave parameterization relationship:

$$M^{(1)} \mathbf{W}_B^{(1)} - M_B^{(2)} \mathbf{W}_B^{(2)} = \mathbf{0}.$$

Depending on the choice of nonmortar side, this will be:

$$\begin{cases} \mathbf{W}_B^{(1)} = R_1 \mathbf{W}_B^{(2)} = M^{(1)^{-1}} M^{(2)} \mathbf{W}_B^{(2)}, & \text{if } j_* = 1 \\ \mathbf{W}_B^{(2)} = R_2 \mathbf{W}_B^{(1)} = M^{(2)^{-1}} M^{(1)} \mathbf{W}_B^{(1)}, & \text{if } j_* = 2, \end{cases}$$

thus defining a map:

$$\begin{aligned} R_1 &: X_{h_2,\tau_2}(B_{12} \times [0,T]) \to X_{h_1,\tau_1}(B_{12} \times [0,T]) \text{ when } j_* = 1 \\ R_2 &: X_{h_1,\tau_1}(B_{12} \times [0,T]) \to X_{h_2,\tau_2}(B_{12} \times [0,T]), \text{ when } j_* = 2. \end{aligned}$$

The formal adjoints will then map:

$$\begin{aligned} R_1^T &: X_{h_1,\tau_1}(B_{12} \times [0,T])' \to X_{h_2,\tau_2}(B_{12} \times [0,T])', \\ R_2^T &: X_{h_2,\tau_2}(B_{12} \times [0,T])' \to X_{h_1,\tau_1}(B_{12} \times [0,T])'. \end{aligned}$$

The flux matching conditions $\frac{\partial w_1}{\partial t} + \mathbf{n} \cdot (a \nabla w_1) = \frac{\partial w_2}{\partial t} + \mathbf{n} \cdot (a \nabla w_2)$ can be *heuristically* discretized on $B_{12} \times [0, T]$ using the adjoint of the preceding maps. For instance, if $j_* = 1$, then $\mathbf{W}_B^{(1)} = R_1 \mathbf{W}_B^{(2)}$ and:

$$R_1^T G_{BI}^{(1)} \mathbf{W}_I^{(1)} + G_{BI}^{(2)} \mathbf{W}_I^{(2)} + \left(R_1^T G_{BB}^{(1)} R_1 + G_{BB}^{(2)} \right) \mathbf{W}_B^{(2)} = R_1^T \mathbf{F}_B^{(1)} + \mathbf{F}_B^{(2)}.$$

When the grids are *nonmatching*, and $j_* = 1$ is the nonmortar side, we may *heuristically* substitute the master-slave expression $\mathbf{W}_B^{(1)} = R_1 \mathbf{W}_B^{(2)}$, and the adjoint map R_1^T to map the flux on $B_{12} \times (0, T)$ from the nonmortar side to the mortar side. Combining all the equations and employing the adjoint map of R_1 will yield the global discretization:

$$\begin{bmatrix} G_{II}^{(1)} & 0 & G_{IB}^{(1)} R_1 \\ 0 & G_{II}^{(2)} & G_{IB}^{(2)} \\ R_1^T G_{BI}^{(1)} & G_{BI}^{(2)} & R_1^T G_{BB}^{(1)} R_1 + G_{BB}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{W}_I^{(1)} \\ \mathbf{W}_I^{(2)} \\ \mathbf{W}_B^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_I^{(1)} \\ \mathbf{F}_I^{(2)} \\ R_1^T \mathbf{F}_B^{(1)} + \mathbf{F}_B^{(2)} \end{bmatrix}.$$

In the case of matching grids, $R_1 = I$ and this will formally reduce to a standard discretization. Convergence results are not known for this scheme.

11.6.3 Least Squares-Control Approach

We conclude our discussion of heuristic methods for discretizing parabolic equations on nonmatching space-time grids by outlining a least squares-control approach. Such formulations have been extensively employed for heterogenous approximation of various time dependent partial differential equations [GL13] but not for discretization on nonmatching space-time grids. For simplicity, our discussion will be *heuristic* and only consider *overlapping* space-time grids.

Let $\{\Omega_l^*\}_{l=1}^p$ form an overlapping decomposition of Ω, obtained by extending a nonoverlapping decomposition $\{\Omega_l\}_{l=1}^p$ of Ω. Let $\Omega_l^* \times [0, T]$ denote overlapping space-time regions. We shall let $B^{(l)} = \partial \Omega_l^* \cap \Omega$ and $B_{[l]} = \partial \Omega_l^* \cap \partial \Omega$, for $1 \leq l \leq p$. We also let $B_{lj} = \partial \Omega_l \cap \partial \Omega_j$ and $\mathcal{I}_*(l)$ denote the indices of *nonmortar* sides. Let $w_l(x, t)$ denotes the restriction of $u(x, t)$ to $\Omega_l^* \times [0, T]$. Then, a least squares-control formulation of (11.113) will seek:

$$J(w_1, \ldots, w_p) = \min_{(v_1, \ldots, v_p) \in \mathcal{K}} J(v_1, \ldots, v_p), \tag{11.124}$$

where $J(v_1, \ldots, v_p)$ is a nonnegative functional defined by:

$$J(v_1, \ldots, v_p) \equiv \sum_{l=1}^p \sum_{j \in \mathcal{I}_*(l)} \|v_l - v_j\|_{L^2(B_{lj} \times (0,T))}^2 \tag{11.125}$$

which measures the difference between the $v_l(.)$ across all nonmortar surfaces, while the constraint set $\mathcal{K}$ is parameterized by the local Dirichlet data g_l:

$$\mathcal{K} = \left\{ (v_1, \ldots, v_p) : \begin{array}{rl} \frac{\partial v_l}{\partial t} + L v_l = f, & \text{on } \Omega_l^* \times (0,T) \\ v_l = g_l, & \text{on } B^{(l)} \times (0,T) \\ v_l = 0, & \text{on } B_{[l]} \times (0,T) \\ v(x, t=0) = u_0(x), & \text{on } \Omega_l^* \end{array} \text{ for } 1 \le l \le p \right\}. \tag{11.126}$$

The minimum value of $J(., \ldots, .)$ within $\mathcal{K}$ will be zero.

On each space-time region, we shall consider a quasiuniform triangulation $\mathcal{T}_{h_l}(\Omega_l^*)$ of Ω_l^* with grid size h_l. A time step $\tau_l = (T/n_l)$ will be employed for $[0,T]$ on $\Omega_l^* \times [0,T]$. For simplicity, we shall assume that the restriction of the quasiuniform triangulation to Ω_l yields a triangulation $\mathcal{T}_{h_l}(\Omega_l)$. On each space-time domain, the parabolic equation $\frac{\partial w_l}{\partial t} + L\, w_l = f$ can be discretized by an *explicit* or *implicit* local scheme and a finite element (or finite difference) method in space. For instance, if a backward Euler scheme is employed in time, and a finite element discretization in space on Ω_l^*, this will yield a system of algebraic equations for $0 \le k \le (n_l - 1)$ and $1 \le l \le p$:

$$\begin{cases} (M_{II}^{(l)} + \tau_l A_{II}^{(l)})\mathbf{w}_I^{(l);k+1} + (M_{IB}^{(l)} + \tau_l A_{IB}^{(l)})\mathbf{w}_B^{(l);k+1} = M^{(l)}\mathbf{w}^{(l);k} + \tau_l \mathbf{f}_I^{(l);k+1} \\ \mathbf{w}_B^{(l);k+1} = \mathbf{g}_B^{(l);k+1}, \end{cases} \tag{11.127}$$

where $\mathbf{g}_B^{(l);k+1}$ denotes the Dirichlet data parameterizing the local solution and $M^{(l)}\mathbf{w}^{(l);k} = M_{II}^{(l)}\mathbf{w}_I^{(l);k} + M_{IB}^{(l)}\mathbf{w}_B^{(l);k}$. The local discretizations can be expressed more compactly using the following block vectors for $1 \le l \le p$:

$$\begin{cases} \mathbf{W}_I^{(l)} = \left(\mathbf{w}_I^{(l);1^T}, \ldots, \mathbf{w}_I^{(l);n_l^T}\right)^T \\ \mathbf{W}_B^{(l)} = \left(\mathbf{w}_B^{(l);1^T}, \ldots, \mathbf{w}_B^{(l);n_l^T}\right)^T \\ \mathbf{g}_B^{(l)} = \left(\mathbf{g}_B^{(l);1^T}, \ldots, \mathbf{g}_B^{(l);n_l^T}\right)^T \\ \mathbf{F}_I^{(l)} = \left(M^{(l)}\mathbf{w}^{(l);0^T} + \tau_l \mathbf{f}_I^{(l);1^T}, \tau_l \mathbf{f}_I^{(l);2^T}, \ldots, \tau_l \mathbf{f}_I^{(l);n_l^T}\right)^T. \end{cases}$$

Define the $n_l \times n_l$ block matrices using $D_{II}^{(l)} = (M_{II}^{(l)} + \tau_l\, A_{II}^{(l)})$:

$$G_{II}^{(l)} = \begin{bmatrix} D_{II}^{(l)} & & \\ -M_{II}^{(l)} & \ddots & \\ & -M_{II}^{(l)} & D_{II}^{(l)} \end{bmatrix}, \quad G_{IB}^{(l)} = \begin{bmatrix} D_{IB}^{(l)} & & \\ -M_{IB}^{(l)} & \ddots & \\ & -M_{IB}^{(l)} & D_{IB}^{(l)} \end{bmatrix},$$

and $D_{BB}^{(l)} = (M_{BB}^{(l)} + \tau_l\, A_{BB}^{(l)})$. Define the following matrices and vectors:

$$G_{II} = \text{blockdiag}(G_{II}^{(1)}, \ldots, G_{II}^{(p)}) \qquad G_{IB} = \text{blockdiag}(G_{IB}^{(1)}, \ldots, G_{IB}^{(p)})$$

$$\mathbf{F}_I = \left(\mathbf{F}_I^{(1)^T}, \ldots, \mathbf{F}_I^{(p)^T}\right)^T \qquad \mathbf{g}_B = \left(\mathbf{g}_B^{(1)^T}, \ldots, \mathbf{g}_B^{(p)^T}\right)^T.$$

Let $\mathbf{V}_I = (\mathbf{v}_I^{(1)^T}, \ldots, \mathbf{v}_I^{(p)^T})^T$, $\mathbf{V}_B = (\mathbf{v}_B^{(1)^T}, \ldots, \mathbf{v}_B^{(p)^T})^T$. System (11.127) is:

$$\begin{cases} G_{II}\mathbf{V}_I + G_{IB}\mathbf{V}_B = \mathbf{F}_I \\ \qquad\qquad\quad \mathbf{V}_B = \mathbf{g}_B, \end{cases} \tag{11.128}$$

where $\mathbf{V}_I = G_{II}^{-1}\left(\mathbf{F}_I - G_{IB}\mathbf{g}_B\right)$ is parameterized by the Dirichlet data $\mathbf{g}_B$. We define the discrete constraint set $\mathcal{K}_{h,\tau}$ as follows:

$$\mathcal{K}_{h,\tau} \equiv \left\{(\mathbf{V}_I, \mathbf{V}_B) : \mathbf{V}_I = G_{II}^{-1}\left(\mathbf{F}_I - G_{IB}\mathbf{V}_B\right)\right\}.$$

Let $(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)}) \in \mathcal{K}_{h,\tau}$ where $\mathbf{v}^{(l)} = (\mathbf{v}_I^{(l)^T}, \mathbf{v}_B^{(l)^T})^T$. We shall let $\mathbf{J}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)})$ denote the discretization of $J(v_1, \ldots, v_p)$:

$$\mathbf{J}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)}) = \frac{1}{2}\sum_{l=1}^{p}\sum_{j\in\mathcal{I}_*(l)} \|R_l\mathbf{v}^{(l)} - R_{lj}\mathbf{v}^{(j)}\|^2_{L^2(B_{12}\times(0,T))}$$

where $R_l\mathbf{v}^{(l)}$ denotes the nodal restriction of $\mathbf{v}^{(l)}$ to nodes on $B^{(l)}$, while R_{lj} denotes a discretization of the *oblique* projection:

$$\int_0^T\int_{B_{lj}} \left(R_{lj}v_j - v_l\right)\,\psi(x,t)\,ds_x\,dt = 0, \quad \forall\psi \in Y_{h_j,\tau_j}(B_{lj}\times(0,T)).$$

The least squares-control approach will seek $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(p)})$ minimizing $\mathbf{J}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(p)})$, subject to constraints (11.128). We omit further details.

12

Heterogeneous Domain Decomposition Methods

A partial differential equation is considered to be of *heterogeneous type* if its classification (as an elliptic, parabolic or hyperbolic equation) changes across subregions. For instance, Tricomi's equation $u_{xx} + y\, u_{yy} = f(x, y)$, is of heterogenous type since it is *elliptic* for $y > 0$ and *hyperbolic* for $y < 0$, see [JO]. An equation will be said to have *heterogeneous character* if it can be accurately approximated by an equation of heterogeneous type.

Heterogeneous domain decomposition methods are computational techniques motivated by perturbation methods [KE5, LA5], which approximate equations of heterogeneous character by equations of heterogeneous type. In applications, equations of heterogeneous type may sometimes be solved numerically at reduced computational cost, and this motivates their use [GL13, GA15, QU5, QU3, AS2, QU4, BO8, LE7, QU6]. Applications include the approximation of the Boltzmann equation by a coupled Boltzmann and Navier-Stokes model, or the approximation of the large Reynolds number Navier-Stokes equation by a coupled Navier-Stokes and Euler model, or the approximation of the Euler equations by a coupled Euler and potential equation model. Although such heterogeneous models will be beyond the scope of these notes, we shall illustrate different methods for constructing an *elliptic-hyperbolic* approximation of an advection dominated elliptic equation.

Our discussion will be organized as follows. In § 12.1, we describe the vanishing viscosity approach of [GA15] for constructing an elliptic-hyperbolic approximation on a non-overlapping decomposition. In § 12.2, we describe an elliptic-hyperbolic approximation on overlapping subdomains, based on a Schwarz hybrid formulation. In § 12.3, we describe an elliptic-hyperbolic approximation based on the least squares-control method [GL13]. In § 12.4, we describe the χ-formulation which adaptively identifies *viscid* and *inviscid* subregions for heterogeneous approximation [BR32, CA29]. This formulation yields a *nonlinear* approximation, even for a linear problem. In § 12.5, we remark on extensions to advection dominated parabolic equations.

12.1 Steklov-Poincaré Heterogeneous Model

In this section, we shall describe a heterogeneous model of [GA15] based on the Steklov-Poincaré formulation. Although this method extends to systems such as Stokes equations, see [QU6], we shall illustrate it for an *elliptic-hyperbolic* approximation of the following advection dominated elliptic equation:

$$\begin{cases} -\epsilon\,\Delta u + \mathbf{b}(x)\cdot\nabla u + c(x)\,u = f(x), & \text{on } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\quad u = 0, & \text{on } \partial\Omega \end{cases} \tag{12.1}$$

where $0 < \epsilon \ll 1$ is a perturbation parameter representing the viscosity. We shall assume $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) \geq \beta > 0$ and that there is a subdomain Ω_1, referred to as an *inviscid* subdomain, on which the following holds:

$$\epsilon\,|\Delta\,u| \ll |\mathbf{b}(x)\cdot\nabla u + c(x)\,u|, \qquad \text{on } \Omega_1.$$

By assumption, on Ω_1 we may approximate $L\,u$ by $L_0\,u$, i.e., $L\,u \approx L_0\,u$, where:

$$L\,u = -\epsilon\Delta u + \mathbf{b}(x)\cdot\nabla u + c(x)\,u \;\text{ and }\; L_0\,u = \mathbf{b}(x)\cdot\nabla u + c(x)\,u.$$

On the complementary subdomain $\Omega_2 = \left(\Omega\backslash\overline{\Omega}_1\right)$, referred to as a *viscid* subdomain, we shall pose the original elliptic equation. To obtain a heterogeneous approximation of (12.1), the vanishing viscosity method of [GA15] employs a Steklov-Poincaré hybrid formulation of (12.1) based on Ω_1 and Ω_2, and considers its formal limit as the "viscosity" coefficient ϵ vanishes on Ω_1.

Subdomain Vanishing Viscosity Approach. Given Ω_1 and Ω_2 as above for (12.1), consider the family of elliptic equations parameterized by η:

$$\begin{cases} L^{\eta}w^{(\eta)} \equiv -\nabla\cdot\left(a_{\eta}(x)\nabla w^{(\eta)}\right) + \mathbf{b}(x)\cdot\nabla w^{(\eta)} + c(x)w^{(\eta)} = f(x), & \text{in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad w^{(\eta)} = 0, & \text{on } \partial\Omega, \end{cases} \tag{12.2}$$

where the coefficient $a_{\eta}(x)$ is defined by:

$$a_{\eta}(x) = \eta, \quad \text{on } \Omega_1 \quad \text{and} \quad a_{\eta}(x) = \epsilon, \text{ on } \Omega_2. \tag{12.3}$$

Equation (12.2) reduces to (12.1) when $\eta = \epsilon$. Define $w_l^{(\eta)}(x) \equiv w^{(\eta)}(x)$ on Ω_l for $l = 1, 2$. Then, a combination of the Dirichlet and Neumann transmission conditions yields the following hybrid formulation of (12.2), see Chap. 1:

$$\begin{cases} L^{\eta}w_1^{(\eta)} = f, & \text{in } \Omega_1 \\ \quad w_1^{(\eta)} = w_2^{(\eta)}, & \text{on } B \\ \quad w_1^{(\eta)} = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} L^{\eta}w_2^{(\eta)} = f, & \text{in } \Omega_2 \\ \mathbf{n}_2\cdot\mathbf{F}_2(w_2^{(\eta)}) = \mathbf{n}_2\cdot\mathbf{F}_1(w_1^{(\eta)}), & \text{on } B \\ w_2^{(\eta)} = 0, & \text{on } B_{[2]} \end{cases} \tag{12.4}$$

where $B = \partial\Omega_1 \cap \partial\Omega_2$ and $B_{[l]} = \partial\Omega_l \cap \partial\Omega$, while $\mathbf{n}_2(x)$ is a unit normal to $\partial\Omega_2$, with $\mathbf{F}_1(w_1^{(\eta)}) = \eta\,\nabla w_1^{(\eta)} - \frac{1}{2}\mathbf{b}\,w_1^{(\eta)}$ and $\mathbf{F}_2(w_2^{(\eta)}) = \epsilon\,\nabla w_2^{(\eta)} - \frac{1}{2}\mathbf{b}\,w_2^{(\eta)}$ denoting local *fluxes*. We shall consider the formal limit of (12.4) as $\eta \to 0$.

The *formal* limit of the Steklov-Poincaré system (12.4) will *not* be well posed as $\eta \to 0$. To see this, let w_1 and w_2 denote the *formal* limiting solutions. Then, given w_2, the limiting problem for w_1 on Ω_1 is:

$$\begin{cases} L_0 w_1 = f, & \text{in } \Omega_1 \\ w_1 = w_2, & \text{on } B \\ w_1 = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} L w_2 = f, & \text{in } \Omega_2 \\ \mathbf{n}_2 \cdot \mathbf{F}_2(w_2) = \mathbf{n}_2 \cdot \mathbf{F}_1(w_1), & \text{on } B \\ w_2 = 0, & \text{on } B_{[2]} \end{cases} \tag{12.5}$$

where $L_0 w_1 = \mathbf{b}(x) \cdot \nabla w_1 + c(x) w_1$. The partial differential equation for w_1 is of *hyperbolic* type, and will *not* be well posed since Dirichlet boundary conditions are imposed on both on *inflow* and *outflow* segments of $\partial\Omega_1$:

$$\begin{aligned} \partial\Omega_{l,in} &= \{x \in \partial\Omega_l \, : \, \mathbf{n}_l(x) \cdot \mathbf{b}(x) < 0\}, \\ \partial\Omega_{l,out} &= \{x \in \partial\Omega_l \, : \, \mathbf{n}_l(x) \cdot \mathbf{b}(x) \geq 0\}, \end{aligned}$$

where $\mathbf{n}_l(x)$ is the exterior normal to $x \in \partial\Omega_l$. Let $B_{[l,in]} = \partial\Omega \cap \partial\Omega_{l,in}$ and $B_{in}^{(l)} = \partial\Omega_{l,in} \cap B$ for $l = 1, 2$. Locally, a *well posed* hyperbolic (inviscid) problem can be obtained for w_1 on Ω_1 (given w_2) by imposing Dirichlet boundary conditions *only* on the *inflow* boundary $\partial\Omega_{1,in} = B_{in}^{(1)} \cup B_{[1,in]}$:

$$\begin{cases} L_0 w_1 = f, & \text{in } \Omega_1 \\ w_1 = w_2, & \text{on } B_{in}^{(1)} \\ w_1 = 0, & \text{on } B_{[1,in]}. \end{cases} \tag{12.6}$$

Interestingly, substituting this modification into the limiting Steklov-Poincaré system yields a *well posed* heterogeneous problem for w_1 and w_2, see [GA15]:

$$\begin{cases} L_0 w_1 = f, & \text{in } \Omega_1 \\ w_1 = w_2, & \text{on } B_{in}^{(1)} \\ w_1 = 0, & \text{on } B_{[1,in]} \end{cases} \quad \text{and} \quad \begin{cases} L w_2 = f, & \text{in } \Omega_2 \\ \mathbf{n}_2 \cdot \tilde{\mathbf{F}}_2(w_2) = \mathbf{n}_2 \cdot \tilde{\mathbf{F}}_1(w_1), & \text{on } B \\ w_2 = 0, & \text{on } B_{[2]} \end{cases} \tag{12.7}$$

where $\tilde{\mathbf{F}}_1(w_1) = -\frac{1}{2}\mathbf{b}\, w_1$ and $\tilde{\mathbf{F}}_2(w_2) = \epsilon \nabla w_2 - \frac{1}{2}\mathbf{b}\, w_2$ are the local *fluxes.* Importantly, $w_1(x)$ and $w_2(x)$ will *not* match on $B \cap \partial\Omega_{1,out}$, resulting in a *discontinuous* solution, however, the fluxes of $w_1(x)$ and $w_2(x)$ match on B.

Proposition 12.1. *Let $w^{(\eta)}$ be the solution to (12.2), where the coefficients satisfy $\left(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(\mathbf{x})\right) \geq \beta > 0$. The following will hold for $f \in L^2(\Omega)$.*

1. *There exists $(w_1, w_2) \in L^2(\Omega_1) \times H^1(\Omega_2)$ such that as $\eta \to 0$:*

$$w^{(\eta)}\Big|_{\Omega_1} \to w_1, \;\; \textit{weakly in } L^2(\Omega_1), \quad w^{(\eta)}\Big|_{\Omega_2} \to w_2, \;\; \textit{weakly in } H^1(\Omega_2).$$

2. *The weak limits $w_1(x)$ and $w_2(x)$ will satisfy heterogeneous system (12.7).*

Proof. See [GA15]. □

Discretization of (12.7). We consider a *heuristic* discretization of the heterogeneous system (12.7), by directly discretizing its component equations for $w_1(x)$ and $w_2(x)$. Since $w_1(x) \neq w_2(x)$ on $(B \cap \partial\Omega_{1,out})$ and since $\epsilon \ll 1$, care must be exercised to ensure that the global discretization is *stable* [JO2]. Let Ω be triangulated by a quasiuniform grid $\mathcal{T}_h(\Omega)$ with mesh size h. We shall decompose $B = (\partial\Omega_1 \cap \partial\Omega_2)$ as $B = B_- \cup B_+$ where:

$$B_- = \{x \in B \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0\} \;\text{ and }\; B_+ = \{x \in B \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) \geq 0\}\,.$$

Note that $B_- = B_{in}^{(1)}$. We shall assume that all nodes on B lie in $(B_- \cup B_+)$.

On subdomain Ω_1, we shall discretize the following *hyperbolic* problem:

$$\begin{cases} L_0\, w_1 = \mathbf{b}(x) \cdot \nabla w_1 + c(x)\, w_1 = f(x), & \text{in } \Omega_1 \\ \qquad\qquad\qquad\qquad\qquad\qquad w_1 = w_2, & \text{on } B_- \\ \qquad\qquad\qquad\qquad\qquad\qquad w_1 = 0, & \text{on } B_{[1,in]}, \end{cases} \tag{12.8}$$

using a *stable* scheme (such as upwind finite difference, Galerkin finite element with weakly enforced boundary conditions, or streamline-diffusion finite element [JO2]). We shall denote the resulting linear system as:

$$\begin{bmatrix} C_{II}^{(1)} & C_{IB_+}^{(1)} & C_{IB_-}^{(1)} \\ C_{B_+I}^{(1)} & C_{B_+B_+}^{(1)} & C_{B_+B_-}^{(1)} \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_{B_+}^{(1)} \\ \mathbf{w}_{B_-}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_{B_+}^{(1)} \\ \mathbf{w}_{B_-}^{(2)} \end{bmatrix}, \tag{12.9}$$

where we have block partitioned the nodal unknowns on Ω_1 as follows. We let $\mathbf{w}_I^{(1)}$, $\mathbf{w}_{B_+}^{(1)}$ and $\mathbf{w}_{B_-}^{(1)}$ denote nodal vectors corresponding to nodal values on the interior nodes in Ω_1, the nodes on B_+ and on B_-, respectively.

On Ω_2, we shall discretize the elliptic equation:

$$\begin{cases} L\, w_2 = -\epsilon\, \Delta w_2 + \mathbf{b}(x) \cdot \nabla w_2 + c(x)\, w_2 = f(x), & \text{in } \Omega_2 \\ \qquad\qquad\quad \mathbf{n}_2 \cdot \left(\epsilon\, \nabla w_2 - \frac{1}{2}\, \mathbf{b}\, w_2\right) = \beta(x), & \text{on } B \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad w_2 = 0, & \text{on } B_{[2]}. \end{cases} \tag{12.10}$$

We denote a stable discretization of elliptic equation (12.10) as:

$$\begin{bmatrix} \epsilon\, A_{II}^{(2)} + C_{II}^{(2)} & \epsilon\, A_{IB_+}^{(2)} + C_{IB_+}^{(2)} & \epsilon\, A_{IB_-}^{(2)} + C_{IB_-}^{(2)} \\ \epsilon\, A_{B_+I}^{(2)} + C_{B_+I}^{(2)} & \epsilon\, A_{B_+B_+}^{(2)} + C_{B_+B_+}^{(2)} & \epsilon\, A_{B_+B_-}^{(2)} + C_{B_+B_-}^{(2)} \\ \epsilon\, A_{B_-I}^{(2)} + C_{B_-I}^{(2)} & \epsilon\, A_{B_-B_+}^{(2)} + C_{B_-B_+}^{(2)} & \epsilon\, A_{B_-B_-}^{(2)} + C_{B_-B_-}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(2)} \\ \mathbf{w}_{B_+}^{(2)} \\ \mathbf{w}_{B_-}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(2)} \\ \tilde{\boldsymbol{\beta}}_{B_+}^{(2)} \\ \tilde{\boldsymbol{\beta}}_{B_-}^{(2)} \end{bmatrix}, \tag{12.11}$$

where the unknowns on Ω_2 were block partitioned as $\mathbf{w}_I^{(2)}$, $\mathbf{w}_{B_+}^{(2)}$ and $\mathbf{w}_{B_-}^{(2)}$, while $A_{XY}^{(2)}$ for $X,\, Y = I,\, B_-,\, B_+$ denotes submatrices of the stiffness matrix when $\mathbf{b}(x) = \mathbf{0}$ and $c(x) = 0$, and $C_{XY}^{(2)}$ denotes the discretization of the first and zeroth order terms. We let $L_{XY}^{(2)} = \epsilon\, A_{XY}^{(2)} + C_{XY}^{(2)}$ for $X,\, Y = I,\, B_+,\, B_-$.

To obtain a global discretization, we shall need a discretization of the transmission condition $\mathbf{n}_1 \cdot \left(-\frac{1}{2}\mathbf{b}(x)\, w_1\right) + \mathbf{n}_2 \cdot \left(\epsilon \nabla w_2 - \frac{1}{2}\mathbf{b}(x)\, w_2\right) = 0$ on B. We shall express this separately on B_- and B_+. Given $\mathbf{w}_I^{(1)}$, $\mathbf{w}_{B_+}^{(1)}$ and $\mathbf{w}_{B_-}^{(1)}$, let $\boldsymbol{\beta}_{B_+}^{(1)}$ and $\boldsymbol{\beta}_{B_-}^{(1)}$ denote the discrete flux $\mathbf{n}_1 \cdot \left(-\frac{1}{2}\mathbf{b}(x)\, w_1\right)$ on B_+ and B_-:

$$\begin{cases} \boldsymbol{\beta}_{B_+}^{(1)} = D^{(1)}_{B_+I}\mathbf{w}_I^{(1)} + D^{(1)}_{B_+B_+}\mathbf{w}^{(1)}_{B_+} + D^{(1)}_{B_+B_-}\mathbf{w}^{(1)}_{B_-} \\ \boldsymbol{\beta}_{B_-}^{(1)} = D^{(1)}_{B_-I}\mathbf{w}_I^{(1)} + D^{(1)}_{B_-B_+}\mathbf{w}^{(1)}_{B_-} + D^{(1)}_{B_+B_-}\mathbf{w}^{(1)}_{B_-}. \end{cases} \tag{12.12}$$

For instance $\boldsymbol{\beta}_{B_+}^{(1)} = M_{B_+B_+}\mathbf{w}^{(1)}_{B_+}$ and $\boldsymbol{\beta}_{B_-}^{(1)} = M_{B_-B_-}\mathbf{w}^{(1)}_{B_-}$, where:

$$\begin{cases} \left(M_{B_+B_+}\right)_{ij} = \int_{B_+} \left(-\mathbf{n}_1(x)\cdot\mathbf{b}(x)\right)\phi_i(x)\,\phi_j(x)\,ds_x \\ \left(M_{B_-B_-}\right)_{ij} = \int_{B_-} \left(-\mathbf{n}_1(x)\cdot\mathbf{b}(x)\right)\phi_i(x)\,\phi_j(x)\,ds_x, \end{cases}$$

where $\{\phi_i(.)\}$ is the finite element basis restricted to B. The discrete flux tranmission condition can then be expressed as:

$$\begin{cases} \quad L^{(2)}_{B_+I}\,\mathbf{w}^{(2)}_I + L^{(2)}_{B_+B_+}\,\mathbf{w}^{(2)}_{B_+} + L^{(2)}_{B_+B_-}\,\mathbf{w}^{(2)}_{B_-} \\ + D^{(1)}_{B_+I}\,\mathbf{w}^{(1)}_I + D^{(1)}_{B_+B_+}\,\mathbf{w}^{(1)}_{B_+} + D^{(1)}_{B_+B_-}\,\mathbf{w}^{(1)}_{B_-} = \mathbf{f}^{(1)}_{B_+} + \mathbf{f}^{(2)}_{B_+} \\ \quad L^{(2)}_{B_-I}\,\mathbf{w}^{(2)}_I + L^{(2)}_{B_-B_+}\,\mathbf{w}^{(2)}_{B_+} + L^{(2)}_{B_-B_-}\,\mathbf{w}^{(2)}_{B_-} \\ + D^{(1)}_{B_-I}\,\mathbf{w}^{(1)}_I + D^{(1)}_{B_-B_+}\,\mathbf{w}^{(1)}_{B_+} + D^{(1)}_{B_-B_-}\,\mathbf{w}^{(1)}_{B_-} = \mathbf{f}^{(1)}_{B_-} + \mathbf{f}^{(2)}_{B_-}. \end{cases}$$

A global discretization of *heterogeneous* problem (12.7) can now be obtained employing the preceding local discretizations.

$$\begin{cases} C^{(1)}_{II}\mathbf{w}^{(1)}_I + C^{(1)}_{IB_+}\mathbf{w}^{(1)}_{B_+} + C^{(1)}_{IB_-}\mathbf{w}^{(1)}_{B_-} = \mathbf{f}^{(1)}_I \\ C^{(1)}_{B_+I}\mathbf{w}^{(1)}_I + C^{(1)}_{B_+B_+}\mathbf{w}^{(1)}_{B_+} + C^{(1)}_{B_+B_-}\mathbf{w}^{(1)}_{B_-} = \mathbf{f}^{(1)}_{B_+} \\ \mathbf{w}^{(1)}_{B_-} = \mathbf{w}^{(2)}_{B_-} \\ L^{(2)}_{II}\,\mathbf{w}^{(2)}_I + L^{(2)}_{IB_+}\,\mathbf{w}^{(2)}_{B_+} + L^{(2)}_{IB_-}\,\mathbf{w}^{(2)}_{B_-} = \mathbf{f}^{(2)}_I \\ L^{(2)}_{B_+I}\,\mathbf{w}^{(2)}_I + L^{(2)}_{B_+B_+}\,\mathbf{w}^{(2)}_{B_+} + L^{(2)}_{B_+B_-}\,\mathbf{w}^{(2)}_{B_-} \\ +D^{(1)}_{B_+I}\,\mathbf{w}^{(1)}_I + D^{(1)}_{B_+B_+}\,\mathbf{w}^{(1)}_{B_+} + D^{(1)}_{B_+B_-}\,\mathbf{w}^{(1)}_{B_-} = \mathbf{f}^{(1)}_{B_+} + \mathbf{f}^{(2)}_{B_+} \\ L^{(2)}_{B_-I}\,\mathbf{w}^{(2)}_I + L^{(2)}_{B_-B_+}\,\mathbf{w}^{(2)}_{B_+} + L^{(2)}_{B_-B_-}\,\mathbf{w}^{(2)}_{B_-} \\ +D^{(1)}_{B_-I}\,\mathbf{w}^{(1)}_I + D^{(1)}_{B_-B_+}\,\mathbf{w}^{(1)}_{B_+} + D^{(1)}_{B_-B_-}\,\mathbf{w}^{(1)}_{B_-} = \mathbf{f}^{(1)}_{B_-} + \mathbf{f}^{(2)}_{B_-}. \end{cases} \tag{12.13}$$

The first three blocks discretize $\mathbf{b}(x)\cdot\nabla w_1 + c(x)\,w_1 = f(x)$ in $\Omega_1 \cup B_+$ with boundary conditions $w_1 = w_2$ on B_-. The fourth block corresponds to a discretization of $L\,w_2 = f$ in Ω_2. The fifth block discretizes the flux matching condition $\mathbf{n}_2 \cdot \left(\epsilon\,\nabla w_2 - \frac{1}{2}\mathbf{b}\,w_2\right) - \frac{1}{2}\,\mathbf{n}_1\cdot\mathbf{b}\,w_1 = 0$ on B_+, while the sixth block discretizes $\mathbf{n}_2 \cdot \left(\epsilon\,\nabla w_2 - \frac{1}{2}\mathbf{b}\,w_2\right) - \frac{1}{2}\,\mathbf{n}_1\cdot\mathbf{b}\,w_1 = 0$ on B_-.

Eliminating $\mathbf{w}_{B_-}^{(1)}$ using $\mathbf{w}_{B_-}^{(1)} = \mathbf{w}_{B_-}^{(2)}$, we may express the preceding system compactly using the unknowns $\mathbf{w}_I^{(1)}$, $\mathbf{w}_{B_+}^{(1)}$, $\mathbf{w}_I^{(2)}$, $\mathbf{w}_{B_+}^{(2)}$ and $\mathbf{w}_{B_-}^{(2)} = \mathbf{w}_{B_-}^{(1)}$:

$$\begin{bmatrix} C_{II}^{(1)} & C_{IB_+}^{(1)} & 0 & 0 & C_{IB_-}^{(1)} \\ C_{B_+I}^{(1)} & C_{B_+B_+}^{(1)} & 0 & 0 & C_{B_+B_-}^{(1)} \\ 0 & 0 & L_{II}^{(2)} & L_{IB_+}^{(2)} & L_{IB_-}^{(2)} \\ D_{B_+I}^{(1)} & D_{B_+B_+}^{(1)} & L_{B_+I}^{(2)} & L_{B_+B_+}^{(2)} & L_{B_+B_-} \\ D_{B_-I}^{(1)} & D_{B_-B_+}^{(1)} & L_{B_-I}^{(2)} & L_{B_-B_+}^{(2)} & L_{B_-B_-} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_{B_+}^{(1)} \\ \mathbf{w}_I^{(2)} \\ \mathbf{w}_{B_+}^{(2)} \\ \mathbf{w}_{B_-}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_{B_+}^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_{B_+} \\ \mathbf{f}_{B_-} \end{bmatrix}, \quad (12.14)$$

where $L_{XY}^{(2)} = \epsilon\, A_{XY}^{(2)} + C_{XY}^{(2)}$ and $L_{XX} = L_{XX}^{(2)} + D_{XX}^{(2)}$ for $X = B_+$ or $X = B_-$, while $\mathbf{f}_{B_+} = \mathbf{f}_{B_+}^{(1)} + \mathbf{f}_{B_+}^{(2)}$ and $\mathbf{f}_{B_-} = \mathbf{f}_{B_-}^{(1)} + \mathbf{f}_{B_-}^{(2)}$.

Using the above block system, we may construct a Schur complement system involving $\mathbf{w}_{B_-}^{(2)}\,(= \mathbf{w}_{B_-}^{(1)})$ and $\mathbf{w}_{B_+}^{(2)}$, by eliminating $\mathbf{w}_I^{(1)}$, $\mathbf{w}_{B_+}^{(1)}$ and $\mathbf{w}_I^{(2)}$ as follows. Given $\mathbf{w}_{B_-}^{(1)}$ and $\mathbf{w}_{B_+}^{(2)}$, solving the first three block equations yields:

$$\begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_{B_+}^{(1)} \\ \mathbf{w}_I^{(2)} \end{bmatrix} = \begin{bmatrix} C_{II}^{(1)} & C_{IB_+}^{(1)} & 0 \\ C_{B_+I}^{(1)} & C_{B_+B_+}^{(1)} & 0 \\ 0 & 0 & L_{II}^{(2)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_I^{(1)} - C_{IB_-}^{(1)}\mathbf{w}_{B_-}^{(1)} \\ \mathbf{f}_{B_+}^{(1)} - C_{B_+B_-}^{(1)}\mathbf{w}_{B_-}^{(1)} \\ \mathbf{f}_I^{(2)} - L_{IB_-}^{(2)}\mathbf{w}_{B_-}^{(1)} - L_{IB_+}^{(2)}\mathbf{w}_{B_+}^{(2)} \end{bmatrix}.$$

Substituting this expression into the fourth and fifth block row equations yields the reduced Schur complement system:

$$S \begin{bmatrix} \mathbf{w}_{B_+}^{(2)} \\ \mathbf{w}_{B_-}^{(2)} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_{B_+} \\ \tilde{\mathbf{f}}_{B_-} \end{bmatrix}, \quad (12.15)$$

where $S = S^{(1)} + S^{(2)}$ is a sum of local Schur complements defined by:

$$\begin{aligned} S^{(1)} &= \left(\begin{bmatrix} D_{B_+B_-}^{(1)} \\ D_{B_-B_-}^{(1)} \end{bmatrix} - \begin{bmatrix} D_{B_+I}^{(1)} & D_{B_+B_+}^{(1)} \\ D_{B_-I}^{(1)} & D_{B_-B_+}^{(1)} \end{bmatrix} \begin{bmatrix} C_{II}^{(1)} & C_{IB_+}^{(1)} \\ C_{B_+I}^{(1)} & C_{B_+B_+}^{(1)} \end{bmatrix}^{-1} \begin{bmatrix} C_{IB_-}^{(1)} \\ C_{B_+B_-}^{(1)} \end{bmatrix} \right) \begin{bmatrix} 0 \\ I \end{bmatrix}^T \\ S^{(2)} &= \begin{bmatrix} L_{B_+B_+}^{(2)} & L_{B_+B_-}^{(2)} \\ L_{B_-B_+}^{(2)} & L_{B_-B_-}^{(2)} \end{bmatrix} - \begin{bmatrix} L_{B_+I}^{(2)} L_{II}^{(2)^{-1}} L_{IB_+}^{(2)} & L_{B_+I}^{(2)} L_{II}^{(2)^{-1}} L_{IB_-}^{(2)} \\ L_{B_-I}^{(2)} L_{II}^{(2)^{-1}} L_{IB_+}^{(2)} & L_{B_-I}^{(2)} L_{II}^{(2)^{-1}} L_{IB_-}^{(2)} \end{bmatrix}, \end{aligned} \quad (12.16)$$

and where the forcing terms $\tilde{\mathbf{f}}_{B_+}$ and $\tilde{\mathbf{f}}_{B_-}$ are defined by:

$$\begin{aligned} \begin{bmatrix} \tilde{\mathbf{f}}_{B_+} \\ \tilde{\mathbf{f}}_{B_-} \end{bmatrix} &= \begin{bmatrix} \mathbf{f}_{B_+} - L_{B_+I}^{(2)} L_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \\ \mathbf{f}_{B_-} - L_{B_-I}^{(2)} L_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \end{bmatrix} \\ &\qquad - \begin{bmatrix} D_{B_+I}^{(1)} & D_{B_+B_+}^{(1)} \\ D_{B_-I}^{(1)} & D_{B_-B_+}^{(1)} \end{bmatrix} \begin{bmatrix} C_{II}^{(1)} & C_{IB_+}^{(1)} \\ C_{B_+I}^{(1)} & C_{B_+B_+}^{(1)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_{B_+}^{(1)} \end{bmatrix} \end{aligned}$$

Each local discretization should be chosen to ensure *stability* of (12.14).

Iterative Solver. Below, we describe a variant of the Dirichlet-Neumann algorithm of [GA15] to solve (12.7) using a relaxation parameter $0 < \theta < 1$.

Algorithm 12.1.1 *(Dirichlet-Neumann Algorithm for Solving (12.7))*
Let $(w_1^{(0)}, w_2^{(0)})$ be a starting iterate, and $(w_1^{(k)}, w_2^{(k)})$ the k'th iterate

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     Solve an hyperbolic equation to determine $w_1^{(k+1)}$:
$$\begin{cases} L_0\, w_1^{(k+1)} = f(x), & \text{in } \Omega_1 \\ \quad w_1^{(k+1)} = 0, & \text{on } B_{[1,in]} \\ \quad w_1^{(k+1)} = \theta\, w_1^{(k)} + (1-\theta)\, w_2^{(k)}, & \text{on } B_{-}. \end{cases}$$
3.     Solve an elliptic equation to determine $w_2^{(k+1)}$:
$$\begin{cases} L w_2^{(k+1)} = f(x), & \text{on } \Omega_2 \\ \mathbf{n}_2 \cdot \left(\epsilon\, \nabla w_2^{(k+1)} - \frac{1}{2}\mathbf{b}\, w_2^{(k+1)}\right) = -\frac{1}{2}\mathbf{n}_2 \cdot \mathbf{b}\, w_1^{(k+1)}, & \text{on } B_{+} \\ \mathbf{n}_2 \cdot \left(\epsilon\, \nabla w_2^{(k+1)} - \frac{1}{2}\mathbf{b}\, w_2^{(k+1)}\right) = -\frac{1}{2}\mathbf{n}_2 \cdot \mathbf{b}\, w_1^{(k+1)}, & \text{on } B_{-} \\ w_2^{(k+1)} = 0, & \text{on } B_{[2]}. \end{cases}$$
4. Endfor
```

In the discrete version, $\mathbf{w}_X^{(l);k}$ will denote the k'th iterate for $\mathbf{w}_X^{(l)}$.

Algorithm 12.1.2 *(Dirichlet-Neumann Matrix Algorithm for (12.14))*

```
1. For $k = 0, 1, \cdots$ until convergence do:
2.     Define $\mathbf{w}_{B_-}^{(1);k+1} = \theta\, \mathbf{w}_{B_-}^{(1);k} + (1-\theta)\, \mathbf{w}_{B_-}^{(2);k}$
3.     Solve the linear system:
$$\begin{bmatrix} C_{II}^{(1)} & C_{IB_+}^{(1)} \\ C_{B_+I}^{(1)} & C_{B_+B_+}^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1);k+1} \\ \mathbf{w}_{B_+}^{(1);k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} - C_{IB_-}^{(1)} \mathbf{w}_{B_-}^{(1);k+1} \\ \mathbf{f}_{B_+}^{(1)} - C_{B_+B_-}^{(1)} \mathbf{w}_{B_-}^{(1);k+1} \end{bmatrix}$$
4.     Update $\tilde{\mathbf{f}}_{B_+}^{(k+1)} = \mathbf{f}_{B_+} - D_{B_+I}^{(1)} \mathbf{w}_I^{(1):k+1} - D_{B_+B_+}^{(1)} \mathbf{w}_{B_+}^{(1):k+1} - D_{B_+B_-}^{(1)} \mathbf{w}_{B_-}^{(2):k+1}$
5.     Update $\tilde{\mathbf{f}}_{B_-}^{(k+1)} = \mathbf{f}_{B_-} - D_{B_-I}^{(1)} \mathbf{w}_I^{(1):k+1} - D_{B_-B_+}^{(1)} \mathbf{w}_{B_+}^{(1):k+1} - D_{B_-B_-}^{(1)} \mathbf{w}_{B_-}^{(2):k+1}$.
6.     Solve the linear system:
$$\begin{bmatrix} L_{II}^{(2)} & L_{IB_+}^{(2)} & L_{IB_-}^{(2)} \\ L_{B_+I}^{(2)} & L_{B_+B_+}^{(2)} & L_{B_+B_-}^{(2)} \\ L_{B_-I}^{(2)} & L_{B_-B_+}^{(2)} & L_{B_-B_-}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(2);k+1} \\ \mathbf{w}_{B_+}^{(2);k+1} \\ \mathbf{w}_{B_-}^{(2);k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(2)} \\ \tilde{\mathbf{f}}_{B_+}^{(k+1)} \\ \tilde{\mathbf{f}}_{B_-}^{(k+1)} \end{bmatrix}$$
7. Endfor
```

Remark 12.2. If $\big(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\big) \geq \beta > 0$ and $0 < \theta < 1$, this algorithm will converge independent of h for appropriate discretizations, see [GA15].

Remark 12.3. The Dirichlet-Neumann algorithm we have described differs from the algorithm described in [GA15, QU6] in two ways. Firstly, the flux transmission condition employed in [GA15, QU6] has the following form:

$$\mathbf{n}_2 \cdot (\epsilon \nabla w_2 - \mathbf{b}(x)\, w_2) = \mathbf{n}_2 \cdot (-\mathbf{b}(x)\, w_1) \quad \text{on} \quad B.$$

Secondly, since $w_1(x) = w_2(x)$ on B_-, this is equivalent to:

$$\mathbf{n}_2 \cdot (\epsilon \nabla w_2) = 0 \;\text{ on }\; B_- \text{ and } \mathbf{n}_2 \cdot (\epsilon \nabla w_2 - \mathbf{b}(x)\, w_2) = \mathbf{n}_2 \cdot (-\mathbf{b}(x)\, w_1) \text{ on } B_+.$$

The transmission conditions we employ are different from the above (though equivalent), but yields coercive local problems.

Remark 12.4. Generally, it is preferable to employ GMRES acceleration when solving either (12.14) or (12.15). In this case, one iteration of the preceding Dirichlet-Neumann algorithm (or its Robin-Robin generalizations) with a *zero* starting guess, can be employed to formulate a preconditioner.

Remark 12.5. If $\big(c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x)\big) \geq \beta > 0$ and $w_1(.)$ is smooth on Ω_1, a stable discretization of (12.8) can be obtained using a Galerkin method with weakly enforced inflow boundary conditions on $\partial\Omega_{1,in}$. A Galerkin method with weakly enforced boundary conditions $w_1(.) = g(.)$ on $\partial\Omega_{1,in}$ will be based on Galerkin approximation of $\mathcal{B}_1(u, v) = \mathcal{F}_1(v)$ where:

$$\begin{cases} \mathcal{B}_1(u, v) \equiv \int_{\Omega_1} (\mathbf{b}(x) \cdot \nabla u + c(x) u)\, v\, dx - \int_{\partial\Omega_{1,in}} \mathbf{n}_1(x) \cdot \mathbf{b}(x)\, u\, v\, ds_x \\ \quad \mathcal{F}_1(v) \equiv \int_{\Omega_1} f(x)\, v\, dx + \int_{\partial\Omega_{1,in}} g(x)\, v ds_x. \end{cases}$$

It can be verified that $\mathcal{B}_1(u, u) \geq \beta \, \|u\|^2_{L^2(\Omega_1} + \frac{1}{2} \int_{\partial\Omega_1} |\mathbf{n}_1(x) \cdot \mathbf{b}(x)|\, u^2\, ds_x$, thereby ensuring stability in the induced norm. Alternatively, a streamline diffusion or upwind difference discretization can be employed [JO2].

Remark 12.6. If the local solution $w_1(.)$ is smooth in Ω_1, the grid size h_1 in Ω_1 can be chosen to be larger than the grid size h_2 in the layer region Ω_2. Depending on the number of unknowns in Ω_1, it may be possible to use a direct solver. Furthermore, if the layer region Ω_2 is of width $O(\epsilon)$, it may be possible to reorder the unknowns and to employ a band solver on Ω_2.

Accuracy of the Heterogeneous Approximation. Generally, the local solutions $w_1(.)$ and $w_2(.)$ to the heterogenous system (12.7) will differ from the solution $u(.)$ to the original advection dominated elliptic equation (12.1) on the subdomains. It is important to estimate the error due to omission of the viscous term in Ω_1 and omission of the conormal derivative on B_-. For a stable discretization of (12.7), we *heuristically* indicate why the elliptic-hyperbolic approximation will introduce an error of magnitude $O(\epsilon)$, provided the subdomains are chosen carefully, and the solution is sufficiently smooth on Ω_1, and the heterogeneous problem is well posed.

Let $H\,\mathbf{w} = \mathbf{f}$ formally denote the linear system (12.14) obtained by discretizing the heterogeneous problem (12.7) (where $\mathbf{w} = \left(\mathbf{w}^{(1)^T}, \mathbf{w}^{(2)^T}\right)^T$ represents the nodal vectors associated with the heterogeneous solution on Ω_1 and Ω_2, respectively). Let $\mathbf{u}$ denote an interpolant or projection of the solution $u(x)$ of (12.1) to the nodes associated with $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$. Then, by formulation of the elliptic-hyperbolic model, it will hold that $H\,\mathbf{u} = \mathbf{f} + \boldsymbol{\mathcal{E}}$, where $\boldsymbol{\mathcal{E}}$ combines the discretization error for the heterogeneous system (12.7) and the error introduced by omission of viscous terms in Ω_1 and conormal derivatives on B. Subtracting the two equations formally yields:

$$H\,(\mathbf{u} - \mathbf{w}) = \boldsymbol{\mathcal{E}} \quad \Longrightarrow \quad \|\mathbf{u} - \mathbf{w}\| \le \|H^{-1}\|\,\|\boldsymbol{\mathcal{E}}\|.$$

The magnitude of the error can be estimated under the following assumptions.

Lemma 12.7. *Suppose the following conditions hold.*

1. *Let the heterogeneous discretization be stable, i.e.,* $\|H^{-1}\| \le c_1$ *for some* $c_1 > 0$ *independent of* h *and* ϵ.
2. *Let* $\|\boldsymbol{\mathcal{E}}\| \le c_2\,h^q + c_3\,\epsilon$ *for* $c_2 > 0$, $c_3 > 0$ *independent of* h *and* ϵ, *where* $\mathbf{u}$ *denotes the restriction of* $u(x)$ *to the grid points associated with* $\mathbf{w}$ *and* $O(h^q)$ *denotes the truncation error of the heterogeneous discretization.*

Then, the following error bound will hold:

$$\|\mathbf{u} - \mathbf{w}\| \le c_1\,(c_2\,h^q + c_3\,\epsilon)\,.$$

Proof. Follows trivially from the assumptions. □

Remark 12.8. The assumption that $\|H^{-1}\| \le c_1$ be independent of h, ϵ will depend on the well posedness of the heterogenous problem as $\epsilon \to 0$, and on the stability of the heterogeneous discretization in the norm $\|\cdot\|$. The term $\boldsymbol{\mathcal{E}}$ will be a sum of the *truncation* error for the discretization of the heterogeneous system, and a viscous term $-\epsilon\,\Delta u$ on interior nodes of Ω_1 and $\epsilon\,\mathbf{n}\cdot\nabla u$ on nodes of B. Provided $u(x)$ is sufficiently smooth on $\overline{\Omega}_1$ (which was our assumption regarding the choice of subdomain Ω_1), the latter term will be $O(\epsilon)$ in magnitude. In this case, the discrete solution to the heterogeneous model will have an accuracy comparable to the discrete solution of the original problem, provided $\epsilon \le O(h^q)$.

Application to a Model Problem. We end this section by illustrating a Steklov-Poincaré heterogeneous approximation of the following one-dimensional elliptic boundary value problem which is explicitly solvable:

$$\begin{cases} -\epsilon\,u''(x) + u'(x) = 1, & \text{for } x \in (0,1) \\ \qquad\qquad u(0) = 0, & \\ \qquad\qquad u(1) = 0. & \end{cases} \tag{12.17}$$

An explicit solution of the above boundary value problem is:

$$u(x) = x - \left(\frac{1 - e^{x/\epsilon}}{1 - e^{1/\epsilon}} \right).$$

Examination of $u(x)$ indicates a *boundary layer* of width $O(\epsilon)$ near $x = 1$. Thus, for some $0 < a < 1$ such that $|1 - a| \gg \epsilon$, we may choose $\Omega_1 = (0, a)$ as the inviscid subdomain and $\Omega_2 = (a, 1)$ as the viscid subdomain. For such a choice of a, the boundary layer will be contained within $\Omega_2 = (a, 1)$. For simplicity assume $a = 1/2$ (though a value of a depending on ϵ would be more appropriate). Since $\mathbf{b}(x) = 1$, the left boundary $x = 0$ will be an *inflow* segment for $\Omega_1 = (0, 0.5)$, while $B = 0.5$ will be an *outflow* boundary for $\Omega_1 = (0, 0.5)$. Thus $B_- = \emptyset$ and $B_+ = B$. Applying the limiting transmission conditions $\mathbf{n}_2 \cdot (\epsilon \nabla w_1 - \mathbf{b}\, w_2) = \mathbf{n}_2 \cdot (-\mathbf{b}\, w_2)$ on B_+ as in [GA15, QU6], and omitting the transmission condition $w_1 = w_2$ on B_- (since B_- is empty), we obtain the following heterogeneous Steklov-Poincaré approximation of (12.17), coupling a viscous problem on Ω_2 with an inviscid problem on Ω_1:

$$\begin{cases} w_1'(x) = 1, \; x \in (0, 0.5) \\ w_1(0) = 0, \end{cases} \quad \begin{cases} -\epsilon w_2''(x) + w_2'(x) = 1, & x \in (0.5, 1) \\ \epsilon\, w_2'(1/2) + w_2(1/2) = w_1(1/2), & \\ w_2(1) = 0. & \end{cases}$$

Since the equation for $w_1(x)$ is decoupled from $w_2(x)$, we may solve for $w_1(x)$, determine its flux on $B = 0.5$ and subsequently solve for $w_2(x)$:

$$\begin{cases} w_1(x) = x, & x \in (0, 1/2) \\ w_2(x) = x - \left(\frac{\epsilon\, e^{\frac{1}{\epsilon}} - 2\, e^{\frac{1}{2\epsilon}}}{e^{\frac{1}{\epsilon}} - 2\, e^{\frac{1}{2\epsilon}}} \right) + \left(\frac{(\epsilon - 1)\, e^{\frac{x}{\epsilon}}}{e^{\frac{1}{\epsilon}} - 2\, e^{\frac{1}{2\,\epsilon}}} \right), & x \in (1/2, 1). \end{cases}$$

Note that $w_1(\frac{1}{2}) \neq w_2(\frac{1}{2})$, i.e., the heterogeneous solution is *discontinuous* at $x = 1/2$. A different heterogeneous approximation will be obtained for the transmission condition $\mathbf{n}_2 \cdot \left(\epsilon \nabla w_1 - \frac{1}{2}\, \mathbf{b}\, w_2\right) = \mathbf{n}_2 \cdot \left(-\frac{1}{2}\, \mathbf{b}\, w_2\right)$ on B_+:

$$\begin{cases} w_1(x) = x, & x \in (0, 1/2) \\ w_2(x) = x - \left(\frac{2\,\epsilon\, e^{\frac{1}{\epsilon}} - 3\, e^{\frac{1}{2\epsilon}}}{e^{\frac{1}{\epsilon}} - 3\, e^{\frac{1}{2\epsilon}}} \right) + \left(\frac{(2\,\epsilon - 1)\, e^{\frac{x}{\epsilon}}}{e^{\frac{1}{\epsilon}} - 3\, e^{\frac{1}{2\,\epsilon}}} \right), & x \in (1/2, 1). \end{cases}$$

Again, $w_1(\frac{1}{2}) \neq w_2(\frac{1}{2})$. Despite the discontinuity across B_+, the maximum norm error will satisfy: $\|w_1 - u\|_{\infty, \Omega_1} + \|w_2 - u\|_{\infty, \Omega_2} = O(\epsilon)$ as $\epsilon \to 0$.

12.2 Schwarz Heterogeneous Models

In this section, we shall describe an *elliptic-hyperbolic* approximation of (12.1) on two *overlapping* subdomains, and based on a Schwarz hybrid formulation of (12.1), see [GL13, GA8, AS2, GA9, BR32, CA29, MA35]. We will assume that $c(x) \geq c_0 > 0$ in (12.1) and let $\Omega_1 \subset \Omega$ denote a subregion such that:

$$|\epsilon\, \Delta u| \ll |\mathbf{b}(x) \cdot \nabla u + (x)\, u|\,, \quad \text{for } x \in \Omega_1. \tag{12.18}$$

Given Ω_1, we shall assume that an overlapping decomposition Ω_1^* and Ω_2^* is chosen such that $\Omega_1^* = \Omega_1$ and $\Omega_2^* \supset (\Omega \backslash \Omega_1)$. The subdomain Ω_1^* will be referred to as the *inviscid* subdomain and Ω_2^* as the *viscid* subdomain. Given Ω_1^* and Ω_2^*, we shall employ a Schwarz hybrid formulation of (12.1) as in (12.18), and take its limit as the viscosity vanishes on Ω_1^*. This will yield an hyperbolic approximation on Ω_1^* and an elliptic approximation on Ω_2^*.

Limiting Schwarz Hybrid Formulation. Given the overlapping subdomains Ω_1^* and Ω_2^*, the advection dominated elliptic equation (12.1) will have the following equivalent Schwarz hybrid formulation. Let $u_l(x) = u(x)$ on Ω_l^* for $1 \leq l \leq 2$. Then, $u_1(x)$ and $u_2(x)$ will solve the following system for $\eta = \epsilon$:

$$\begin{cases} L_1^{(\eta)}\, u_1 = f, & \text{on } \Omega_1^* \\ \quad u_1 = u_2, & \text{on } B^{(1)} \\ \quad u_1 = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} L_2^{(\eta)}\, u_2 = f, & \text{on } \Omega_2^* \\ \quad u_2 = u_1, & \text{on } B^{(2)} \\ \quad u_2 = 0, & \text{on } B_{[2]}. \end{cases} \tag{12.19}$$

Here $B_{[l]} = \partial\Omega_l^* \cap \partial\Omega$ and $B^{(l)} = \partial\Omega_l^* \cap \Omega$, and the local elliptic operators $L_l^{(\eta)}$ are defined by:

$$\begin{cases} L_1^{(\eta)} u_1 = -\eta\, \Delta u_1 + \mathbf{b}(x) \cdot \nabla u_1 + c(x)\, u_1, & \text{on } \Omega_1^* \\ L_2^{(\eta)} u_2 = -\epsilon\, \Delta u_2 + \mathbf{b}(x) \cdot \nabla u_2 + c(x)\, u_2, & \text{on } \Omega_2^* \end{cases} \tag{12.20}$$

In the vanishing viscosity Schwarz approach, an heterogeneous approximation of (12.1)is obtained by letting the viscosity parameter $\eta \to 0$ in Ω_1^*. This will yield the following *formal* heterogeneous limiting system, where $w_l(.) \approx u_l(.)$:

$$\begin{cases} L_0\, w_1 = f, & \text{on } \Omega_1^* \\ \quad w_1 = w_2, & \text{on } B^{(1)} \\ \quad w_1 = 0, & \text{on } B_{[1]} \end{cases} \quad \text{and} \quad \begin{cases} L\, w_2 = f, & \text{on } \Omega_2^* \\ \quad w_2 = w_1, & \text{on } B^{(2)} \\ \quad w_2 = 0, & \text{on } B_{[2]}, \end{cases} \tag{12.21}$$

where $L_0 w_1 = \mathbf{b}(x) \cdot \nabla w_1 + c(x)\, w_1$ and $L w_2 = -\epsilon \Delta w_2 + \mathbf{b}(x) \cdot \nabla w_2 + c(x)\, w_2$. Unfortunately, since $L_0\, w_1 = f$ is of *hyperbolic* type, the hyperbolic problem on Ω_1^* will *not* be well posed due to the Dirichlet boundary conditions being imposed on all of $\partial\Omega_1$. However, this can be modified.

The hyperbolic problem on Ω_1^* can be made well posed *locally* by imposing Dirichlet conditions only on its *inflow* boundary segment $\partial\Omega_{1,in} \subset \partial\Omega_1$:

$$\partial\Omega_{1,in} = \{x \in \partial\Omega_1^* \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0\}\,,$$

where $\mathbf{n}_1(x)$ denotes the unit exterior normal to $x \in \partial\Omega_1$. Substituting this modification into (12.21) yields the following coupled heterogeneous system:

$$\begin{cases} L_0\, w_1 = f, & \text{on } \Omega_1^* \\ \quad w_1 = w_2, & \text{on } B^{(1)} \cap \partial\Omega_{1,in} \\ \quad w_1 = 0, & \text{on } B_{[1]} \cap \partial\Omega_{1,in} \end{cases} \text{ and } \begin{cases} L\, w_2 = f, & \text{on } \Omega_2^* \\ \quad w_2 = w_1, & \text{on } B^{(2)} \\ \quad w_2 = 0, & \text{on } B_{[2]}. \end{cases} \tag{12.22}$$

When $c(x) \geq c_0 > 0$, a discretization of this coupled system satisfying a discrete maximum principle can be shown to be *well posed*, see Chap. 15.

Discretization of the Schwarz Heterogeneous System (12.22). To obtain a finite difference discretization of (12.22), we consider a stable finite difference discretization of the Schwarz *hybrid* formulation (12.19), with $\eta = \epsilon$:

$$\begin{cases} L_{II}^{(1)}\mathbf{w}_I^{(1)} + L_{IB}^{(1)}\mathbf{w}_B^{(1)} = \mathbf{f}_I^{(1)} \\ \qquad\qquad\qquad \mathbf{w}_B^{(1)} = \mathcal{I}_1\mathbf{w}^{(2)} \end{cases} \text{ and } \begin{cases} L_{II}^{(2)}\mathbf{w}_I^{(2)} + L_{IB}^{(2)}\mathbf{w}_B^{(2)} = \mathbf{f}_I^{(2)} \\ \qquad\qquad\qquad \mathbf{w}_B^{(2)} = \mathcal{I}_2\mathbf{w}^{(1)}, \end{cases} \tag{12.23}$$

where $L_{XY}^{(1)} = \eta\, A_{XY}^{(1)} + C_{XY}^{(1)}$ and $L_{XY}^{(2)} = \epsilon\, A_{XY}^{(2)} + C_{XY}^{(2)}$ for $X,\, Y = I,\, B$, and $\mathbf{w}^{(l)} = \left(\mathbf{w}_I^{(l)^T}, \mathbf{w}_B^{(l)^T}\right)^T$. Here $\mathcal{I}_l$ denotes an interpolation defining the nodal values of $w_l(x)$ on the nodes of $B^{(l)}$, using $w_j(x)$ for $j \neq l$. Setting the viscosity $\eta = 0$ in Ω_1^* and imposing boundary conditions only on $B_{in}^{(1)} = B^{(1)} \cap \partial\Omega_{1,in}$ will yield the following discretization of (12.22).

$$\begin{cases} C_{II}^{(1)}\mathbf{w}_I^{(1)} + C_{IB_{in}}^{(1)}\mathbf{w}_{B_{in}}^{(1)} = \mathbf{f}_I^{(1)} \\ \qquad\qquad\qquad \mathbf{w}_{B_{in}}^{(1)} = \mathcal{I}_{1,in}\mathbf{w}^{(2)} \end{cases} \text{ and } \begin{cases} L_{II}^{(2)}\mathbf{w}_I^{(2)} + L_{IB}^{(2)}\mathbf{w}_B^{(2)} = \mathbf{f}_I^{(2)} \\ \qquad\qquad\qquad \mathbf{w}_B^{(2)} = \mathcal{I}_2\mathbf{w}^{(1)}, \end{cases} \tag{12.24}$$

where $\left(\epsilon\, A_{II}^{(2)} + C_{II}^{(2)}\right)\mathbf{w}_I^{(2)} + \left(\epsilon\, A_{IB}^{(2)} + C_{IB}^{(2)}\right)\mathbf{w}_B^{(2)} = \mathbf{f}_I^{(2)}$ denotes a discretization of $Lw_2 = f$ on Ω_2^* and $C_{II}^{(1)}\mathbf{w}_I^{(1)} + C_{IB_{in}}^{(1)}\mathbf{w}_{B_{in}}^{(1)} = \mathbf{f}_I^{(1)}$ denotes an upwind finite difference discretization of $L_0 w_1 = f$ on Ω_1^*.

Remark 12.9. When $c(x) \geq c_0 > 0$ and both subdomain discretizations satisfy a discrete maximum principle, and each interpolation map $\mathcal{I}_l$ has its maximum norm bounded by one, and when the subdomains have sufficient overlap, then the above discretization will be *stable* in the maximum norm, see Chap. 15.

Iterative Solvers. Under appropriate assumptions, system (12.24) can be solved using an unaccelerated sequential Schwarz algorithm. Typically, it will converge robustly provided $c(x) \geq c_0 > 0$. Below, we summarize the discrete Schwarz algorithm using $\mathbf{w}_X^{(l);k}$ to denote the k'th iterate approximating $\mathbf{w}_X^{(l)}$ for $X = I,\, B^{(l)},\, B_{in}^{(l)}$. We let $\mathbf{w}_X^{(l);0}$ denote starting iterates.

Algorithm 12.2.1 *(Multiplicative Schwarz for (12.24))*

```
Let $\mathbf{w}^{(1);0} = \left(\mathbf{w}_I^{(1);0^T}, \mathbf{w}_{B_{in}}^{(1);0^T}\right)^T$ and $\mathbf{w}^{(2);0} = \left(\mathbf{w}_I^{(2);0^T}, \mathbf{w}_B^{(2);0^T}\right)^T$
1. For $k = 0, 1, \cdots$ until convergence do:
2.     Solve the linear system:
           $\begin{cases} C_{II}^{(1)}\,\mathbf{w}_I^{(1);k+1} + C_{IB_{in}}^{(1)}\,\mathbf{w}_{B_{in}}^{(1);k+1} = \mathbf{f}_I^{(1)} \\ \mathbf{w}_{B_{in}}^{(1);k+1} = \mathcal{I}_{1,in}\mathbf{w}^{(2);k} \end{cases}$
3.     Solve the linear system:
           $\begin{cases} L_{II}^{(2)}\,\mathbf{w}_I^{(2);k+1} + L_{IB}^{(2)}\,\mathbf{w}_B^{(2);k+1} = \mathbf{f}_I^{(2)} \\ \mathbf{w}_B^{(2);k+1} = \mathcal{I}_2\mathbf{w}^{(1);k+1} \end{cases}$
4. Endfor
```

Remark 12.10. When $c(x) \geq c_0 > 0$, when both subdomain discretizations satisfy a discrete maximum principle, when each interpolation map $\mathcal{I}_l$ has maximum norm bounded by one, and when the subdomains have sufficient overlap, then it can be shown that the above iterates converge to the solution of (12.24) geometrically at a rate independent of h and ϵ. The parallel version of the Schwarz algorithm will also converge at a rate independent of h and ϵ, see Chap. 15. In practice, Krylov acceleration should be employed.

Accuracy of the Schwarz Heterogeneous Approximation. Here, we indicate heuristically the effect of omitting the viscous term $-\epsilon\Delta u$ in Ω_1^*, on the accuracy of a stable discretization of the heterogeneous system. Let $H\mathbf{w} = \mathbf{f}$ formally denote the Schwarz heterogeneous linear system (12.24). Let $\mathbf{u}$ denote the restriction of the solution $u(x)$ of (12.1) to the gridpoints associated with $\mathbf{w}$. Then, $H\,\mathbf{u} = \mathbf{f} + \mathcal{E}$, where $\mathcal{E}$ is a sum of the local truncation error of the heterogeneous Schwarz discretization and a term of the form $-\epsilon\Delta u$ restricted to the grids points associated with $\mathbf{w}$. If the Schwarz discretization is *stable*, then the following error bound can be obtained.

Lemma 12.11. *Suppose the following conditions hold.*

1. *Let the Schwarz heterogeneous discretization be stable, i.e.,* $\|H^{-1}\| \leq c_1$ *for some* $c_1 > 0$ *independent of* h *and* ϵ.
2. *Let* $\|\boldsymbol{\mathcal{E}}\| \leq c_2\,h^q + c_3\,\epsilon$ *for* $c_2 > 0$, $c_3 > 0$ *independent of* h *and* ϵ, *where* $\mathbf{u}$ *denotes the restriction of* $u(x)$ *to the grid points associated with* $\mathbf{w}$ *and* $O(h^q)$ *denotes the truncation error of the heterogeneous discretization.*

Then, the following error bound will hold:

$$\|\mathbf{u} - \mathbf{w}\| \leq c_1\,(c_2\,h^q + c_3\,\epsilon)\,.$$

Proof. Follows trivially from the assumptions. □

Remark 12.12. The assumption that $\|H^{-1}\| \leq c_1$ be independent of h, ϵ will depend on the *stability* of the heterogenous discretization (12.24) as $\epsilon \to 0$. The term $\boldsymbol{\mathcal{E}}$ will be a sum of the *truncation* error for the discretization of the heterogeneous system and a viscous term $-\epsilon\, \Delta u$ on interior nodes of Ω_1^*. Provided $u(x)$ is sufficiently smooth on $\overline{\Omega}_1^*$ (which was our assumption on the choice of subdomain Ω_1^*), the latter term will be $O(\epsilon)$ in magnitude, and the Schwarz heterogeneous approximation can be employed when $\epsilon \leq O(h^q)$.

Application to a Model Problem. Below, we illustrate how to construct a Schwarz heterogeneous approximation of a model one dimensional elliptic boundary value problem of heterogeneous character for $0 < \epsilon \ll 1$:

$$\begin{cases} -\epsilon\, u''(x) + u'(x) = 1, \;\text{ for } x \in (0,1) \\ \qquad\qquad\quad u(0) = 0, \\ \qquad\qquad\quad u(1) = 0. \end{cases}$$

An explicit solution of the above boundary value problem is:

$$u(x) = x - \left(\frac{1 - e^{x/\epsilon}}{1 - e^{1/\epsilon}} \right).$$

The exact solution $u(x)$ will have a *boundary layer* of width $O(\epsilon)$ near $x = 1$. Let $\Omega_1^* = (0, b)$ denote the inviscid subdomain and $\Omega_2^* = (a, 1)$ the viscid subdomain. The parameter $a < b$ should be chosen so that the boundary layer of width $O(\epsilon)$ is contained within $\Omega_2 = (a, 1)$. To be specific, we shall choose $a = 1/2$ and $b = 3/4$ (though the values of a and b should ideally depend on ϵ). Since $\mathbf{b}(x) = 1$, the left boundary $x = 0$ of Ω_1^* is an *inflow* segment, while boundary $B^{(1)} = 3/4$ is an *outflow* boundary, with $\partial\Omega_{1,in} = \{0\}$.
As a result, the Schwarz heterogeneous system is:

$$\begin{cases} w_1'(x) = 1, \;\; x \in (0, 3/4) \\ w_1(0) = 0, \end{cases} \quad \text{and} \quad \begin{cases} -\epsilon\, w_2''(x) + w_2'(x) = 1, & x \in (1/2, 1) \\ \qquad\quad w_2(1/2) = w_1(1/2), & \\ \qquad\quad\;\; w_2(1) = 0. & \end{cases}$$

Since $w_1(x)$ is decoupled from $w_2(x)$, we obtain the following solutions:

$$w_1(x) = x \quad \text{ and } \quad w_2(x) = x - \left(\frac{1 - e^{\frac{(2\,x-1)}{2\,\epsilon}}}{1 - e^{\frac{1}{2\,\epsilon}}} \right).$$

A *continuous* approximation $w(x)$ of $u(x)$ can be obtained as:

$$w(x) = \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x),$$

for any choice of partition of unity functions $\chi_1(x)$, $\chi_2(x)$ subordinate to $(0, 3/4)$ and $(1/2, 1)$. As $\epsilon \to 0^+$, it can be verified that $\|w - u\|_\infty = O(\epsilon)$.

12.3 Least Squares-Control Heterogeneous Models

In this section, we describe an alternative *elliptic-hyperbolic* approximation of the advection dominated elliptic equation (12.1), based on a least squares-control formulation of (12.1), see [GL13]. This general approach can be applied given an overlapping or nonoverlapping decomposition, however, for simplicity we shall only illustrate it for a *two overlapping* subdomain decomposition.

Let $\Omega_1^* \subset \Omega$ denote a *inviscid* subregion such that:

$$|\epsilon\, \Delta u| \ll |\mathbf{b}(x) \cdot \nabla u + c(x)\, u|\,, \qquad \text{on } \Omega_1^*.$$

Thus, on Ω_1^* we may omit the diffusion term to obtain an approximate solution. We shall assume that Ω_2^* is an overlapping *viscid* subregion such that Ω_1^* and Ω_2^* form an overlapping decomposition of Ω. To obtain a heterogeneous approximation of (12.1), we shall employ a least squares-control formulation of (12.1) and take a zero viscosity limit on Ω_1^*, resulting in a *hyperbolic* equation, while employing the original *elliptic* equation on Ω_2^*, see [GL13].

Least Squares-Control Heterogeneous Limit. To obtain a hybrid least squares-control formulation of (12.1) based on the overlapping subdomains Ω_1^* and Ω_2^*, let $u_l(x) = u(x)$ on Ω_l^* for $l = 1, 2$. Define $\Omega_{12}^* = \Omega_1^* \cap \Omega_2^*$. Then, $u_1(x)$ and $u_2(x)$ will trivially minimize the following square norm:

$$J(w_1, w_2) = \frac{1}{2}\, \|w_1 - w_2\|^2_{L^2(\Omega_{12}^*)}, \tag{12.25}$$

subject to the constraints $(w_1, w_2) \in \mathcal{K}_{\epsilon,\epsilon}$, where:

$$\mathcal{K}_{\eta,\epsilon} = \left\{ (w_1, w_2) \,:\, \begin{pmatrix} L_l^{(\eta,\epsilon)}\, w_l = f, & \text{on } \Omega_l^* \\ w_l = g_l, & \text{on } B^{(l)} \\ w_l = 0, & \text{on } B_{[l]} \end{pmatrix} \text{ for } 1 \le l \le 2 \right\}, \tag{12.26}$$

where $g_l(.)$ are Dirichlet data and $B_{[l]} = \partial\Omega_l^* \cap \partial\Omega$ and $B^{(l)} = \partial\Omega_l^* \cap \Omega$, with:

$$L_l^{(\eta,\epsilon)} w_l = \begin{cases} -\eta\, \Delta w_1 + \mathbf{b}(x) \cdot \nabla w_1 + c(x)\, w_1, & \text{if } l = 1 \\ -\epsilon\, \Delta w_2 + \mathbf{b}(x) \cdot \nabla w_2 + c(x)\, w_2, & \text{if } l = 2. \end{cases}$$

By construction $(u_1, u_2) \in \mathcal{K}_{\epsilon,\epsilon}$ and will satisfy:

$$J(u_1, u_2) = \min_{(w_1,w_2) \in \mathcal{K}_{\epsilon,\epsilon}} J(w_1, w_2). \tag{12.27}$$

To obtain a *heterogeneous* approximation of (12.1), $J(.,.)$ can be minimized over $\mathcal{K}_{0,\epsilon}$, the limit of $\mathcal{K}_{\eta,\epsilon}$ as $\eta \to 0^+$ on Ω_1^*. Unfortunately, the formal limit $\mathcal{K}_{0,\epsilon}$ will *not* be well defined on Ω_1^*, since $\mathbf{b}(x) \cdot \nabla w_1 + c(x)\, w_1 = f$ is *hyperbolic* and since Dirichlet boundary conditions cannot be imposed on all of $\partial\Omega_1$. However, this limiting problem can be made locally well posed by imposing Dirichlet boundary conditions only on $\partial\Omega_{1,in} = \{x \in \partial\Omega_1^* : \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0\}$.

Motivated by this observation, we define:

$$\mathcal{K}_* = \left\{ (w_1, w_2) : \begin{pmatrix} L_0\, w_1 = f, & \text{on } \Omega_1^* \\ w_1 = g_1, & \text{on } B_{in}^{(1)} \\ w_1 = 0, & \text{on } B_{[1],in} \end{pmatrix} \text{ and } \begin{pmatrix} L\, w_2 = f, & \text{on } \Omega_2^* \\ w_2 = g_2, & \text{on } B^{(2)} \\ w_2 = 0, & \text{on } B_{[2]} \end{pmatrix} \right\}, \tag{12.28}$$

where $L\,w_2 = -\epsilon \Delta w_2 + \mathbf{b} \cdot \nabla w_2 + c\, w_2$ and $L_0\, w_1 = \mathbf{b} \cdot \nabla w_1 + c\, w_1$, with $B_{in}^{(1)} = \partial \Omega_{1,in} \cap \Omega$ and $B_{[1],in} = \partial \Omega_{1,in} \cap \partial \Omega$. We may now formulate the *heterogeneous* least squares-control problem as seeking $(w_1, w_2) \in \mathcal{K}_*$:

$$J(w_1, w_2) = \min_{(v_1, v_2) \in \mathcal{K}_*} J(v_1, v_2). \tag{12.29}$$

Here $g_1(.)$ and $g_2(.)$ are *unknown* boundary data parameterizing $\mathcal{K}_*$.

Discretization of (12.29). A *heuristic* discretization of (12.29) can be obtained as follows. On subdomain Ω_1^* let:

$$\begin{cases} C_{II}^{(1)} \mathbf{v}_I^{(1)} + C_{IB_{in}}^{(1)} \mathbf{v}_{B_{in}}^{(1)} = \mathbf{f}_I^{(1)} \\ \mathbf{v}_{B_{in}}^{(1)} = \mathbf{g}_{1,in}, \end{cases} \tag{12.30}$$

denote a stable discretization of $L_0\, v_1 = f$ with $v_1 = g_1$ on $B_{in}^{(1)}$. Similarly, on subdomain Ω_2^* let:

$$\begin{cases} L_{II}^{(2)} \mathbf{v}_I^{(2)} + L_{IB}^{(2)} \mathbf{v}_B^{(2)} = \mathbf{f}_I^{(2)} \\ \mathbf{v}_B^{(2)} = \mathbf{g}_2, \end{cases} \tag{12.31}$$

denote a stable discretization of $L\, v_2 = f$ with $v_2 = g_2$ on $B^{(2)}$, where matrices $L_{XY}^{(2)} = \epsilon\, A_{XY}^{(2)} + C_{XY}^{(2)}$ for X, $Y = I$, B. If the triangulations of Ω_1^* and Ω_2^* match on $\overline{\Omega}_{12}^*$, let M denote the mass matrix on $\overline{\Omega}_{12}^*$. Then, in the discrete case the square norm $J(v_1, v_2)$ can be represented as $\mathbf{J}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$:

$$\mathbf{J}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}) = \frac{1}{2} \begin{bmatrix} R_1 \mathbf{v}^{(1)} \\ R_2 \mathbf{v}^{(2)} \end{bmatrix}^T \begin{bmatrix} M & -M \\ -M & M \end{bmatrix} \begin{bmatrix} R_1 \mathbf{v}^{(1)} \\ R_2 \mathbf{v}^{(2)} \end{bmatrix}, \tag{12.32}$$

where $\mathbf{v}^{(1)} = \left(\mathbf{v}_I^{(1)^T}, \mathbf{v}_{B_{in}}^{(1)^T} \right)^T$ and $\mathbf{v}^{(2)} = \left(\mathbf{v}_I^{(2)^T}, \mathbf{v}_B^{(2)^T} \right)^T$ are nodal vectors associated with the finite element solutions v_1 and v_2, while R_1, R_2 restrict $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}$, respectively, onto nodes on Ω_{12}^*. We then obtain:

$$\int_{\Omega_{12}^*} v_i\, v_j\, dx = \mathbf{v}^{(i)^T} \left(R_i^T M R_j \right) \mathbf{v}^{(j)}, \ \text{ for } 1 \le i, j \le 2.$$

A discretization of constrained minimization (12.28) will seek:

$$\mathbf{J}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) = \min_{(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}) \in \mathcal{K}_*^h} \mathbf{J}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}). \tag{12.33}$$

where $\mathcal{K}_*^h$ is defined by (12.30) and (12.31).

The discretized *constrained* minimization problem (12.33) can be expressed as an *unconstrained* minimization problem, provided we can parameterize $\mathcal{K}_*^h$ defined by (12.30) and (12.31) in terms of the boundary data $\mathbf{g}_1$ and $\mathbf{g}_2$:

$$\mathbf{v}^{(1)} = T_1 \mathbf{g}_1 \quad \text{and} \quad \mathbf{v}^{(1)} = T_2 \mathbf{g}_2,$$

where

$$T_1 \mathbf{g}_1 \equiv \begin{bmatrix} C_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - C_{IB_{in}}^{(1)} \mathbf{g}_1 \right) \\ \mathbf{g}_1 \end{bmatrix}, \quad T_2 \mathbf{g}_2 \equiv \begin{bmatrix} L_{II}^{(2)^{-1}} \left(\mathbf{f}_I^{(2)} - L_{IB}^{(2)} \mathbf{g}_2 \right) \\ \mathbf{g}_2 \end{bmatrix}.$$

A modified functional $\tilde{\mathbf{J}}(.,.)$ can be defined as follows:

$$\tilde{\mathbf{J}}(\mathbf{g}_1, \mathbf{g}_2) \equiv \mathbf{J}(T_1 \mathbf{g}_1, T_2 \mathbf{g}_2).$$

The minimum of $\mathbf{J}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$ in $\mathcal{K}_*^h$ can be sought by minimizing $\tilde{\mathbf{J}}(\mathbf{g}_1, \mathbf{g}_2)$. The latter is an unconstrained minimization problem. Its solution can be sought by solving the linear system corresponding to the first order critical point of $\tilde{\mathbf{J}}(\mathbf{g}_1, \mathbf{g}_2)$ relative to $\mathbf{g}_1$ and $\mathbf{g}_2$. This will yield:

$$\begin{bmatrix} E_1^T M E_1 & -E_1^T M E_2 \\ -E_2^T M E_1 & E_2^T M E_2 \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix}, \tag{12.34}$$

where:

$$E_1 = R_1 \begin{bmatrix} -C_{II}^{(1)^{-1}} C_{IB_{in}}^{(1)} \\ I \end{bmatrix} \quad \text{and} \quad E_2 = R_2 \begin{bmatrix} -L_{II}^{(2)^{-1}} L_{IB}^{(2)} \\ I \end{bmatrix}, \tag{12.35}$$

and:

$$\begin{cases} \boldsymbol{\gamma}_1 = \quad E_1^T M\, R_1 \begin{bmatrix} C_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} \\ \mathbf{0} \end{bmatrix} - E_1^T M R_2 \begin{bmatrix} L_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \\ \mathbf{0} \end{bmatrix}, \\ \boldsymbol{\gamma}_2 = -E_2^T M\, R_1 \begin{bmatrix} C_{II}^{(1)^{-1}} \mathbf{f}_I^{(1)} \\ \mathbf{0} \end{bmatrix} + E_2^T M R_2 \begin{bmatrix} L_{II}^{(2)^{-1}} \mathbf{f}_I^{(2)} \\ \mathbf{0} \end{bmatrix}. \end{cases} \tag{12.36}$$

System (12.34) will be symmetric positive definite.

Remark 12.13. More generally, instead of $J(v_1, v_2) = \frac{1}{2} \|v_1 - v_2\|_{0,\Omega_{12}^*}^2$, we may employ a similar square norm on some *carefully* chosen subset of Ω_{12}^*. For instance, if Ω_1^* and Ω_2^* were constructed by extension of nonoverlapping subdomains Ω_1 and Ω_2, then we may replace Ω_{12}^* by $B = \partial\Omega_1 \cap \partial\Omega_2$ and employ $J(v_1, v_2) = \frac{1}{2} \|v_1 - v_2\|_{0,B}^2$. In particular, if Ω_1^* and Ω_2^* are *nonoverlapping* subdomains, then Dirichlet boundary conditions on $B^{(l)}$ must be replaced by transmission condition enforcing a common flux $g(.)$ on $B = \partial\Omega_1 \cap \partial\Omega_2$.

Iterative Solvers. Under appropriate assumptions, system (12.34) will be symmetric and positive definite, and can be solved using a PCG method:

$$K_0^{-1}\begin{bmatrix} E_1^T M E_1 & -E_1^T M E_2 \\ -E_2^T M E_1 & E_2^T M E_2 \end{bmatrix}\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = K_0^{-1}\begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix},$$

where K_0 is a preconditioner. Effective preconditioners have not been studied extensively. *Heuristically,* $K_0 = \text{blockdiag}(E_1^T M E_1, E_2^T M E_2)$ may be an effective preconditioner. When $\mathbf{b}(x) = \mathbf{0}$, matrix $E_l^T M E_l$ may also be spectrally equivalent to the inverse of a scaled square-root of the Laplace-Beltrami operator on $B^{(l)}$ with zero boundary conditions. We omit further discussion.

Accuracy of the Heterogeneous Approximation. When the solution $u(x)$ to (12.1) is sufficiently smooth on Ω_1^*, the magnitude of the omitted term $-\epsilon\,\Delta u$ in Ω_1^* will be $O(\epsilon)$. Consequently, if the heterogeneous least squares-control problem is well-posed, we *heuristically* expect its solution $(w_1(x), w_2(x))$ to approximate $u(x)$ to $O(\epsilon)$. Below, we outline how to heuristically estimate the accuracy of the discrete heterogeneous least-squares control solution. Formally denote the saddle point system associated with the discretized least squares-control heterogeneous problem as $H\tilde{\mathbf{w}} = \mathbf{f}$, where:

$$\tilde{\mathbf{w}} = \left(\mathbf{w}^{(1)^T}, \mathbf{w}^{(2)^T}, \boldsymbol{\mu}^{(1)^T}, \boldsymbol{\mu}^{(2)^T}\right)^T,$$

with $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ denoting the discrete solutions on Ω_1^* and Ω_2^*, respectively, and $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ denoting Lagrange multiplier variables enforcing the discrete constraints for $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, respectively. If $\tilde{\mathbf{u}}$ denotes the nodal values of the solution $u(x)$ and the Lagrange multiplier function $p(x)$ restricted to the nodes associated with $\tilde{\mathbf{w}}$, it should hold that $H\tilde{\mathbf{u}} = \mathbf{f} + \mathcal{E}$ where $\mathcal{E}$ corresponds to a sum of local truncation error and the omitted viscous term $\epsilon\,\Delta w_1$. When $u(.)$ is smooth in Ω_1^*, it should hold that $\|\mathcal{E}\| \le (c_2\,h^q + c_3\,\epsilon)$ for c_2, $c_3 > 0$ independent of h and ϵ. If H is stable (i.e., $\|H^{-1}\| \le c_1$ where c_1 is independent of h and ϵ), then the error will satisfy the bound $\|\tilde{\mathbf{u}} - \tilde{\mathbf{w}}\| \le c_1\,\|\mathcal{E}\|$.

Application to a Model Problem. We end this section by illustrating a least squares-control heterogeneous approximation of the following model one dimensional elliptic boundary value problem:

$$\begin{cases} -\epsilon\,u''(x) + u'(x) = 1, \;\; \text{for } x \in (0,1) \\ \qquad\qquad\qquad u(0) = 0, \\ \qquad\qquad\qquad u(1) = 0. \end{cases}$$

An explicit solution of the above boundary value problem is:

$$u(x) = x - \left(\frac{1 - e^{x/\epsilon}}{1 - e^{1/\epsilon}}\right).$$

The exact solution $u(x)$ will have a *boundary layer* of width $O(\epsilon)$ near $x = 1$, and motivated by this, we shall choose $\Omega_1^* = (0, 3/4)$ as the inviscid subdomain and $\Omega_2^* = (1/2, 1)$ as the viscid subdomain. Since $\mathbf{b}(x) = 1$, it follows that $\partial\Omega_{1,in} = \{0\}$ for the hyperbolic problem on $\Omega_1^* = (0, 3/4)$.

The resulting *heterogenous* elliptic-hyperbolic model will seek (w_1, w_2):

$$\begin{cases} w_1'(x) = 1, \text{ in } (0, 3/4) \\ w_1(0) = 0, \text{ on } \partial\Omega_{1,in} \end{cases} \quad \text{and} \quad \begin{cases} -\epsilon w_2''(x) + w_2'(x) = 1, & \text{in } (1/2, 1) \\ w_2(1) = 0, & \text{on } B_{[2]} \\ w_2(1/2) = \beta_2, & \text{on } B^{(2)}. \end{cases} \tag{12.37}$$

Solving for $w_1(x)$ (which is decoupled from $w_2(x)$) yields $w_1(x) = x$.

The general solution to heterogeneous system (12.37) will thus be:

$$w_1(x) = x \quad \text{and} \quad w_2(x) = x + c_1 + c_2\, e^{x/\epsilon}, \tag{12.38}$$

where c_1 and c_2 can be determined by imposing the boundary conditions involving the *control* parameter β_2 (there is no β_1 in this problem):

$$c_1 = \left(\frac{(\frac{1}{2} - \beta_2)\, e^{\frac{1}{2\epsilon}} - 1}{1 - e^{\frac{1}{2\epsilon}}} \right) \quad \text{and} \quad c_2 = \left(\frac{(\frac{1}{2} + \beta_2)\, e^{-\frac{1}{2\epsilon}}}{1 - e^{\frac{1}{2\epsilon}}} \right).$$

The parameter β_2 must be chosen to minimize $\tilde{J}(\beta_2)$:

$$\tilde{J}(\beta_2) = \int_{1/2}^{3/4} \left(w_1(x) - w_2(x)\right)^2 \, dx.$$

Substituting for $w_1(x)$ and $w_2(x)$ in terms of $c_1(\beta_2)$ and $c_2(\beta_2)$, and minimizing $\tilde{J}(\beta_2)$ with respect to β_2 yields the following optimal value β_2^*:

$$\beta_2^* = -\frac{1}{2} \left(\frac{-8\, e^{\frac{1}{4\,\epsilon}} + (2 + 2\,\epsilon)\, e^{\frac{1}{2\,\epsilon}} - e^{\frac{1}{\epsilon}} + 6\,\epsilon}{-8\,\epsilon\, e^{\frac{3}{4\,\epsilon}} + 10\,\epsilon\, e^{\frac{1}{2\,\epsilon}} + e^{\frac{1}{\epsilon}} - 2\,\epsilon} \right).$$

Note that $\beta_2^* \to -1$ as $\epsilon \to 0$. Substituting $c_1(\beta_2^*)$, $c_2(\beta_2^*)$ into (12.38) using β_2^* above, a continuous approximation $w(x) \approx u(x)$ can be obtained given a partition of unity:

$$w(x) = \chi_1(x)\, w_1(x) + \chi_2(x)\, w_2(x).$$

It can be verified that $\|u - w\| \to 0$ as $\epsilon \to 0$, in some appropriate norm.

Remark 12.14. If $\partial\Omega_{1,in} \cap B^{(1)} = \emptyset$, i.e., $\mathbf{n}_1(x) \cdot \mathbf{b}(x) \geq 0$ on $B^{(1)}$, then the local solution $w_1(x)$ will *not* depend on β_1. In this case $\tilde{J}(g_1, g_2) = \tilde{J}(g_2)$, and the bulk of the computation will involve determining $w_2(x)$.

12.4 χ-Formulation

In the preceding sections, it was assumed that subdomains Ω_1 (or Ω_1^*) and Ω_2 (or Ω_1^*) were known in advance for (12.1) such that:

$$|\epsilon \Delta u| \ll |\mathbf{b}(x) \cdot \nabla u + c(x)\, u| \qquad \text{on } \Omega_1 \;\; (\text{or } \Omega_1^*).$$

When such regions are not known in advance, they may be determined adaptively. The χ-formulation ("chi"-formulation) of [BR32, CA29] provides a framework for adaptively determining an *inviscid* region Ω_1 and a *viscid* region Ω_2, and for building an *heterogeneous* model approximating a viscous problem. Unlike the Steklov-Poincaré vanishing viscosity approach [GA15], the χ-formulation yields a continuous solution, but the resulting χ-equation is *nonlinear* even when the original problem is linear.

In this section, we shall illustrate the χ-formulation for approximating the model advection-diffusion equation:

$$\begin{cases} -\epsilon\, \Delta u + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\quad\;\; u = 0, & \text{on } \partial\Omega. \end{cases} \tag{12.39}$$

Our discussion will be organized as follows. We shall first describe a *nonlinear* heterogeneous χ-approximation of (12.39). We shall reformulate the χ-equation as an hyperbolic-elliptic system on two *nonoverlapping* subdomains. Following that, we shall describe an *overlapping* subdomain based hyperbolic-elliptic system. Both formulations yield iterative algorithms. We then remark on discretizations and linearizations of the χ-formulation, and the error associated with a χ-approximation.

The χ-Formulation of (12.39). The χ-formulation constructs an *heterogeneous* approximation of (12.39) by replacing the diffusion term Δu by a *nonlinear* term $\chi\,(\Delta u)$ which vanishes whenever $|\Delta\, u| \leq \delta_1$. More precisely, the nonlinear function $\chi(\cdot)$ is the following user chosen *odd* and *monotone* increasing scalar function, see Fig. 12.1, satisfying:

$$\begin{cases} \chi(s) &= u, & \text{for } \delta_2 \leq s \\ \chi(s) &= \frac{\delta_2}{\delta_2 - \delta_1}\,(s - \delta_1)\,, & \text{for } s \in [\delta_1, \delta_2] \\ \chi(s) &= 0, & \text{for } 0 \leq s \leq \delta_1, \\ \chi(-s) &= -\chi(s), & \qquad 0 \leq s, \end{cases}$$

where $0 < \delta_1 < \delta_2 < 1$ are user chosen parameters (that may depend on the viscous coefficient ϵ). Alternatively, any smooth monotone increasing function having a similar profile can be employed. By construction, $\chi(\Delta u)$ vanishes when $|\Delta u| \leq \delta_1$ and $\chi(\Delta u) = \Delta u$ when $|\Delta u| \geq \delta_2$.

The χ-formulation seeks an approximation $w(x)$ of $u(x)$ by solving the following *degenerate* nonlinear elliptic equation for $w(x)$:

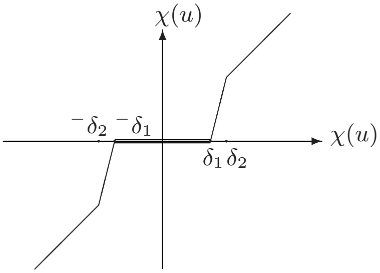


Fig. 12.1. Profile of a $\chi(\cdot)$ function

$$\begin{cases} L_\chi(w) \equiv -\epsilon\,\chi(\Delta w) + \mathbf{b}(x)\cdot\nabla w + c(x)\,w = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad w = 0, & \text{on } \partial\Omega_\chi. \end{cases} \tag{12.40}$$

Here $\partial\Omega_\chi \subset \partial\Omega$ denotes a boundary segment on which Dirichlet boundary conditions can be enforced, depending on $w(x)$. For instance, if $\chi(\Delta w(x)) \neq 0$ on $\partial\Omega$, then $\partial\Omega_\chi = \partial\Omega$, while $\partial\Omega_\chi = \{x : \mathbf{n}(x)\cdot\mathbf{b}(x) < 0\}$ if $\chi(\Delta w(x)) = 0$ on $\partial\Omega$, where $\mathbf{n}(x)$ denotes the unit exterior normal at $x \in \partial\Omega$. More generally:

$$\partial\Omega_\chi = \{x \in \partial\Omega \,:\, \chi(\Delta w(x)) \neq 0 \,\text{or}\, \chi(\Delta w(x)) = 0 \text{ and } \mathbf{n}(x)\cdot\mathbf{b}(x) < 0\}\,.$$

When $\chi(.)$ is monotone increasing, i.e., $\chi'(\cdot) \geq 0$, equation (12.40) will be *well posed*, see [BR32, CA29]. By choice of $\chi(.)$, the term $\chi(\Delta w(x)) = 0$ wherever $|\Delta w(x)| \leq \delta_1$ so that $L_\chi(w)$ becomes hyperbolic:

$$L_\chi(w) = L_0\,w = \mathbf{b}(x)\nabla w + c(x)\,w = f(x), \qquad \text{when} \qquad |\Delta w(x)| \leq \delta_1.$$

On the other hand, $L_\chi(\cdot)$ will be uniformly elliptic when $|\Delta w(x)| \geq \delta_2$, since:

$$L_\chi(w) = -\epsilon\,\Delta w + \mathbf{b}(x)\nabla w + c(x)\,w = f(x), \qquad \text{when} \qquad |\Delta w(x)| \geq \delta_2.$$

Thus, even though the original problem (12.39) is linear, the χ-equation (12.40) is *nonlinear*. Consequently, the χ-formulation will be more expensive to solve than the original linear problem, unless additional structure of the χ-equation is used to advantage. In accordance, equation (12.40) can be reformulated as a coupled *heterogeneous* system, either based on two *nonoverlapping* subdomains Ω_1 and Ω_2 or based on two *overlapping* subdomains Ω_1^* and Ω_2^*.

Remark 12.15. Well posedness of (12.40) will be guaranteed under smoothness assumptions on the coefficients $\mathbf{b}(x)$, $c(x)$, $f(x)$ and $\partial\Omega$, see [BR32, CA29]. We shall require that $c(x) \geq c_0 > 0$ and $c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x) \geq c_0 > 0$ and that there exists a sufficiently smooth function $\phi_1(x)$ such that:

$$-\Delta\phi_1(x) \,\geq\, 0, \quad \text{in } \overline{\Omega} \quad \text{and} \quad \mathbf{b}(x)\cdot\nabla\phi_1(x) \,\geq\, \alpha_0 > 0, \quad \text{in } \overline{\Omega}.$$

Under such smoothness assumptions, the χ-equation (12.40) will have a solution $w(x) \in C^1(\overline{\Omega})$. In particular, $w(\cdot)$ will be *continuous*, see [BR32, CA29].

Nonoverlapping Heterogeneous Formulation of (12.40). To obtain a heterogeneous formulation of (12.40), let Ω_1 denote an *open* subregion:

$$\Omega_1 \subset \{ x : |\chi(\Delta w(x))| = 0 \},$$

with the property that $\chi(\Delta w(x)) = 0$ on $B^{(1)} = \partial\Omega_1 \cap \Omega$. Choose Ω_2 as its complementary subregion so that Ω_1 and Ω_2 form a nonoverlapping decomposition of Ω. Let $w_l(x) = w(x)$ on Ω_l denote the restriction of $w(x)$ to Ω_l, and $B = \partial\Omega_1 \cap \partial\Omega_2$. Define $\partial\Omega_{1,in}$ and $\partial\Omega_{1,out}$ as follows:

$$\begin{aligned} \partial\Omega_{1,in} &= \{x \in \partial\Omega_1 \,:\, \mathbf{n}_1(x)\cdot\mathbf{b}(x) < 0\} \\ \partial\Omega_{1,out} &= \{x \in \partial\Omega_1 \,:\, \mathbf{n}_1(x)\cdot\mathbf{b}(x) > 0\}, \end{aligned} \tag{12.41}$$

where $\mathbf{n}_1(x)$ denotes the unit exterior normal to $x \in \partial\Omega_1$. We shall assume that $\partial\Omega_1 = \overline{\partial\Omega}_{1,in} \cup \overline{\partial\Omega}_{1,out}$.

A *heterogeneous* reformulation of the χ-equation (12.40) can be obtained analogous to the Steklov-Poincaré heterogeneous formulation of [GA15, QU6] based on the two nonoverlapping subdomains Ω_1 and Ω_2.

- By choice of Ω_1, the local solution $w_1(x) = w(x)$ on Ω_1 will satisfy:

  $$\mathbf{b}(x)\cdot\nabla w_1 + c(x)\, w_1 = f(x), \quad \text{in } \Omega_1.$$

  since $\chi(\Delta w_1) = 0$ in Ω_1. Additionally, continuity of $w(\cdot)$ for sufficiently smooth $\mathbf{b}(.)$, $c(.)$, $f(.)$ and boundary $\partial\Omega$, requires $w_1(x)$ and $w_2(x)$ to match on $B = \partial\Omega_1 \cap \partial\Omega_2$, see[BR32, CA29]:

  $$w_1(x) = w_2(x) \quad \text{on } B.$$

  As a result, *inflow* boundary conditions can be posed for w_1 locally:

  $$w_1 = w_2, \quad \text{on } B \cap \partial\Omega_{1,in} \quad \text{and} \quad w_1 = 0, \quad \text{on } \partial\Omega \cap \partial\Omega_{1,in}.$$

- On Ω_2, the component $w_2(x) = w(x)$ will satisfy:

  $$-\epsilon\, \chi(\Delta w_2) + \mathbf{b}(x)\cdot\nabla w_2 + c(x) w_2 = f(x), \quad \text{in } \Omega_2.$$

  Since continuity of $w(.)$ was not enforced on $\partial\Omega_{1,out}$, we require:

  $$w_2(x) = w_1(x), \quad \text{on} \quad B \cap \partial\Omega_{1,out}.$$

  By selection of Ω_1, it will hold that $\chi(\Delta w) = 0$ on $B = B^{(1)}$, so that $\mathbf{b}(x)\cdot\nabla w_2 + c(x)\, w_2 = f(x)$ on B. It will be sufficient to require:

  $$\mathbf{b}(x)\cdot\nabla w_2(x) + c(x)\, w_2(x) = f(x) \quad \text{on } B \cap \partial\Omega_{1,in}.$$

  The above first order boundary condition on $B \cap \partial\Omega_{1,in}$ is referred to as an "oblique derivative" boundary condition. Additional boundary conditions for $w_2(.)$ must be enforced on $\partial\Omega_2 \cap \partial\Omega$, as in (12.40).

The preceding *heuristic* observations suggest the following *nonoverlapping* subdomain based equivalent reformulation of the χ-equation (12.40).

Lemma 12.16. *Suppose the following conditions hold.*

1. *Let $w(x)$ be a solution of:*

$$\begin{cases} -\epsilon\,\chi(\Delta w) + \mathbf{b}(x)\cdot\nabla w + c(x)\,w = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad w = 0, & \text{on } \partial\Omega_\chi. \end{cases}$$

2. *Let Ω_1 and Ω_2 form a nonoverlapping subdomain decomposition with $\chi(\Delta w) = 0$ on $\Omega_1 \cup B$, with $w_i(x) = w(x)$ on Ω_i for $1 \le i \le 2$.*

Then $(w_1(x), w_2(x))$ will solve:

$$\begin{cases} \mathbf{b}(x)\cdot\nabla w_1 + c(x)\,w_1 = f(x), & \text{in } \Omega_1 \\ w_1 = w_2, & \text{on } B \cap \partial\Omega_{1,in} \\ w_1 = 0, & \text{on } \partial\Omega_1 \cap \partial\Omega_{1,in} \\ -\epsilon\,\chi(\Delta w_2) + \mathbf{b}(x)\cdot\nabla w_2 + c(x)\,w_2 = f(x), & \text{in } \Omega_2 \\ \mathbf{b}(x)\cdot\nabla w_2 + c(x)\,w_2 = f(x), & \text{on } B \cap \partial\Omega_{1,in} \\ w_2 = w_1, & \text{on } B \cap \partial\Omega_{1,out} \\ w_2 = 0, & \text{on } \partial\Omega_2 \cap \partial\Omega_\chi. \end{cases} \tag{12.42}$$

Proof. See [BR32, CA29]. □

In the preceding formulation, Ω_1 and Ω_2 were assumed to be given. In applications, given $w^k(x) \approx w(x)$, Ω_1 can be estimated using $\chi(\Delta w^k)$, see Chap. 12.4.1. Below, we describe a Dirichlet-Neumann algorithm for (12.42).

Algorithm 12.4.1 *(Dirichlet-Oblique Derivative Algorithm for (12.42))*
Let $w_2^{(0)}$ be a starting iterate

1. **For** $k = 1, 2, \ldots$ *until convergence* **do**
2. *Solve the local hyperbolic equation on Ω_1 for $w_1^{(k)}$:*

$$\begin{cases} \mathbf{b}(x)\cdot\nabla w_1^{(k)} + c(x)\,w_1^{(k)} = f(x), & \text{in } \Omega_1 \\ w_1^{(k)} = w_2^{(k-1)}, & \text{on } B \cap \partial\Omega_{1,in} \\ w_1^{(k)} = 0, & \text{on } \partial\Omega_1 \cap \partial\Omega_{1,in} \end{cases}$$

3. *Solve the following elliptic equation on Ω_2 for $w_2^{(k)}$:*

$$\begin{cases} -\epsilon\,\chi(\Delta w_2^{(k)}) + \mathbf{b}(x)\cdot\nabla w_2^{(k)} + c(x)\,w_2^{(k)} = f(x), & \text{in } \Omega_2 \\ w_2^{(k)} = w_1^{(k)}, & \text{on } B \cap \partial\Omega_{1,out} \\ \mathbf{b}(x)\cdot\nabla w_2^{(k)} + c(x)\,w_2^{(k)} = f(x), & \text{on } B \cap \partial\Omega_{1,in} \\ w_2^{(k)} = 0, & \text{on } \partial\Omega_2 \cap \partial\Omega_\chi \end{cases}$$

4. **Endfor**

Under appropriate assumptions, the iterates $w_l^{(k)}$ can be shown to converge geometrically to the exact solution $w_l(x)$ of (12.42), see [BR32, CA29].

Overlapping Subdomain Heterogeneous Formulation of (12.40). We shall next indicate a reformulation of the χ-equation based on *overlapping* subdomains. Let $\Omega_1^* = \Omega_1$ as before, and let Ω_2^* denote an overlapping domain. Let $\partial\Omega_{1,in}^*$ and $\partial\Omega_{1,out}^*$ denote segments as in (12.41) with Ω_1^* replacing Ω_1. Due to *continuity* of the solution $w(.)$ of the χ-equation, we obtain the following reformulation [BR32, CA29] of the χ-equation.

Lemma 12.17. *Suppose the following conditions hold.*

1. *Let w be the solution of the χ-equation.*
2. *Let (w_1, w_2) satisfy:*

$$\begin{cases} \mathbf{b}(x)\cdot\nabla w_1 + c(x)\, w_1 = f(x), & \text{in } \Omega_1^* \\ w_1 = w_2, & \text{on } \partial\Omega_{1,in}^* \cap \Omega \\ w_1 = 0, & \text{on } \partial\Omega \cap \partial\Omega_{1,in}^* \\ -\epsilon\,\chi\,(\Delta w_2) + \mathbf{b}(x)\cdot\nabla w_2 + c(x)\, w_2 = f(x), & \text{in } \Omega_2^* \\ w_2 = w_1, & \text{on } \partial\Omega_2^* \cap \Omega \\ w_2 = 0, & \text{on } \partial\Omega_2^* \cap \partial\Omega_\chi. \end{cases} \tag{12.43}$$

Then the following result will hold:

$$w = w_1 \text{ on } \partial\Omega_1^*, \qquad \text{and} \qquad w = w_2 \text{ on } \partial\Omega_2^*.$$

Proof. See [BR32, CA29]. □

The above reformulation suggests a Schwarz iterative algorithm.

Algorithm 12.4.2 *(Sequential Schwarz Algorithm for (12.43))*
Let $\left(w_1^{(0)}, w_2^{(0)}\right)$ denote starting iterates

```
1. For $k = 1, \ldots$ until convergence do
2.     Solve the local hyperbolic equation for $w_1^{(k)}$ on $\Omega_1^*$:
           $\begin{cases} \mathbf{b}(x)\cdot\nabla w_1^{(k)} + c(x)\, w_1^{(k)} = f(x), & \text{in } \Omega_1^* \\ w_1^{(k)} = w_2^{(k-1)}, & \text{on } \partial\Omega_{1,in}^* \cap \Omega \\ w_1^{(k)} = 0, & \text{on } \partial\Omega \cap \partial\Omega_{1,in}^* \end{cases}$
3.     Solve the local nonlinear elliptic equation for $w_2^{(k)}$ on $\Omega_2^*$:
           $\begin{cases} -\epsilon\,\chi(\Delta w_2^{(k)}) + \mathbf{b}(x)\cdot\nabla w_2^{(k)} + c(x)\, w_2^{(k)} = f(x), & \text{in } \Omega_2^* \\ w_2^{(k)} = w_1^{(k)}, & \text{on } \partial\Omega_2^* \cap \Omega \\ w_2^{(k)} = 0, & \text{on } \partial\Omega_2^* \cap \partial\Omega_\chi \end{cases}$
4. Endfor
```

Under appropriate assumptions, the above iterates can be shown to converge geometrically to the local solution (w_1, w_2) of the χ-equation.

12.4.1 Discretization of the χ-Formulation

We shall now describe a *heuristic finite difference* discretization of (12.40). Let Ω be triangulated by a finite difference grid $\mathcal{T}_h(\Omega)$ of size h. We shall denote the vector of *interior* nodal unknowns as $\mathbf{u}_{\mathcal{I}}$ and the vector of *boundary* nodal unknowns as $\mathbf{u}_{\mathcal{B}}$, and denote the linear system corresponding to a finite difference discretization of (12.39) as:

$$\begin{cases} \epsilon \left(A_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}} \right) + \left(C_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + C_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}} \right) = \mathbf{f}_{\mathcal{I}} \\ \hfill \mathbf{u}_{\mathcal{B}} = \mathbf{0}, \end{cases} \tag{12.44}$$

where $A_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}}$ denotes a discretization of $-\Delta u$ and $C_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + C_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}}$ denotes an upwind discretization of $\mathbf{b}(x) \cdot \nabla u + c(x)\, u$.

To obtain a discretization of the χ-equation, we shall formally apply the $\chi(\cdot)$-function to each row entry $\left(A_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}} \right)_i$ and denote the resulting row vector as $\boldsymbol{\chi} \left(A_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}} \right)$, where:

$$\boldsymbol{\chi} \left(A_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}} \right)_i = \chi \left((A_{\mathcal{I}\mathcal{I}} \mathbf{u}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{u}_{\mathcal{B}})_i \right). \tag{12.45}$$

Thus, a discretization of the χ-equation can be obtained formally as:

$$\begin{cases} \epsilon \, \boldsymbol{\chi} \left(A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{w}_{\mathcal{B}} \right) + \left(C_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}} + C_{\mathcal{I}\mathcal{B}} \mathbf{w}_{\mathcal{B}} \right) = \mathbf{f}_{\mathcal{I}} \\ \hfill \mathbf{w}_{\mathcal{B}} = \mathbf{0}, \end{cases} \tag{12.46}$$

where $\boldsymbol{\chi} \left(A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{w}_{\mathcal{B}} \right)$ corresponds to a discretization of $\chi \left(-\Delta w \right)$ and $C_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}} + C_{\mathcal{I}\mathcal{B}} \mathbf{w}_{\mathcal{B}}$ to a discretization of $\mathbf{b}(x) \cdot \nabla w + c(x)\, w$.

Remark 12.18. If one or more row entries of $\boldsymbol{\chi} \left(A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{w}_{\mathcal{B}} \right)$ are zero, then it is possible that not all boundary nodal values $\mathbf{w}_{\mathcal{B}}$ influence the interior solution $\mathbf{w}_{\mathcal{I}}$. Indeed, suppose $\boldsymbol{\chi} \left(A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}} + A_{\mathcal{I}\mathcal{B}} \mathbf{w}_{\mathcal{B}} \right)_i = 0$, then at the associated node x_i, the stencil will depend solely on the stencil of matrix C, where the latter will have a domain of dependence primarily on the *inflow* boundary $\partial \Omega_{in}$, i.e., $x \in \partial \Omega$ such that $\mathbf{n}(x) \cdot \mathbf{b}(x) < 0$. Furthermore, if A and C are M-matrices, and $\chi'(\cdot) \geq 0$, then the linearized system (12.47) will yield an M-matrix, and a discretize maximum principle will hold.

Remark 12.19. The nonlinear χ-equations (12.46) can be solved by applying a Newton iteration. Given the k'th Newton iterate $\mathbf{w}_{\mathcal{I}}^k \approx \mathbf{w}_{\mathcal{I}}$, a new iterate $\mathbf{w}_{\mathcal{I}}^{k+1}$ can be computed by solving the following system for $(\mathbf{w}_{\mathcal{I}}^{k+1} - \mathbf{w}_{\mathcal{I}}^k)$:

$$H_{\mathcal{I}\mathcal{I}}^k \left(\mathbf{w}_{\mathcal{I}}^{k+1} - \mathbf{w}_{\mathcal{I}}^k \right) = \mathbf{f}_{\mathcal{I}} - \epsilon \, \boldsymbol{\chi}(A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}}^k) - C_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}}^k, \tag{12.47}$$

where $H_{\mathcal{I}\mathcal{I}}^k = \epsilon \, \mathrm{diag} \left(\boldsymbol{\chi}'(A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}}^k) \right) A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}$ and $\boldsymbol{\chi}'(\mathbf{y}_{\mathcal{I}}^k)$ denotes the vector with entries $\chi' \left((A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}}^k)_i \right)$. Here, since $\mathrm{diag} \left(\boldsymbol{\chi}(\mathbf{y}_{\mathcal{I}}^k) \right)$ is a diagonal matrix, the i'th row of $\epsilon \, \mathrm{diag} \left(\boldsymbol{\chi}(\mathbf{y}_{\mathcal{I}}^k) \right) A_{\mathcal{I}\mathcal{I}}$ corresponds to i'th row of $A_{\mathcal{I}\mathcal{I}}$ multiplied by $\epsilon \, \chi \left((A_{\mathcal{I}\mathcal{I}} \mathbf{w}_{\mathcal{I}}^k)_i \right)$. Each Newton iteration requires the solution of system (12.47).

Below, we summarize the Newton iteration to solve (12.46).

Algorithm 12.4.3 *(Newton Iteration to solve (12.46))*
Let $w_{\mathcal{I}}^{0} = \mathbf{0}$ *denote a starting iterate*

```
1. For $k = 1, \ldots$ until convergence do
2.     Assemble $H_{\mathcal{I}\mathcal{I}}^{k} = \left(\epsilon\, \mathrm{diag}\left(\boldsymbol{\chi}'(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^{k})\right) A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)$
3.     Compute $\mathbf{r}_{\mathcal{I}}^{k} \equiv \mathbf{f}_{\mathcal{I}} - \epsilon\, \boldsymbol{\chi}(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^{k}) - C_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^{k}$ and solve:

           $H_{\mathcal{I}\mathcal{I}}^{k}\left(\mathbf{w}_{\mathcal{I}}^{k+1} - \mathbf{w}_{\mathcal{I}}^{k}\right) = \mathbf{r}_{\mathcal{I}}^{k}$

4. Endfor
```

Assembly of matrix $H_{\mathcal{I}\mathcal{I}}^{k}$ is not expensive, since only the diagonal matrix $\mathrm{diag}\left(\boldsymbol{\chi}'(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^{k})\right)$ multiplying $A_{\mathcal{I}\mathcal{I}}$ depends on $\mathbf{w}_{\mathcal{I}}^{k}$. However, an efficient solver will be required to solve $H_{\mathcal{I}\mathcal{I}}^{k}\left(\mathbf{w}_{\mathcal{I}}^{k+1} - \mathbf{w}_{\mathcal{I}}^{k}\right) = \mathbf{r}_{\mathcal{I}}^{k}$ in step 3 above. We shall describe two iterative solvers for $H_{\mathcal{I}\mathcal{I}}^{k}$, yielding *inner* and *outer* iterations.
Non-Overlapping Subdomains Based Algorithm. Here, we describe a discrete linearized version of Alg. 12.4.1. Given $\mathbf{w}_{\mathcal{I}}^{k} \approx \mathbf{w}_{\mathcal{I}}$, define the set:

$$I_1^k = \left\{ i : \chi((A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^{k})_i) = 0 \right\}. \tag{12.48}$$

Given I_1^k, let Ω_1^k denote an *open* region such that for each $i \in I_1^k$, it holds that node $x_i \in \overline{\Omega}_1^k \cap \Omega$, i.e., $\chi\left(\left(A_{II}\mathbf{w}_I^k\right)_i\right) = 0$. Let Ω_2^k denote a subregion complementary to Ω_1^k with interface $B^k = \partial\Omega_1^k \cap \partial\Omega_2^k$. For simplicity, we shall assume that the subregions are consistent with the cells of the grid on Ω. Given this decomposition, we partition $\mathbf{y}_{\mathcal{I}} = \left(\mathbf{y}_I^{k;(1)^T}, \mathbf{y}_I^{k;(2)^T}, \mathbf{y}_B^{k^T}\right)^T$ corresponding to nodal values of $\mathbf{y}_{\mathcal{I}}$ in Ω_1^k, Ω_2^k and B^k, respectively. We let $R_I^{k;(1)}$, $R_I^{k;(2)}$ and R_B^k denote the restriction map onto nodal vectors on Ω_1^k, Ω_2^k and B^k, respectively. A Then, the Newton system $H_{\mathcal{I}\mathcal{I}}^{k}\left(\mathbf{w}_{\mathcal{I}}^{k+1} - \mathbf{w}_{\mathcal{I}}^{k}\right) = \tilde{\mathbf{f}}_{\mathcal{I}}^{k}$, can be block partitioned using this $\mathbf{w}_{\mathcal{I}}^{k}$ dependent decomposition:

$$\begin{bmatrix} H_{II}^{k;(1)} & 0 & H_{IB}^{k;(1)} \\ 0 & H_{II}^{k;(2)} & H_{IB}^{k;(2)} \\ H_{BI}^{k;(1)} & H_{BI}^{k;(2)} & H_{BB}^{k} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{k+1;(1)} - \mathbf{w}_I^{k;(1)} \\ \mathbf{w}_I^{k+1;(2)} - \mathbf{w}_I^{k;(2)} \\ \mathbf{w}_B^{k+1} - \mathbf{w}_B^{k} \end{bmatrix} = \begin{bmatrix} R_I^{k;(1)}\mathbf{r}_{\mathcal{I}}^{k} \\ R_I^{(k:(2)}\mathbf{r}_{\mathcal{I}}^{k} \\ R_B^{k}\mathbf{r}_{\mathcal{I}}^{k} \end{bmatrix},$$

where $H_{XY}^{k;(l)} = \epsilon\, A_{XY}^{k;(l)} + C_{XY}^{(l)}$ for X, $Y, = I$, B. Our choice of Ω_l^k and B^k yields $H_{II}^{k;(1)} = C_{II}^{(1)}$, $H_{IB}^{k;(1)} = C_{IB}^{(1)}$, $H_{BB}^{k} = C_{BB}$ and $H_{BI}^{k;(l)} = C_{BI}^{(l)}$ for $l = 1, 2$. Substituting this into the above system yields the Newton system:

$$\begin{bmatrix} C_{II}^{(1)} & 0 & C_{IB}^{(1)} \\ 0 & H_{II}^{k;(2)} & H_{IB}^{k;(2)} \\ C_{BI}^{(1)} & C_{BI}^{(2)} & C_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{k+1;(1)} - \mathbf{w}_I^{k;(1)} \\ \mathbf{w}_I^{k+1;(2)} - \mathbf{w}_I^{k;(2)} \\ \mathbf{w}_B^{k+1} - \mathbf{w}_B^{k} \end{bmatrix} = \begin{bmatrix} R_I^{k;(1)}\mathbf{r}_{\mathcal{I}}^{k} \\ R_I^{k:(2)}\mathbf{r}_{\mathcal{I}}^{k} \\ R_B^{k}\mathbf{r}_{\mathcal{I}}^{k} \end{bmatrix}. \tag{12.49}$$

Applying a block Gauss-Seidel iteration to the above system yields a discrete linearized version of the Dirichlet-Oblique derivative algorithm (12.4.1).

Although matrix $C_{\mathcal{I}\mathcal{I}}$ does not depend on $\mathbf{w}_{\mathcal{I}}^k$, its submatrices $C_{XY}^{(l)}$ depend on the partition induced by Ω_1^k, Ω_2^k and B^k. Despite this, for notational simplicity we have omitted the k dependence of $C_{XY}^{(l)}$. The first block row in (12.49) corresponds to a discretized hyperbolic equation in Ω_1^k, while the second block row corresponds to a linearization of the discrete χ-equation in Ω_2^k, and the third block row corresponds to a a discretized oblique derivative condition on B^k. Below, we describe a discrete and linearized version of Alg. 12.4.1 using N_* *inner* iterations.

Algorithm 12.4.4 *(Newton-Dirichlet-Oblique Algorithm for (12.46))*
Let $w_{\mathcal{I}}^0 = \mathbf{0}$ denote a starting iterate

```
1.  For $k = 1, \ldots$ until convergence do
2.      Assemble $H_{\mathcal{I}\mathcal{I}}^k = \left(\epsilon \,\mathrm{diag}\left(\boldsymbol{\chi}'(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^k)\right) A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)$
3.      Compute the residual $\mathbf{r}_{\mathcal{I}}^k \equiv \left(\mathbf{f}_{\mathcal{I}} - \epsilon\, \boldsymbol{\chi}(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^k) - C_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^k\right)$
4.      Define $\mathbf{x}_{\mathcal{I}}^1 \equiv \mathbf{0}$
5.      For $l = 1, \ldots, N_*$
6.      Update the solution:
            $\mathbf{x}_{\mathcal{I}}^{l+\frac{1}{2}} = \mathbf{x}_{\mathcal{I}}^{l} + R_I^{k;(1)^T} C_{II}^{(1)^{-1}} R_I^{k;(1)T} \left(\mathbf{r}_{\mathcal{I}}^k - H_{\mathcal{I}\mathcal{I}}^k \mathbf{x}_{\mathcal{I}}^l\right)$
7.      Update the solution:
            $\mathbf{x}_{\mathcal{I}}^{l+1} = \mathbf{x}_{\mathcal{I}}^{l+\frac{1}{2}} + \begin{bmatrix} R_I^{k;(2)} \\ R_B^k \end{bmatrix}^T \begin{bmatrix} H_{II}^{k;(2)} & H_{IB}^{k;(2)} \\ C_{BI}^{(2)} & C_{BB} \end{bmatrix}^{-1} \begin{bmatrix} R_I^{k;(2)} \\ R_B^k \end{bmatrix} \left(\mathbf{r}_{\mathcal{I}}^k - H_{\mathcal{I}\mathcal{I}}^k \mathbf{x}_{\mathcal{I}}^{l+\frac{1}{2}}\right)$
8.      Endfor
9.      Update $\mathbf{w}_{\mathcal{I}}^{k+1} \equiv \mathbf{w}_{\mathcal{I}}^k + \mathbf{x}_{\mathcal{I}}^{N_*+1}$
10. Endfor
```

Remark 12.20. The preceding algorithm applies a two-step block Gauss-Seidel iteration to solve each Newton system. Step 6 above corresponds to a discrete and linearized version of step 2 in Alg. 12.4.1, while step 7 above corresponds to a discrete and linearized version of step 3 in Alg. 12.4.1. If matrix $A_{\mathcal{I}\mathcal{I}}$ is an M-matrix, then matrix $H_{\mathcal{I}\mathcal{I}}^k$ will also be an M-matrix, and the block Gauss-Seidel *inner iteration* will be convergent provided $c(x) \geq c_0 > 0$. More rapid convergence can be obtained by increasing the *overlap* between the subregions. More generally, Krylov acceleration should be employed.

Overlapping Subdomains Based Algorithm. We next outline a discrete version of Alg. 12.4.2 to solve (12.46). Given an iterate $\mathbf{w}_{\mathcal{I}}^k \approx \mathbf{w}_{\mathcal{I}}$, we employ the index set I_1^k in (12.48) and let $\Omega_1^{k;*} = \Omega_1^k$ denote an open subregion containing all the nodes x_i for $i \in I_1^k$. A subdomain $\Omega_2^{k;*} \subset \Omega$ is chosen so that it overlaps with $\Omega_1^{k;*}$. On each subdomain, we let $B^{k;(l)} \equiv \left(\partial\Omega_l^{k;*} \cap \Omega\right)$ denote its interior boundary. On subdomain $\Omega_l^{k;*}$ we let $R_I^{k;(l)}$ and $R_B^{k;(l)}$ denote restriction maps which map a nodal vector on Ω into vectors of nodal values on $\Omega_1^{k;*}$ and $B^{k;(l)}$, respectively. Additionally, given $\mathbf{y}_{\mathcal{I}}$, we let $\mathbf{y}_I^{k;(1)} = R_I^{k;(l)}\mathbf{y}_{\mathcal{I}}$ and $\mathbf{y}_B^{k;(1)} = R_B^{k;(l)}\mathbf{y}_{\mathcal{I}}$ denote its nodal vectors associated with $\Omega_l^{k;*}$ and $B^{k;(l)}$, respectively, and we block partition the submatrix of $H_{\mathcal{I}\mathcal{I}}^k$ corresponding to nodes in $\Omega_l^{k;*}$ and $B^{k;(l)}$ with sub-blocks $H_{XY}^{k;(l)} = \epsilon\,\tilde{A}_{XY}^{k;(l)} + C_{XY}^{(l)}$, where the index sets X, $Y = I$, B. By our choice of $\Omega_1^{k;*}$, it will hold that $\tilde{A}_{XY}^{k;(l)} = 0$. Below, we summarize the Schwarz alternating method to solve (12.46) using Newton linearization and a fixed number N_* of inner iterations.

Algorithm 12.4.5 *(Newton-Schwarz Alternating Algorithm for (12.46))*
Let $w_{\mathcal{I}}^0$ denote a starting iterate

```
1.  For $k = 1, \ldots$ until convergence do
2.      Assemble $H_{\mathcal{I}\mathcal{I}}^k = \left(\epsilon\,\mathrm{diag}\left(\boldsymbol{\chi}'(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^k)\right)\, A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)$.
3.      Compute the residual $\mathbf{r}_{\mathcal{I}}^k \equiv \left(\mathbf{f}_{\mathcal{I}} - \epsilon\,\boldsymbol{\chi}(\mathbf{y}_{\mathcal{I}}^k) - C_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}^k\right)$
4.      Define $\mathbf{x}_{\mathcal{I}}^1 \equiv \mathbf{0}$
5.      For $l = 1, \ldots, N_*$
6.      Update the solution in $\Omega_1^{k;*}$:
            $$\mathbf{x}_{\mathcal{I}}^{l+\frac{1}{2}} = \mathbf{x}_{\mathcal{I}}^l + R_I^{k;(1)^T} C_{II}^{(1)^{-1}} R_I^{k;(1)} \left(\mathbf{r}_{\mathcal{I}}^k - H_{\mathcal{I}\mathcal{I}}^k \mathbf{x}_{\mathcal{I}}^l\right)$$
7.      Update the solution in $\Omega_2^{k;*}$ by solving:
            $$\mathbf{x}_{\mathcal{I}}^{l+1} = \mathbf{x}_{\mathcal{I}}^{l+\frac{1}{2}} + \begin{bmatrix} R_I^{k;(2)} \\ R_B^k \end{bmatrix} \begin{bmatrix} H_{II}^{k;(2)} & H_{IB}^{k;(2)} \\ H_{BI}^{k;(2)} & H_{BB}^k \end{bmatrix}^{-1} \begin{bmatrix} R_I^{k;(2)} \\ R_B^k \end{bmatrix} \left(\mathbf{r}_{\mathcal{I}}^k - H_{\mathcal{I}\mathcal{I}}^k \mathbf{x}_{\mathcal{I}}^{l+\frac{1}{2}}\right)$$
8.      Endfor
9.      Define $\mathbf{w}_{\mathcal{I}}^{k+1} = \mathbf{w}_{\mathcal{I}}^k + \mathbf{x}_{\mathcal{I}}^{N_*+1}$
10. Endfor
```

Remark 12.21. Steps 6 and 7 in the preceding unaccelerated Newton-Schwarz algorithm correspond to steps 2 and 3 in Alg.12.4.2. If matrix $A_{\mathcal{I}\mathcal{I}}$ is an M-matrix, then matrix $H_{\mathcal{I}\mathcal{I}}^k$ will also be an M-matrix. Furthermore, if the coefficient $c(x) \geq c_0 > 0$ and the *overlap* between the two subdomains is sufficiently large, then the rate of convergence of the *inner iteration* will be independent of h and ϵ, see Chap. 15. Krylov acceleration can be employed.

Accuracy of the $\boldsymbol{\chi}$-Formulation. We end this section by heuristically estimating the accuracy of the discrete χ-formulation (12.46). Let $\mathbf{u}_{\mathcal{I}}$ denote the original discretization of (12.39):

$$\epsilon\, A_{\mathcal{I}\mathcal{I}}\mathbf{u}_{\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\mathbf{u}_{\mathcal{I}} = \mathbf{f}_{\mathcal{I}}.$$

By comparison, the solution $\mathbf{w}_{\mathcal{I}}$ to the discrete χ-formulation will satisfy:

$$\epsilon\, \boldsymbol{\chi}\left(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}\right) + C_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}} = \mathbf{f}_{\mathcal{I}}.$$

Subtracting the two yields:

$$\left(\epsilon\, A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)\left(\mathbf{u}_{\mathcal{I}} - \mathbf{w}_{\mathcal{I}}\right) = \epsilon\, \left(\boldsymbol{\chi}(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}) - A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}\right).$$

Formally applying matrix norm bounds yields the estimates:

$$\begin{aligned} \|\mathbf{u}_{\mathcal{I}} - \mathbf{w}_{\mathcal{I}}\| &\le \|\left(\epsilon\, A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)^{-1}\|\, \epsilon\, \|\boldsymbol{\chi}\left(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}\right) - \left(A_{\mathcal{I}\mathcal{I}}\mathbf{w}_{\mathcal{I}}\right)\| \\ &\le \|\left(\epsilon\, A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)^{-1}\|\, \epsilon\, \delta_2, \end{aligned}$$

using the definition of $\chi(\cdot)$. So, if the original discretization is *stable*, i.e., $\|\left(\epsilon\, A_{\mathcal{I}\mathcal{I}} + C_{\mathcal{I}\mathcal{I}}\right)^{-1}\| \le C$ independent of h (and ϵ), then $\|\mathbf{u}_{\mathcal{I}} - \mathbf{w}_{\mathcal{I}}\| = O(\epsilon)$.

12.5 Applications to Parabolic Equations

Our discussion in this section will be *heuristic* and brief. Given an *advection dominated* parabolic equation, we outline how alternative *hyperbolic-parabolic* approximations of the parabolic equation can be formulated based on two space-time subregions. The resulting heterogeneous system may subsequently be discretized on matching or non-matching space-time grids and iterative algorithms can be formulated for its solution.

We consider the following advection dominated parabolic equation:

$$\begin{cases} u_t + Lu = f, & \text{in } \Omega \times (0,T) \\ u = 0, & \text{on } \partial\Omega \times (0,T) \\ u(x,0) = u_0(x), & \text{on } \Omega \text{ when } t = 0, \end{cases} \tag{12.50}$$

where the underlying advection-diffusion elliptic operator is:

$$L\, u = -\epsilon\, \Delta u + \mathbf{b}(x)\cdot\nabla u + c(x)\, u, \tag{12.51}$$

for $0 < \epsilon \ll 1$ and $c(x) \ge c_0 > 0$ with $\left(c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x)\right) \ge \beta > 0$. We shall assume there is a subdomain $\Omega_1 \subset \Omega$ (or $\Omega_1 = \Omega_1^* \subset \Omega$) such that:

$$\epsilon\, |\Delta u| \ll |u_t + \mathbf{b}(x)\cdot\nabla u + c(x)u| \quad \text{for} \quad (x,t) \in \Omega_1 \times (0,T). \tag{12.52}$$

When this holds, a *parabolic-hyperbolic* approximation of (12.50) can be obtained by constructing a hybrid formulation of (12.50) involving $\Omega_1 \times (0,T)$ and another space-time region, and by omitting $-\epsilon\, \Delta u$ on $\Omega_1 \times (0,T)$. The resulting problem can subsequently be discretized.

Steklov-Poincaré Approximation. Given $\Omega_1 \subset \Omega$ satisfying (12.52), let Ω_2 be a complementary region with interface $B = \partial\Omega_1 \cap \partial\Omega_2$. Analogous to the elliptic case, a Steklov-Poincaré hybrid formulation of (12.50) can be formulated based on $\Omega_1 \times (0,T)$ and $\Omega_2 \times (0,T)$ with *transmission* conditions on $B \times (0,T)$, see [GA15, QU3, QU4, QU6]. If we omit $-\epsilon\,\Delta u$ in $\Omega_1 \times (0,T)$ and appropriately modify the transmission conditions, we shall obtain the following *hyperbolic-parabolic* system for $w_l(x,t) \approx u(x,t)$ on $\Omega_l \times (0,T)$:

$$\begin{cases} \frac{\partial w_1}{\partial t} + L_0 w_1 = f, & \text{in } \Omega_1 \times (0,T) \\ w_1 = w_2, & \text{on } (B \cap \partial\Omega_{1,in}) \times (0,T) \\ w_1 = 0, & \text{on } \left(B_{[1]} \cap \partial\Omega_{1,in}\right) \times (0,T) \\ w_1(x,0) = u_0(x), & \text{on } \Omega_1 \text{ when } t = 0 \\ \frac{\partial w_2}{\partial t} + L w_2 = f, & \text{in } \Omega_2 \times (0,T) \\ \mathbf{n}_1 \cdot \mathbf{F}_2(w_2) = \mathbf{n}_1 \cdot \mathbf{F}_1(w_1), & \text{on } B \times (0,T) \\ w_2 = 0, & \text{on } B_{[2]} \times (0,T) \\ w_2(x,0) = u_0(x), & \text{on } \Omega_1 \text{ when } t = 0, \end{cases} \tag{12.53}$$

where $\mathbf{F}_1(w_1) = -\frac{1}{2}\,\mathbf{b}\,w_1$ and $\mathbf{F}_2(w_2) = \epsilon\,\nabla w_2 - \frac{1}{2}\mathbf{b}\,w_2$ are the local *fluxes*, with $L_0 w_1 = \mathbf{b}(x)\cdot\nabla w_1 + c(x)\,w_1$ and $\partial\Omega_{1,in} = \{x \in \partial\Omega_1 : \mathbf{n}_1(x)\cdot\mathbf{b}(x) < 0\}$.

A *heuristic* discretization of this hyperbolic-parabolic system can be obtained by discretizing the evolution equation on each region $\Omega_l \times (0,T)$ by a locally stable scheme, and by carefully discretizing the transmission conditions. Let $\mathbf{w}_I^{(1)}$ denote the vector of nodal unknowns associated with $w_1(.,.)$ on $\Omega_1 \times (0,T)$ and let $\mathbf{w}_{B_\pm}^{(1)}$ denote the vector of nodal unknowns associated with $w_1(.,.)$ on $B_\pm \times (0,T)$. Similarly, let $\mathbf{w}_I^{(2)}$ denote the vector of nodal unknowns associated with $w_2(.,.)$ on $\Omega_2 \times (0,T)$ and $\mathbf{w}_{B_\pm}^{(2)}$ the vector of nodal unknowns associated with $w_2(.,.)$ on $B_\pm \times (0,T)$, where:

$$\begin{aligned} B_- &= \{x \in B \,:\, \mathbf{n}_1(x)\cdot\mathbf{b}(x) < 0\} \\ B_+ &= \{x \in B \,:\, \mathbf{n}_1(x)\cdot\mathbf{b}(x) \geq 0\}\,. \end{aligned}$$

Then, a discretization of the hyperbolic-parabolic system will yield a large system of equations with the block structure (12.14), where $\mathbf{w}_{B_-}^{(1)} = \mathbf{w}_{B_-}^{(2)}$. Heuristic Schur complement solvers can be formulated for the resulting system. Importantly, since $-\epsilon\,\Delta w_1$ is omitted in $\Omega_1 \times (0,T)$, we may employ a stable *explicit* scheme in $\Omega_1 \times (0,T)$ and a stable *implicit* scheme in $\Omega_2 \times (0,T)$, with a larger time step $\tau_1 \gg \tau_2$, where τ_l denotes the time step on $\Omega_l \times (0,T)$, see § 11.6 (appropriately extended). We omit further details.

Schwarz Approximation. Given $\Omega_1^* = \Omega_1$ satisfying (12.52), let Ω_2^* denote an overlapping subdomain. Let $B_{[l]} = \partial\Omega_l^* \cap \partial\Omega$ and $B^{(l)} = \partial\Omega_l^* \cap \Omega$. By analogy with the elliptic case, we may construct a hybrid Schwarz formulation of (12.50) based on $\Omega_1^* \times (0,T)$ and and $\Omega_2^* \times (0,T)$, and omit $-\epsilon\Delta u$ to obtain a *hyperbolic* equation on $\Omega_1^* \times (0,T)$, see [BR32, CA29, MA35]. If $w_l(x,t)$ denotes the heterogeneous solution on $\Omega_l^* \times (0,T)$, then appropriately modifying the boundary conditions on $\partial\Omega_1^*$, we obtain the following hyperbolic-parabolic Schwarz approximation, as in § 11.6:

$$\begin{cases} \frac{\partial w_1}{\partial t} + L_0\, w_1 = f, & \text{on } \Omega_1^* \times (0,T) \\ w_1 = w_2, & \text{on } \left(B^{(1)} \cap \partial\Omega_{1,in}^*\right) \times (0,T) \\ w_1 = 0, & \text{on } \left(B_{[1]} \cap \partial\Omega_{1,in}^*\right) \times (0,T) \\ w_1(x,0) = u_0(x), & \text{on } \Omega_1^* \text{ when } t=0 \\ \frac{\partial w_2}{\partial t} + L\, w_2 = f, & \text{on } \Omega_2^* \times (0,T) \\ w_2 = w_1, & \text{on } B^{(2)} \times (0,T) \\ w_2 = 0, & \text{on } B_{[2]} \times (0,T) \\ w_2(x,0) = u_0(x), & \text{on } \Omega_2^* \text{ when } t=0. \end{cases} \tag{12.54}$$

Here $\partial\Omega_{1,in} = \{x \in \partial\Omega_1^* : \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0\}$ and $L_0\, w_1 = \mathbf{b}(x) \cdot \nabla w_1 + c(x)\, w_1$. This heterogeneous system can be discretized on a matching or non-matching space-time grid, see § 11.6, using a stable scheme for each evolution equation on $\Omega_l^* \times (0,T)$ and inter-subdomain interpolation. Importantly, since $-\epsilon\,\Delta w_1$ is omitted in $\Omega_1^* \times (,0,T)$, we may employ a time step $\tau_1 \gg \tau_2$, where τ_l denotes the time step in $\Omega_l^* \times (,0,T)$ without adversely affecting local stability.

If $\mathbf{w}_I^{(l)}$ denotes the vector of nodal values of $w_l(.,.)$ on $\Omega_l^* \times (0,T)$ and $\mathbf{w}_B^{(l)}$ the vector of nodal values associated with $w_l(.,.)$ on $B^{(l)} \times (0,T)$, we will obtain a large linear system having the same block structure as (12.24). Schwarz iterative algorithms can be formulated to solve this system, see [MA35]. We omit further details.

Least Squares-Control Approximation. An heterogeneous least squares-control approximation of (12.50) can be constructed based on *overlapping* or *nonoverlapping* subdomains, see [GL13]. For simplicity, we shall consider overlapping subdomains. Let $\Omega_1^* = \Omega_1$ denote a subdomain on which (12.52) holds, and let Ω_2^* denote an overlapping subdomain. Let $B_{[l]} = \partial\Omega_l^* \cap \partial\Omega$ and $B^{(l)} = \partial\Omega_l^* \cap \Omega$. By analogy with the vanishing viscosity least squares-control formulation for advection dominated *elliptic* equations, we define:

$$\mathcal{K}_* = \{\, (v_1, v_2) \,:\, v_1,\, v_2 \text{ satisfies (12.56)}\}\,, \tag{12.55}$$

where:

$$\begin{cases} \dfrac{\partial v_1}{\partial t} + L_0\, v_1 = f, & \text{on } \Omega_1^* \times (0,T) \\ v_1 = g_1, & \text{on } \left(B^{(1)} \cap \partial\Omega_{1,in}\right) \times (0,T) \\ v_1 = 0, & \text{on } \left(B_{[1]} \cap \partial\Omega_{1,in}\right) \\ v_1(x,0) = u_0(x), & \text{in } \Omega_1^* \text{ for } t=0 \\ \dfrac{\partial v_2}{\partial t} + L\, v_2 = f, & \text{on } \Omega_2^* \times (0,T) \\ v_2 = g_2, & \text{on } B^{(2)} \times (0,T) \\ v_2 = 0, & \text{on } B_{[2]} \times (0,T) \\ v_2(x,0) = u_0(x), & \text{in } \Omega_2^* \text{ for } t=0. \end{cases} \tag{12.56}$$

Here $v_l(x,t)$ denotes a local solution on $\Omega_l^* \times (0,T)$ and $\partial\Omega_{1,in}$ is:

$$\partial\Omega_{1,in} = \{\, x \in \partial\Omega_1^* \,:\, \mathbf{n}_1(x) \cdot \mathbf{b}(x) < 0 \} \,,$$

and $L_0\, v_1 = \mathbf{b}(x) \cdot \nabla v_1 + c(x)\, v_1$. We have approximated (12.50) by a *hyperbolic* equation on $\Omega_1^* \times (0,T)$, by omitting $-\epsilon \Delta u$, see [GL13].

An heterogeneous least squares-control approximation of (12.50) will seek:

$$J(w_1, w_2) = \min_{(v_1,v_2) \in \mathcal{K}_*} J(v_1, v_2). \tag{12.57}$$

where $\mathcal{K}_*$ is as defined by (12.56). An unconstrained minimization formulation of (12.55) can be obtained by using a parametric representation of $\mathcal{K}_*$ using the Dirichlet boundary values $g_1(x,t)$ and $g_2(x,t)$ as *control* data. This will formally yield $v_l = E_l\, g_l$ for $l = 1$, 2 where $E_l\, g_l$ denotes a formal *affine linear* extension of the Dirichlet boundary data as defined above.

To obtain an unconstrained minimization formulation of (12.57), define:

$$\tilde{J}(g_1, g_2) \equiv J\left(E_1\, g_1, E_2\, g_2\right).$$

By construction, we may equivalently seek g_1 and g_2 which minimizes $\tilde{J}(.,.)$.. A discretization of the least squares-control problem can be constructed using matching on non-matching space-time grids. Importantly, since we have a hyperbolic equation on $\Omega_1^* \times (0,T)$, we may employ a larger time step τ_1 on $\Omega_1^* \times (0,T)$, provided the local stability conditions are satisfied. We shall let $\tau_2 > \tau_1$ denote the time step on $\Omega_2^* \times (0,T)$. This will yield a nonmatching space-time grid, and least squares-control problem (12.57) can be discretized *heuristically* as outlined in § 11.6. Iterative solvers can be formulated for the resulting linear system. We omit further details.

χ-formulation. As in the elliptic case, since we may not know Ω_1 *a priori*, it can be estimated iteratively using the χ-formulation [BR32, CA29]. If $w(x,t)$ denotes the χ-approximation of $u(x,t)$, then we shall obtain a *nonlinear* parabolic equation (even though the original problem is linear):

$$\begin{cases} w_t - \epsilon\, \chi(\Delta w) + \mathbf{b}(x) \cdot \nabla w + c(x)\, w = f, & \text{in } \Omega \times (0,T) \\ \qquad\qquad\qquad\qquad\qquad\qquad w = 0, & \text{on } \partial\Omega \times (0,T) \\ \qquad\qquad\qquad\qquad\qquad w(x,0) = u_0(x), & \text{on } \Omega \text{ when } t = 0, \end{cases} \tag{12.58}$$

where $\chi(.)$ is as defined in Chap. 12.4. The above system can be discretized on a matching space-time grid (for instance finite difference in space and a θ-scheme in time). Given an approximation $w_k(\cdot) \approx w(\cdot)$, we may employ a Newton *linearization* of the discrete equations and solve the resulting system by methods analogous to those in Chap. 12.4.

13

Fictitious Domain and Domain Imbedding Methods

Fictitious domain and domain imbedding methods are methods for *imbedding* an elliptic equation within a family of elliptic equations posed on an extended or *fictitious* domain. A solution to the original elliptic equation is sought based on solving the associated elliptic equation on the extended domain. Unlike the divide and conquer strategies employed in domain decomposition methods, fictitious domain methods employ an *imbed and conquer* strategy, and it can be advantageous when the underlying domain is irregular or when the boundary conditions are complex. However, such imbedding may reduce computational costs only if the extended problem can be solved efficiently.

In this chapter, we describe fictitious domain iterative methods for solving discretizations of elliptic equations on irregular domains. Early literature on such methods, which pre-date domain decomposition methodology, focused on block matrix preconditioners [BU, PR2, AS4, OL, LE12, MA28, FI3, BO5] based on the extended domain. Recent literature has also included formulations which are based on the Lagrange multiplier and least squares-control frameworks [AT, NE7, DI2, GL12, PR, GL4, GL11]. Chap. 13.1 describes heuristic examples motivating fictitious domain methods, and formulates a useful matrix lemma for constructing preconditioners on extended domains. Chap. 13.2 describes a fictitious domain preconditioner for Neumann data problems, while Chap. 13.3 describes a similar preconditioner for Dirichlet problems. Chap. 13.4 describes several fictitious domain solvers based on the Lagrange multiplier and least squares-control formulations.

13.1 Background

Consider a self adjoint and coercive elliptic equation on a domain Ω:

$$L\,u = -\nabla \cdot (a(x)\nabla u) = f, \quad \text{ in } \Omega, \tag{13.1}$$

with either *Dirichlet* or *Neumann* boundary conditions imposed on $\partial\Omega$. The use of *fictitious domains* to *approximate* its solution $u(\cdot)$ can be motivated by regarding the above as a heat conduction problem to determine the *steady state* temperature $u(x)$ in a conductor Ω, see [LE12, MA28, FI3, NE7].

Suppose a constant temperature T_0 is imposed on $\partial\Omega$, then an *approximate* steady state temperature $w_\epsilon(x) \approx u(x)$ in Ω can be sought by extending the conductor to $\Omega_* \supset \Omega$, using a material of high conductivity $1/\epsilon$ in $\Omega_* \setminus \Omega$ (where $\epsilon \to 0^+$), see Fig. 13.1. If the temperature on the boundary $\partial\Omega_*$ is T_0, then due to the high conductivity in $\Omega_* \setminus \Omega$ we *heuristically* expect the steady state temperature $w_\epsilon(x) \to T_0$ on $\Omega_* \setminus \Omega$, yielding $w_\epsilon(x) \approx T_0$ on $\partial\Omega$. As a result, we heuristically expect $w_\epsilon(x) \approx u(x)$ on Ω. Formally, $u(x)$ and $w(_\epsilon(x)$ will solve:

$$\begin{cases} L\,u = f, & \text{in } \Omega \\ \quad u = T_0, & \text{on } \partial\Omega, \end{cases} \quad \text{and} \quad \begin{cases} L_\epsilon\, w_\epsilon = f_\epsilon, & \text{in } \Omega_* \\ \quad w_\epsilon = T_0, & \text{on } \partial\Omega_*, \end{cases} \tag{13.2}$$

where $L\,u = -\nabla \cdot (a(x)\nabla u)$ and $L_\epsilon\, w_\epsilon = -\nabla \cdot (\alpha_\epsilon(x)\nabla u)$ with:

$$\alpha_\epsilon(x) = \begin{cases} a(x) & \text{in } \Omega \\ \frac{1}{\epsilon} & \text{in } \Omega_* \setminus \Omega, \end{cases} \quad \text{and} \quad f_\epsilon(x) = \begin{cases} f(x), & \text{in } \Omega \\ 0, & \text{in } \Omega_* \setminus \Omega, \end{cases} \tag{13.3}$$

and $w_\epsilon \to T_0$ on $\Omega_* \setminus \Omega$, as $\epsilon \to 0^+$, yielding that $w_\epsilon \to u$ on Ω.

A similar approximation can be constructed when zero Neumann (flux) boundary conditions are imposed on $\partial\Omega$. However, in this case the conductivity on the extended region $\Omega_* \setminus \Omega$ should be small, say $\epsilon \to 0^+$. Then, zero flux boundary conditions on $\partial\Omega_*$ and low conductivity within $\Omega_* \setminus \Omega$ will ensure that $w_\epsilon(x)$ has approximately zero flux on $\partial\Omega$, yielding that $w_\epsilon(x) \approx u(x)$ within Ω. Formally, $u(x)$ and $w_\epsilon(x)$ will satisfy:

$$\begin{cases} L\,u = f, & \text{in } \Omega \\ \mathbf{n} \cdot (a\nabla u) = 0, & \text{on } \partial\Omega, \end{cases} \quad \text{and} \quad \begin{cases} L_\epsilon\, w_\epsilon = f_\epsilon, & \text{in } \Omega_* \\ \mathbf{n} \cdot (\epsilon\nabla w_\epsilon) = 0, & \text{on } \partial\Omega_*, \end{cases} \tag{13.4}$$

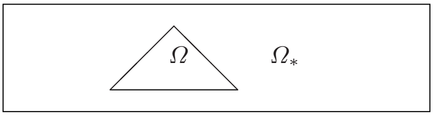


Fig. 13.1. A triangular domain Ω imbedded inside a rectangle Ω_*

where $L\,u = -\nabla \cdot (a(x)\nabla u)$ and $L_\epsilon\, w_\epsilon = -\nabla \cdot (\alpha_\epsilon(x)\nabla u)$ with:

$$\alpha_\epsilon(x) = \begin{cases} a(x) & \text{in } \Omega \\ \epsilon & \text{in } \Omega_* \setminus \Omega, \end{cases} \quad \text{and} \quad f_\epsilon(x) = \begin{cases} f(x), & \text{in } \Omega \\ 0, & \text{in } \Omega_* \setminus \Omega. \end{cases} \tag{13.5}$$

As $\epsilon \to 0^+$ the flux $\mathbf{n} \cdot (\epsilon\, \nabla w_\epsilon) \to 0$ on $\partial\Omega$, yielding $w_\epsilon(x) \approx u(x)$ on Ω.

The preceding heat conduction example illustrates that an approximate solution can be sought on Ω by solving an associated elliptic equation on an extended domain $\Omega_* \supset \Omega$. More generally, by varying the boundary conditions on $\partial\Omega_*$, it may be possible to obtain an exact solution to the original problem on Ω by solving an extended problem on Ω_*. However, such an approach will be computationally advantageous only if the extended problem can be solved efficiently. We next state a matrix result for preconditioning on a domain Ω based on an extended domain Ω_*.

A Fictitious Domain Preconditioning Lemma. Several of the block matrix fictitious domain preconditioners can be analyzed using a common preconditioning lemma [NE7]. Let $A = A^T > 0$ denote a stiffness matrix of size n arising form the discretization of a self adjoint elliptic equation on Ω. Let $K = K^T > 0$ denote a stiffness matrix of size $m > n$, arising from the discretization of an extended elliptic equation on an extended domain $\Omega_* \supset \Omega$. From a matrix viewpoint, the action M^{-1} of the inverse of several fictitious domain preconditioners M for A have the form $M^{-1} = RK^{-1}R^T$, where R is a rectangular *restriction* matrix of size $n \times m$. The following lemma states appropriate conditions for estimating the condition number of such a preconditioned system, see [NE7]. We shall let V and V_* denote Euclidean spaces of nodal vectors defined on Ω and Ω_*, respectively. We let R denote a restriction map $R : V_* \to V$ and $E : V \to V_*$ an extension map such that $R\,E = I_V$, where I_V denotes the identity matrix of size $n < m$.

Lemma 13.1. *Suppose the following conditions hold.*

1. *Suppose $\gamma_R > 0$ exists, such that:*
$$\left(\mathbf{v}_*^T R^T A R \mathbf{v}_*\right) \le \gamma_R \left(\mathbf{v}_*^T K \mathbf{v}_*\right), \qquad \forall \mathbf{v}_* \in V_*.$$
2. *Suppose $\gamma_E > 0$ exists, such that:*
$$\gamma_E \left(\mathbf{v}^T E^T K E \mathbf{v}\right) \le \left(\mathbf{v}^T A \mathbf{v}\right), \qquad \forall \mathbf{v} \in V.$$

Then, the following estimate will hold for $\mathrm{cond}\left(RK^{-1}R^T, A^{-1}\right)$*:*

$$\gamma_E \le \frac{\mathbf{v}^T R K^{-1} R^T \mathbf{v}}{\mathbf{v}^T A^{-1} \mathbf{v}} \le \gamma_R, \qquad \forall \mathbf{v} \in V.$$

Proof. We omit the proof, see [NE7]. □

Remark 13.2. The preconditioning lemma above formally generalizes the equivalence between the Schur complement energy and the energy of the underlying matrix on a space of homogeneous solutions. Indeed, let $\mathbf{v} = \mathbf{v}_B \in V$ and $\mathbf{v}_* = (\mathbf{v}_I^T, \mathbf{v}_B^T)^T \in V_*$ denote nodal vectors. Let $R(\mathbf{v}_I^T, \mathbf{v}_B^T)^T = \mathbf{v}_B$ and let matrix K be block partitioned based on $\mathbf{v}_I$ and $\mathbf{v}_B$. Furthermore, if $E\mathbf{v}_B = (\mathbf{v}_I^T, \mathbf{v}_B^T)^T$ where $\mathbf{v}_I = -K_{II}^{-1} K_{IB} \mathbf{v}_B$, then the assumptions in the preceding lemma are equivalent to requiring that matrix A be spectrally equivalent to the Schur complement $(K_{BB} - K_{IB}^T K_{II}^{-1} K_{IB})$ of K.

13.2 Preconditioners for Neumann Problems

Consider the following Neumann problem on $\Omega \subset \Omega_*$ for $c(x) \geq 0$:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)u = f, & \text{in } \Omega \\ \mathbf{n} \cdot (a\,\nabla u) = g, & \text{on } \partial\Omega. \end{cases} \tag{13.6}$$

If $c(x) = 0$, then we shall assume the *compatibility* condition:

$$\int_\Omega f(x)\,dx + \int_{\partial\Omega} g(x)\,ds_x = 0.$$

We *imbed* Ω within a rectangular or periodic domain Ω_* and define $\Omega_1 = \Omega$, $\Omega_2 = (\Omega_* \setminus \Omega_1)$, $B = \partial\Omega_1 \cap \partial\Omega_2$. A discretization of (13.6) then yields:

$$A_{\mathrm{N}} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_B \end{bmatrix} \equiv \begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} \\ A_{IB}^{(1)^T} & A_{BB}^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B^{(1)} \end{bmatrix}, \tag{13.7}$$

where $\mathbf{u}_I^{(1)}$ and $\mathbf{u}_B$ correspond to nodal vectors of unknowns in the *interior* of $\Omega = \Omega_1$ and on $B = \partial\Omega_1 \cap \partial\Omega_2$, respectively.

On the extended region Ω_*, we consider the extended elliptic equation:

$$\begin{cases} -\nabla \cdot (a_*(x)\nabla u) + c_*(x)u = f, & \text{in } \Omega_* \\ u = 0, & \text{on } \partial\Omega_*, \end{cases} \tag{13.8}$$

where the extended coefficients $a_*(x)$ and $c_*(x)$ have the form:

$$a_*(x) = \begin{cases} a(x), & \text{for } x \in \Omega \\ \tilde{a}(x), & \text{for } x \in \Omega_* \setminus \Omega \end{cases} \quad \text{and} \quad c_*(x) = \begin{cases} c(x), & \text{for } x \in \Omega \\ \tilde{c}(x), & \text{for } x \in \Omega_* \setminus \Omega \end{cases} \tag{13.9}$$

with $\tilde{a}(x)$ and $\tilde{c}(x)$ chosen so that (13.8) has an efficient solver.

We denote a discretization of the extended elliptic equation (13.8) as:

$$\begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B \end{bmatrix}, \tag{13.10}$$

where the unknowns are partitioned according to nodes in the *interior* of Ω_1, *interior* of $\Omega_2 = (\Omega_* \setminus \Omega_1)$ and in $B = \partial\Omega_1 \cap \partial\Omega_2$. By construction, it should hold that $K_{II}^{(1)} = A_{II}^{(1)}$, $K_{IB}^{(1)} = A_{IB}^{(1)}$, $K_{BB} = A_{II}^{(1)} + K_{BB}^{(2)}$, and $\mathbf{f}_B = \mathbf{f}_B^{(1)} + \mathbf{f}_B^{(2)}$.

We shall consider a fictitious domain preconditioner M_{N} for the coefficient matrix A_{N} in (13.7), such that its inverse has the form $M_{\mathrm{N}}^{-1} = RK^{-1}R^T$, and which yields optimal order convergence with respect to mesh size h, see [BO5]. Matrix R will be a restriction and K the stiffness matrix in (13.10):

$$M_{\mathrm{N}}^{-1} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_3 \end{bmatrix} \equiv \begin{bmatrix} I & 0 \\ 0 & 0 \\ 0 & I \end{bmatrix}^T \begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_3 \end{bmatrix}.$$

Computing the action of M_{N}^{-1}, thus involves the solution of a linear system of the form (13.10) with right hand side $\mathbf{f}_I^{(1)} = \mathbf{g}_1$, $\mathbf{f}_I^{(2)} = \mathbf{0}$ and $\mathbf{f}_B = \mathbf{g}_3$. The following can be easily verified using a block factorization of K:

$$M_{\mathrm{N}} = \begin{bmatrix} K_{II}^{(1)} & K_{IB}^{(1)} \\ K_{IB}^{(1)^T} & K_{BB}^{(1)} + S^{(2)} \end{bmatrix},$$

where $S^{(2)} = (K_{BB}^{(2)} - K_{IB}^{(2)^T} K_{II}^{(2)^{-1}} K_{IB}^{(2)})$ denotes the Schur complement with respect to nodes in Ω_2 and B. Since $K_{II}^{(1)} = A_{II}^{(1)}$, $K_{IB}^{(1)} = A_{IB}^{(1)}$ and since $K_{BB}^{(1)} = A_{BB}^{(1)}$, preconditioner M_{N} will correspond to a modification of matrix A_{N}, by addition of the Schur complement $S^{(2)}$ to its lower diagonal block. The convergence rate will be of optimal order with respect to h.

Theorem 13.3. *The exists $C > 0$, independent of h, such that*

$$\mathrm{cond}(M_{\mathrm{N}}, A_N) \le C.$$

Proof. See [BO5]. □

Remark 13.4. In practice, there are computational issues for choosing a grid on Ω_1 which allows a fast solver on Ω_*. This is discussed at length in [BO5], where a triangulation algorithm is also described. Additionally, exact solvers for the extended stiffness matrix K can be replaced by *inexact* solvers based on a topologically equivalent grid.

13.3 Preconditioners for Dirichlet Problems

We shall next outline two preconditioners for the Dirichlet problem:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad u = 0, & \text{on } \partial\Omega. \end{cases} \tag{13.11}$$

We *imbed* Ω in a rectangular or periodic domain Ω_* and define $\Omega_1 = \Omega$, $\Omega_2 = (\Omega_* \setminus \Omega_1)$, $B = \partial\Omega_1 \cap \partial\Omega_2$. A discretization of (13.11) will then be:

$$A_{II}^{(1)} \mathbf{u}_I^{(1)} = \mathbf{f}_I^{(1)},$$

where we shall denote $A_{\mathrm{D}} = A_{II}^{(1)}$. On the extended region Ω_*, we shall pose elliptic equation (13.8) and employ its discretization (13.10). By construction, it will hold that $K_{II}^{(1)} = A_{\mathrm{D}}$.

An obvious first choice of fictitious domain preconditioner $\tilde{M}_{\mathrm{D}}$ for A_{D} is:

$$\tilde{M}_{\mathrm{D}}^{-1}\mathbf{f}_1 \equiv \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}^T \begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix}^{-1} \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} \mathbf{f}_1,$$

obtained by analogy with the Neumann preconditioner M_{N}. *Unfortunately*, $\mathrm{cond}(\tilde{M}_{\mathrm{D}}, A_{\mathrm{D}})$ grows as $\mathcal{O}(h^{-1})$, see [BO5]. Instead, we describe an alternative fictitious domain preconditioner [PR] for A_{D}. The Dirichlet preconditioner M_{D} of [PR] is motivated by the following block matrix identity.

Lemma 13.5. *Consider block matrix K in (13.10) with $K_{II}^{(1)} = A_{II}^{(1)} = A_{\mathrm{D}}$ and define a Schur complement matrix S as:*

$$S = K_{BB} - K_{IB}^{(1)^T} K_{II}^{(1)^{-1}} K_{IB}^{(1)} - K_{IB}^{(2)^T} K_{II}^{(2)^{-1}} K_{IB}^{(2)}.$$

Then the following identity will hold:

$$A_{\mathrm{D}}^{-1} = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}^T K^{-1} \left(\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S \end{bmatrix} K^{-1} \right) \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}, \qquad (13.12)$$

where

$$S^{-1} = \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}^T \begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}.$$

Proof. Follows from the block matrix factorization of K. □

A preconditioner M_{D} for A_{D} can be obtained by replacing the *action* of S by a *scaled* preconditioner M_{S} of S. We express it in symmetric form as:

$$M_{\mathrm{D}}^{-1} = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}^T K^{-1} \left(\begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M_{\mathrm{S}} \end{bmatrix} \right) K^{-1} \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}.$$

From the preceding expression, to ensure that $M_{\mathrm{D}} \geq 0$ we shall assume that $M_{\mathrm{S}} \leq S$ (which can be obtained by scaling M_{S}). Choices for M_{S} which do not employ inversion of subdomain stiffness matrices include square root of the discrete Laplace-Beltrami matrix on B and multilevel approximations. Below, we summarize an algorithm for computing the action of M_{D}^{-1}.

Algorithm 13.3.1 *(Capacitance Matrix Preconditioner $M_{\mathrm{D}}^{-1}\mathbf{f}_1$)*

1. *Solve:*
$$\begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_I^{(2)} \\ \mathbf{w}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$
2. *Compute* $\mathbf{g}_B = M_{\mathrm{S}}\mathbf{w}_B$
3. *Solve:*
$$\begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{g}_B \end{bmatrix}$$
4. *Define:* $M_{\mathrm{D}}^{-1}\mathbf{f}_1 \equiv \mathbf{w}_I^{(1)} - \mathbf{v}_I^{(1)}$

Theorem 13.6. *If $M_{\mathrm{S}} \asymp S$ and $M_{\mathrm{S}} \leq S$, it will hold that:*
$$\mathrm{cond}(M_{\mathrm{D}}, A_{\mathrm{D}}) \leq c,$$
where $c > 0$ is independent of h.

Proof. See [PR]. □

Remark 13.7. Another fictitious domain preconditioner M_{D} for $A_{\mathrm{D}} = A_{II}^{(1)}$ having the form $M_{\mathrm{D}} = R\,K^{-1}R^T$ is described in [NE7] for R defined as:
$$R \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_I^{(2)} \\ \mathbf{w}_B \end{bmatrix} = \mathbf{w}_I^{(1)} - C_{II}^{(1)^{-1}} C_{IB}^{(1)} \mathbf{w}_B,$$
where $C \asymp K$. An explicit and computationally efficient algorithm is described for computing the action $-C_{II}^{(1)^{-1}} C_{IB}^{(1)} \mathbf{w}_B$ which approximates a discrete harmonic extension of $\mathbf{w}_B$ into Ω_1 without requiring the inversion of $K_{II}^{(1)}$. Nonsymmetric capacitance matrix preconditioners for A_{D} are described in [BO5].

13.4 Lagrange Multiplier and Least Squares-Control Solvers

The Lagrange multiplier and least squares-control formulations provide useful frameworks for constructing fictitious domain solvers for elliptic equations [AT, NE7, DI2, DE4, GL12, PR, GL4, GL11]. The Lagrange multiplier formulation we describe will be applicable only to a self adjoint coercive elliptic equation with Dirichlet boundary conditions, while the least squares-control formulation will be applicable to general elliptic boundary value problems. Our discussions, however, will be restricted to the *discrete* case. We consider simply connected domains Ω_1, and omit discussion of *exterior* problems.

The Dirichlet problem we consider will be of the form:

$$\begin{cases} -\nabla \cdot (a(x)\nabla) + c(x)\, u = f, & \text{in } \Omega \\ \qquad\qquad\qquad\qquad u = g_{\mathrm{D}}, & \text{on } \partial\Omega, \end{cases} \tag{13.13}$$

for $c(x) \geq 0$. We shall denote its discretization on $\Omega_1 = \Omega$ as:

$$\begin{cases} A_{II}^{(1)} \mathbf{u}_I^{(1)} + A_{IB}^{(1)} \mathbf{u}_B = \mathbf{f}_I^{(1)} \\ \qquad\qquad\qquad \mathbf{u}_B = \mathbf{g}_{\mathrm{D}}, \end{cases} \tag{13.14}$$

where $\mathbf{g}_{\mathrm{D}}$ denotes a discretization of the Dirichlet boundary data.

The Neumann problem we consider will be of the form:

$$\begin{cases} -\nabla \cdot (a(x)\nabla u) + c(x)\, u = f, & \text{in } \Omega \\ \qquad\qquad \mathbf{n} \cdot a(x)\nabla u = g_{\mathrm{N}}, & \text{on } \partial\Omega. \end{cases} \tag{13.15}$$

When $c(x) = 0$, we shall assume $\int_\Omega f(x)\, dx + \int_{\partial\Omega} g_{\mathrm{N}}(x)\, ds_x = 0$, for compatability. We denote a discretization of the Neumann problem as:

$$\begin{bmatrix} A_{II}^{(1)} & A_{IB}^{(1)} \\ A_{IB}^{(1)^T} & A_{BB}^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_B^{(1)} \end{bmatrix}. \tag{13.16}$$

As before, $\Omega = \Omega_1$ will be *imbedded* within an extended region Ω_*, with $\Omega_2 \equiv \Omega_* \setminus \Omega_1$. On the extended domain Ω_*, we pose the elliptic equation:

$$\begin{cases} -\nabla \cdot (a_*(x)\nabla) + c_*(x)\, u = f_*(x), & \text{in } \Omega_* \\ \qquad\qquad\qquad\qquad\qquad u = 0, & \text{on } \partial\Omega_*, \end{cases} \tag{13.17}$$

where $a_*(x)$ and $c_*(x)$ are defined as in (13.9), while:

$$f_*(x) = \begin{cases} f(x), & \text{for } \; x \in \Omega \\ \tilde{f}(x), & \text{for } \; x \in \Omega_* \setminus \Omega. \end{cases} \tag{13.18}$$

A discretization of this extended elliptic equation will be denoted:

$$\begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B \end{bmatrix}, \qquad (13.19)$$

where $K_{II}^{(1)} = A_{II}^{(1)}$, $K_{IB}^{(1)} = A_{IB}^{(1)}$ and $K_{BB} = A_{BB}^{(1)} + K_{BB}^{(2)}$.

Lagrange Multiplier Formulation. Our description of the Lagrange multiplier formulation for fictitious domain problems will focus only on a *matrix version* of an algorithm of [DI2]. We will seek the solution to the discretized *Dirichlet* problem (13.14) on Ω_1, by imbedding this system within a larger linear system involving the discretization (13.19) of the extended equation (13.17) on $\Omega_* \supset \Omega_1$. The imbedding may be motivated by the observation that if $\mathbf{u}_B = \mathbf{g}_D$ in (13.19), then the first block row of (13.19) yields $\mathbf{u}_I^{(1)} = K_{II}^{(1)^{-1}} \left(\mathbf{f}_I^{(1)} - K_{IB}^{(1)} \mathbf{g}_D \right)$ which is the desired solution to (13.14) for $K_{II}^{(1)} = A_{II}^{(1)}$ and $K_{IB}^{(1)} = A_{IB}^{(1)}$. We now describe a *constrained* minimization problem which yields a system with $\mathbf{u}_B = \mathbf{g}_D$.

Accordingly, consider the energy $J(\cdot)$ associated with system (13.19):

$$J(\mathbf{v}) = \frac{1}{2} \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B \end{bmatrix}^T \begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B \end{bmatrix} - \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B \end{bmatrix}^T \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B \end{bmatrix},$$

where $\mathbf{v} = \left(\mathbf{v}_I^{(1)^T}, \mathbf{v}_I^{(2)^T}, \mathbf{v}_B^T \right)^T$. We shall assume K and $A_{II}^{(1)}$ are symmetric positive definite matrices of size m and $n < m$, respectively. Seeking the minimum of $J(\mathbf{v})$ subject to a *constraint* of the form $M\,\mathbf{v}_B = M\,\mathbf{g}_D$, (where M denotes a mass or identity matrix of size n) will yield a saddle point linear system as in Chap. 10. Indeed, define a Lagrangian functional:

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = J(\mathbf{w}) + \boldsymbol{\lambda}^T M \left(\mathbf{w}_B - \mathbf{g}_D \right),$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ denotes a vector of Lagrange multipliers which enforce the constraint $M \mathbf{w}_B = \mathbf{g}_D$. As in Chap. 10, the saddle point of $\mathcal{L}(.,.)$ will solve:

$$\begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} & 0 \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} & 0 \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} & M^T \\ 0 & 0 & M & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_B \\ M \mathbf{g}_D \end{bmatrix}. \qquad (13.20)$$

Solving the fourth block row of (13.20) yields $\mathbf{v}_B = \mathbf{g}_D$, and substituting this into the first block row of (13.20) yields $K_{IB}^{(1)} \mathbf{v}_I^{(1)} + K_{IB}^{(1)} \mathbf{g}_D = \mathbf{f}_I^{(1)}$. This has the same solution $\mathbf{v}_I^{(1)}$ as (13.14), and can be sought by solving (13.20).

Saddle point system (13.20) can be solved using any of the iterative algorithms from Chap. 10. Such algorithms require either an efficient solver for K or an efficient preconditioner for K (this will hold by our assumption on the choice of fictitious domain), and an efficient preconditioner for the saddle point Schur complement matrix T:

$$T = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ M \end{bmatrix}^T \begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ M \end{bmatrix}.$$

It is easily verified that $T \equiv M^T S^{-1} M$, where:

$$S = \left(K_{BB} - K_{IB}^{(1)^T} K_{II}^{(1)^{-1}} K_{IB}^{(1)} - K_{IB}^{(2)^T} K_{II}^{(2)^{-1}} K_{IB}^{(2)} \right),$$

denotes a two subdomain (domain decomposition) Schur complement matrix. Since M is a mass or identity matrix, it will follow that T is spectrally equivalent to S^{-1}, and thus, the action of T^{-1} may be approximated by multiplication by any suitable preconditioner M_S for S. We omit the details. See [GL4, GL11] for alternative Lagrange multiplier methods.

Least Squares-Control Formulation. The preceding method does not generalize to Neumann problems (in any obvious way), nor to non-self adjoint problems. However, a least squares-control method can be formulated for such problems. We outline *matrix versions* of two families of least squares-control fictitious domain methods [DE4, GL12]. The first least squares-control fictitious domain method we describe corresponds to a *weighted* residual least squares method, while the second method to a constrained least squares method involving a weighted residual on the subdomain boundary. We describe both algorithms in matrix terms, and assume that the matrices are *nonsingular*.

Weighted Residual Least Squares. The algorithm of [DE4] seeks the solution to a linear system $A\mathbf{u} = \mathbf{f}$ of size n, corresponding to the discretization of an elliptic equation on Ω_1 with *Dirichlet* or *Neumann* boundary conditions on $\partial\Omega_1$, by a weighted residual method. Let K denote the extended stiffness matrix of size $m > n$, associated with the elliptic equation on $\Omega_* \supset \Omega_1$. We assume that an efficient solver is available for K. Let $E : \mathbb{R}^n \to \mathbb{R}^m$ denote an extension map which extends a nodal vector on Ω_1 by *zero* to $\Omega_* \setminus \Omega_1$. Let $\mathbf{y} \in \mathbb{R}^m$ denote a *weighted* residual:

$$K\,\mathbf{y} + E\,(A\mathbf{u} - \mathbf{f}) = \mathbf{0},$$

so that when $A\mathbf{u} = \mathbf{f}$, variable $\mathbf{y} = \mathbf{0}$. Here $\mathbf{u} \in \mathbb{R}^n$ corresponds to a *control* variable and $\mathbf{y} \in \mathbb{R}^m$ to a *state* variable.

Using the preceding, a solution to $A\mathbf{u} = \mathbf{f}$ may be sought by minimizing a nonnegative quadratic functional $J(\mathbf{y}) = \mathbf{y}^T H \mathbf{y}$ (for any matrix $H = H^T > 0$ of size m) subject to the following constraint:

$$J(\mathbf{y}_*) = \min_{\mathbf{y} \in \mathcal{V}} J(\mathbf{y}) \tag{13.21}$$

where

$$\mathcal{V} = \{\mathbf{y} \in \mathbb{R}^m \,:\, K\,\mathbf{y} + E\,(A\mathbf{v} - \mathbf{f}) = \mathbf{0},\ \forall \mathbf{v} \in \mathbb{R}^n\}\,.$$

By construction, the minimum will occur when $\mathbf{y}_* = \mathbf{0}$ and $A\mathbf{v} = \mathbf{f}$. We may *parameterize* the space $\mathcal{V}$ in terms of the control variables as follows:

$$\mathcal{V} = \left\{ K^{-1}E\,(\mathbf{f} - A\mathbf{v}) \,:\, \text{for } \mathbf{v} \in \mathbb{R}^n \right\}.$$

Substituting this parameterization, we may reduce the *constrained* minimization problem (13.21) to the following *unconstrained* minimization problem:

$$J_*(\mathbf{u}) = \min_{\mathbf{v} \in \mathbb{R}^n} J_*(\mathbf{v}) \tag{13.22}$$

where $J_*(\mathbf{v}) \equiv J\left(K^{-1}E(\mathbf{f} - A\mathbf{v})\right)$. At the minimum, $A\mathbf{u} = \mathbf{f}$. Applying the first order derivative test $\nabla J_* = \mathbf{0}$ for a minimum yields:

$$\left(A^T E^T K^{-T} H K^{-1} E A\right) \mathbf{u} = \mathbf{f}_*, \tag{13.23}$$

where $\mathbf{f}_* \equiv \left(A^T E^T K^{-T} H K^{-1} E\right) \mathbf{f}$. This system will be symmetric positive definite and can be solved by a preconditioned conjugate gradient method. Each matrix vector product with matrix $\left(A^T E^T K^{-T} H K^{-1} E A\right)$ will require computing the action of K^{-1} and K^{-T}. Here, matrix $H = H^T > 0$ of size m can be regarded as a preconditioner to reduce the condition number.

Remark 13.8. In applications, block matrix K in (13.19) can be used for *Dirichlet* and *Neumann* boundary value problems. In this case, the matrices $A = A_{\mathrm{D}}$, E_{D}, $A = A_{\mathrm{N}}$ and E_{N} can be chosen as follows:

$$A_{\mathrm{D}} = K_{II}^{(1)},\ E_{\mathrm{D}} = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad A_{\mathrm{N}} = \begin{bmatrix} K_{II}^{(1)} & K_{IB}^{(1)} \\ K_{IB}^{(1)^T} & K_{BB}^{(1)} \end{bmatrix},\ E_{\mathrm{N}} = \begin{bmatrix} I & 0 \\ 0 & 0 \\ 0 & I \end{bmatrix},$$

to solve the discretized *Dirichlet* problem (13.14) on Ω_1 and the discretized *Neumann* problem (13.16) on Ω_1, respectively. The vectors $\mathbf{f} = \mathbf{f}_I^{(1)}$ for the Dirichlet problem and $\mathbf{f} = (\mathbf{f}_I^{(1)^T}, \mathbf{f}_B^{(1)^T})^T$ for the Neumann problem.

Boundary Residual Least Squares. We next describe the matrix version of a control algorithm of [GL12] involving a *boundary functional*. It can be formulated to solve either a discretized *Dirichlet* or *Neumann* problem, but to be specific, we shall consider the following discretized Dirichlet problem:

$$\begin{cases} A^{(1)}_{II}\mathbf{u}^{(1)}_I + A^{(1)}_{IB}\mathbf{u}_B = \mathbf{f}^{(1)}_I \\ \mathbf{u}_B = \mathbf{g}_{\mathrm{D}}. \end{cases} \tag{13.24}$$

The matrix version of such algorithms employ constraints of the form:

$$\begin{bmatrix} K^{(1)}_{II} & 0 & K^{(1)}_{IB} \\ 0 & K^{(2)}_{II} & K^{(2)}_{IB} \\ K^{(1)^T}_{IB} & K^{(2)^T}_{IB} & K_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}^{(1)}_I \\ \mathbf{v}^{(2)}_I \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)}_I \\ \boldsymbol{\beta}^{(2)}_I \\ \boldsymbol{\beta}_B \end{bmatrix}, \tag{13.25}$$

where $\boldsymbol{\beta}^{(2)}_I = \mathbf{0}$ and $K^{(1)}_{II} = A^{(1)}_{II}$, $K^{(1)}_{IB} = A^{(1)}_{IB}$ and $K^{(1)}_{BB} = A^{(1)}_{BB}$ in stiffness matrix K associated with the extended elliptic equation.

A control algorithm for solving (13.24) can be motivated as follows. Let $\mathbf{f}^{(1)}_I$ in (13.25) be the same as in system (13.24). Then, if $\mathbf{v}_B = \mathbf{g}_{\mathrm{D}}$, the first block row of (13.25) yields $\mathbf{v}^{(1)}_I = \mathbf{u}^{(1)}_I$. This suggests minimizing $\|\mathbf{v}_B - \mathbf{g}_{\mathrm{D}}\|^2$ subject to a constraint of the form (13.25). Then, by construction, at the minimum we should obtain $\mathbf{v}^{(1)}_I = \mathbf{u}^{(1)}_I$ and $\mathbf{v}_B = \mathbf{g}_{\mathrm{D}}$.

More specifically, let n_1 be the size of $\mathbf{v}^{(1)}_I$, n_2 the size of $\mathbf{v}^{(2)}_I$, and n_3 the size of $\mathbf{v}_B$, with $n = (n_1 + n_2 + n_3)$. Given an extended nodal vector $\mathbf{v} = \left(\mathbf{v}^{(1)^T}_I, \mathbf{v}^{(2)^T}_I, \mathbf{v}^T_B\right)^T$ and a symmetric positive definite matrix H of size n_3, define a *boundary energy* functional $J(\mathbf{v})$:

$$J(\mathbf{v}) = (\mathbf{v}_B - \mathbf{g}_B)^T H (\mathbf{v}_B - \mathbf{g}_B).$$

The preceding least squares control formulation of (13.24) will seek:

$$J(\mathbf{u}) = \min_{\mathbf{v}\in\mathcal{V}} J(\mathbf{v}) \tag{13.26}$$

where the constraint set $\mathcal{V}$ is defined as:

$$\mathcal{V} = \left\{ \mathbf{v} \in \mathbb{R}^n \,:\, (\mathbf{v}^{(1)^T}_I, \mathbf{v}^{(2)^T}_I, \mathbf{v}^T_B)^T \text{ satisfies (13.25)} \right\}.$$

Since (13.24) is solvable, we should obtain $J(\cdot) = 0$ as the minimum, yielding $\mathbf{u}_B = \mathbf{g}_{\mathrm{D}}$. A solution to (13.24) can thus be obtained by solving (13.26). Importantly, we may set $\boldsymbol{\beta}^{(2)}_I = \mathbf{0}$, as it will only yield extra control parameters. Then, $\mathcal{V}$ can be parameterized in terms of $\boldsymbol{\beta}_B \in \mathbb{R}^{n_3}$ as:

$$\mathcal{V} = \left\{ \begin{bmatrix} \mathbf{v}^{(1)}_I \\ \mathbf{v}^{(2)}_I \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} K^{(1)}_{II} & 0 & K^{(1)}_{IB} \\ 0 & K^{(2)}_{II} & K^{(2)}_{IB} \\ K^{(1)^T}_{IB} & K^{(2)^T}_{IB} & K_{BB} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}^{(1)}_I \\ \mathbf{0} \\ \boldsymbol{\beta}_B \end{bmatrix} \,:\, \forall \boldsymbol{\beta}_B \in \mathbb{R}^{n_3} \right\}.$$

An *unconstrained* minimization problem equivalent to (13.26) will thus be:

$$J_*(\mathbf{u}_B) = \min_{\boldsymbol{\beta}_B \in \mathbb{R}^{n_3}} J_*(\boldsymbol{\beta}_B) \tag{13.27}$$

where $J_*(\boldsymbol{\beta}_B) = J(F\,\boldsymbol{\beta}_B + \boldsymbol{\gamma})$ is defined using the parametric representation:

$$\begin{bmatrix} \mathbf{v}_I^{(1)} \\ \mathbf{v}_I^{(2)} \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} K_{II}^{(1)} & 0 & K_{IB}^{(1)} \\ 0 & K_{II}^{(2)} & K_{IB}^{(2)} \\ K_{IB}^{(1)^T} & K_{IB}^{(2)^T} & K_{BB} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{0} \\ \boldsymbol{\beta}_B \end{bmatrix} \equiv F\,\boldsymbol{\beta}_B + \boldsymbol{\gamma},$$

with

$$F\,\boldsymbol{\beta}_B \equiv K^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{\beta}_B \end{bmatrix} \quad \text{and} \quad \boldsymbol{\gamma} \equiv \begin{bmatrix} \boldsymbol{\gamma}_I^{(1)} \\ \boldsymbol{\gamma}_I^{(2)} \\ \boldsymbol{\gamma}_B \end{bmatrix} \equiv K^{-1} \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

Here F will be a matrix of size $n \times n_3$ and $\boldsymbol{\gamma}$ a vector of size n. Substituting the above parameterization and applying a first order derivative test for determining the minimum of $J_*(\cdot)$ will yield the linear system:

$$\left(F^T R^T H R F\right) \mathbf{u}_B = \left(F^T R^T H\right) \mathbf{g}_{\mathrm{D}}. \tag{13.28}$$

Here R denotes a *restriction* matrix $R = \begin{bmatrix} 0 & 0 & I \end{bmatrix}$. A conjugate gradient algorithm may be employed to solve system (13.28), using a suitable preconditioner H. It is easily verified that $RF = RK^{-1}R^T = S^{-1}$ where:

$$S = \left(K_{BB} - K_{IB}^{(1)^T} K_{II}^{(1)^{-1}} K_{IB}^{(1)} - K_{IB}^{(2)^T} K_{II}^{(2)^{-1}} K_{IB}^{(2)}\right),$$

is a two subdomain Schur complement. Thus, H may be chosen as a discrete Laplace-Beltrami matrix on B. We omit further details.

Remark 13.9. A constrained minimization problem similar to (13.26) can also be developed for solving the discretized Neumann problem (13.16). In this case, the boundary functional could be defined as:

$$J(\mathbf{v}) = \left\| \left(A_{IB}^{(1)^T} \mathbf{v}_I^{(1)} + A_{BB}^{(1)} \mathbf{v}_B\right) - \mathbf{g}_{\mathrm{N}} \right\|_H^2,$$

for a suitably chosen positive definite matrix H of size n_3. The least squares-control problem will then seek the minimum of $J(\cdot)$ subject to the constraint $\mathcal{V}$ as before. Using the same parameterization of $\mathcal{V}$, an unconstrained minimization problem can be obtained, yielding a linear system of the form (13.28) with matrix R replaced by:

$$R = \begin{bmatrix} A_{IB}^{(1)^T} & 0 & A_{BB}^{(1)} \end{bmatrix}.$$

We omit further details.

14

Variational Inequalities and Obstacle Problems

In this chapter, we describe traditional and domain decomposition Schwarz algorithms for iteratively solving obstacle problems. In an obstacle problem, an elliptic (or parabolic) equation or inequality is posed on a domain, however, the desired solution is constrained to lie above a specified function, referred to as an obstacle [CR2, GL, FR6, KI4]. Applications arise in elasticity theory [GL10, GL], heat conduction (Stefan problems) and mathematical finance (option pricing) [CR2, EL, WI10, WI11]. Even when the underlying elliptic (or parabolic) equation is linear, an obstacle problem is *nonlinear* due to the unknown region of *contact* between the solution and the obstacle. However, once the contact set is known, the problem is linear on its complementary set.

Our discussion will focus on variational inequalities which arise from *scalar* elliptic equations with obstacle constraints. Its discretization yields a *linear complementarity* algebraic problem. We describe algorithms for solving such problems iteratively, both when the elliptic equation is self adjoint and coercive and when it is non-self adjoint. Chap. 14.1 describes properties of variational inequalities and their discretizations, and a projection theorem onto *convex* sets. Chap. 14.2 describes the *gradient* and *relaxation* methods for iteratively solving linear complementarity problems. Schwarz linear complementarity algorithms, see [LI6, HO3, KU12, KU13, ZE, LI10, BA10, TA2, TA3], are described in Chap. 14.3. Chap. 14.4. discusses the maximum norm convergence of Schwarz linear complementarity algorithms. Chap. 14.5 briefly discusses extensions to parabolic variational inequalities.

14.1 Background

Let L denote an elliptic operator defined on a domain Ω:

$$L\,u \equiv -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u, \tag{14.1}$$

where the coefficients $a(x) \geq a_0 > 0$, $\mathbf{b}(x)$ and $c(x) \geq 0$ are smooth. For simplicity, when $\mathbf{b}(x) \neq \mathbf{0}$, we assume $c(x) \geq c_0 > 0$. Below, we describe background on the *continuous* and *discrete* versions of an obstacle problem.

Continuous Obstacle Problem. Let $\psi(x) \in C^1(\overline{\Omega})$ be a given function, referred to as an *obstacle function*, and let $f(x) \in C(\overline{\Omega})$, $g(x) \in C(\partial\Omega)$ denote forcing and boundary terms, respectively. Then, an obstacle problem formally seeks $u(x) \in C^2(\overline{\Omega})$ satisfying:

$$\begin{cases} L\,u(x) - f(x) \geq 0, & \text{a.e. in } \Omega \\ u(x) - \psi(x) \geq 0, & \text{a.e. in } \Omega \\ (u(x) - \psi(x))\,(Lu(x) - f(x)) = 0, & \text{a.e. in } \Omega \\ u(x) = g(x), & \text{on } \partial\Omega. \end{cases} \tag{14.2}$$

Since $(L\,u(x) - f(x)) \geq 0$ and since $(u(x) - \psi(x)) \geq 0$, the requirement that:

$$(L\,u(x) - f(x))\,(u(x) - \psi(x)) = 0, \qquad \text{on } \Omega,$$

can be equivalently stated that, *if* $u(x) > \psi(x)$, *then* $L\,u(x) = f(x)$. For compatability of the boundary data with the obstacle, we shall require that $\psi|_{\partial\Omega} \leq g(x)$. However, for simplicity we assume $g(x) = 0$, in which case this reduces to $\psi|_{\partial\Omega} \leq 0$. See Fig. 14.1 for an illustration of $u(x)$ and $\psi(x)$.

Definition 14.1. *Given a solution $u(x)$ to (14.2), we define its contact set as $G \equiv \{x \in \Omega : u(x) = \psi(x)\}$ for the obstacle problem. The boundary ∂G of the contact set is referred to as a free boundary, and is generally unknown. However, when G is known, $u(x)$ can formally be determined by solving:*

$$L\,u = f, \quad \text{in} \quad (\Omega \backslash G) \quad \text{with } u = \psi \ \text{ on } \ (\partial G \cap \Omega) \ \text{ and } u = g \ \text{ on } \partial\Omega.$$

Let $\mathcal{K}$ denote a *closed, convex* subset of $H_0^1(\Omega)$ satisfying:

$$\mathcal{K} \equiv \left\{ u \in H_0^1(\Omega) : u(x) \geq \psi(x), \quad \text{a.e. in } \Omega \right\}. \tag{14.3}$$

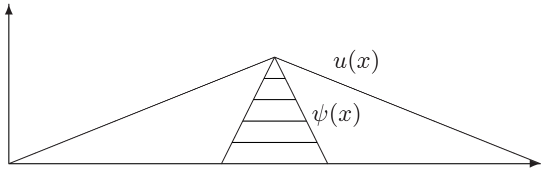


Fig. 14.1. A triangular obstacle function $\psi(x)$ on a one dimensional domain

We note that the second requirement in (14.2) justifies seeking $u(\cdot) \in \mathcal{K}$. By definition of the contact set G, for any $v(\cdot) \in \mathcal{K}$ it must hold that $(v(x) - u(x)) \geq 0$ on G. Multiplying $(L\, u(x) - f(x))$ by $(v(x) - u(x))$ and integrating yields:

$$\begin{aligned} \int_{\Omega}(Lu - f)(v - u)\, dx &= \int_{G}(Lu - f)(v - u)\, dx + \int_{\Omega\backslash G}(Lu - f)(v - u)\, dx \\ &= \int_{G}(Lu - f)(v - u)\, dx \, \geq \, 0, \end{aligned}$$

since $(L\, u(x) - f(x)) \geq 0$ on Ω and since $(L\, u(x) - f(x)) = 0$ on $(\Omega \backslash G)$.

Integrating the preceding expression by parts yields a weak version of the obstacle problem, which seeks $u(x) \in \mathcal{K}$ satisfying:

$$a(u, v - u) - (f, v - u) \geq 0, \qquad \forall v \in \mathcal{K}, \tag{14.4}$$

where

$$\begin{cases} a(u, v) \equiv \int_{\Omega} \left(a(x)\nabla u \cdot \nabla v + \mathbf{b}(x) \cdot \nabla u\, v + c(x)\, u\, v\right)\, dx, \\ \;(f, w) \equiv \int_{\Omega} f(x)\, w(x)\, dx. \end{cases}$$

Inequality (14.4) is referred to as a *variational inequality* [GL, KI4]. When $\mathbf{b}(x) = 0$ and $c(x) \geq 0$, the variational inequality (14.4) can be shown to be equivalent to the following *constrained* minimization problem [GL, KI4]:

$$J(u) = \min_{w \in \mathcal{K}} J(w) \tag{14.5}$$

where the energy is defined $J(w) \equiv \frac{1}{2}\, a(w, w) - (f, w)$ for $w \in H^1(\Omega)$. The existence and uniqueness of solutions to this constrained optimization problem can be found in [GL, KI4]. In the following, we verify that the differential version (14.2) of the obstacle problem can be derived from (14.4).

Lemma 14.2. *Let $u \in \mathcal{K}$ be a sufficiently smooth solution of the variational inequality (14.4). Then $u(x)$ will satisfy the following:*

1. *$u(x) \geq \psi(x)$ a.e. in Ω.*
2. *$(L\, u(x) - f(x)) \geq 0$ a.e. in Ω.*
3. *$(u(x) - \psi(x))\; (L\, u(x) - f(x)) = 0$ a.e. in Ω.*

Proof. Since $u \in \mathcal{K}$, the first inequality holds by definition of $\mathcal{K}$. To prove the second inequality, choose a *nonnegative* function $\phi \in H_0^1(\Omega)$ and define:

$$v(x) = u(x) + \phi(x), \qquad \text{where} \quad \phi(x) \geq 0, \quad x \in \Omega.$$

Since $u(x) \geq \psi(x)$ and $\phi(x) \geq 0$ it follows that $v(x) = u(x) + \phi(x) \geq \psi(x)$ and so $v(\cdot) \in \mathcal{K}$. Substituting $v(\cdot)$ into the variational inequality, we obtain:

$$\begin{aligned} 0 &\leq a(u, v - u) - (f, v - u) \\ &= a(u, \phi) - (f, \phi) \\ &= \int_{\Omega} \left(L\, u(x) - f(x)\right)\, \phi(x)\, dx, \end{aligned}$$

where the last line follows by integration by parts when $u(x)$ is sufficiently smooth, since $\phi(\cdot) \in H_0^1(\Omega)$. Thus, $\int_\Omega (L\,u(x) - f(x))\,\phi(x)\,dx \geq 0$ for smooth nonnegative test functions $\phi(x)$, yielding $(L\,u(x) - f(x)) \geq 0$ a.e. in Ω.

To prove the third item, suppose that $u(x) > \psi(x)$ for $x \in N(x_0)$, some open set containing x_0. Then, choose any sufficiently smooth nonpositive test function $0 \geq \phi(x) \in H_0^1(\Omega)$ with support in $N(x_0)$ and *negative* at x_0, yet such that $v(x) = u(x) + \phi(x) \geq \psi(x)$. For this choice of $\phi(\cdot)$, we obtain:

$$\begin{aligned} 0 &\leq a(u, v-u) - (f, v-u) \\ &= a(u,\phi) - (f,\phi) \\ &= \int_{N(x_0)} (L\,u(x) - f(x))\,\phi(x)\,dx \leq 0, \end{aligned}$$

since $(L\,(x) - f(x)) \geq 0$ and $\phi(x) \leq 0$. Thus $a(u,\phi) - (f,\phi) = 0$ and since $\phi(x_0) < 0$, this can hold only if $(L\,u(x) - f(x)) = 0$ at x_0. Thus, we have established that if $u(x_0) > \psi(x_0)$ then $(L\,u(x_0) - f(x_0)) = 0$. □

Discrete Obstacle Problem. A discretization of the obstacle problem can be obtained using either a finite element or finite difference method. For definiteness, we consider a *finite difference* discretization. Let $\mathcal{T}_h(\Omega)$ denote a triangulation of Ω with grid size h and interior nodes $x_1, \ldots, x_n$. We shall denote the linear system corresponding to $L\,u = f$ in Ω with $u = 0$ on $\partial\Omega$ as $A\mathbf{u} = \mathbf{f}$, where $\mathbf{u} \in \mathbb{R}^n$ denotes a vector of nodal unknowns with $(\mathbf{u})_i$ approximating $u(x_i)$ and $\mathbf{f} \in \mathbb{R}^n$ denoting the discrete forcing term. We let $\boldsymbol{\psi} \in \mathbb{R}^n$ denote a nodal vector corresponding to the obstacle function at the interior nodes, with $(\boldsymbol{\psi})_i = \psi(x_i)$.

A discretization of the obstacle problem can be obtained either by discretization of (14.2) or by discretization of (14.4). The obstacle constraint $u(x) \geq \psi(x)$ will be discretized as $\mathbf{u} \geq \boldsymbol{\psi}$, where the inequality between the two nodal vectors apply *component wise*, i.e., with $(\mathbf{u})_i \geq (\boldsymbol{\psi})_i$ for $1 \leq i \leq n$. A discretization of (14.2) will seek $\mathbf{u} \in \mathbb{R}^n$ satisfying:

$$\begin{cases} \mathbf{u} - \boldsymbol{\psi} \geq \mathbf{0} \\ A\mathbf{u} - \mathbf{f} \geq \mathbf{0} \\ (\mathbf{u} - \boldsymbol{\psi})^T (A\,\mathbf{u} - \mathbf{f}) = 0. \end{cases} \tag{14.6}$$

This problem is referred to as a *linear complementarity* problem. Since $\mathbf{u} \geq \boldsymbol{\psi}$ and $(A\mathbf{u} - \mathbf{f}) \geq \mathbf{0}$ are taken *component wise*, it will follow that if $\mathbf{u}_i > \boldsymbol{\psi}_i$, then $(A\mathbf{u})_i = \mathbf{f}_i$ and that if $\mathbf{u}_i = \boldsymbol{\psi}_i$, then $(A\mathbf{u})_i \geq \mathbf{f}_i$.

Definition 14.3. *The discrete constraint set $\mathcal{K}_h$ is defined as:*

$$\mathcal{K}_h = \{\mathbf{v} \in \mathbb{R}^n \,:\, (\mathbf{v})_i \geq (\boldsymbol{\psi})_i, \quad \textit{for } 1 \leq i \leq n\},$$

and it is easily verified to be a closed, convex set.

Remark 14.4. We leave it to the reader to verify that linear complementarity problem (14.6) can be expressed equivalently as seeking $\mathbf{u} \in \mathcal{K}_h$ such that:

$$(\mathbf{v} - \mathbf{u})^T (A\,\mathbf{u} - \mathbf{f}) \geq 0, \quad \forall \mathbf{v} \in \mathcal{K}_h. \tag{14.7}$$

This corresponds to a discretization of the variational inequality (14.4). When matrix $A = A^T > 0$ (if $\mathbf{b}(x) = \mathbf{0}$ and $c(x) \geq 0$), the linear complementarity problem (14.7) will be equivalent to a constrained minimization problem.

Lemma 14.5. *Suppose the following conditions hold.*

1. *Let A be a symmetric positive definite matrix of size n and let $J(\mathbf{v})$ denote the quadratic function associated with $A\mathbf{u} = \mathbf{f}$ for $\mathbf{f} \in \mathbb{R}^n$:*

$$J(\mathbf{v}) \equiv \frac{1}{2}\mathbf{v}^T A \mathbf{v} - \mathbf{v}^T \mathbf{f}. \tag{14.8}$$

2. *Let $\mathbf{u}$ denote the minimum of $J(\cdot)$ within the closed, convex set $\mathcal{K}_h$:*

$$J(\mathbf{u}) = \min_{\mathbf{v} \in \mathcal{K}_h} J(\mathbf{v}) \tag{14.9}$$

Then, the following result will hold.

$$(\mathbf{v} - \mathbf{u})^T (A\mathbf{u} - \mathbf{f}) \geq 0, \quad \forall \mathbf{v} \in \mathcal{K}_h.$$

Proof. Given $\mathbf{u} \in \mathcal{K}_h$ satisfying (14.9), choose $\mathbf{v} \in \mathcal{K}_h$. By convexity of $\mathcal{K}_h$:

$$(1-t)\,\mathbf{u} + t\,\mathbf{v} = \mathbf{u} + t\,(\mathbf{v} - \mathbf{u}) \in \mathcal{K}_h, \quad \text{for} \;\; t \in [0,1].$$

By assumption, $J(\mathbf{u}) \leq J(\mathbf{u} + t\,(\mathbf{v} - \mathbf{u}))$ for $0 \leq t \leq 1$, so we must have:

$$0 \leq \left. \frac{dJ(\mathbf{u} + t(\mathbf{v} - \mathbf{u}))}{dt} \right|_{t=0} = (\mathbf{v} - \mathbf{u})^T A\mathbf{u} - (\mathbf{v} - \mathbf{u})^T \mathbf{f}, \quad \forall \mathbf{v} \in \mathcal{K}_h.$$

This yields the desired result. □

The preceding result shows that when A is symmetric positive definite, the variational inequality version (14.7) of the linear complementarity problem (14.6) is equivalent to the constrained minimization version (14.9). The following result concerns the uniqueness of the solution in this case.

Lemma 14.6. *Let $A = A^T > 0$ be of size n and let $\mathbf{w} \in \mathcal{K}_h$ satisfy:*

$$(\mathbf{v} - \mathbf{w})^T (A\mathbf{w} - \mathbf{f}) \geq 0, \quad \forall \mathbf{v} \in \mathcal{K}_h.$$

Then $\mathbf{w}$ is unique.

Proof. To show uniqueness of the solution, suppose there were two distinct solutions $\mathbf{u}$, $\mathbf{w} \in \mathcal{K}_h$ of the discrete variational inequality above. Given solution $\mathbf{u}$, substitute $\mathbf{v} = \mathbf{w}$ to obtain:

$$(A\mathbf{u} - \mathbf{f}, \mathbf{w} - \mathbf{u}) \geq 0.$$

When $\mathbf{w}$ is the solution, substitute $\mathbf{v} = \mathbf{u}$ and reverse signs to obtain:

$$(-A\mathbf{w} + \mathbf{f}, \mathbf{w} - \mathbf{u}) \geq 0.$$

Adding this with the preceding expression and reversing signs yields:

$$(\mathbf{w} - \mathbf{u})^T A (\mathbf{w} - \mathbf{u}) \leq 0.$$

Since $A = A^T > 0$ this yields that $\|\mathbf{u} - \mathbf{w}\|^2 = 0$, establishing uniqueness. □

Remark 14.7. When matrix A is not symmetric positive definite, the linear complementarity problem (14.6) will *not* have a minimization interpretation.

Projection Theorem onto Convex Sets. We next state an abstract *projection* theorem, onto a *closed, convex* set $\mathcal{K}$ in a Hilbert space V, see [CI4]. This result will be used when formulating the projected gradient method.

Theorem 14.8. *Suppose the following conditions hold.*

1. *Let V denote a Hilbert space with inner product $(\cdot,\cdot)_V$ and norm $\|\cdot\|_V$.*
2. *Let $\mathcal{K}$ be a closed convex set in V.*
3. *Given $w \in V$, let $d(w,\mathcal{K})$ denote the distance between w and $\mathcal{K}$:*

$$d(w,\mathcal{K}) \equiv \inf_{v \in \mathcal{K}} \|v - w\|_V.$$

Then, the following results will hold:

1. *For any $w \in V$, there exists a unique element $P_{\mathcal{K}}w \in \mathcal{K}$ closest to w:*

$$\|w - P_{\mathcal{K}}w\|_V = \inf_{v \in \mathcal{K}} \|v - w\|_V.$$

2. *The element $P_{\mathcal{K}}w \in \mathcal{K}$ is characterized by the variational inequality:*

$$(P_{\mathcal{K}}w - w, v - P_{\mathcal{K}}w)_V \geq 0, \quad \forall v \in \mathcal{K}. \tag{14.10}$$

3. *The projection operator: $P_{\mathcal{K}} : V \to \mathcal{K}$ is nonexpansive, i.e.,*

$$\|P_{\mathcal{K}}w - P_{\mathcal{K}}v\|_V \leq \|w - v\|_V.$$

4. *$P_{\mathcal{K}}$ is linear if and only if $\mathcal{K}$ is a subspace of V.*

Proof. For a complete proof, see [CI4]. We shall only prove 2 (the variational inequality characterization of $P_{\mathcal{K}}w$). Given $w \in V$, let $w_* \in \mathcal{K}$ be the closest vector in $\mathcal{K}$ to w, i.e.,

$$\|w - w_*\|_V = \inf_{v \in \mathcal{K}} \|v - w\|_V.$$

The existence of a closest element is proved in [CI4], and will be valid only when $\mathcal{K}$ is closed). The vector w_* can be expressed as:

$$J_w(w_*) = \min_{v \in \mathcal{K}} J_w(v),$$

where $J_w(v) \equiv \frac{1}{2}(v - w, v - w)_V$ is a quadratic functional. Given $w_* \in \mathcal{K}$ and $v \in \mathcal{K}$, consider $\theta v + (1-\theta)w_* = w_* + \theta(v - w_*) \in \mathcal{K}$ for $\theta \in [0,1]$, since $\mathcal{K}$ is convex. Since the minimum of $J_w(\cdot)$ in $\mathcal{K}$ occurs at w_*, we obtain:

$$J_w(w_*) \leq J_w(w_* + \theta(v - w_*)), \quad \text{for } \theta \in [0,1].$$

Using the definition of $J_w(\cdot)$, we obtain:

$$J_w(w_* + \theta(v - w_*)) = \frac{1}{2}\, \|w_* - w\|_V^2 + \theta\, (w_* - w, v - w_*)_V + \frac{\theta^2}{2}\, \|v - w_*\|_V^2 .$$

Requiring that:

$$\frac{dJ_w(w_* + \theta(v - w_*))}{d\theta}|_{\theta=0} \geq 0,$$

yields that:

$$(w_* - w, v - w_*)_V \geq 0.$$

This is valid for each $v \in \mathcal{K}$. Thus w_* (which we denoted by $P_{\mathcal{K}}w$) satisfies:

$$(P_{\mathcal{K}}w - w, v - P_{\mathcal{K}}w) \geq 0, \quad \forall v \in \mathcal{K}.$$

For a complete proof, see [CI4]. □

Remark 14.9. The projection $P_{\mathcal{K}}$ onto a closed convex set $\mathcal{K}$ will not be linear, unless unless $\mathcal{K}$ is a linear space. For general convex sets $\mathcal{K}$, it may be computationally expensive to determine $P_{\mathcal{K}}w$ given w. However, for the type of convex sets occurring in obstacle problems, it will be inexpensive to compute $P_{\mathcal{K}}w$. Indeed, let $V = \mathbb{R}^n$ be equipped with the Euclidean inner product. Given a discrete obstacle vector $\boldsymbol{\psi}$, define the closed convex set $\mathcal{K}_h$ as:

$$\mathcal{K}_h = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v}_i \geq \boldsymbol{\psi}_i, \quad \text{for } i = 1, \cdots, n\}.$$

Then, the projection $P_{\mathcal{K}}\mathbf{w}$ of $\mathbf{w} \in \mathbb{R}^n$ onto $\mathcal{K}_h$ can be verified to satisfy:

$$(P_{\mathcal{K}}\mathbf{w})_i = \max\{\mathbf{w}_i, \boldsymbol{\psi}_i\}, \quad \text{for} \quad 1 \leq i \leq n.$$

This can be derived using the variational characterization of the projection.

Remark 14.10. The preceding projection theorem applies when minimizing the *distance* defined using the Hilbert norm $\| \cdot \|_V$. The minimum of a more general quadratic functional $J(\cdot)$:

$$J(v) = \frac{1}{2}\, a(v,v) - (f,v),$$

within $\mathcal{K}$ can be sought, if desired, by applying the preceding projection theorem, provided $a(.,.)$ is a symmetric, coercive bilinear form on V equivalent to $(.,.)_V$ and $(f,\cdot)$ is a bounded linear functional on V. Indeed, if $u \in V$ solves:

$$a(u,v) = (f,v), \qquad \forall v \in V,$$

then it can be verified that:

$$J(v) = \frac{1}{2}\, a(v-u, v-u) - \frac{1}{2}\, a(u,u).$$

Since $a(\cdot,\cdot)$ is equivalent to $(\cdot,\cdot)_V$, we may apply the projection theorem by employing the inner product induced by $a(.,.)$ and by translating $\mathcal{K}$ by u. We omit the details.

14.2 Projected Gradient and Relaxation Algorithms

In this section, we shall describe two traditional iterative methods, see [CI4], for solving the linear complementarity problem (14.6). The *projected gradient method* will be applicable primarily when A is symmetric and positive definite, while the *relaxation method* will be applicable more generally (even when matrix A is nonsymmetric, provided A is a diagonally dominant M-matrix). Readers are referred to [CI4] for a detailed exposition and analysis of these algorithms. The iterative methods described here can be used to solve local problems when Schwarz algorithms are employed.

Projected Gradient Method. The projected gradient method seeks to iteratively determine the minimum $\mathbf{u} \in \mathcal{K}_h$ of a functional $J(\cdot) : \mathcal{K}_h \to \mathbb{R}$ by applying a fixed point iteration:

$$J(\mathbf{u}) = \min_{\mathbf{u} \in \mathcal{K}_h} J(\mathbf{v}).$$

We assume that $J(\mathbf{v})$ is a sufficiently smooth *elliptic* and *Lipschitz* functional defined on a closed, convex set $\mathcal{K}_h \subset \mathbb{R}^n$.

Definition 14.11. *A sufficiently smooth functional $J : \mathbb{R}^n \to \mathbb{R}$ will be said to be elliptic if there exists $\alpha > 0$ such that:*

$$(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v}), \mathbf{u} - \mathbf{v}) \geq \alpha \, \|\mathbf{u} - \mathbf{v}\|^2,$$

where $\| \cdot \|$ and $(\cdot,\cdot)$ denotes the Euclidean norm and inner product on $\mathbb{R}^n$.

Definition 14.12. *A sufficiently smooth functional* $J : \mathbb{R}^n \to \mathbb{R}$ *will be said to be Lipschitz if there exists* $M > 0$ *such that:*

$$\|\nabla J(\mathbf{u}) - \nabla J(\mathbf{v})\| \leq M \, \|\mathbf{u} - \mathbf{v}\|,$$

where $\|\cdot\|$ *denotes the Euclidean norm.*

The projected gradient method can be motivated as follows. If $\mathbf{u} \in \mathcal{K}_h$ is a solution to the constrained minimization problem:

$$J\left(\mathbf{u}\right) = \min_{\mathbf{v} \in \mathcal{K}_h} J\left(\mathbf{v}\right),$$

then the first order optimality conditions at $\mathbf{u}$ requires:

$$(\nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u}) \geq 0, \quad \forall \mathbf{v} \in \mathcal{K}_h,$$

since the derivative of $J\left(\mathbf{u} + \theta(\mathbf{v} - \mathbf{u})\right)$ with respect to θ must be nonnegative when $\theta = 0$. Next, note that for any $\rho > 0$, the above optimality condition will be equivalent to find $\mathbf{u} \in \mathcal{K}_h$ such that:

$$\begin{cases} \rho\left(\nabla J(\mathbf{u}), \mathbf{v} - \mathbf{u}\right) \geq 0, & \forall \mathbf{v} \in \mathcal{K}_h \\ \left(\mathbf{u} - \left(\mathbf{u} - \rho \nabla J(\mathbf{u})\right), \mathbf{v} - \mathbf{u}\right) \geq 0, & \forall \mathbf{v} \in \mathcal{K}_h. \end{cases}$$

The preceding characterization is identical to characterization (14.10) of the projection onto $\mathcal{K}_h$. As a result, the preceding seeks $\mathbf{u} \in \mathcal{K}_h$:

$$P_{\mathcal{K}}\left(\mathbf{u} - \rho \nabla J(\mathbf{u})\right) = \mathbf{u}.$$

Thus, the solution $\mathbf{u}$ of the constrained minimization problem will be a *fixed point* of the mapping $T : \mathcal{K}_h \to \mathcal{K}_h$ defined by:

$$T(w) \equiv P_{\mathcal{K}}\left(\mathbf{w} - \rho \nabla J(\mathbf{w})\right),$$

for any $\rho > 0$. If the mapping T is a *contraction* for a suitable choice of parameter $\rho > 0$, then given a starting guess $\mathbf{u}^{(0)} \in \mathcal{K}_h$, we can determine the solution to the variational inequality by iterating the contraction mapping:

$$\mathbf{u}^{(k+1)} = T\Big(\mathbf{u}^{(k)}\Big), \qquad \text{for} \quad k = 0, 1, \ldots$$

We show below that when $J\left(\cdot\right)$ is elliptic and $\nabla J\left(\cdot\right)$ is Lipschitz, then $T(\cdot)$ will be a *contraction* for appropriate $\rho > 0$, with geometric convergence.

Algorithm 14.2.1 *(Projected Gradient Method)*
Given a starting iterate $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ *and* $\rho > 0$

```
1. For $k = 0, \cdots$, until convergence do:
2.     Compute:
           $\mathbf{u}^{(k+1)} \equiv P_{\mathcal{K}}\left(\mathbf{u}^{(k)} - \rho \nabla J(\mathbf{u}^{(k)})\right)$
3. Endfor
```

Remark 14.13. Substituting for $P_{\mathcal{K}}$ and ∇J in the obstacle problem yields:

$$\mathbf{u}_i^{(k+1)} = \max\{\mathbf{u}_i^{(k)} - \rho\,(A\mathbf{u}^{(k)} - \mathbf{f})_i, \boldsymbol{\psi}_i\} \quad \text{for } 1 \le i \le n,$$

for step 2 above. The following convergence result will hold.

Proposition 14.14. *Let $J : \mathbb{R}^n \to \mathbb{R}$ be elliptic with parameter $\alpha > 0$ and Lipschitz with parameter $M > 0$. Then, there exists an interval:*

$$0 < \delta_1 \le \rho \le \delta_2 \le \frac{2\alpha}{M^2},$$

on which the map $T(w) \equiv P_{\mathcal{K}}\,(\mathbf{w} - \rho\,\nabla J(\mathbf{w}))$ is a contraction, i.e., there exists $0 < \gamma < 1$ such that:

$$\|T(\mathbf{w}) - T(\mathbf{v})\| \le \gamma\,\|\mathbf{u} - \mathbf{w}\|, \quad \forall \mathbf{w},\, \mathbf{v} \in \mathcal{K}_h.$$

Furthermore, the iterates $\{\mathbf{u}^{(k)}\}$ of the projected gradient method will satisfy:

$$\|\mathbf{u}^{(k)} - \mathbf{u}\| \le \gamma^k\,\|\mathbf{u}^{(0)} - \mathbf{u}\|.$$

Proof. We follow the proof in [CI4]. Using the definition of $T(\cdot)$ and the nonexpansive property of the projection $P_{\mathcal{K}}$ map, we obtain:

$$\begin{aligned}
\|T(\mathbf{u}) - T(\mathbf{v})\|^2 &= \|P_{\mathcal{K}}(\mathbf{u}) - P_{\mathcal{K}}(\mathbf{v})\|^2 \\
&\le \|\,(\mathbf{u} - \rho\,\nabla J(\mathbf{u})) - (\mathbf{v} - \rho\,\nabla J(\mathbf{v}))\,\|^2 \\
&= \|(\mathbf{u} - \mathbf{v}) - \rho\,(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v}))\|^2 \\
&= (\mathbf{u} - \mathbf{v}, \mathbf{u} - \mathbf{v}) - 2\rho(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v}), \mathbf{u} - \mathbf{v}) \\
&\quad + \rho^2\,\left(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v}), \nabla J(\mathbf{u}) - \nabla J(\mathbf{v})\right) \\
&\le \|\mathbf{u} - \mathbf{v}\|^2 - 2\,\rho\,\alpha\,\|\mathbf{u} - \mathbf{v}\|^2 + \rho^2\,M^2\,\|\mathbf{u} - \mathbf{v}\|^2 \\
&= \left(1 - 2\,\rho\,\alpha + \rho^2\,M^2\right)\|\mathbf{u} - \mathbf{v}\|^2,
\end{aligned}$$

where we have used ellipticity of $J(\cdot)$ and Lipschitz continuity of ∇J. Thus:

$$\|T(\mathbf{u}) - T(\mathbf{u})\| \le \left(1 - 2\,\rho\,\alpha + \rho^2 M^2\right)^{1/2}\|\mathbf{u} - \mathbf{v}\|,$$

which yields that for $0 < \rho < \frac{2\alpha}{M^2}$ the above is a contraction. □

Remark 14.15. When $A = A^T > 0$, the optimal parameter ρ will be:

$$\rho = \frac{2}{\lambda_{min}(A) + \lambda_{max}(A)}.$$

When A is *nonsymmetric*, the projected gradient algorithm can be shown to converge, *provided* A is a strictly diagonally dominant M-matrix. In this case, the algorithm will correspond to a Richardson method for an associated parabolic variational inequality. The projection $P_{\mathcal{K}}$ can also be shown to be *nonexpansive* in the maximum norm, and $\|I - \rho A\|_\infty < 1$ for appropriate choices of ρ. For each $0 < \rho$ the contraction factor will satisfy:

$$\|T\mathbf{u} - T\mathbf{v}\|_\infty \leq \left(\max_i \{ (1 - \rho A_{ii}), \, \rho \max_{j \neq i} |A_{ij}|\}\right) \|\mathbf{u} - \mathbf{v}\|_\infty,$$

and an optimal choice of ρ can be selected based on the preceding expression.

Relaxation Methods. Gauss-Seidel and Jacobi relaxation algorithms can be formulated to solve variational inequalities, see [CR4, CI4]. We shall describe versions of these relaxation methods, applicable when matrix A is either a symmetric positive definite matrix or a strictly diagonally dominant M-matrix (possibly nonsymmetric):

$$\begin{cases} A_{ii} > 0, & \text{for } 1 \leq i \leq n \\ A_{ij} \leq 0, & \text{for } j \neq i, \, 1 \leq i \leq n \\ \sum_j A_{ij} > 0, & \text{for } 1 \leq i \leq n. \end{cases} \tag{14.11}$$

When matrix A is symmetric positive definite, these relaxation algorithms will have interpretations in terms of the minimization of functional $J(\cdot)$ along one dimensional subspaces. We shall use $\mathbf{e}_i \in \mathbb{R}^n$ to denote the i'th column of the identity matrix I of size n. We shall require the starting guess $\mathbf{u}^{(0)}$ in the relaxation algorithms to solve (14.6) to satisfy $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ and $A\,\mathbf{u}^{(0)} \geq \mathbf{f}$. Such a starting guess can be constructed as follows. Solve the linear system $A\mathbf{w}^{(0)} = (\mathbf{f} + \mathbf{1})$, for $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$, and compute:

$$\delta = \min_{1 \leq i \leq n} \{\mathbf{w}_i^{(0)} - \boldsymbol{\psi}_i\}.$$

Then define $\mathbf{u}^{(0)} \equiv (\mathbf{w}^{(0)} - \delta\,\mathbf{1})$. It can easily be verified that $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ and $A\mathbf{u}^{(0)} \geq \mathbf{f}$, when A is strictly diagonally dominant.

Algorithm 14.2.2 *(Gauss-Seidel Linear Complementarity Relaxation)*
Choose a starting guess $\mathbf{u}^{(0)}$ satisfying $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ and $(A\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}$

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $i = 1, \cdots, n$ in sequence do:
3.         Update i'th component of current iterate
               $\left(\mathbf{u}^{(k+\frac{i}{n})}\right)_i = \max\left\{\mathbf{u}_i^{(k+\frac{i-1}{n})} + \frac{\mathbf{f}_i - \left(A\mathbf{u}^{(k+\frac{i-1}{n})}\right)_i}{A_{ii}}, \boldsymbol{\psi}_i\right\}$
4.     Endfor
5. Endfor
```

Output: $\mathbf{u}^{(k)}$

Remark 14.16. When matrix A is symmetric positive definite, step 3 above will have the following minimization interpretation:

$$J(\mathbf{u}^{(k+\frac{i}{n})}) = \min_{\{\alpha:\mathbf{u}^{(k+\frac{i-1}{n})}+\alpha\,\mathbf{e}_i\geq\boldsymbol{\psi}\}} J(\mathbf{u}^{(k+\frac{i-1}{n})} + \alpha\,\mathbf{e}_i).$$

An over-relaxation parameter can also be introduced, see [CR4, WI10].

We next describe a Jacobi relaxation method for the iterative solution of the linear complementarity problem (14.6). Let $0 < \alpha_i < 1$ denote user chosen parameters for $1 \leq i \leq n$ satisfying:

$$\sum_{i=1}^{n} \alpha_i = 1.$$

For instance, $\alpha_i = (1/n)$ for $1 \leq i \leq n$. (Alternatively, we may choose the parameters $0 < \alpha_i < 1$ for $1 \leq i \leq n$ such that $\sum_{i=1}^{n} \alpha_i < 1$.)

Algorithm 14.2.3 *(Jacobi Linear Complementarity Relaxation Method)*
Choose a starting guess $\mathbf{u}^{(0)}$ *satisfying* $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ *and* $(A\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}$

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $i = 1, \ldots, n$ in parallel update:
$$\left(\mathbf{u}^{(k+1)}\right)_i = (1-\alpha_i)\mathbf{u}_i^{(k)} + \alpha_i \max\left\{\mathbf{u}_i^{(k)} + \frac{\mathbf{f}_i - \left(A\mathbf{u}^{(k)}\right)_i}{A_{ii}}, \boldsymbol{\psi}_i\right\}$$
3.     Endfor
4. Endfor
```

Remark 14.17. When A is symmetric positive definite, each Jacobi update in step 2 will have the form $\mathbf{u}^{(k+1)} = \sum_{i=1}^{p} \alpha_i \left(\mathbf{u}^{(k)} + \xi_i^k\,\mathbf{e}_i\right)$ and will have the following minimization interpretation:

$$J\left(\mathbf{u}^{(k)} + \xi_i^k\,\mathbf{e}_i\right) = \min_{\{\xi:\mathbf{u}^{(k)}+\xi\,\mathbf{e}_i\geq\boldsymbol{\psi}\}} J\left(\mathbf{u}^{(k)} + \xi\,\mathbf{e}_i\right).$$

Convexity of $J(\cdot)$ will yield that:

$$J\left(\mathbf{u}^{(k+1)}\right) = J\left(\sum_{i=1}^{n} \alpha_i(\mathbf{u}^{(k)} + \xi_i^k\mathbf{e}_i)\right) \leq \sum_{i=1}^{n} \alpha_i\, J\left(\mathbf{u}^{(k)} + \xi_i^k\mathbf{e}_i\right) \leq J\left(\mathbf{u}^{(k)}\right).$$

Remark 14.18. Unlike the projected gradient method, the Gauss-Seidel and Jacobi complementarity *relaxation* algorithms may *not* converge on general convex sets [CI4]. For instance, if we seek the minimum of $J\left(x_1, x_2\right) = x_1^2 + x_2^2$ on the convex set $\mathcal{K}_* = \left\{(x_1, x_2) : x_1 + x_2 \geq 2\right\}$, then, each point along the line $x_1 + x_2 = 2$ will be a fixed point of the preceding Gauss-Seidel and Jacobi iterations, but only $(1, 1)$ will correspond to the true minimum.

14.3 Schwarz Algorithms for Variational Inequalities

Schwarz algorithms [LI6, HO3, KU12, KU13, ZE, LI10, BA10, TA2, TA3] can be formulated for *variational inequalities* by generalizing point relaxation methods to involve blocks of unknowns based on subdomains. We consider the iterative solution of linear complementarity problem (14.6), and describe sequential and parallel Schwarz complementarity algorithms (without and with coarse grid correction). These algorithms can be employed either when A is symmetric and positive definite or when A is a strictly diagonally dominant M-matrix (possibly *nonsymmetric*). For a description of multilevel algorithms for variational inequalities, see [MA10, KO5, TA2, TA3].

Let $x_1, \ldots, x_n$ denote the interior nodes in Ω, and let $\Omega_1, \ldots, \Omega_p$ denote an *nonoverlapping* decomposition of Ω into subdomains of size h_0. We let $\Omega_1^*, \ldots, \Omega_p^*$ denote an overlapping subdomain decomposition of Ω, obtained be extending each Ω_l to Ω_l^* by including a points within a distance of $\beta\, h_0 > 0$. Given an overlapping covering $\Omega_1^*, \ldots, \Omega_p^*$ of Ω, we let $\mathcal{I}_l^*$ denote the set of indices of nodes $x_i \in \Omega_l^*$. We shall also let $\mathcal{I}_l$ for $1 \leq l \leq p$ denote a *partition* of all indices, so that each node on $\partial\Omega_l \cap \partial\Omega_j$ is assigned either to $\mathcal{I}_l$ or to $\mathcal{I}_j$, and such that if $i \in \mathcal{I}_l$, then $x_i \in \overline{\Omega}_l$. Thus, $\mathcal{I}_l$ will correspond to a partition of the nodes based on subdomains with minimal overlap. We shall let n_l and n_l^* denote the number of nodes in $\mathcal{I}_l$ and $\mathcal{I}_l^*$, respectively.

Define $V = \mathbb{R}^n$ and equip it with the Euclidean inner product. Corresponding to index sets $\mathcal{I}_l$ and $\mathcal{I}_l^*$, define subspaces V_l and V_l^* of V as:

$$V_l = \text{span}\,\{\mathbf{e}_j \,:\, j \in \mathcal{I}_l\} \quad \subset \quad V_l^* = \text{span}\,\{\mathbf{e}_j \,:\, j \in \mathcal{I}_l^*\}\,.$$

This yields $V = V_1 + \cdots + V_p = V_1^* + \cdots + V_p^*$. We employ the *notation*:

- Let R_l denote an $n_l \times n$ *restriction* matrix whose rows form a basis for V_l, consisting of *elementary* vectors $\{\mathbf{e}_j\}$ for indices $j \in \mathcal{I}_l$. By construction, matrix R_l will have 0 or 1 entries, with orthogonal rows.
- Let $R_{l,*}$ denote an $n_l^* \times n$ *restriction* matrix whose rows form a basis for V_l^*, consisting of *elementary* vectors $\{\mathbf{e}_j\}$ for indices $j \in \mathcal{I}_l^*$. By construction, matrix $R_{l,*}$ will have 0 or 1 entries, with orthogonal rows.
- Let $A_l = R_l A R_l^T$ and $A_{l,*} = R_{l,*} A R_{l,*}^T$ denote submatrices of A associated with indices $\mathcal{I}_l$ and $\mathcal{I}_l^*$.

If A is a strictly diagonally dominant M-matrix, we shall choose a starting guess $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ satisfying $(A\,\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}$ component wise, constructed as described earlier. In the following, we summarize the sequential Schwarz linear complementarity algorithm for solving (14.6), without a coarse grid.

Algorithm 14.3.1 *(Sequential Schwarz Complementarity Algorithm)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ *satisfy* $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ *and* $(A\,\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}$

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $l = 1, \ldots, p$ in sequence determine $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l^*}$ on $\Omega_l^*$ s.t:
           $\begin{cases} \mathbf{d}^{(l)} \,\geq\, R_{l,*}\left(\boldsymbol{\psi} - \mathbf{u}^{(k+\frac{l-1}{p})}\right) \\ A_{l,*}\mathbf{d}^{(l)} \,\geq\, R_{l,*}\left(\mathbf{f} - A\mathbf{u}^{(k+\frac{l-1}{p})}\right) \\ \left(\mathbf{d}^{(l)} - R_{l,*}(\boldsymbol{\psi} - \mathbf{u}^{(k+\frac{l-1}{p})})\right)^T \left(A_{l,*}\mathbf{d}^{(l)} - R_{l,*}(\mathbf{f} - A\,\mathbf{u}^{(k+\frac{l-1}{p})})\right) \,=\, \mathbf{0}, \end{cases}$
       Define $\mathbf{u}^{(k+\frac{l}{p})} = \mathbf{u}^{(k+\frac{l-1}{p})} + R_{l,*}^T\mathbf{d}^{(l)}$
3.     Endfor
4. Endfor
```

Output: $\mathbf{u}^{(k)}$

Remark 14.19. When $A = A^T > 0$, step 2 above will correspond to choosing $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l^*}$ which solves the local constrained minimization problem:

$$J(\mathbf{u}^{(k+\frac{l-1}{p})} + R_{l,*}^T\mathbf{d}^{(l)}) = \min_{\left\{\mathbf{v}^{(l)}:\, \mathbf{u}^{(k+\frac{l-1}{p})} + R_{l,*}^T\mathbf{v}^{(l)} \geq \boldsymbol{\psi}\right\}} J(\mathbf{u}^{(k+\frac{l-1}{p})} + R_{l,*}^T\mathbf{v}^{(l)}).$$

Remark 14.20. This sequential algorithm can be parallelized by coloring the subdomains $\Omega_1^*, \ldots, \Omega_p^*$ into a minimal number of *colors*, so that problems on *disjoint* subdomains of the same color can be solved in parallel. Below, we list a parallel Schwarz algorithm for (14.6).

Algorithm 14.3.2 *(Parallel Schwarz Complementarity Algorithm)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ *satisfy* $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ *and* $(A\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}$
Let $0 < \alpha_i < 1$ *satisfy* $\sum_{i=1}^p \alpha_i = 1$

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $l = 1, \ldots, p$ in sequence determine $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l^*}$ satisfying:
           $\begin{cases} \mathbf{d}^{(l)} \,\geq\, R_{l,*}\left(\boldsymbol{\psi} - \mathbf{u}^{(k)}\right) \\ A_{l,*}\mathbf{d}^{(l)} \,\geq\, R_{l,*}\left(\mathbf{f} - A\mathbf{u}^{(k)}\right) \\ \left(\mathbf{d}^{(l)} - R_{l,*}(\boldsymbol{\psi} - \mathbf{u}^{(k)})\right)^T \left(A_{l,*}\mathbf{d}^{(l)} - R_{l,*}(\mathbf{f} - A\mathbf{u}^{(k)})\right) = 0 \end{cases}$
3.     Endfor
4.     Define $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \sum_{l=1}^p \alpha_l\, R_{l,*}^T\mathbf{d}^{(l)}$
5. Endfor
```

Output: $\mathbf{u}^{(k)}$

Remark 14.21. A variant of the preceding parallel Schwarz algorithm can be constructed as follows. Employ the index sets $\mathcal{I}_l$ having *minimal* overlap, and replace the local solve in step 2 above by seeking $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}$ satisfying:

$$\begin{cases} \mathbf{d}^{(l)} \geq R_l\left(\boldsymbol{\psi} - \mathbf{u}^k\right) \\ A_l\mathbf{d}^{(l)} \geq R_l\left(\mathbf{f} - A\mathbf{u}^k\right) \\ \left(\mathbf{d}^{(l)} - R_l(\boldsymbol{\psi} - \mathbf{u}^k)\right)^T\left(A_l\mathbf{d}^l - R_l(\mathbf{f} - A\mathbf{u}^l)\right) = 0, \end{cases}$$

and replace step 4 by $\mathbf{u}^{k+1} = \mathbf{u}^k + \sum_{l=1}^{p} R_l^T\mathbf{d}^{(l)}$. However, $A\mathbf{u}^{k+1} \not\geq \mathbf{f}$.

Remark 14.22. If A is a strictly diagonally dominant M-matrix, the submatrices $A_{l,*} = R_{l,*}AR_{l,*}^T$ and $A_l = R_lAR_l^T$ of A corresponding to the indices in $\mathcal{I}_l^*$ or $\mathcal{I}_l$, will be M-matrices.

Lemma 14.23. *When $A = A^T > 0$ and the overlap between the subdomains is $\beta\, h_0$, the rate of convergence of the sequential and parallel Schwarz algorithms without a coarse grid will be independent of h (but dependent on h_0).*

Proof. See [BA10, TA2, TA3]. □

Remark 14.24. As for elliptic equations without obstacle constraints, the use of a *coarse grid* of size h_0 can speed up the convergence of Schwarz complementarity algorithms, eliminating the dependence on h_0. However, care must be exercised when enforcing obstacle *constraints* on the coarse grid, as this can adversely affect convergence. From an algorithmic viewpoint, such an algorithm may associate the coarse grid problem with $l = 0$ (so that the indices in step 2 involve $0 \leq l \leq p$, instead of $1 \leq l \leq p$). Let n_0 denote the number of interior nodes on the coarse grid, and let $R_{0,*}$ denote an $n_0 \times n$ matrix corresponding to a *restriction* onto the coarse grid with coarse grid matrix $A_{0,*} = R_{0,*}AR_{0,*}^T$ of size n_0. The difficulty arises when enforcing the obstacle constraints $\mathbf{u}_i \geq \boldsymbol{\psi}_i$ on the coarse grid (since only $n_0 \ll n$ constraints can be employed for the coarse grid problem, unlike the n fine grid obstacle constraints). One approach is to employ a *nonlinear* interpolation map I_0^- when enforcing obstacle constraints on the coarse grid. More specifically, if $z_1, \ldots, z_{n_0}$ denotes the coarse grid interior nodes with associated corse grid basis functions $\phi_1^{(0)}(\cdot), \ldots, \phi_{n_0}^{(0)}(\cdot)$, define:

$$\mathcal{N}(z_i) = \{j : x_j \in \text{support}\left(\phi_i^{(0)}(\cdot)\right).$$

Given a nodal vector $\mathbf{v} \in \mathbb{R}^n$ define $I_0^-\mathbf{v} \in \mathbb{R}^{n_0}$ as follows:

$$\left(I_0^-\mathbf{v}\right)_i \equiv \min_{\{j \in \mathcal{N}(z_i)\}} (\mathbf{v})_j\,, \qquad \text{for } 1 \leq i \leq n_0.$$

Then, given a current iterate $\mathbf{w} \in \mathbb{R}^n$, the linear complementarity problem on the coarse grid should seek an update $\mathbf{d}^{(0)} \in \mathbb{R}^{n_0}$ enforcing the constraints:

$$\mathbf{d}^{(0)} \geq I_0^-\left(\boldsymbol{\psi} - \mathbf{w}\right),$$

instead of $\mathbf{d}^{(0)} \geq R_{0,*}\left(\boldsymbol{\psi} - \mathbf{w}\right)$. Analysis of coarse grid Schwarz algorithms yields a convergence rate independent of h and h_0, see [BA10, TA2, TA3].

14.4 Monotone Convergence of Schwarz Algorithms

We next describe selected results on the maximum norm convergence of Schwarz complementarity algorithms for (14.6), without coarse grid correction, see [BA9, KU12, KU13, ZE]. We assume that A is a strictly diagonally dominant M-matrix as in (14.11). For energy based convergence results, readers are referred to [LI6, HO3, BA10, TA2, TA3]. The following preliminary result describes the monotone nature of *subdomain updates.*

Lemma 14.25. *Suppose the following conditions hold.*

1. *Let A be a strictly diagonally dominant M-matrix.*
2. *Let $\mathbf{w} \in \mathbb{R}^n$ satisfy $\mathbf{w} \geq \boldsymbol{\psi}$ and $(A\,\mathbf{w} - \mathbf{f}) \geq \mathbf{0}$ component wise.*
3. *Let $R_{l,*}$ denote a restriction matrix of size $n_l^* \times n$ corresponding to index set $\mathcal{I}_l^*$ associated with Ω_l^*. Let $A_{l,*} = R_{l,*} A R_{l,*}^T$ and let $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l^*}$ satisfy:*

$$\begin{cases} \mathbf{d}^{(l)} \;\geq\; R_l\,(\boldsymbol{\psi} - \mathbf{w}) \\ A_{l,*}\,\mathbf{d}^{(l)} \;\geq\; R_{l,*}\,(\mathbf{f} - A\mathbf{w}) \\ \left(\mathbf{d}^{(l)} - R_{l,*}(\boldsymbol{\psi} - \mathbf{w})\right)^T \left(A_{l,*}\,\mathbf{d}^{(l)} - R_{l,*}(\mathbf{f} - A\mathbf{w})\right) = 0, \end{cases}$$

4. *Let $\mathbf{u} \in \mathcal{K}_h$ denote the solution of the linear complementarity problem:*

$$(\mathbf{v} - \mathbf{u})^T\; A\;(\mathbf{u} - \mathbf{f}) \geq 0, \quad \forall \mathbf{v} \in \mathcal{K}.$$

Then, the following results will hold:

$$\begin{cases} \mathbf{d}^{(l)} \leq \mathbf{0} \\ \mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} \geq \mathbf{u} \\ A\left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)}\right) - \mathbf{f} \geq \mathbf{0}, \end{cases}$$

and update $\mathbf{w} + R_{l,*}^T\,\mathbf{d}^{(l)}$ will be "sandwiched" $\mathbf{u} \;\leq\; \left(\mathbf{w} + R_{l,*}^T\,\mathbf{d}^{(l)}\right) \;\leq\; \mathbf{w}$. Furthermore, it will hold that $A\left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)}\right) \geq \mathbf{f}$.

Proof. We shall first show that $\mathbf{d}^{(l)} \leq \mathbf{0}$ component wise. Suppose, for contradiction, that there is at least one component of $\mathbf{d}^{(l)}$ which is positive:

$$0 < \left(\mathbf{d}^{(l)}\right)_j = \max_k \left(\mathbf{d}^{(l)}\right)_k.$$

Then, j cannot be a *contact* node (i.e., $\left(\mathbf{d}^{(l)}\right)_j > \left(R_{l,*}(\boldsymbol{\psi} - \mathbf{w})\right)_j$ must hold, since $\boldsymbol{\psi} - \mathbf{w} \leq \mathbf{0}$ and $R_{l,*}(\boldsymbol{\psi} - \mathbf{w}) \leq \mathbf{0}$). Then, the *linear complementarity* equation will yield:

$$\left(A_{l,*}\mathbf{d}^{(l)}\right)_j + \left(R_{l,*}(A\mathbf{w} - \mathbf{f})\right)_j = 0.$$

Since A is a strictly diagonally dominant matrix M-matrix, $A_{l,*}$ will also be a strictly diagonally dominant M-matrix. Since $R_{l,*}(A\mathbf{w} - \mathbf{f}) \geq \mathbf{0}$, we obtain:

$$\begin{aligned} (A_{l,*})_{jj}\ \left(\mathbf{d}^{(l)}\right)_j &= -\left(\sum_{k\neq j} (A_{l,*})_{jk} \,(\mathbf{d}^{(l)})_k\right) - \left(R_{l,*}(A\mathbf{w} - \mathbf{f})\right)_j \\ &\leq -\left(\sum_{k\neq j} (A_{l,*})_{jk} \,(\mathbf{d}^{(l)})_k\right) \\ &\leq -\left(\sum_{k\neq j} (A_{l,*})_{jk} \ (\mathbf{d}^{(l)})_j\right) \\ &= -\left(\sum_{k\neq j} (A_{l,*})_{jk}\right) \,(\mathbf{d}^{(l)})_j \\ &< (A_{l,*})_{jj} \ (\mathbf{d}^{(l)})_j, \end{aligned}$$

which is a contradiction. Thus, $\mathbf{d}^{(l)} \leq \mathbf{0}$ and $\left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)}\right) \leq \mathbf{w}$.

Next, we shall show that $\mathbf{w}+R_{l,*}^T \mathbf{d}^{(l)} \geq \mathbf{u}$. Again, suppose for contradiction that there is an index j such that:

$$\left(\mathbf{u} - \mathbf{w} - R_{l,*}^T \mathbf{d}^{(l)}\right)_j = \max_k \left(\mathbf{u} - \mathbf{w} - R_{l,*}^T \mathbf{d}^{(l)}\right)_k > 0.$$

Then, j cannot be a contact index of $\mathbf{u}$ with the obstacle $\boldsymbol{\psi}$. This is because, $(\mathbf{u})_j > \mathbf{w}_j + \left(R_{l,*}\mathbf{d}^{(l)}\right)_j \geq \boldsymbol{\psi}_j$. Thus, the following expressions must hold:

$$\left(A\,(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)}) - \mathbf{f}\right)_j \geq \mathbf{0} \quad \text{and} \quad (A\,\mathbf{u} - \mathbf{f})_j = 0.$$

Subtracting the two equations and using that $A_{jk} \leq 0$ for $j \neq k$ yields:

$$\begin{aligned} 0 &\leq \left(A\,\mathbf{w} + A\,R_{l,*}^T \mathbf{d}^{(l)} - A\,\mathbf{u}\right)_j \\ &= \sum_{k=1}^n A_{jk} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_k \\ &= A_{jj} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j + \sum_{k\neq j} A_{jk} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_k \\ &\leq A_{jj} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j + \sum_{k\neq j} A_{jk} \left(\min_k \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_k\right) \\ &= A_{jj} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j - \sum_{k\neq j} A_{jk} \left(\max_k \left(\mathbf{u} - \mathbf{w} - R_{l,*}^T \mathbf{d}^{(l)}\right)_k\right) \\ &= A_{jj} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j - \sum_{k\neq j} A_{jk} \left(\mathbf{u} - \mathbf{w} - R_{l,*}^T \mathbf{d}^{(l)}\right)_j \\ &= A_{jj} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j + \sum_{k\neq j} A_{jk} \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j \\ &= \left(\sum_{k=1}^n A_{jk}\right) \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j. \end{aligned}$$

Now, since $\sum_k A_{jk} > 0$ and $\left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)} - \mathbf{u}\right)_j < 0$, we obtain:

$$0 \leq \left(\sum_{k=1}^n A_{jk}\right) \left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{i} - \mathbf{u}\right)_j < 0,$$

which is a contradiction. Thus, we must have $\left(\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)}\right) \geq \mathbf{u}$.

Next, we shall verify that $A\left(\mathbf{w}+R_{l,*}^T\mathbf{d}^{(l)}\right)-\mathbf{f}\geq\mathbf{0}$. To do this, note that by construction of $\mathbf{d}^{(l)}$, it will hold that $\left(A(\mathbf{w}+R_{l,*}^T\mathbf{d}^{(l)})-\mathbf{f}\right)_j\geq\mathbf{0}$ for $j\in\mathcal{I}_l^*$. When $j\notin\mathcal{I}_l^*$, since $\left(R_{l,*}^T\mathbf{d}^{(l)}\right)_j=0$, we will obtain:

$$\begin{aligned}\left(A(\mathbf{w}+R_{l,*}^T\mathbf{d}^{(l)})-\mathbf{f}\right)_j &= (A\mathbf{w}-\mathbf{f})_j+\left(AR_{l,*}^T\mathbf{d}^{(l)}\right)_j\\ &\geq\left(AR_{l,*}^T\mathbf{d}^{(l)}\right)_j\\ &\geq\textstyle\sum_{k\neq j}A_{jk}\left(R_{l,*}^T\mathbf{d}^{(l)}\right)_k\geq 0,\end{aligned}$$

since $A_{jk}\leq 0$ for $j\neq k$ and since $R_{l,*}^T\mathbf{d}^{(l)}\leq\mathbf{0}$. □

Remark 14.26. The preceding result yields that if A is a strictly diagonally dominant M-matrix, with $\mathbf{w}\geq\boldsymbol{\psi}$ and $A\,\mathbf{w}\geq\mathbf{f}$, then the subdomain updates $\mathbf{w}+R_{l,*}^T\mathbf{d}^{(l)}$ will lie "sandwiched" between the true solution $\mathbf{u}$ and the approximation $\mathbf{w}$. Thus, given an iterate $\mathbf{u}^{(0)}$ satisfying $A\,\mathbf{u}^{(0)}\geq\mathbf{f}$ and $\mathbf{u}^{(0)}\geq\boldsymbol{\psi}$, a *monotone* decreasing sequence of iterates can be constructed using subdomain solves, with each iterate lying above the desired solution $\mathbf{u}$.

To analyze the convergence of $\mathbf{u}^{(k)}$, we employ a *convex* set $\mathcal{H}\subset\mathbb{R}^n$:

$$\mathcal{H}\equiv\{\mathbf{v}\in\mathbb{R}^n\,:\,\mathbf{v}\geq\boldsymbol{\psi}\}\cap\{\mathbf{v}\in\mathbb{R}^n\,:\,A\mathbf{v}-\mathbf{f}\geq\mathbf{0}\}.\tag{14.12}$$

We first focus on the convergence of the parallel Schwarz complementarity algorithm, for which $T\left(\mathbf{u}^{(k)}\right)=\mathbf{u}^{(k+1)}$ for $\mathbf{u}^{(k+1)}\in\mathcal{H}$ with:

$$T(\mathbf{v})=\mathbf{v}+\sum_{l=1}^p\alpha_l\,R_{l,*}^T\mathbf{d}^{(l)},\tag{14.13}$$

where

$$\begin{cases}\mathbf{d}^{(l)}\;\geq\;R_{l,*}\,(\boldsymbol{\psi}-\mathbf{v})\\ A_{l,*}\mathbf{d}^{(l)}\;\geq\;R_{l,*}\,(\mathbf{f}-A\mathbf{v})\\ \left(\mathbf{d}^{(l)}-R_{l,*}(\boldsymbol{\psi}-\mathbf{v})\right)^T\left(A_{l,*}\mathbf{d}^{(l)}-R_{l,*}(\mathbf{f}-A\mathbf{v})\right)=0.\end{cases}\tag{14.14}$$

By construction, it is clear that if $\mathbf{u}$ is a solution of the linear complementarity problem (14.6), then $T(\mathbf{u})=\mathbf{u}$, i.e., $\mathbf{u}$ will be a *fixed point* of T. It will be shown later that $T:\mathcal{H}\to\mathcal{H}$ is a *contraction.* The following result shows that a fixed point of T will solve the linear complementarity problem (14.6).

Lemma 14.27. *Suppose the following conditions hold.*

1. *Let A be a strictly diagonally dominant M-matrix and let $T\mathbf{v}$ be defined by (14.13) and (14.14) for $\mathbf{v}\in\mathcal{H}$.*
2. *Let $\mathbf{w}_*\in\mathcal{H}$ denote a fixed point of T, with $T(\mathbf{w}_*)=\mathbf{w}_*$.*

Then $\mathbf{w}_*=\mathbf{u}$ will be a solution to linear complementarity problem (14.6).

Proof. To verify that $\mathbf{w}_* \in \mathcal{K}_h$ is a solution to (14.6), we need to show that:

$$(\mathbf{v} - \mathbf{w}_*)^T (A\,\mathbf{w}_* - \mathbf{f}) \geq 0, \qquad \forall \mathbf{v} \in \mathcal{K}_h.$$

Let $\{\chi_l(x)\}_{l=1}^p$ denote a partition of unity, subordinate to $\{\Omega_l^*\}_{l=1}^p$:

$$0 \leq \chi_l(x) \leq 1 \quad \text{for } 1 \leq l \leq p \quad \text{and} \quad \sum_{l=1}^p \chi_l(x) = 1.$$

Using the $\chi_l(\cdot)$, given $\mathbf{v} \in \mathcal{K}_h$, we decompose $(\mathbf{v} - \mathbf{w}_*) = \sum_{l=1}^p \chi_l(\mathbf{v} - \mathbf{w}_*)$:

$$(\chi_l(\mathbf{v} - \mathbf{w}_*))_j \equiv \left(R_{l,*}^T \mathbf{e}^{(l)}\right)_j \equiv \chi_l(x_j)\,(\mathbf{v} - \mathbf{w}_*)_j\,, \quad \text{for } 1 \leq j \leq n,$$

where x_j denotes the node corresponding to index j. By construction:

$$\sum_{l=1}^p \chi_l(\mathbf{v} - \mathbf{w}_*) = \sum_{l=1}^p R_{l,*}^T \mathbf{e}^{(l)} = \mathbf{v} - \mathbf{w}_*.$$

Substituting this decomposition yields:

$$\begin{aligned} (\mathbf{v} - \mathbf{w}_*)^T (A\mathbf{w}_* - \mathbf{f}) &= \left(\textstyle\sum_{l=1}^p R_{l,*}^T \mathbf{e}^{(l)}\right)^T (A\mathbf{w}_* - \mathbf{f}) \\ &= \textstyle\sum_{l=1}^p \mathbf{e}^{(l)^T} R_{l,*} (A\mathbf{w}_* - \mathbf{f})\,. \end{aligned} \tag{14.15}$$

We may express $\mathbf{e}^{(l)} = (\mathbf{v}^{(l)} - R_{l,*}\mathbf{w}_*)$ where $\mathbf{v}^{(l)} \equiv R_{l,*}(\mathbf{w}_* + \chi_l(\mathbf{v} - \mathbf{w}_*))$. Since $\mathbf{v} \geq \boldsymbol{\psi}$ and since $0 \leq \chi_l(x_i) \leq 1$, it will hold that:

$$(R_{l,*}^T \mathbf{v}^{(l)})_i \geq (\mathbf{w}_*)_i + \chi_l(x_i)(\boldsymbol{\psi}_i - (\mathbf{w}_*)_i) \;\geq\; \boldsymbol{\psi}_i$$

Now, we may apply local optimality of $\mathbf{w}_*$ to obtain $\mathbf{e}^{(l)^T} R_{l,*} (A\mathbf{w}_* - \mathbf{f}) \geq \mathbf{0}$. Substituting this into (14.15) yields $(\mathbf{v} - \mathbf{w}_*)^T (A\,\mathbf{w}_* - \mathbf{f}) \geq 0$ for $\mathbf{v} \in \mathcal{K}_h$. □

The preceding results can be employed to show that the parallel Schwarz complementarity iterates converge monotonically.

Proposition 14.28. *Suppose the following conditions hold.*

1. *Let A be a strictly diagonally dominant M-matrix.*
2. *Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ satisfy the following, component wise:*

$$\mathbf{u}^{(0)} \geq \boldsymbol{\psi} \quad \textit{and} \quad (A\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}.$$

Then the following results will hold for the iterates $\mathbf{u}^{(k+1)} = T(\mathbf{u}^{(k)})$:

1. *The iterates will be monotonically decreasing:*

$$\mathbf{u}^{(0)} \geq \mathbf{u}^{(1)} \geq \cdots \geq \mathbf{u}^{(k)} \geq \cdots$$

2. Each iterate $\mathbf{u}^{(k)}$ *will satisfy:*

$$(A\mathbf{u}^{(k)} - \mathbf{f}) \geq \mathbf{0}, \quad k = 0, 1, 2, \ldots$$

3. The iterates will converge to a fixed point $\mathbf{u}$ *of* $T(\cdot)$*:*

$$\lim_{k \to \infty} \mathbf{u}^{(k)} = \mathbf{u} = T(\mathbf{u}).$$

Proof. First, we shall show monotonicity of the iterates. Recall that:

$$\mathbf{u}^{(k+1)} = T\left(\mathbf{u}^{(k)}\right) = \mathbf{u}^{(k)} + \sum_{l=1}^{p} \alpha_l \, R_{l,*}^T \mathbf{d}^{(l)},$$

where, by Lemma 14.25 each $\mathbf{d}^{(l)} \leq \mathbf{0}$, so that $\mathbf{u}^{(k)} \geq \mathbf{u}^{(k+1)} \geq \mathbf{u}$.

Next, we shall show that $(A\mathbf{u}^{(k)} - \mathbf{f}) \geq \mathbf{0}$ for each k, using induction. Assume it holds for k, we shall it holds for $k+1$. By Lemma 14.25, it will hold that $A\left(\mathbf{u}^{(k)} + R_{l,*}^T \mathbf{d}^{(l)}\right) - \mathbf{f} \geq \mathbf{0}$ for $1 \leq l \leq p$. Since $\sum_{l=1}^{p} \alpha_l = 1$, we obtain:

$$(A\mathbf{u}^{(k+1)} - \mathbf{f}) = \sum_{l=1}^{p} \alpha_l \left(A(\mathbf{u}^{(k)} + R_{l,*}^T \mathbf{d}^{(l)}) - \mathbf{f}\right) \geq \mathbf{0}.$$

Thus, each iterate will satisfy $(A\mathbf{u}^{(k)} - \mathbf{f}) \geq \mathbf{0}$, provided $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$.

Since $\left(\mathbf{u}^{(k)}\right)_j$ will be *monotone* decreasing as $k \to \infty$ and bounded below by $\boldsymbol{\psi}_j$, it must be convergent to some limit $(\mathbf{u})_j$. It will thus hold:

$$\min_{k \to \infty} \mathbf{u}^k = \mathbf{u},$$

in any norm in $\mathbb{R}^n$. Since $\mathbf{u}^{(k+1)} = T\left(\mathbf{u}^{(k)}\right)$, as $u^{(k)} \to \mathbf{u}$ it will hold that:

$$T\left(\mathbf{u}\right) = \mathbf{u},$$

yielding that $\mathbf{u}$ is a fixed point of T. Since $\mathbf{u}^{(k)} \geq \boldsymbol{\psi}$, it will follow that $\mathbf{u} \geq \boldsymbol{\psi}$. Furthermore, it will also hold that $(A\mathbf{u} - \mathbf{f}) \geq \mathbf{0}$. □

When A is a strictly diagonally dominant M-matrix, the parallel Schwarz iteration map $T : \mathcal{H} \to \mathcal{H}$ will be a *contraction.* The following result estimates a *subdomain* contraction factor ρ_l in the maximum norm. Let $\{\Omega_l^*\}_{l=1}^{p}$ denote an overlapping decomposition obtained from a non-overlapping decomposition $\{\Omega_l\}_{l=1}^{p}$ where $\Omega_l^* = \{x \in \Omega \, : \, \text{dist}(x, \Omega_l) < \beta h_0\}$. Let $\mathcal{I}_l^*$ and $\mathcal{B}_l^*$ denote sets of nodal indices associated with *interior* nodes in Ω_l^* and on $B^{(l)} = \partial\Omega_l^* \cap \Omega$, respectively. We shall let $A_{II}^{(l)}$ and $A_{IB}^{(l)}$ denote submatrices of A coupling indices within $\mathcal{I}_l^*$, and between $\mathcal{I}_l^*$ and $\mathcal{B}_l^*$, respectively. We also let $\overline{\mathcal{I}}_l$ denote the indices of nodes in $\overline{\Omega}_l$.

Definition 14.29. *Let $\boldsymbol{\gamma}_B^{(l)} = (1, \ldots, 1)^T$ denote a vector associated with subdomain boundary values on $B^{(l)} = \partial\Omega_l^* \cap \Omega$ and solve for the vector $\boldsymbol{\gamma}_I^{(l)}$ associated with interior nodal values on Ω_l^*:*

$$A_{II}^{(l)} \boldsymbol{\gamma}_I^{(l)} + A_{IB}^{(l)} \boldsymbol{\gamma}_B^{(l)} = \mathbf{0}.$$

Define a local contraction factor $0 < \rho_l < 1$ as:

$$\rho_l \equiv \max_{\{i \,:\, x_i \in \overline{\Omega}_l\}} \left(\boldsymbol{\gamma}_I^{(l)}\right)_i = \max_{\{i \in \overline{\mathcal{I}}_l\}} \left(\boldsymbol{\gamma}_I^{(l)}\right)_i \tag{14.16}$$

In the following, we shall estimate the contraction factor of T in terms of ρ_l.

Lemma 14.30. *Suppose A is a strictly diagonally dominant M-matrix.*

1. *Let $R_{l,*}$ denote a restriction matrix of size $n_l^* \times n$ corresponding to index set $\mathcal{I}_l^*$ associated with Ω_l^* and let $A_{l,*} = R_{l,*} A R_{l,*}^T$.*
2. *Given $\mathbf{w} \in \mathcal{H}$ define $\tilde{\mathbf{w}} = (\mathbf{w} + R_{l,*}^T \mathbf{d}^{(l)}) \geq \boldsymbol{\psi}$ where $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l^*}$ satisfies:*

$$\begin{cases} \mathbf{d}^{(l)} \geq R_l\,(\boldsymbol{\psi} - \mathbf{w}) \\ A_{l,*}\,\mathbf{d}^{(l)} \geq R_{l,*}\,(\mathbf{f} - A\mathbf{w}) \\ \left(\mathbf{d}^{(l)} - R_{l,*}(\boldsymbol{\psi} - \mathbf{w})\right)^T \left(A_{l,*}\,\mathbf{d}^{(l)} - R_{l,*}(\mathbf{f} - A\mathbf{w})\right) = 0, \end{cases}$$

3. *Given $\mathbf{v} \in \mathcal{H}$ define $\tilde{\mathbf{v}} = \mathbf{v} + R_{l,*}^T \mathbf{e}^{(l)} \geq \boldsymbol{\psi}$ where $\mathbf{e}^{(l)} \in \mathbb{R}^{n_l^*}$ satisfies:*

$$\begin{cases} \mathbf{e}^{(l)} \geq R_l\,(\boldsymbol{\psi} - \mathbf{v}) \\ A_{l,*}\,\mathbf{e}^{(l)} \geq R_{l,*}\,(\mathbf{f} - A\mathbf{v}) \\ \left(\mathbf{e}^{(l)} - R_{l,*}(\boldsymbol{\psi} - \mathbf{v})\right)^T \left(A_{l,*}\,\mathbf{e}^{(l)} - R_{l,*}(\mathbf{f} - A\mathbf{v})\right) = 0, \end{cases}$$

Then, the following result will hold in the maximum norm:

$$\|\tilde{\mathbf{w}} - \tilde{\mathbf{v}}\|_{\infty, \overline{\mathcal{I}}_l} \equiv \max_{\{j \in \overline{\mathcal{I}}_l\}} |\tilde{\mathbf{w}}_j - \tilde{\mathbf{v}}_j| \leq \rho_l\, \|\mathbf{w} - \mathbf{v}\|_\infty.$$

Proof. By Lemma 14.25 it will follow that $\mathbf{d}^{(l)} \leq \mathbf{0}$ and $\mathbf{e}^{(l)} \leq \mathbf{0}$. To show that $|\tilde{\mathbf{w}}_j - \tilde{\mathbf{v}}_j| \leq \rho_l\, \|\mathbf{w} - \mathbf{v}\|_\infty$ for $j \in \mathcal{I}_l^*$, there will be three cases to consider.

Case A. If j is *not* a contact index for both vectors, then $\tilde{\mathbf{w}}_j > \boldsymbol{\psi}_j$ and $\tilde{\mathbf{v}}_j > \boldsymbol{\psi}_j$. In this case, by linear complementarity, the following will hold:

$$\begin{cases} (A\tilde{\mathbf{w}})_j = \mathbf{f}_j \\ (A\tilde{\mathbf{v}})_j = \mathbf{f}_j \end{cases}$$

Subtracting the two, and using that $A_{jk} \leq 0$ for $j \neq k$ yields:

$$\begin{aligned} (A(\tilde{\mathbf{w}} - \tilde{\mathbf{v}}))_j = 0 &\Longrightarrow A_{jj}\,(\tilde{\mathbf{w}} - \tilde{\mathbf{v}})_j = -\textstyle\sum_{k \neq j} A_{jk}\,(\tilde{\mathbf{w}} - \tilde{\mathbf{v}})_k \\ &\Longrightarrow A_{jj}\left|(\tilde{\mathbf{w}} - \tilde{\mathbf{v}})_j\right| \leq -\textstyle\sum_{k \neq j} A_{jk}\,|(\tilde{\mathbf{w}} - \tilde{\mathbf{v}})_k|\,. \end{aligned}$$

Case B. If j is a contact index for only one of the vectors, say $\tilde{\mathbf{w}}_j = \boldsymbol{\psi}_j$. In this case, by linear complementarity, the following will hold:

$$\begin{cases} (A\tilde{\mathbf{v}})_j = \mathbf{f}_j \\ (A\tilde{\mathbf{w}})_j \geq \mathbf{f}_j \end{cases}$$

Subtracting the two equations and using that $A_{jk} \leq 0$ for $j \neq k$ yields:

$$\begin{aligned} (A(\tilde{\mathbf{v}} - \tilde{\mathbf{w}}))_j \leq 0 &\Longrightarrow A_{jj}\ (\tilde{\mathbf{v}} - \tilde{\mathbf{w}})_j \leq -\textstyle\sum_{k\neq j} A_{jk}\,(\tilde{\mathbf{v}} - \tilde{\mathbf{w}})_k \\ &\Longrightarrow A_{jj}\ \left|(\tilde{\mathbf{v}} - \tilde{\mathbf{w}})_j\right| \leq -\textstyle\sum_{k\neq j} A_{jk}\,|(\tilde{\mathbf{v}} - \tilde{\mathbf{w}})_k|\,. \end{aligned}$$

Case C. If j is a contact index for both vectors, then $\tilde{\mathbf{w}}_j = \tilde{\mathbf{v}}_j = \boldsymbol{\psi}_j$. In this case, the following bound will hold trivially:

$$A_{jj}\left|(\tilde{\mathbf{w}} - \tilde{\mathbf{v}})_j\right| \leq -\sum_{k\neq j} A_{jk}\ |(\tilde{\mathbf{w}} - \tilde{\mathbf{v}})_k|\,.$$

To estimate the local contraction factor, define a nodal vector $\mathbf{m} \in \mathrm{I\!R}^n$:

$$0 \leq (\mathbf{m})_j = |\tilde{\mathbf{v}}_j - \tilde{\mathbf{w}}_j|\,,$$

with nonnegative entries. Then, combining the three preceding cases yields:

$$A_{jj}\mathbf{m}_j + \sum_{k\neq j} \mathbf{m}_k \leq 0, \qquad \text{for } j \in \mathcal{I}_l^*.$$

Using comparison Thm. 15.20, we may estimate the majorants $\mathbf{m}_j$ for $j \in \overline{\mathcal{I}}_l$ in terms of $\beta\boldsymbol{\gamma}_I^{(l)}$ and $\beta\boldsymbol{\gamma}_B^{(l)}$ for the choice $\beta = \|\tilde{\mathbf{v}} - \tilde{\mathbf{w}}\|_{\infty,B^{(l)}}$. It yields:

$$\|\tilde{\mathbf{w}} - \tilde{\mathbf{v}}\|_{\infty,\overline{\mathcal{I}}_l} = \|\mathbf{m}\|_{\infty,\overline{\mathcal{I}}_l} \leq \rho_l\,\|\beta\gamma_B\|_{\infty,\mathcal{B}_l^*} = \rho_l\,\|\mathbf{v} - \mathbf{w}\|_{\infty,\mathcal{B}_l^*}.$$

We omit the details. □

The following result estimates the global contraction factor of T.

Lemma 14.31. *Let A be a strictly diagonally dominant M-matrix, and let $\alpha_1 = \cdots = \alpha_p = (1/p)$. Then, the parallel Schwarz iteration $T : \mathcal{H} \to \mathcal{H}$ will be a contraction map satisfying:*

$$\|T(\mathbf{v}) - T(\mathbf{w})\|_\infty \leq \frac{1}{p}\left((p-1) + \max_l \rho_l\right)\|\mathbf{v} - \mathbf{w}\|_\infty.$$

Proof. Follows by an application of the preceding result and convexity. □

Remark 14.32. For an elliptic operator L with $c(x) \geq c_0 > 0$, one can obtain estimates for the local contraction factors ρ_l independent of the mesh parameters, provided there is sufficient overlap between the subregions, see Chap. 15.

Remark 14.33. The above contraction property may also be used to establish the *solvability* of the linear complementarity problem, provided the local subproblems are solvable. When V_l is one-dimensional, we can easily show that each one-dimensional linear complementarity problem is solvable (using the strict diagonal dominance of A). We omit the details.

We next describe the convergence estimate for the sequential version of the Schwarz complementarity algorithm to solve (14.6).

Lemma 14.34. *Let A be a strictly diagonally dominant M-matrix and let ρ_l denote the subdomain contraction factor (14.16) for subdomain Ω_l^*. Given a starting iterate $\mathbf{u}^{(0)}$ satisfying $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ and $A\mathbf{u}^{(0)} \geq \mathbf{f}$, the sequential Schwarz iterates $\{\mathbf{u}^{(k)}\}$ will converge monotonically down to $\mathbf{u}$, satisfying:*

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}\|_\infty \leq \left(\max_l \rho_l\right) \|\mathbf{u}^{(k)} - \mathbf{u}\|_\infty.$$

Proof. The monotone nature of the fractional iterates $\mathbf{u} \leq \mathbf{u}^{(k+\frac{l}{p})} \leq \mathbf{u}^{(k+\frac{l-1}{p})}$ follows by Lemma 14.25. Employing the monotonicity of the iterates yields:

$$\begin{aligned} 0 \leq \left(\mathbf{u}^{(k+1)}(x) - \mathbf{u}(x)\right) &\leq \min_l \left(\mathbf{u}^{(k+\frac{l}{p})}(x) - \mathbf{u}(x)\right) \\ &\leq (\max_l \rho_l) \left(\mathbf{u}^{(k)}(x) - \mathbf{u}(x)\right) \\ &\leq (\max_l \rho_l) \|\mathbf{u}^{(k)} - \mathbf{u}\|_\infty \end{aligned}$$

where the contraction bounds follow by Lemma 14.30 for $x \in \overline{\Omega}_l$. □

A Two Sided Approximation of the Free Boundary ∂G. We end our discussion of monotone convergence results by considering a *two-sided* approximation of the *contact* set G in a variational inequality, see [KU12, KU13]. The contact set G is defined as the set where the continuous solution $u(x)$ is in contact with the obstacle $\psi(x)$:

$$G = \{x \in \Omega \, : u(x) = \psi(x)\}.$$

From a computational viewpoint, knowledge of the contact set G can be useful, since a linear complementarity problem becomes *linear* on $\left(\Omega\backslash\overline{G}\right)$, and linear solvers (which are cheaper) can be employed on such regions.

In the discrete case, a contact set G_h is defined as the index set:

$$G_h = \left\{j \, : \, (\mathbf{u})_j = \boldsymbol{\psi}_j \right\}.$$

A two-sided approximation of G_h can be determined using the *monotone* nature of Schwarz iterates [KU12, KU13]. A sequence of index sets $\{G_k\}$ and $\{\hat{G}_k\}$ can be constructed, "sandwiching" the discrete contact set G_h:

$$\cdots G_k \subset G_{k+1} \subset G_h \subset \hat{G}_{k+1} \subset \hat{G}_k \cdots$$

An *inner* approximation of G_h can be constructed based on the following observation. If the starting guess $\mathbf{u}^{(0)} \in \mathcal{K}_h$ satisfies $(A\,\mathbf{u}^{(0)} - \mathbf{f}) \geq \mathbf{0}$, then subsequent Schwarz iterates will satisfy:

$$\mathbf{u} \leq \cdots \leq \mathbf{u}^{(k+1)} \leq \mathbf{u}^{(k)} \leq \cdots \leq \mathbf{u}^{(0)}.$$

Thus, if we define:

$$G_k \equiv \left\{ j : \left(\mathbf{u}^{(k)}\right)_j = \boldsymbol{\psi}_j \right\},$$

then since $\mathbf{u}^{(k)}$ converges monotonically down to $\mathbf{u}$, it will hold that:

$$G_k \subset G_{k+1} \subset \cdots G_h.$$

Formally, $\{\partial G_k\}$ will approach ∂G monotonically.

Given an *inner* contact region G_k associated with Schwarz iterate $\mathbf{u}^{(k)}$, an *outer* approximation $\hat{G}_k \subset G_h$ can be constructed as follows. Note that $(\mathbf{u})_j = \left(\mathbf{u}^{(k)}\right)_j = \boldsymbol{\psi}_j$ for $j \in G_k$. Define $\hat{\mathbf{u}}^{(k)}$ as the solution to:

$$\begin{cases} \left(A\hat{\mathbf{u}}^{(k)}\right)_j = \mathbf{f}_j, & \text{for } j \notin G_k \\ \left(\hat{\mathbf{u}}^{(k)}\right)_j = \mathbf{u}_j^{(k)}, & \text{for } j \in G_k. \end{cases}$$

Then, by construction, $\left(\mathbf{u} - \hat{\mathbf{u}}^{(k)}\right)$ will satisfy:

$$\begin{cases} \left(A(\mathbf{u} - \hat{\mathbf{u}}^{(k)})\right)_j \geq 0, & \text{for } j \notin G_k \\ \left(\mathbf{u} - \hat{\mathbf{u}}^{(k)}\right)_j = 0, & \text{for } j \in G_k. \end{cases}$$

Applying the discrete maximum principle (since A is a strictly diagonally dominant M-matrix) yields $\hat{\mathbf{u}}^{(k)} \leq \mathbf{u}$. Thus, if we define:

$$\hat{G}_k \equiv \left\{ j \,:\, \left(\hat{\mathbf{u}}^{(k)}\right)_j \leq \boldsymbol{\psi}_j \right\},$$

then $G_h \subset \hat{G}_k$. Similarly, it can be shown that: $\hat{\mathbf{u}}^{(k+1)} \geq \hat{\mathbf{u}}^{(k)}$, from which it follows that:

$$G_h \subset \cdots \subset \hat{G}_{k+1} \subset \hat{G}_k \cdots$$

Thus, a *two-sided* approximation of G_h can be constructed as $G_k \subset G_h \subset \hat{G}_k$.

14.5 Applications to Parabolic Variational Inequalities

A parabolic variational inequality [EL, CR2, FR6, WI10] seeks a sufficiently smooth solution $u(x,t)$ on $\Omega \times (0,T)$ satisfying:

$$\begin{cases} (u_t + L\,u - f) \geq 0, & \text{in } \Omega \times (0,T) \\ (u - \psi) \geq 0, & \text{in } \Omega \times (0,T) \\ (u - \psi)\,(u_t + L\,u - f) = 0, & \text{in } \Omega \times (0,T) \\ u = g, & \text{on } \partial\Omega \times (0,T) \\ u(x,0) = u_0(x), & \text{in } \Omega, \end{cases} \tag{14.17}$$

where $L\,u = \ \nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u$ is an elliptic operator with sufficiently smooth coefficients, and $\psi(x,t)$ is a sufficiently smooth obstacle function. For compatibility of the data, we shall assume that $g(x,t) \geq \psi(x,t)$ on $\partial\Omega \times (0,T)$ and that $u_0(x) \geq \psi(x,0)$ on Ω. For simplicity, let $g(x,t) = 0$.

A finite difference discretization of (14.17) in *space* yields:

$$\begin{cases} (\mathbf{u}_t + A\mathbf{u} - \mathbf{f}) \geq \mathbf{0}, & \text{for } 0 < t < T \\ (\mathbf{u} - \boldsymbol{\psi}) \geq \mathbf{0}, & \text{for } 0 < t < T \\ (\mathbf{u} - \boldsymbol{\psi})^T\,(\mathbf{u}_t + A\mathbf{u} - \mathbf{f}) = 0, & \text{for } 0 < t < T \\ \mathbf{u}(0) = \mathbf{u}_0. \end{cases} \tag{14.18}$$

Here A denotes the matrix discretization of L. Discretization of the preceding system in time, using an *implicit* time stepping scheme, such as backward Euler, with step size $\tau = (T/N)$, will yield:

$$\begin{cases} \left((I + \tau\,A)\mathbf{u}^{(k+1)} - (\mathbf{u}^{(k)} + \tau\mathbf{f}^{(k+1)})\right) \geq \mathbf{0}, \\ \qquad \text{for } 0 \leq k \leq (N-1) \\ \left(\mathbf{u}^{(k+1)} - \boldsymbol{\psi}^{(k+1)}\right) \geq \mathbf{0}, \\ \qquad \text{for } 0 \leq k \leq (N-1) \\ \left(\mathbf{u}^{(k+1)} - \boldsymbol{\psi}^{(k+1)}\right)^T \left((I + \tau\,A)\mathbf{u}^{(k+1)} - (\mathbf{u}^{(k)} + \tau\mathbf{f}^{(k+1)})\right) = 0, \\ \qquad \text{for } 0 \leq k \leq (N-1) \\ \mathbf{u}(0) = \mathbf{u}_0. \end{cases} \tag{14.19}$$

At discrete time $(k+1)\,\tau$, given $\mathbf{u}^{(k)}$, $\mathbf{f}^{(k+1)}$ and the discrete obstacle $\boldsymbol{\psi}^{(k+1)}$, the preceding system of equations will require the solution of the following *linear complementarity* problem to determine $\mathbf{u}^{(k+1)}$:

$$\begin{cases} \left((I + \tau\,A)\mathbf{u}^{(k+1)} - (\mathbf{u}^{(k)} + \tau\mathbf{f}^{(k+1)})\right) \geq \mathbf{0} \\ \left(\mathbf{u}^{(k+1)} - \boldsymbol{\psi}^{(k+1)}\right) \geq \mathbf{0} \\ \left(\mathbf{u}^{(k+1)} - \boldsymbol{\psi}^{(k+1)}\right)^T \left((I + \tau\,A)\mathbf{u}^{(k+1)} - (\mathbf{u}^{(k)} + \tau\mathbf{f}^{(k+1)})\right) = 0. \end{cases} \tag{14.20}$$

Methods described in preceding sections can be employed to solve such linear complementarity problems, including Schwarz algorithms. In particular, if Schwarz algorithms are employed with subdomains of size h_0, then *heuristics* suggest that a coarse space may not be necessary when $\tau \leq C\,h_0^2$.

Discretization of (14.18) by an *explicit* scheme in time, such as forward Euler, will yield:

$$\begin{cases} \left(\mathbf{u}^{(k+1)} - (I - \tau\, A)\mathbf{u}^{(k)} - \tau \mathbf{f}^{(k)}\right) \geq \mathbf{0}, \\ \quad \text{for } 0 \leq k \leq (N-1) \\ \left(\mathbf{u}^{(k+1)} - \boldsymbol{\psi}^{(k+1)}\right) \geq \mathbf{0}, \\ \quad \text{for } 0 \leq k \leq (N-1) \\ \left(\mathbf{u}^{(k+1)} - \boldsymbol{\psi}^{(k+1)}\right)^T \left(\mathbf{u}^{(k+1)} - (I - \tau\, A)\mathbf{u}^{(k)} - \tau \mathbf{f}^{(k)}\right) = 0, \\ \quad \text{for } 0 \leq k \leq (N-1) \\ \mathbf{u}(0) = \mathbf{u}_0. \end{cases} \tag{14.21}$$

In this case, at discrete time $(k+1)\,\tau$, given $\mathbf{u}^{(k)}$, $\mathbf{f}^{(k+1)}$ and the discrete obstacle $\boldsymbol{\psi}^{(k+1)}$, the preceding system of equations yields an immediate solution to the linear complementarity problem:

$$\mathbf{u}_j^{(k+1)} = \max\left\{\boldsymbol{\psi}_j^{(k+1)}, \left(\mathbf{u}^{(k)} - \tau\, A\mathbf{u}^{(k)} - \tau\, \mathbf{f}^{(k)}\right)_j\right\}, \quad \text{for } 1 \leq j \leq n.$$

This is similar to a projected gradient method. When matrix $A = A^T > 0$, convergence and stability can be guaranteed if the time step parameter $\tau > 0$ is chosen so that $\|I - \tau\, A\| \leq \rho < 1$. We summarize the resulting algorithm.

Algorithm 14.5.1 *(Explicit Time Stepping Algorithm)*
Given a starting iterate satisfying $\mathbf{u}^{(0)} \geq \boldsymbol{\psi}$ *and* $A\,\mathbf{u}^{(0)} - \mathbf{f} \geq \boldsymbol{\psi}$

```
1. For $k = 0, 1, \ldots,$ until convergence do:
2.       For $j = 1, \ldots, n$ update:
3.            $\mathbf{u}_j^{(k+1)} = \max\{\left(\mathbf{u}^{(k)} - \tau\, A\,\mathbf{u}^{(k)} - \tau\, \mathbf{f}^{(k)}\right)_j, \boldsymbol{\psi}_j^{(k+1)}\}$
4.       Endfor
5. Endfor
```

Readers are referred to [CR2, EL, WI10, WI11] for applications.

15

Maximum Norm Theory

In this chapter, we describe theoretical results on the *maximum norm* stability and convergence of Schwarz domain decomposition methods. The tools we describe are continuous and discrete versions of maximum principles and comparison theorems for elliptic equations, as well as methods for estimating the maximum norm error *contraction* factor on subdomains. From a matrix viewpoint, these tools are applicable when the discretization of the given elliptic equation results in a strictly diagonally dominant *M-matrix.*

We shall consider an elliptic equation of the form:

$$\begin{cases} L\,u = -\nabla \cdot (a(x)\,\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u = f, \text{ in } \Omega \\ \hfill u = g, \text{ on } \partial\Omega \end{cases} \tag{15.1}$$

with coefficients $a(x) \geq 0$ and $c(x) \geq c_0 > 0$. In some applications $a(x) = 0$. We consider a finite difference discretization of (15.1) which yields the system:

$$\begin{cases} A_{II}\mathbf{u}_I + A_{IB}\mathbf{u}_B = \mathbf{f}_I \\ \hfill \mathbf{u}_B = \mathbf{g}_B, \end{cases} \tag{15.2}$$

where $\mathbf{u}_I$, $\mathbf{u}_B$ are nodal vectors corresponding to the *interior* nodes of Ω and boundary nodes $\partial\Omega$. Matrix A_{II} will be assumed to be an M-matrix and A_{IB} to have nonpositive entries. Various approximations, based on *upwind* discretization of the convection term $\mathbf{b}(x) \cdot \nabla u$, satisfy these properties.

Our discussion of maximum norm theory for Schwarz algorithms is based on [LI7, CH6, MI, ST, GA8, BL3, CH22, GA10, MA33, CA17, GA12, MA35]. Related literature includes [CI, CI5, OH, HA6, IK, MI3, YA3, KA2, TA8], [SU, FE3, FR7, FR8]. Chap. 15.1 describes maximum principles and comparison theorems Chap. 15.2 describes the well posedness of the Schwarz hybrid formulation. Chap. 15.3 describes the maximum norm convergence of Schwarz iterative algorithms that do not employ coarse space correction. In Chap. 15.4 nonmatching grid Schwarz discretizations are analyzed. Chap. 15.5 analyzes the well posedness and accuracy of Schwarz heterogeneous approximations. In Chap. 15.6, we briefly indicate extensions to parabolic equations.

15.1 Maximum Principles and Comparison Theorems

In this section, we introduce theoretical tools that will be employed to analyze the convergence and stability of Schwarz methods in the maximum norm. We describe continuous and discrete *maximum principles* and *comparison theorems*, and indicate how comparison theorems can be employed to estimate the *contraction factor* of harmonic functions on subregions. The stability and rate of convergence of Schwarz hybrid formulations depend on such contraction factors on individual subdomains, and such factors can be estimated by constructing comparison (or "barrier") functions. Under appropriate assumptions, this methodology applies to elliptic, hyperbolic and parabolic equations.

Maximum Principles, Comparison Theorems, Contraction Factors. Let $\Omega_1, \ldots, \Omega_p$ form a *nonoverlapping* decomposition of Ω with subdomains of size h_0. We shall construct an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ where each Ω_l^* is obtained by extension of Ω_l by $\beta\, h_0$. We will employ the notation $B^{(l)} = \partial\Omega_l^* \cap \Omega$ and $B_{[l]} = \partial\Omega_l^* \cap \partial\Omega$. Additionally, $\|w\|_{\infty,\mathcal{S}}$ will denote the maximum norm of $w(x)$ on the set $\mathcal{S}$. Below, we state a weak version of the maximum principle for the elliptic operator L defined in (15.1) with $c(x) \geq 0$.

Lemma 15.1. *Suppose $a(x) \geq a_0 > 0$. Let $L\, w(x) = 0$ in Ω_l^* with Dirichlet data $w(x) = g(x)$ on $\partial\Omega_l^*$. Then, the following results will hold:*

$$\begin{cases} w(x) \leq \max_{\{\tilde{x} \in \partial\Omega_l^*\}} g(\tilde{x}) & \text{when } c(\cdot) = 0 \\ w(x) \geq \min_{\{\tilde{x} \in \partial\Omega_l^*\}} g(\tilde{x}) & \text{when } c(\cdot) = 0 \\ \|w\|_{\infty,\overline{\Omega}_l^*} \leq \|g\|_{\infty,\partial\Omega_l^*} & \text{when } c(\cdot) \geq 0. \end{cases} \tag{15.3}$$

When $c(x) \geq c_0 > 0$, a stronger result can be shown on $\Omega_l \subset \Omega_l^*$:

$$\|w\|_{\infty,\Omega_l} \leq \rho_l\, \|g\|_{\infty,\partial\Omega_l^*}, \tag{15.4}$$

for some $0 \leq \rho_l < 1$ (referred to as the local contraction factor).

Proof. See [JO, SM7, GI] . □

Remark 15.2. A modified result can be shown to hold when $a(x) = 0$. In this case operator L will be *hyperbolic*. Let $\Gamma_{l,in}$ denote the *inflow* boundary segment of $\partial\Omega_l^*$ and let $\mathbf{n}(x)$ denote the unit exterior normal to Ω_l^*:

$$\Gamma_{l,in} \equiv \{x \in \partial\Omega_l^* \,:\, \mathbf{n}(x) \cdot \mathbf{b}(x) < 0\}. \tag{15.5}$$

Then, the following can be shown to hold:

$$\begin{cases} w(x) \leq \max_{\{\tilde{x} \in \Gamma_{l,in}\}} g(\tilde{x}) & \text{when } c(\cdot) = 0 \\ \|w\|_{\infty,\Omega_l^*} \leq \|g\|_{\infty,\Gamma_{l,in}} & \text{when } c(\cdot) \geq 0. \end{cases} \tag{15.6}$$

When $c(x) \geq c_0 > 0$, a stronger result can be shown:

$$\|w\|_{\infty,\Omega_l} \leq \rho_l \, \|g\|_{\infty,\Gamma_{l,*}}, \tag{15.7}$$

for some $0 \leq \rho_l < 1$. These results can be proved formally by employing the method of characteristics, *provided* the characteristic curves fill Ω_l^*, and provided the coefficients and boundary are smooth.

We next describe a *comparison* theorem for elliptic equations. Such a result makes a pointwise comparison between two sufficiently smooth solutions to an elliptic equation (or inequality) on Ω_l^* with Dirichlet data.

Lemma 15.3. *Let $a(x) \geq a_0 > 0$ and $c(x) \geq 0$ on Ω_l^*. Suppose that:*

$$\begin{cases} L\,w_1(x) = f_1(x), \text{ in } \Omega_l^* \\ \quad w_1(x) = g_1(x), \text{ on } \partial\Omega_l^*. \end{cases} \text{and} \begin{cases} L\,w_2(x) = f_2(x), \text{ in } \Omega_l^* \\ \quad w_2(x) = g_2(x), \text{ on } \partial\Omega_l^*. \end{cases}$$

Then, if $f_1(x) \geq f_2(x)$ and $g_1(x) \geq g_2(x)$, the following result will hold:

$$w_1(x) \geq w_2(x), \text{ in } \Omega_l^*. \tag{15.8}$$

Proof. See [JO, SM7, GI]. □

We next associate a *contraction factor* $0 \leq \rho_l \leq 1$ with an elliptic operator on a domain. It will represent the reduction in the magnitude of a *homogeneous* solution to an elliptic equation within an *interior* region $\overline{\Omega}_l \subset \Omega_l^*$.

Definition 15.4. *Let $a(x) \geq a_0 > 0$ and $c(x) \geq c_0 > 0$, and let $w_*(x)$ solve:*

$$\begin{cases} L\,w_*(x) = 0, & \text{for } x \in \Omega_l^* \\ \quad w_*(x) = 1, & \text{on } B^{(l)} \\ \quad w_*(x) = 0, & \text{on } B_{[l]} \end{cases} \tag{15.9}$$

Then, the contraction factor ρ_l of L from Ω_l^* to Ω_l is defined as:

$$\rho_l \equiv \max_{\overline{\Omega}_l} w_*(x) = \|w_*\|_{\infty,\Omega_l}, \tag{15.10}$$

within the interior region Ω_l of Ω_l^*, where $\operatorname{dist}\left(B^{(l)}, \partial\Omega_l \cap \Omega\right) \geq \beta\, h_0$.

Remark 15.5. An application of the comparison theorem yields $w_*(x) \geq 0$. Since $c(x) \geq c_0 > 0$, by the strong maximum principle it will hold that $0 \leq \rho_l < 1$. Subdomain contraction factors will be an important parameter in maximum norm convergence theory, for estimating the maximum norm stability of Schwarz hybrid formulations (in the discrete case) and the convergence of Schwarz iterative algorithms.

The next result shows how a contraction factor and the comparison theorem, enables estimating the *interior* maximum norm of a L-harmonic function.

Lemma 15.6. *Suppose $a(x) \geq a_0 > 0$ and $c(x) \geq c_0 > 0$, and let $v(x)$ be a sufficiently smooth solution to $L\, v(x) = 0$ in Ω_l^* satisfying $v(x) = g(x)$ on $B^{(l)}$ and $v(x) = 0$ on $B_{[l]}$. Then, the following bound will hold on $\Omega_l \subset \Omega_l^*$:*

$$\begin{aligned} v(x) &\leq \left(\|g\|_{\infty, B^{(l)}}\right) w_*(x) \quad \text{in } \overline{\Omega}_l \\ \|v\|_{\infty, \Omega_l} &\leq \rho_l \, \|g\|_{\infty, B^{(l)}}, \end{aligned}$$

where $w_*(x)$ is the comparison function defined in (15.9).

Proof. Apply the comparison theorem using $w_1(x) = \left(\|g\|_{\infty, B^{(l)}}\right) w_*(x)$, where $w_*(x)$ is defined by (15.9), and $w_2(x) = v(x)$. By construction, it will hold that $L\, w_1(x) = f_1(x) = 0 \geq f_2(x) = 0$ in Ω_l^*. Additionally, it will hold that $w_1(x) = \|g\|_{\infty, B^{(l)}} \geq w_2(x) = g(x)$, on $B^{(l)}$, with $w_1(x) = 0 \geq w_2(x) = 0$ on $B_{[l]}$. Repeat the arguments using $-v(x)$. □

We next describe the construction of *comparison functions* (or "barrier functions") to explicitly estimate the contraction factor associated with an elliptic operator L on a domain Ω_l^*. We shall assume that $a(x) \geq a_0 > 0$ and $c(x) \geq c_0 > 0$. Additionally, to simplify our discussion, we shall assume there exits a sufficiently smooth function $\mathrm{d}(x) \geq 0$ satisfying:

$$\begin{cases} \mathrm{d}(x) = 0, & \text{on } B^{(l)} \\ \mathrm{d}(x) \geq \operatorname{dist}\left(x, B^{(l)}\right) & \text{for } x \in \overline{\Omega}_l, \end{cases} \tag{15.11}$$

where $\operatorname{dist}(x, B^{(l)})$ denotes the minimal distance between x and $B^{(l)}$. Indeed, if $\operatorname{dist}(x, B^{(l)})$ is sufficiently smooth, choose $\mathrm{d}(x) = \operatorname{dist}(x, B^{(l)})$. Otherwise $\operatorname{dist}(x, B^{(l)})$ will need to be appropriately mollified.

Lemma 15.7. *Let $w_*(x)$ satisfy (15.9) and let $\mathrm{d}(x) \geq 0$ satisfy (15.11). Also, let $c(x) \geq c_0 > 0$. Then, there will exist $\gamma > 0$ such that:*

$$0 \leq w_*(x) \leq e^{-\gamma\, \mathrm{d}(x)}.$$

Proof. We shall outline the proof sketched in [LI7] (see also [MA33, MA35]). Substitute $z(x) = e^{-\gamma\, \mathrm{d}(x)}$ into $L\, z(x)$ to obtain:

$$\begin{cases} L\, z(x) = \quad \gamma \left(\nabla a(x) \cdot \nabla d(x) + a(x)\, \Delta d(x)\right) z(x) \\ \qquad\qquad - \gamma \left(\gamma\, a(x)\, |\nabla d(x)|^2 + \mathbf{b}(x) \cdot \nabla d(x)\right) z(x) + c(x)\, z(x), \text{ in } \Omega_l^* \\ \quad z(x) \geq 1, \text{ on } B^{(l)} \\ \quad z(x) \geq 0, \text{ on } B_{[l]}. \end{cases}$$

When the coefficients are sufficiently smooth, select $\gamma > 0$ such that:

$$\gamma \left(|\nabla a \cdot \nabla d| + |a\, \Delta d| + a\, |\nabla d|^2 + |\nabla d \cdot \mathbf{b}| \right) \leq \frac{1}{2} c_0, \text{ in } \Omega_l^*.$$

For such a choice of γ we obtain $L\, z(x) \geq \frac{1}{2}\, c_0\, z(x) \geq 0$ in Ω_l^* (parameter γ will depend on $\mathrm{d}(x)$, $a(x)$ and $\mathbf{b}(x)$). Applying the comparison theorem using $w_1(x) = z(x)$ and $w_2(x) = w_*(x)$ yields the desired result. □

Remark 15.8. When $c(x) \geq c_0 > 0$, the preceding estimate shows that the *contraction factor* ρ_l decreases with increasing distance between Ω_l and $B^{(l)}$:

$$\rho_l \leq \max_{\{x \in \overline{\Omega}_l\}} \left(e^{-\gamma\, \mathrm{d}(x)} \right) = e^{-\gamma\, \mathrm{d}_l}, \tag{15.12}$$

where $d_l = \mathrm{dist}(B^{(l)}, \partial\Omega_l \cap \Omega)$. Formally, this result should hold even when $a(x) = 0$. It will be shown later that for sufficiently small grid size h, a similar contraction factor will apply in the discrete case.

In applications involving *singularly* perturbed elliptic equations, often better estimates can be obtained for the contraction factors. The following result describes the construction of a comparison function for a singularly perturbed elliptic equation obtained by time stepping a parabolic equation using a time step τ, see [KU3, KU6, MA33]. The result shows that ρ_l decreases rapidly as $\tau \to 0^+$. Heuristically, this may be expected, since as $\tau \to 0^+$ the homogeneous solution to the formal limiting equation approaches zero.

Lemma 15.9. *Let* $0 < \tau \ll 1$ *denote a time step and let* $w_*(x)$ *solve:*

$$\begin{cases} L\, w_*(x) = -\tau\, \nabla \cdot (\tilde{a}(x) \nabla w_*) + \tau\, \tilde{\mathbf{b}}(x) \cdot \nabla w_*(x) + c(x)\, w_*(x) = 0, \text{ in } \Omega_l^* \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad w_*(x) = 1, \text{ on } B^{(l)} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad w_*(x) = 0, \text{ on } B_{[l]}, \end{cases} \tag{15.13}$$

where $c(x) \geq c_0 > 0$. *Then, there exists* $\gamma > 0$ *such that:*

$$\rho_l = \max_{\{x \in \overline{\Omega}_l\}} w_*(x) \;\leq\; e^{-\frac{\gamma}{\sqrt{\tau}} \mathrm{d}_l}, \tag{15.14}$$

where $\mathrm{d}_l = \mathrm{dist}\left(B^{(l)}, \partial\Omega_l \cap \Omega\right)$.

Proof. Apply the comparison theorem using $w_1(x) = z(x) = e^{-\frac{\gamma}{\sqrt{\tau}} \mathrm{d}(x)}$ and $w_2(x) = w_*(x)$, where $\mathrm{d}(x)$ satisfies (15.11). Computing $L\, z(x)$ yields:

$$\begin{cases} L\, z(x) = \quad \gamma\, (\sqrt{\tau}\, \nabla\tilde{a}(x) \cdot \nabla d(x) + \sqrt{\tau}\, \tilde{a}(x)\, \Delta d(x))\; z(x) \\ \qquad\qquad - \gamma \left(\gamma\, \tilde{a}(x)\, |\nabla d(x)|^2 + \sqrt{\tau}\, \nabla d(x) \cdot \tilde{\mathbf{b}}(x) \right) z(x) + c(x)\, z(x), \text{ in } \Omega_l^* \\ \quad z(x) \geq 1, \text{ on } B^{(l)} \\ \quad z(x) \geq 0, \text{ on } B_{[l]}. \end{cases}$$

When the coefficients of L are smooth, select $\gamma > 0$ such that:

$$\gamma \left(|\nabla\tilde{a} \cdot \nabla d| + |\tilde{a}\, \Delta d| + \tilde{a}\; |\nabla d|^2 + \left| \tilde{\mathbf{b}} \cdot \nabla d \right| \right) \leq \frac{1}{2}\, c_0.$$

Then, for $0 \leq \tau \leq 1$, an application of the comparison theorem will yield the estimate $0 \leq w_*(x) \leq e^{-\frac{1}{\sqrt{\tau}}\, \mathrm{d}(x)}$. The desired result follows. □

The preceding contraction factor decreases exponentially as $\tau \to 0^+$. We shall next estimate the contraction factor of an *advection dominated* elliptic equation on Ω_l^*, in the special case in which the interior boundary segment $B^{(l)}$ is contained in the *outflow* boundary $\Gamma_{l,out}$ of the limiting advection (hyperbolic) equation. We shall estimate the contraction factor ρ_l of the following homogeneous advection dominated elliptic equation for $0 < \epsilon \ll 1$:

$$\begin{cases} L_\epsilon \, w_* = -\epsilon \, \nabla \cdot (\tilde{a}(x) \nabla w_*) + \mathbf{b}(x) \cdot \nabla w_* + c(x) \, w_* = 0, \text{ in } \Omega_l^* \\ \hfill w_*(x) = 1, \text{ on } B^{(l)} \\ \hfill w_*(x) = 0, \text{ on } B_{[l]}, \end{cases} \tag{15.15}$$

in the special case where $B^{(l)} \subset \Gamma_{l,out}$, where $\Gamma_{l,out}$ denotes the *outflow* boundary segment of the limiting operator $L_0 \, w_* = \mathbf{b}(x) \cdot \nabla w_* + c(x) \, w_*$:

$$\Gamma_{l,out} = \left\{ x \in \partial \Omega_l^* \; : \; \mathbf{n}(x) \cdot \mathbf{b}(x) > 0 \right\}, \tag{15.16}$$

where $\mathbf{n}(x)$ denotes the unit exterior normal to $\partial \Omega_l^*$. The following result describes a distance-like function $q(x) \geq 0$ defined in Ω_l^*, constructed using characteristic curves from x to $B^{(l)}$.

Lemma 15.10. *Let Ω_l^* be smooth of diameter h_0, with $B^{(l)} \subset \Gamma_{l,out}$.*

1. *Let the Euclidean norm $\|\mathbf{b}(x)\| \geq b_0$ for $x \in \Omega_l^*$.*
2. *Let the characteristic curves of $L_* \, q \equiv -\left(\frac{\mathbf{b}(x)}{\|\mathbf{b}(x)\|} \right) \cdot \nabla q$ fill Ω_l^*.*
3. *Let $0 \leq \psi(x) \leq h_0$ denote a smooth function on $\Gamma_{l,out}$ satisfying $\psi(x) = 0$ if $x \in B^{(1)}$ and $\psi(x) = h_0$ if $\operatorname{dist}(x, B^{(l)}) > \eta$ for some $0 < \eta \ll \beta \, h_0$.*

Then, there exists a function $q(x) \geq \operatorname{dist}(x, B^{(l)})$ defined in Ω_l^* satisfying:

$$\begin{cases} -\left(\frac{\mathbf{b}(x)}{\|\mathbf{b}(x)\|} \right) \cdot \nabla q(x) = 1, \qquad \text{in } \Omega_l^* \\ \hfill q(x) = \psi(x), \text{ on } \Gamma_{l,out}. \end{cases} \tag{15.17}$$

Proof. Solve the following hyperbolic equation for $q(x)$:

$$\begin{cases} -\frac{\mathbf{b}(x)}{\|\mathbf{b}(x)\|} \cdot \nabla q = 1, \qquad \text{in } \Omega_l^* \\ \hfill q(x) = \psi(x), \text{ on } \Gamma_{l,out}. \end{cases} \tag{15.18}$$

Employ the method of characteristics using $x(s)$ to denote the characteristic curve and $q(x(s))$ to denote the solution along $x(s)$. This yields the following coupled ordinary differential equations for each $x_0 \in \Gamma_{l,out}$:

$$\begin{cases} \dfrac{dq}{ds} = 1 \\ \dfrac{dx}{ds} = \left(\frac{-\mathbf{b}(x)}{\|\mathbf{b}(x)\|} \right) \end{cases} \quad \text{with initial conditions} \quad \begin{cases} q(s = 0) = \psi(x_0) \\ x(s = 0) = x_0. \end{cases}$$

Since $\mathbf{n}(x_0) \cdot \mathbf{b}(x_0) > 0$, the characteristic curve $x(s)$ will traverse *into* Ω_l^*. By assumption on $\mathbf{b}(x)$, these curves will fill Ω_l^*. Since $\mathbf{b}(x)/\|\mathbf{b}(x)\|$ is a unit vector field, parameter s will represent arclength. If $\psi(x_0) = 0$, then $q(s) = s$ will denote the arclength distance from x_0 to $x(s)$ along the characteristic curve. The desired result now follows. □

Remark 15.11. Since $dq/ds = 1$ and $\psi(x_0) = 0$ for $x_0 \in B^{(l)}$, and $\psi(x_0) = h_0$ for $x_0 \in B^{(l)}$ it follows that $q(x(s)) = \psi(x_0) + s \geq \operatorname{dist}(x, B^{(l)})$, since s represents arclength distance along a characteristic curve, by construction.

Below, employing $q(x)$, we construct a comparison function for (15.15).

Lemma 15.12. *Let $q(x)$ defined by (15.17) satisfy $q(x) \geq \operatorname{dist}(x, B^{(l)})$. Also, let $c(x) \geq c_0 > 0$. Then, there exists $\gamma > 0$ for which the following holds:*

$$0 \leq w_*(x) \leq e^{-\frac{\gamma}{\epsilon} q(x)}.$$

Proof. Substitute the ansatz $z(x) = e^{-\frac{\gamma}{\epsilon} q(x)}$ to obtain:

$$\begin{cases}
\begin{aligned} L\, z(x) &= \gamma\,(\nabla \tilde{a} \cdot \nabla q + \tilde{a}\, \Delta q)\, z(x) \\ &\quad + \left(c(x) - \tfrac{\gamma}{\epsilon} \mathbf{b} \cdot \nabla q - \tfrac{\gamma^2}{\epsilon}\, \tilde{a}\, |\nabla q|^2 \right) z(x) \end{aligned} & \text{in } \Omega_l^* \\
\begin{aligned} &= \gamma\,(\nabla \tilde{a} \cdot \nabla q + \tilde{a}\, \Delta q)\, z(x) \\ &\quad + \left(c(x) + \tfrac{\gamma}{\epsilon} \|\mathbf{b}\| - \tfrac{\gamma^2}{\epsilon}\, \tilde{a}\, |\nabla q|^2 \right) z(x) \end{aligned} & \text{in } \Omega_l^* \\
\begin{aligned} &\geq \gamma\,(\nabla \tilde{a} \cdot \nabla q + \tilde{a}\, \Delta q)\; z(x) \\ &\quad + \left(c(x) + \tfrac{\gamma}{\epsilon}\,(b_0 - \gamma\, \tilde{a}\, |\nabla q|^2) \right) z(x) \end{aligned} & \text{in } \Omega_l^* \\
z(x) \geq 1, & \text{on } B^{(l)} \\
z(x) \geq 0, & \text{on } B_{[l]}.
\end{cases}$$

Since $b_0 > 0$, we may choose γ sufficiently small, so that:

$$\gamma\,(|\nabla \tilde{a} \cdot \nabla q| + \tilde{a}\; |\Delta q|) \leq \frac{1}{2} c_0 \qquad \text{and} \qquad \gamma\, \tilde{a}\, |\nabla q|^2 \leq b_0.$$

This will yield $L\, z(x) \geq \frac{c_0}{2} z(x) \geq 0$. Applying the comparison theorem using $w_1(x) = z(x)$ and $w_2(x) = w_*(x)$ yields the desired result. □

Remark 15.13. The preceding result shows that if $B^{(l)}$ is contained within the *outflow* segment $\Gamma_{l,out}$ of the limiting hyperbolic problem on Ω_l^* as $\epsilon \to 0^+$, the contraction factor ρ_l associated with $w_*(x)$ decreases exponentially:

$$\rho_l \leq e^{-\frac{\gamma}{\epsilon}\, \mathrm{d}_l},$$

where $\mathrm{d}_l = \operatorname{dist}(B^{(l)}, \partial\Omega_l \cap \Omega)$. A heuristic explanation for this rapid contraction is that as $\epsilon \to 0^+$, the homogeneous solution $w_*(x)$ of (15.15) approaches zero in Ω_l^* due to zero boundary conditions on $B_{[l]}$, which contains the *inflow* boundary of the limiting hyperbolic equation.

Remark 15.14. Similar estimates may be obtained for *anisotropic* problems. Consider $L_\epsilon\, u = -\epsilon\, u_{x_1 x_1} - u_{x_2 x_2} + c(x)\, u = f(x)$ on $\Omega \subset \mathbb{R}^2$ for $0 < \epsilon \ll 1$. For such anisotropy, choose $\Omega_l^* \equiv \Omega \cap \{\theta_l < x_1 < \theta_{l+1}\}$ as strip like subdomains. A barrier function of the form $z(x) = e^{-\frac{\gamma}{\sqrt{\epsilon}}\, q(x)}$ can be constructed, where $0 \le q(x) \equiv q(x_1)$ depends only on x_1 (for instance $q(x_1, x_2) = \min\{|x_1 - y_1|\}$ for $y = (y_1, y_2) \in B^{(l)} = \partial\Omega_l^* \cap \Omega$). We omit the details.

Discrete Maximum, Comparison and Contraction Principles. We shall now describe *discrete* analogs of maximum principles, comparison theorems and contraction factors. These results will be employed to analyze the stability and convergence of Schwarz discretizations, iterative algorithms and heterogeneous approximations in the maximum norm. Our focus will be on *finite difference* discretizations in which the coefficient matrix is a strictly diagonally dominant M-matrix, see [VA9, SA2].

Definition 15.15. *A matrix K of size n is said to be an M-matrix if:*

1. $K_{ii} > 0$ *for each* $1 \le i \le n$.
2. $K_{ij} \le 0$ *for* $i \ne j$.
3. K *is nonsingular with* $\left(K^{-1}\right)_{ij} \ge 0$ *for each* i, j.

It is easily shown that if $K_{ij} \le 0$ for $j \ne i$, and if K is a strictly diagonally dominant matrix with: $K_{ii} > \sum_{j \ne i} |K_{ij}|$, for $1 \le i \le n$, then K will be an M-matrix, see [SA2].

We shall assume that the finite difference discretization (15.2) of elliptic equation (15.1) in which $c(x) \ge c_0 > 0$ results in a matrix A_{II} which is a strictly diagonally dominant M-matrix and yields $\left(A_{IB}\right)_{ij} \le 0$:

$$\begin{cases} A_{II}\mathbf{u}_I + A_{IB}\mathbf{u}_B = \mathbf{f}_I \\ \qquad\qquad\qquad \mathbf{u}_B = \mathbf{g}_B, \end{cases}$$

Here $\mathbf{u}_I$ denotes a nodal vector of size n corresponding to nodal unknowns at *interior* grid points $x_1, \ldots, x_n$ of Ω, while $\mathbf{u}_B$ denotes a nodal vector of size m corresponding to nodal unknowns on the *boundary* $\partial\Omega$.

Notation. Given nodes x_i in $\overline{\Omega}$, let $\mathcal{I}_l^*$ and $\mathcal{I}_l$ denote the *global* indices of nodes in $\overline{\Omega}_l^*$ and $\overline{\Omega}_l$, respectively. The nodes in a subdomain will be given a local ordering, with $\mathcal{J}_l^*$ and $\mathcal{J}_l$ denoting the *local* indices of nodes in $\overline{\Omega}_l^*$ and $\overline{\Omega}_l$, respectively. These local nodes in $\overline{\Omega}_l^*$ will be denoted as $\{y_i^{(l)}\}$. By construction, $\{y_j^{(l)}\}_{j \in \mathcal{J}_l}$ and $\{y_j^{(l)}\}_{j \in \mathcal{J}_l^*}$ should correspond to the same nodes as $\{x_i\}_{i \in \mathcal{I}_l}$ and $\{x_i\}_{i \in \mathcal{I}_l^*}$.

On each subdomain, let nodal vectors $\mathbf{u}_I^{(l)} \in \mathbb{R}^{n_l}$ and $\mathbf{u}_B^{(l)} \in \mathbb{R}^{m_l}$ satisfy:

$$\begin{cases} A_{II}^{(l)}\mathbf{u}_I^{(l)} + A_{IB}^{(l)}\mathbf{u}_B^{(l)} = \mathbf{f}_I^{(l)} \\ \qquad\qquad\qquad\quad \mathbf{u}_B^{(l)} = \mathbf{g}_B^{(l)}, \end{cases} \tag{15.19}$$

corresponding to a discretization of:

$$\begin{cases} L\,u = -\nabla \cdot (a(x)\,\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u = f, & \text{in } \Omega_l^* \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\; u = g, & \text{on } \partial\Omega_l^*. \end{cases} \tag{15.20}$$

Matrices $A_{II}^{(l)}$ and $A_{IB}^{(l)}$ will be submatrices of $[A_{II} \;\; A_{IB}]$ for indices in $\mathcal{I}_l^*$. Matrix $A_{II}^{(l)}$ will be a strictly diagonally dominant M-matrix, and $A_{IB}^{(l)} \leq 0$.

The following result describes a *discrete maximum principle*. We consider a rectangular matrix $A = \left[A_{II} \;\; A_{IB} \right]$, where A_{II} is of size n and A_{IB} of size $n \times m$, corresponding to the coefficient matrix for a Dirichlet problem. We shall assume that matrix A_{II} is *irreducible* (i.e., the incidence matrix associated with A is connected) and that $A_{IB} \neq 0$.

Lemma 15.16. *Let $A = \left[A_{II} \;\; A_{IB} \right]$ be of size $n \times (n+m)$ satisfy $A_{ii} > 0$ for each i and $A_{ij} \leq 0$ for $i \neq j$. Let A be weakly diagonally dominant with zero row sums, i.e., $A_{ii} = \sum_{j \neq i} |A_{ij}|\,, \quad \text{for each } i$, and let A_{II} be irreducible. Let $\mathcal{N}(i) \equiv \{j \,:\, A_{ij} \neq 0, \text{ where } j \neq i\}$ denote the index set of neighbours of node i. Then, the following will hold:*

1. *If $\mathbf{u} = \left(\mathbf{u}_I^T, \mathbf{u}_B^T\right)^T \in \mathbb{R}^{n+m}$ satisfies $(A\mathbf{u})_i \leq 0$ for some $1 \leq i \leq n$, then:*
$$\mathbf{u}_i \;\leq\; \max_{j \in \mathcal{N}(i)} \mathbf{u}_j. \tag{15.21}$$
2. *If $(A\mathbf{u})_i = 0$ for some $1 \leq i \leq n$, then:*
$$|\mathbf{u}_i| \;\leq\; \max_{j \in \mathcal{N}(i)} |\mathbf{u}_j|\,. \tag{15.22}$$
3. *If $A\,\mathbf{u} = \mathbf{0}$, then either $\mathbf{u}_I = \alpha\,(1, \ldots, 1)^T \in \mathbb{R}^n$ for some $\alpha \in \mathbb{R}$ (i.e., each entry of $\mathbf{u}_I$ will be identical), or $\|\mathbf{u}_I\|_\infty \leq \|\mathbf{u}_B\|_\infty$.*

Proof. Since $A_{ii} > 0$, the condition that $(A\mathbf{u})_i \leq 0$ reduces to:

$$\mathbf{u}_i \leq \sum_{j \in \mathcal{N}(i)} \left(\frac{-A_{ij}}{A_{ii}} \right) \mathbf{u}_j.$$

Weak diagonal dominance of A yields zero row sums, with $A_{ii} = -\sum_{j \neq i} A_{ij}$, since $A_{ij} \leq 0$ for $j \neq i$, with $\left(\frac{-A_{ij}}{A_{ii}}\right) \geq 0$ and $\sum_{j \neq i} \left(\frac{-A_{ij}}{A_{ii}}\right) = 1$. This shows that $\sum_{j \neq i} \left(\frac{-A_{ij}}{A_{ii}}\right) \mathbf{u}_j$ is a *convex* combination of $\mathbf{u}_j$ for $j \in \mathcal{N}(i)$, and yields:

$$\mathbf{u}_i \leq \sum_{j \in \mathcal{N}(i)} \left(\frac{-A_{ij}}{A_{ii}} \right) \mathbf{u}_j \leq \sum_{j \in \mathcal{N}(i)} \left(\frac{-A_{ij}}{A_{ii}} \right) \left(\max_{\tilde{j} \in \mathcal{N}(i)} \mathbf{u}_{\tilde{j}} \right) \leq \max_{\{\tilde{j} \in \mathcal{N}(i)\}} \mathbf{u}_{\tilde{j}}.$$

To show part 2, apply the preceding result for $\mathbf{u}$ and $-\mathbf{u}$ (since $A(-\mathbf{u}) = \mathbf{0}$). The desired result will follow since $\mathbf{u}_{\tilde{j}} \leq |\mathbf{u}_{\tilde{j}}|$. To show part 3, suppose that

$1 \leq i_* \leq n$ denotes an index such that $|\mathbf{u}_{i_*}| = \|\mathbf{u}\|_\infty$. Since $(A\,\mathbf{u})_{i_*} = 0$, apply weak diagonal dominance of A to obtain:

$$\begin{aligned}
\mathbf{u}_{i_*} &= \sum_{j \in \mathcal{N}(i_*)} \left(\frac{-A_{i_* j}}{A_{i_* i_*}} \right) \mathbf{u}_j \leq \sum_{j \in \mathcal{N}(i_*)} \left(\frac{-A_{i_* j}}{A_{i_* i_*}} \right) |\mathbf{u}_j| \\
&\leq \sum_{j \in \mathcal{N}(i_*)} \left(\frac{-A_{i_* j}}{A_{i_* i_*}} \right) \mathbf{u}_{i_*} = \mathbf{u}_{i_*}.
\end{aligned}$$

Thus, we must have $\mathbf{u}_j = \mathbf{u}_{i_*}$ for all nodes $j \in \mathcal{N}(i_*)$, and the irreducibility of A_{II} will yield $\mathbf{u}_I = \alpha\,(1, \ldots, 1)^T$ for $\alpha = \mathbf{u}_{i_*}$. Thus, if the maximum occurs in the interior, $\mathbf{u}_I$ must be constant, otherwise $\|\mathbf{u}_I\| \leq \|\mathbf{u}\|_\infty$. □

Remark 15.17. When matrix A is *strictly diagonally dominant* (such as when $c(x) \geq c_0 > 0$), the following changes can be made to the preceding results:

$$\begin{cases}
\text{if } \mathbf{u}_i \neq 0 \text{ and } (A\mathbf{u})_i \leq 0, \text{ then} & \mathbf{u}_i < \max_{\{j \in \mathcal{N}(i)\}} \mathbf{u}_j \\
\text{if } \mathbf{u}_i \neq 0 \text{ and } (A\mathbf{u})_i = 0, \text{ then} & |\mathbf{u}_i| < \max_{\{j \in \mathcal{N}(i)\}} |\mathbf{u}_j| \\
\text{if } \mathbf{u} \neq \mathbf{0} \text{ and } A\mathbf{u} = 0, \text{ then} & \|\mathbf{u}_I\|_\infty < \|\mathbf{u}_B\|_\infty, \text{ if } A_{IB} \neq 0,
\end{cases}$$

since $\sum_{j \neq i} \left(\frac{-A_{ij}}{A_{ii}} \right) < 1$. We next describe a discrete comparison principle.

Lemma 15.18. *Suppose the following conditions hold.*

1. *Let $A_{II}^{(l)}$ be an M-matrix and let $A_{IB}^{(l)} \leq 0$ entry wise.*
2. *Let $\mathbf{u}_I^{(l)}$, $\mathbf{u}_B^{(l)}$ and $\mathbf{v}_I^{(l)}$, $\mathbf{v}_B^{(l)}$ satisfy:*

$$\begin{cases} A_{II}^{(l)} \mathbf{u}_I^{(l)} + A_{IB}^{(l)} \mathbf{u}_B^{(l)} = \mathbf{f}_I^{(l)} \\ \mathbf{u}_B^{(l)} = \mathbf{g}_B^{(l)} \end{cases} \quad \text{and} \quad \begin{cases} A_{II}^{(l)} \mathbf{v}_I^{(l)} + A_{IB}^{(l)} \mathbf{v}_B^{(l)} = \tilde{\mathbf{f}}_I^{(l)} \\ \mathbf{v}_B^{(l)} = \tilde{\mathbf{g}}_B^{(l)} \end{cases} \tag{15.23}$$

If $\mathbf{f}_I^{(l)} \geq \tilde{\mathbf{f}}_I^{(l)}$ and $\mathbf{g}_B^{(l)} \geq \tilde{\mathbf{g}}_B^{(l)}$ component wise, then $\mathbf{u}_I^{(l)} \geq \mathbf{v}_I^{(l)}$.

Proof. Substitute $\mathbf{u}_B^{(l)} = \mathbf{g}_B^{(l)}$ and $\mathbf{v}_B^{(l)} = \tilde{\mathbf{g}}_B^{(l)}$ in the above equations, and subtract the resulting expressions to obtain:

$$A_{II}^{(l)} \left(\mathbf{u}_I^{(l)} - \mathbf{v}_I^{(l)} \right) = \left(\mathbf{f}_I^{(l)} - \tilde{\mathbf{f}}_I^{(l)} \right) - A_{IB}^{(l)} \left(\mathbf{g}_B^{(l)} - \tilde{\mathbf{g}}_B^{(l)} \right).$$

Since $\left(\mathbf{f}_I^{(l)} - \tilde{\mathbf{f}}_I^{(l)} \right) \geq \mathbf{0}$, $\left(\mathbf{g}_B^{(l)} - \tilde{\mathbf{g}}_B^{(l)} \right) \geq \mathbf{0}$, and $\left(-A_{IB}^{(l)} \right) \geq 0$ component wise, it will hold that $A_{II}^{(l)} \left(\mathbf{u}_I^{(l)} - \mathbf{v}_I^{(l)} \right) \geq \mathbf{0}$. Since $A_{II}^{(l)}$ is an M-matrix with $\left(A_{II}^{(l)} \right)^{-1} \geq 0$, multiplying both sides of $A_{II}^{(l)} \left(\mathbf{u}_I^{(l)} - \mathbf{v}_I^{(l)} \right) \geq \mathbf{0}$ by $\left(A_{II}^{(l)} \right)^{-1}$ will preserve the inequality, yielding:

$$\left(\mathbf{u}_I^{(l)} - \mathbf{v}_I^{(l)} \right) \geq \mathbf{0}.$$

The desired result follows. □

Definition 15.19. *Let $\mathbf{w}_I^{(l)}$ denote the solution to the homogenous system:*

$$A_{II}^{(l)}\mathbf{w}_I^{(l)} + A_{IB}^{(l)}\mathbf{w}_B^{(l)} = \mathbf{0}, \quad \text{in } \Omega_l^* \tag{15.24}$$

where the boundary data $\mathbf{w}_B^{(l)}$ satisfies:

$$\left(\mathbf{w}_B^{(l)}\right)_i = \begin{cases} 1, & \text{if } y_i^{(l)} \in B^{(l)} \\ 0, & \text{if } y_i^{(l)} \in B_{[l]}. \end{cases} \tag{15.25}$$

We define a discrete contraction factor $\rho_{h,l}$ for A from Ω_l^* to Ω_l as:

$$\rho_{h,l} \equiv \max_{\{j\in\mathcal{J}_l\}} \left|\left(\mathbf{w}_I^{(l)}\right)_j\right| = \max_{x_j\in\overline{\Omega}_l} \left|\left(\mathbf{w}_I^{(l)}\right)_j\right|.$$

The following is a consequence of the discrete comparison principle.

Lemma 15.20. *Let $A_{II}^{(l)}$ be an M-matrix and let $A_{IB}^{(l)} \le 0$ component wise.*

1. *Let $\mathbf{w}_I^{(l)}$ and $\mathbf{w}_B^{(l)}$ satisfy (15.24) and (15.25).*
2. *Let $\mathbf{v}_I^{(l)}$ and $\mathbf{v}_B^{(l)}$ satisfy:*

$$\begin{cases} A_{II}^{(l)}\mathbf{v}_I^{(l)} + A_{IB}^{(l)}\mathbf{v}_B^{(l)} = \mathbf{0} \\ \mathbf{v}_B^{(l)} = \mathbf{g}_B^{(l)}. \end{cases} \tag{15.26}$$

Then, it will hold that:

$$\|\mathbf{v}_I^{(l)}\|_{\infty,\overline{\Omega}_l} = \max_{\{x_i\in\overline{\Omega}_l\}} \left|\left(\mathbf{v}_I^{(l)}\right)_i\right| \le \rho_{h,l}\,\|\mathbf{g}_B^{(l)}\|_\infty.$$

Proof. Follows by an application of the discrete comparison theorem, using $\mathbf{u}_I^{(l)} = (\|\mathbf{g}_B\|_\infty)\,\mathbf{w}_I^{(l)}$ and $\mathbf{u}_B^{(l)} = (\|\mathbf{g}_B\|_\infty)\,\mathbf{w}_B^{(l)}$. □

The next result shows that, for a consistent discretization of (15.1), the discrete contraction factor $\rho_{h,l} \le \rho_l$ for sufficiently small h with $h \le h_0$.

Lemma 15.21. *Suppose $L\,u = -\nabla\cdot(a(x)\,\nabla u) + \mathbf{b}(x)\cdot\nabla u + c(x)\,u$ and:*

1. *Let $z(x) = e^{-\gamma\,\mathrm{d}(x)}$ denote a comparison function satisfying:*

$$L\,z(x) \ge \frac{1}{2}c_0 z(x), \text{ in } \Omega_l^*, \quad z(x) = 1, \text{ on } B^{(l)}, \quad z(x) \ge 0, \text{ on } B_{[l]}.$$

2. *Let $A_{II}^{(l)}$ be a strictly diagonally dominant M-matrix, and let $A_{IB}^{(l)} \le 0$.*
3. *Let $|\mathcal{E}_i^{(l)}| \le C(z)h^r$ be the truncation error of $z(x)$ at the grid point $y_i^{(l)}$:*

$$(L\,z)\,(y_i^{(l)}) = \left(A_{II}^{(l)}\mathbf{z}_I^{(l)} + A_{IB}^{(l)}\mathbf{z}_B^{(l)}\right)_i + \mathcal{E}_i^{(l)},$$

where $\mathbf{z}_I^{(l)} = \boldsymbol{\pi}_I^{(l)}z$, $\mathbf{z}_B^{(l)} = \boldsymbol{\pi}_B^{(l)}z$ are interpolants of $z(\cdot)$ onto the grids.

Then, the discrete contraction factor $\rho_{h,l}$ will satisfy:

$$\rho_{h,l} \le \rho_l = e^{-\gamma\, \mathrm{d}_l},$$

for sufficiently small $h \le h_0$, where $\mathrm{d}_l = \mathrm{dist}(B^{(l)}, \partial\Omega_l \cap \Omega)$.

Proof. Given $z(x)$, let the local discretization error $\mathcal{E}_i^{(l)}$ at node $y_i^{(l)}$ satisfy:

$$\left|\mathcal{E}_i^{(l)}\right| \le C(z)\, h^r,$$

for some $r \ge 1$. If we choose h small so that $\left|\mathcal{E}_i^{(l)}\right| \le C(z)\, h^r \le \frac{c_0}{4}\, z(y_i^{(l)})$, holds for each $y_i^{(l)} \in \Omega_l^*$, then we will obtain that:

$$\left(A_{II}^{(l)}\mathbf{z}_I^{(l)} + A_{IB}^{(l)}\mathbf{z}_B^{(l)}\right)_i \ge 0, \quad \forall y_i^{(l)} \in \Omega_l^* \text{ or } i \in \mathcal{J}_l^*.$$

Apply the discrete comparison Lemma 15.18 using $\mathbf{u}_I^{(l)} = \mathbf{z}_I^{(l)}$, $\mathbf{u}_B^{(l)} = \mathbf{z}_B^{(l)}$ and $\mathbf{v}_I^{(l)} = \mathbf{w}_I^{(l)}$, $\mathbf{v}_B^{(l)} = \mathbf{w}_B^{(l)}$, for $\mathbf{w}_I^{(l)}$, $\mathbf{w}_B^{(l)}$ satisfying (15.24) and (15.25), to obtain $\mathbf{u}_I^{(l)} = \mathbf{z}_I^{(l)} \ge \mathbf{v}_I^{(l)} = \mathbf{w}_I^{(l)}$:

$$\rho_{h,l} = \max_{\{y_i^{(l)} \in \overline{\Omega}_l\}} \left(\mathbf{w}_I^{(l)}\right)_i \le z(y_i^{(l)}) \le \max_{\{x \in \overline{\Omega}_l\}} z(x) \le e^{-\gamma\, \mathrm{d}_l},$$

since $z(x) = e^{-\gamma\, \mathrm{d}(x)}$. □

Remark 15.22. In practical computations, the discrete contraction factor $\rho_{h,l}$ can be computed *numerically*, by solving (15.24). This requires one solve per subdomain. This contraction factor will *decrease* with *increasing overlap*. For time stepped problems and advection dominated elliptic equations, $\rho_{h,l}$ may be estimated in terms of $e^{-\frac{\gamma}{\sqrt{\tau}}\mathrm{d}_l}$ and $e^{-\frac{\gamma}{\epsilon}\mathrm{d}_l}$, under suitable assumptions, for sufficiently small h, see [GA10, MA33, GA12, CA17, MA35].

Remark 15.23. Explicit estimates of contraction factors can be constructed in special cases in which matrix A_{II} is Toeplitz [GA12, MA35]. For instance, if matrix A is tridiagonal and Toeplitz with $A = \mathrm{tridiag}(\theta, \alpha, \mu)$, then a homogeneous system of the form (15.24) can be solved analytically using the ansatz $\mathbf{w}_i = c_1\kappa_1^i + c_2\kappa_2^i$, where κ_1 and κ_2 are roots of $\theta + \alpha\, r + \mu\, r^2 = 0$. Boundary conditions can be enforced and the contraction factors estimated.

15.2 Well Posedness of the Schwarz Hybrid Formulation

In this section, we *heuristically* study the well posedness of the Schwarz hybrid formulation (11.22) of elliptic equation (15.1), see [CA17]. Many of the results in later sections are based on the contraction mapping described here. We shall construct an overlapping decomposition $\{\Omega_l^*\}_{l=1}^p$ of Ω based on a non-overlapping decomposition $\{\Omega_l\}_{l=1}^p$ of Ω, with subdomains of size h_0. For $0 < \beta < 1$, define an overlapping decomposition with overlap βh_0 as:

$$\Omega_l^* \equiv \Omega_l^{\beta h_0} \equiv \{x \in \Omega \, : \, \mathrm{dist}(x, \Omega_l) < \beta h_0\}, \quad \text{for} \quad 1 \le l \le p. \tag{15.27}$$

Choose $0 < \epsilon < \beta$ and let $\{\chi_l(x)\}_{l=1}^p$ denote a *smooth* partition of unity subordinate to $\{\Omega_l^{\epsilon h_0}\}_{l=1}^p$. Since $\Omega_l^{\epsilon h_0} \subset \Omega_l^*$ for $1 \le l \le p$, the partition of unity will also be subordinate to $\{\Omega_l^*\}_{l=1}^p$. Further, it will satisfy:

$$\sum_{j \ne l} \chi_j(x) = 1, \quad \text{on} \quad B^{(l)}, \;\; \text{for } 1 \le l \le p,$$

since each $\chi_l(x) = 0$ on $B^{(l)}$, where $B^{(l)} \equiv \partial\Omega_l^* \cap \Omega$ and $B_{[l]} \equiv \partial\Omega_l^* \cap \partial\Omega$.

The Schwarz hybrid formulation of (15.1) seeks $w_1(x), \ldots, w_p(x)$ solving:

$$\begin{cases} L\, w_l(x) = f(x), & \text{in } \Omega_l^* \\ \quad w_l(x) = \sum_{j \ne l} \chi_j(x)\, w_j(x), & \text{on } B^{(l)} \quad \text{for} \quad 1 \le l \le p. \\ \quad w_l(x) = 0, & \text{on } B_{[l]} \end{cases} \tag{15.28}$$

We shall *heuristically* indicate why (15.28) will be well posed in the maximum norm, provided $\epsilon \ll \beta$ and $\beta > 0$ is sufficiently large, and provided the subdomains and $f(x)$ are sufficiently smooth, to ensure the existence and regularity of the local solutions.

We shall employ a *formal* metric space $\mathcal{H}$ defined as follows:

$$\mathcal{H} = \left\{ (v_1, \ldots, v_p) \, : \, v_l \in C^2\left(\overline{\Omega}_l^*\right), \, L\, v_l = f, \text{ in } \Omega_l^*, v_l = 0 \, \text{on } B_{[l]}, \; 1 \le l \le p \right\}. \tag{15.29}$$

Given $v = (v_1, \ldots, v_p) \in \mathcal{H}$, we define $\|v\|_\infty = \max_{1 \le l \le p} \|v_l\|_{\infty, \Omega_l^*}$. We define the metric $d(u, v) = \|u - v\|_\infty$ for $u = (u_1, \ldots, u_p)\,,\, v = (v_1, \ldots, v_p) \in \mathcal{H}$.

Definition 15.24. *Given $v = (v_1, \ldots, v_p) \in \mathcal{H}$, define $T\,v = \tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_p)$:*

$$\begin{cases} L\, \tilde{v}_l(x) = f(x), & \text{in } \Omega_l^* \\ \quad \tilde{v}_l(x) = \sum_{j \ne l} \chi_j(x)\, v_j(x), & \text{on } B^{(l)} \quad \text{for} \quad 1 \le l \le p. \\ \quad \tilde{v}_l(x) = 0, & \text{on } B_{[l]} \end{cases} \tag{15.30}$$

The map $T : \mathcal{H} \rightarrow \mathcal{H}$ will be pivotal to the analysis of the Schwarz hybrid formulation, since by construction, if a solution $w \in \mathcal{H}$ to (15.28) exists, it will be a *fixed point* of T with $T\,w = w$. The map T will be a *contraction*.

Definition 15.25. *We define a contraction factor* $\rho_{l,\epsilon,\beta}$ *as:*

$$\rho_{l,\epsilon,\beta} \equiv \|v_l\|_{\infty,\Omega_l^{\epsilon h_0}}, \tag{15.31}$$

where $Lv_l = 0$ *in* $\Omega_l^{\beta h_0}$ *with* $v_l = 0$ *on* $\partial\Omega_l^{\beta h_0} \cap \partial\Omega$ *and* $v_l = 1$ *on* $\partial\Omega_l^{\beta h_0} \cap \Omega$.

Lemma 15.26. *Suppose* $c(x) \geq c_0 > 0$ *and let the overlap* $\beta\, h_0$ *between the subdomains* $\Omega_l^{\beta h_0}$ *be large. Let the partition of unity be based on* $\Omega_l^{\epsilon h_0}$ *for* $\epsilon \ll \beta$ *and let* $\rho_{l,\epsilon,\beta}$ *denote the contraction factor into* $\Omega_l^{\epsilon h_0}$, *associated with* L*-harmonic solutions on* $\Omega_l^{\beta h_0}$. *Then* $T : \mathcal{H} \to \mathcal{H}$ *will be a contraction:*

$$d(Tu, Tv) = \|Tu - Tv\|_\infty \leq \left(\max_{1\leq l\leq p} \rho_{l,\epsilon,\beta}\right) \|u - v\|_\infty = \left(\max_{1\leq l\leq p} \rho_{l,\epsilon,\beta}\right) d(u, v).$$

Proof. We outline the proof, assuming sufficiently smooth subdomains and forcing terms. Given $u,\, v \in \mathcal{H}$, let $\tilde{u} = T\,u$ and $\tilde{v} = T\,v$. Then, by definition of T, the components of $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_p)$ and $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_p)$ will solve:

$$\begin{cases} L\,\tilde{u}_l = f, & \text{in } \Omega_l^* \\ \quad \tilde{u}_l = \sum_{j\neq l} \chi_j\, u_j, & \text{on } B^{(l)} \\ \quad \tilde{v}_l = 0, & \text{on } B_{[l]} \end{cases} \text{ and } \begin{cases} L\,\tilde{v}_l = f, & \text{in } \Omega_l^* \\ \quad \tilde{v}_l = \sum_{j\neq l} \chi_j\, v_j, & \text{on } B^{(l)} \\ \quad \tilde{v}_l = 0, & \text{on } B_{[l]}. \end{cases}$$

Subtracting the two will yield:

$$\begin{cases} L\ (\tilde{u}_l - \tilde{v}_l) = 0, & \text{in } \Omega_l^* \\ \quad (\tilde{u}_l - \tilde{v}_l) = \sum_{j\neq l} \chi_j\ (u_j - v_j), & \text{on } B^{(l)} \\ \quad (\tilde{u}_l - \tilde{v}_l) = 0, & \text{on } B_{[l]}. \end{cases}$$

An application of the maximum principle for homogenous solutions yields:

$$\|\tilde{u}_l - \tilde{v}_l\|_{\infty,\Omega_l^*} \leq \|\sum_{j\neq l} \chi_j\ (u_j - v_j)\,\|_{\infty,B^{(l)}}. \tag{15.32}$$

Since $\sum_{j\neq l} \chi_j(x)\ (u_j(x) - v_j(x))$ is a *convex* combination of $(u_j(x) - v_j(x))$ for each $x \in B^{(l)}$, and since $\chi_j(\cdot)$ has support on $\Omega_j^{\epsilon h_0}$, we obtain:

$$\begin{aligned} \|\textstyle\sum_{j\neq l} \chi_j\ (u_j - v_j)\,\|_{\infty,B^{(l)}} &\leq \max_{j\neq l} \|u_j - v_j\|_{\infty,B^{(l)}\cap\Omega_j^{\epsilon h_0}} \\ &\leq \max_{j\neq l}\ \rho_{j,\epsilon,\beta}\, \|u_j - v_j\|_{\infty,B^{(j)}} \\ &\leq \max_{j\neq l}\ \rho_{j,\epsilon,\beta}\, \|u - v\|_\infty. \end{aligned} \tag{15.33}$$

The second and third inequalities above follow by Lemma 15.6 and by the definition of $\rho_{j,\epsilon,\beta}$, since $(u_j - v_j)$ is also L-harmonic in Ω_j^* for $u,\, v \in \mathcal{H}$:

$$\begin{cases} L\ (u_j - v_j) = 0, & \text{in } \Omega_j^* \\ \quad (u_j - v_j) = (u_j - v_j)\,, & \text{on } B^{(j)} \\ \quad (u_j - v_j) = 0. & \text{on } B_{[j]}. \end{cases}$$

Combining (15.32) and (15.33) yields $\|\tilde{u} - \tilde{v}\|_\infty \leq (\max_j \rho_{j,\epsilon,\beta})\ \|u - v\|_\infty$, which shows that T is a *contraction*, provided $(\max_j \rho_{j,\epsilon,\beta}) < 1$. □

Since $T : \mathcal{H} \to \mathcal{H}$ is a contraction, a unique fixed point $w = Tw \in \mathcal{H}$ will exist, which will *formally* solve (15.28). It can be obtained by *Picard iteration.* Each computation $v^{k+1} = T\, v^k$ will correspond to a parallel Schwarz iteration. We next *heuristically* outline the well posedness of system (15.28).

Lemma 15.27. *Let* $(w_1, \ldots, w_p) \in \mathcal{H}$ *solve (15.28) and where* $\Omega_l^* = \Omega_l^{\beta h_0}$ *for* $1 \leq l \leq p$*. Let* $\{\chi_l(.)\}_{l=1}^p$ *be a smooth partition of unity subordinate to* $\{\Omega_l^{\epsilon h_0}\}_{l=1}^p$*. Assume also that* $\rho = (\max_l\, \rho_{l,\epsilon,\beta}) < 1$*. Then:*

$$\|w\|_\infty \;\leq\; \left(\frac{C}{1-\rho}\right) \|f\|_\infty.$$

Proof. We outline a heuristic proof. Choose $v^{(0)} = (v_1^{(0)}, \ldots, v_p^{(0)}) \in \mathcal{H}$ as:

$$\begin{cases} Lv_l^{(0)} = f, \text{ in } \Omega_l^* \\ \;\;v_l^{(0)} = 0, \text{ on } B^{(l)} \quad \text{for} \quad 1 \leq l \leq p. \\ \;\;v_l^{(0)} = 0, \text{ on } B_{[l]}. \end{cases}$$

For sufficiently smooth $f(.)$ and smooth subdomains, maximum norm estimates for elliptic equations [GI] yield:

$$\|v_l^{(0)}\|_{\infty,\Omega_l^*} \;\leq\; \|f\|_{\infty,\Omega_l^*}.$$

By construction $v^{(0)} \in \mathcal{H}$. Apply Picard iteration to define $v^{(k)} \equiv T^k v^{(0)} \in \mathcal{H}$ which converges to $w = Tw$. Estimates for Picard iteration [AR3] yields:

$$d(v^{(0)}, w) \;\leq\; \left(\frac{\rho}{1-\rho}\right) d(v^{(0)}, Tv^{(0)}).$$

Since $d(u, v) = \|u - v\|_\infty$, this yields:

$$\begin{aligned} \|w\|_\infty &\leq \|v^{(0)}\|_\infty + \|w - v^{(0)}\|_\infty \\ &= \|v^{(0)}\|_\infty + \left(\tfrac{\rho}{1-\rho}\right) d(v^{(0)}, Tv^{(0)}). \end{aligned}$$

The term $d(v^{(0)}, Tv^{(0)}) = \|v^{(0)} - Tv^{(0)}\|_\infty$ may be estimated using maximum norm estimates for L-harmonic functions, since the components $v_l^{(0)} - (Tv^{(0)})_l$ are L-harmonic. We omit further details. □

15.3 Convergence of Schwarz Iterative Algorithms

Continuous Case. We shall now describe the maximum norm convergence of the continuous versions of the sequential and parallel Schwarz algorithms (without coarse spaces). Given overlapping subdomains $\Omega_1^*, \ldots, \Omega_p^*$, the continuous version of the multiplicative Schwarz algorithm updates the solution on each subdomain Ω_l^* in sequence as follows.

Algorithm 15.3.1 *(Multiplicative Schwarz Algorithm)*
Let $u^{(0)}$ be a starting iterate

```
1. For $k = 0, 1, \ldots,$ until convergence do:
2.     For $l = 1, \ldots, p$ compute
           $$\begin{cases} L\, w^{(k,l)} = f, & \text{in } \Omega_l^* \\ w^{(k,l)} = u^{(k+\frac{l-1}{p})}, & \text{on } B^{(l)} \\ w^{(k,l)} = 0, & \text{on } B_{[l]} \end{cases}$$
       Update:
           $$u^{(k+\frac{l}{p})} \equiv \begin{cases} w^{(k,l)}, & \text{in } \Omega_l^* \\ u^{(k+\frac{l-1}{p})}, & \text{in } \Omega \setminus \Omega_l^* \end{cases}$$
3.     Endfor
4. Endfor
```

If the starting iterate satisfies $u^{(0)} \geq u$ and $L\, u^{(0)} - f \geq 0$, subsequent iterates will converge *monotonically* downwards to the true solution u.

Lemma 15.28. *Suppose $u^{(0)}$ satisfies $u^{(0)} \geq u$ and $L\, u^{(0)} - f \geq 0$. On Ω_l^*, let ρ_l denote the contraction factor into Ω_l for a L-harmonic solution. Then:*

1. *The iterates $u^{(k+\frac{l}{p})}$ will satisfy $u \leq u^{(k+\frac{l}{p})} \leq u^{(k+\frac{l-1}{p})} \leq \cdots$*
2. *Each iterate will satisfy $L\, u^{(k+\frac{l}{p})} - f \geq 0$.*
3. *The maximum norm of the error will satisfy:*

$$\|u - u^{(k+1)}\|_{\infty,\Omega} \leq \left(\max_{\{1 \leq l \leq p\}} \rho_l \right) \|u - u^{(k)}\|_{\infty,\Omega}.$$

Proof. The properties $u \leq u^{(k+\frac{l}{p})} \leq u^{(k+\frac{l-1}{p})} \leq \cdots$ and $L\, u^{(k+\frac{l}{p})} - f \geq 0$ will be verified only in the discrete case, see [LI6, LI7] and (15.31). To estimate the error reduction, note that by construction $u^{(k+\frac{l}{p})}$ satisfies $L\, u^{(k+\frac{l}{p})} = f$ in Ω_l^*, with $u^{(k+\frac{l}{p})} = u^{(k+\frac{l-1}{p})}$ on $B^{(l)}$. Since u satisfies $L\, u = f$, subtracting the two equations yields:

$$\begin{cases} L\,(u^{(k+\frac{l}{p})} - u) = 0, & \text{in } \Omega_l^* \\ (u^{(k+\frac{l}{p})} - u) = (u^{(k+\frac{l-1}{p})} - u), & \text{on } B^{(l)} \\ (u^{(k+\frac{l-1}{p})} - u) = 0, & \text{on } B_{[l]}. \end{cases}$$

Applying Lemma 15.3 with $w_1(x) = u^{(k+\frac{l-1}{p})}(x) - u(x)$ and $w_2(x) = 0$ in Ω_l^* and $g_1(x) = u^{(k+\frac{l-1}{p})}(x) - u(x) \geq 0$ and $g_2(x) = 0$ on $\partial\Omega_l^*$ yields that $w_1(x) \geq 0$. Thus $u^{(k+\frac{l-1}{p})}(x) \geq u(x)$ in Ω_l^* and outside too. Lemma 15.6 with $v(x) = u^{(k+\frac{l}{p})}(x) - u(x)$ in Ω_l^* and $g(x) = u^{(k+\frac{l-1}{p})}(x) - u(x)$ on $B^{(l)}$ yields:

$$
\begin{aligned}
0 \le \left(u^{(k+\frac{l}{p})}(x) - u(x)\right) &\le \left(\|u - u^{(k+\frac{l-1}{p})}\|_{\infty,B^{(l)}}\right) w_{l,*}(x) \text{ in } \Omega_l \\
&\le \left(\|u - u^{(k+\frac{l-1}{p})}\|_{\infty,\Omega}\right) w_{l,*}(x) \quad \text{in } \Omega_l \\
&\le \left(\|u - u^{(k)}\|_{\infty,\Omega}\right) w_{l,*}(x) \qquad \text{in } \Omega_l
\end{aligned}
$$

where $w_{l,*}(x)$ is the comparison function (15.9) on Ω_l^*. Since the iterates $u^{(k+\frac{l}{p})}$ decrease *monotonically* for each x, and since $w_{l,*}(x) \le \rho_l$ on Ω_l:

$$
\begin{aligned}
0 \le \left(u^{(k+1)}(x) - u(x)\right) &\le \left(u^{(k+\frac{l}{p})}(x) - u(x)\right) \text{ in } \Omega_l \\
&\le \rho_l \, \|u - u^{(k)}\|_{\infty,\Omega} \quad \text{in } \Omega_l.
\end{aligned}
$$

Combining the bounds over Ω yields:

$$
\|u^{(k+1)} - u\|_{\infty,\Omega} \le \left(\max_{1 \le l \le p} \rho_l\right) \|u^{(k)} - u\|_{\infty,\Omega}.
$$

The desired result follows. □

For *two subdomain* decompositions, a sharper bound can be obtained for the convergence of the *sequential Schwarz* algorithm [LI7]. Indeed, suppose Ω_1^* and Ω_2^* are obtained by extending two non-overlapping subdomains Ω_1 and Ω_2. Then $B^{(1)} \subset \overline{\Omega}_2$ and $B^{(2)} \subset \overline{\Omega}_1$ and it will be sufficient to estimate the maximum norm of the error on these segments. Applying Lemma 15.6 yields:

$$
\begin{aligned}
\|u^{(k+1)} - u\|_{\infty,B^{(1)}} &\le \|u^{(k+1)} - u\|_{\infty,\overline{\Omega}_2} \le \rho_2 \, \|u^{(k+\frac{1}{2})} - u\|_{\infty,B^{(2)}} \\
\|u^{(k+\frac{1}{2})} - u\|_{\infty,B^{(2)}} &\le \|u^{(k+\frac{1}{2})} - u\|_{\infty,\overline{\Omega}_1} \le \rho_1 \, \|u^{(k)} - u\|_{\infty,B^{(1)}}.
\end{aligned}
$$

Combining the two bounds yields:

$$
\|u^{(k+1)} - u\|_{\infty,B^{(1)}} \le (\rho_2 \, \rho_1) \, \|u^{(k)} - u\|_{\infty,B^{(1)}}.
$$

Since this is a product of two contraction factors (in contrast to the maximum of the same contraction factors), this error bound is sharper.

We next analyze the continuous version of a parallel partition of unity Schwarz algorithm where each $\chi_j(x)$ has support in a neighborhood of $\overline{\Omega}_j$.

Algorithm 15.3.2 *(Parallel Schwarz Algorithm)*
Let $u^{(0)}$ be a starting iterate

```
1. For $k = 0, 1, \ldots,$ until convergence do:
2.     For $l = 1, \ldots, p$ in parallel compute:
           $\begin{cases} L\, w^{(k,l)} = f, & \text{in } \Omega_l^* \\ w^{(k,l)} = u^{(k)}, & \text{on } B^{(l)} \\ w^{(k,l)} = 0, & \text{on } B_{[l]} \end{cases}$
3.     Endfor
4.     Update $u^{(k+1)}(x) \equiv \sum_{l=1}^{p} \chi_l(x)\, w^{(k,l)}(x)$
5. Endfor
```

The following result describes the convergence of the preceding algorithm.

Lemma 15.29. *Let $\Omega_l^* = \Omega_l^{\beta h_0}$ be overlapping subdomains for $1 \leq l \leq p$, where $\beta < 1$. Let $\{\chi_j(\cdot)\}_{j=1}^p$ be a partition of unity subordinate to $\{\Omega_l^{\epsilon h_0}\}_{l=1}^p$ where $\epsilon \ll \beta$. Let $\rho_{l,\epsilon,\beta}$ be contraction factors defined by (15.31). Then:*

1. *Each local iterate $w^{(k,l)}(x)$ will satisfy:*

$$\|u - w^{(k,l)}\|_{\infty,\Omega_l^{\epsilon h_0}} \leq \rho_{l,\epsilon,\beta}\, \|u - u^{(k)}\|_{\infty,\Omega_l^*}.$$

2. *The maximum norm of the error will satisfy:*

$$\|u - u^{(k+1)}\|_{\infty,\Omega} \leq \left(\max_{1 \leq l \leq p} \rho_{l,\epsilon,\beta} \right) \|u - u^{(k)}\|_{\infty,\Omega}.$$

Proof. Each $w^{(k,l)}(x)$ satisfies $L\, w^{(k,l)} = f$ in Ω_l^* with $w^{(k,l)} = u^{(k)}$ on $B^{(l)}$, while the exact solution satisfies $L\, u = f$ in Ω_l^*. Subtracting them yields:

$$\begin{cases} L\, \left(u - w^{(k,l)}\right) = 0, & \text{in } \Omega_l^* \\ \quad \left(u - w^{(k,l)}\right) = \left(u - u^{(k)}\right), & \text{on } B^{(l)} \\ \quad \left(u - w^{(k,l)}\right) = 0, & \text{on } B_{[l]}. \end{cases}$$

An application of Lemma 15.6 using $v = u - w^{(k,l)}$ and $g = u - u^{(k)}$ yields:

$$\|u - w^{(k,l)}\|_{\infty,\Omega_l^{\epsilon h_0}} \leq \rho_{l,\epsilon,\beta}\, \|u - u^{(k)}\|_{\infty,\Omega}.$$

Provided each $\chi_l(x)$ has support within $\Omega_l^{\epsilon h_0}$. we may combine the estimates:

$$\|u - u^{(k+1)}\|_{\infty,\Omega} \leq \left(\max_{1 \leq l \leq p} \rho_l \right) \|u - u^{(k)}\|_{\infty,\Omega}.$$

This yields the desired result. □

Discrete Case. We now consider the convergence of discrete versions of the multiplicative and parallel Schwarz algorithms to solve system (15.2) obtained by the discretization of (15.1). For simplicity, we rewrite the discrete system $A_{II}\mathbf{u}_I = (\mathbf{f}_I - A_{IB}\mathbf{g}_B)$ as $A\,\mathbf{u} = \mathbf{f}$, where $\mathbf{g}_B$ is the discretization of the boundary data $g(.)$ on $\partial\Omega$. We shall employ the notation $R_B^{(l)}\mathbf{v}$ and $R_I^{(l)}\mathbf{v}$ to denote the restriction of a nodal vector $\mathbf{v} = \mathbf{v}_I$ to nodes on $B^{(l)}$ and in Ω_l^*. For each node $x_i \in \overline{\Omega}_l^*$, we shall let $\tilde{i}$ denote its local index, such that $x_i = y_{\tilde{i}}^{(l)}$. Below, the discrete version of the multiplicative Schwarz algorithm is summarized, where n_l denotes the number of interior nodes in Ω_l^*.

Algorithm 15.3.3 *(Multiplicative Schwarz Algorithm)*
Let $\mathbf{u}^{(0)}$ *be a starting iterate*

```
1. For $k = 0, 1, \ldots,$ until convergence do:
2.     For $l = 1, \ldots, p$ solve for $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}$:
           $R_I^{(l)} \left( A(\mathbf{u}^{(k+\frac{l-1}{p})} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right) = \mathbf{0}$
       Update:
           $\mathbf{u}^{(k+\frac{l}{p})} = \mathbf{u}^{(k+\frac{l}{p})} + R_I^{(l)^T} \mathbf{d}^{(l)}$
3.     Endfor
4. Endfor
```

The following result concerns the *monotonicity* of Schwarz updates.

Lemma 15.30. *Let A be a strictly diagonally dominant M-matrix.*

1. *Let $\mathbf{v} \in \mathbb{R}^n$ satisfy $A\,\mathbf{v} - \mathbf{f} \geq \mathbf{0}$.*
2. *Let $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}$ be chosen such that $R_I^{(l)} \left(A(\mathbf{v} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right) = \mathbf{0}$.*

Then, it will hold that $\mathbf{d}^{(l)} \leq \mathbf{0}$ and $A \left(\mathbf{v} + R_I^{(l)^T} \mathbf{d}^{(l)} \right) - \mathbf{f} \geq \mathbf{0}$.

Proof. By choice of $\mathbf{d}^{(l)}$, it will hold that $\left(A(\mathbf{v} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right)_i = 0$ for each $x_i \in \Omega_l^*$. Since $(A\mathbf{v} - \mathbf{f})_i \geq 0$, subtracting the two equations will yield $\left(A\, R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i \leq 0$. Applying the discrete maximum principle (15.16) will yield $\left(R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i \leq 0$ for each $x_i \in \Omega_l^*$. Since $R_I^{(l)^T}$ is an extension matrix, it will thus hold that $\mathbf{d}^{(l)} \leq \mathbf{0}$.

We shall next show that $A \left(\mathbf{v} + R_I^{(l)^T} \mathbf{d}^{(l)} \right) - \mathbf{f} \geq \mathbf{0}$. By choice of $\mathbf{d}^{(l)}$, for each $x_i \in \Omega_l^*$ it will hold that $\left(A\,(\mathbf{v} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right)_i = 0$. Thus, we only need consider $x_i \notin \Omega_l^*$. Since $(A\mathbf{v} - \mathbf{f})_i \geq 0$, we will obtain:

$$\begin{aligned} \left(A(\mathbf{v} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right)_i &= (A\mathbf{v} - \mathbf{f})_i + \left(A R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i \\ &\geq \left(A R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i . \end{aligned}$$

For $x_i \notin \Omega_l^*$, it will hold that $\left(R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i = 0$. Using this, and noting that $A_{ij} \leq 0$ for $j \neq i$ and that $\left(R_I^{(l)^T} \mathbf{d}^{(l)} \right)_j \leq 0$ for $x_j \in \Omega_l^*$, we obtain that $\left(A R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i = \sum_j A_{ij} (R_I^{(l)^T} \mathbf{d}^{(l)})_j \geq 0$. This yields the desired result. □

We now consider the multiplicative Schwarz algorithm.

Lemma 15.31. *Let A be a strictly diagonally dominant M-matrix. Suppose that $\mathbf{u}^{(0)} \in \mathbb{R}^n$ satisfies $A\,\mathbf{u}^{(0)} - \mathbf{f} \geq \mathbf{0}$, Then, the following results will hold.*

1. *The multiplicative Schwarz iterates $\mathbf{u}^{(k+\frac{l}{p})}$ will satisfy:*
$$\mathbf{u} \leq \mathbf{u}^{(k+\frac{l}{p})} \leq \mathbf{u}^{(k+\frac{l-1}{p})} \leq \cdots$$
2. *Each iterate $\mathbf{u}^{(k+\frac{1}{p})}$ will satisfy $A\,\mathbf{u}^{(k+\frac{1}{p})} - \mathbf{f} \geq \mathbf{0}$.*
3. *The maximum norm error will satisfy:*
$$\|\mathbf{u} - \mathbf{u}^{(k+1)}\|_\infty \leq \left(\max_{1 \leq l \leq p} \rho_{h,l} \right) \|\mathbf{u} - \mathbf{u}^{(k)}\|_\infty.$$

Proof. We shall prove by induction on k and l. Suppose $A\,\mathbf{u}^{(k+\frac{l-1}{p})} \geq \mathbf{f}$. Then the discrete comparison theorem will yield $A\left(\mathbf{u}^{(k+\frac{l-1}{p})} - \mathbf{u}\right) \geq \mathbf{0}$, yielding that $\mathbf{u}^{(k+\frac{l-1}{p})} - \mathbf{u} \geq \mathbf{0}$ (since $A^{-1} \geq 0$). Next, since iterate $\mathbf{u}^{(k+\frac{l}{p})}$ is constructed so that $R_I^{(l)} \left(A\,(\mathbf{u}^{(k+\frac{l-1}{p})} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right) = \mathbf{0}$, we may apply the preceding lemma to obtain that $\mathbf{d}^{(l)} \leq \mathbf{0}$ and that $\left(A\,\mathbf{u}^{(k+\frac{l}{p})} - \mathbf{f} \right) \geq \mathbf{0}$. Since $\mathbf{d}^{(l)} \leq \mathbf{0}$, it will hold that $\mathbf{u}^{(k+\frac{l}{p})} \leq \mathbf{u}^{(k+\frac{l-1}{p})}$. Since $\left(A\,\mathbf{u}^{(k+\frac{l}{p})} - \mathbf{f} \right) \geq \mathbf{0}$, it will hold that $\mathbf{u}^{(k+\frac{l}{p})} \geq \mathbf{u}$. This proves parts 1 and 2.

To obtain an estimate for the reduction in error in the maximum norm, we shall employ the monotone nature of the Schwarz iterates, and employ the reduction in error on each $\overline{\Omega}_l$ by the discrete contraction factor $\rho_{h,l}$. The *monotone* nature of the iterates yields the following for each i:
$$0 \leq \left(\mathbf{u}^{(k+1)} - \mathbf{u} \right)_i \leq \min_{1 \leq l \leq p} \left(\mathbf{u}_i^{(k+\frac{l-1}{p})} - \mathbf{u}_i \right) \leq \left(\mathbf{u}_i^{(k)} - \mathbf{u}_i \right).$$

Applying the discrete contraction factor $\rho_{h,l}$ in (15.25), we estimate:
$$0 \leq \left(\mathbf{u}_i^{(k)} - \mathbf{u}_i \right) \leq \rho_{h,l} \left(\max_{x_j \in B^{(l)}} (\mathbf{u}^{(k)} - \mathbf{u})_j \right), \qquad \text{for} \quad x_i \in \overline{\Omega}_l.$$

Since $(\mathbf{u}^{(k)} - \mathbf{u}) \geq \mathbf{0}$ componentwise, we obtain:
$$\max_{x_j \in B^{(l)}} (\mathbf{u}^{(k)} - \mathbf{u})_j = \|\mathbf{u}^{(k)} - \mathbf{u}\|_{\infty, B^{(l)}} \leq \|\mathbf{u}^{(k)} - \mathbf{u}\|_\infty.$$

Taking the maximum over each $x_i \in \Omega$ and using the preceding yields:
$$\|\mathbf{u}^{(k+1)} - \mathbf{u}\|_{\infty,\Omega} \leq \left(\max_{1 \leq l \leq p} \rho_{h,l} \right) \|\mathbf{u}^{(k)} - \mathbf{u}\|_\infty$$

since $\overline{\Omega}_1, \ldots, \overline{\Omega}_p$ covers Ω. □

Remark 15.32. In the two subdomain case, sharper estimates can be obtained for the rate of convergence of the multiplicative Schwarz algorithm (regardless of whether $A\mathbf{u}^{(0)} \geq \mathbf{f}$ holds). The global contraction factor will be $\rho_{h,1}\,\rho_{h,2}$ instead of $\max\{\rho_{h,1},\, \rho_{h,2}\}$. The proof will be analogous to the proof in the continuous case described earlier.

We next consider the convergence of the discrete version of the parallel Schwarz algorithm. The estimates in the discrete case will be analogous to the continuous case. We summarize the discrete parallel Schwarz algorithm.

Algorithm 15.3.4 *(Parallel Partition of Unity Schwarz Algorithm)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ be a starting iterate

```
For $k = 0, 1, \ldots,$ until convergence do:
    For $l = 1, \ldots, p$ in parallel determine $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}$:
        $R_I^{(l)} \left( A\,(\mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right) = \mathbf{0}$
    Endfor
    Update $\mathbf{u}_i^{(k+1)} \equiv \sum_{l=1}^{p} \chi_l(x_i) \left( \mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i$ for $1 \leq i \leq n$
Endfor
```

We have the following convergence bound.

Lemma 15.33. *Let A be a strictly diagonally dominant M-matrix. Then, the following will hold for the parallel Schwarz iterates.*

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}\|_\infty \leq \left(\max_{1 \leq l \leq p} \right) \|\mathbf{u}^{(k)} - \mathbf{u}\|_\infty .$$

Proof. Since $R_I^{(l)} \left(A\,(\mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)}) - \mathbf{f} \right) = \mathbf{0}$ and $R_I^{(l)} \left(A\,\mathbf{u} - \mathbf{f} \right) = \mathbf{0}$, subtracting the two and applying the discrete maximum principle, using the discrete contraction factor on Ω_l (from Ω_l^*) will yield the estimate:

$$\left| \left(\mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)} \right)_i - \mathbf{u}_i \right| \leq \rho_{h,l}\, \|\mathbf{u}^{(k)} - \mathbf{u}\|_\infty, \quad \text{for } x_i \in \overline{\Omega}_l.$$

Next, decompose the error using the partition of unity to obtain:

$$\begin{aligned} \left(\mathbf{u} - \mathbf{u}^{(k+1)} \right)_i &= \textstyle\sum_{l=1}^{p} \chi_l(x_i)\, \left(\mathbf{u} - \mathbf{u}^{(k+1)} \right)_i \\ &= \textstyle\sum_{l=1}^{p} \chi_l(x_i)\, \left(\mathbf{u} - (\mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)}) \right)_i . \end{aligned}$$

Estimating $\left(\mathbf{u} - (\mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)}) \right)_i$ using the contraction factor yields:

$$\begin{aligned} \left| \left(\mathbf{u} - \mathbf{u}^{(k+1)} \right)_i \right| &= \textstyle\sum_{l=1}^{p} \chi_l(x_i)\, \left| \left(\mathbf{u} - \mathbf{u}^{(k+1)} \right)_i \right| \\ &= \textstyle\sum_{l=1}^{p} \chi_l(x_i)\, \left| \left(\mathbf{u} - (\mathbf{u}^{(k)} + R_I^{(l)^T} \mathbf{d}^{(l)}) \right)_i \right| \\ &\leq \textstyle\sum_{l=1}^{p} \chi_l(x_i)\, \rho_{h,l}\, \|\mathbf{u} - \mathbf{u}^{(k)}\|_\infty \\ &\leq \left(\max_{\{1 \leq l \leq p\}} \rho_{h,l} \right) \textstyle\sum_{l=1}^{p} \chi_l(x_i)\, \|\mathbf{u} - \mathbf{u}^{(k)}\|_\infty \\ &\leq \left(\max_{\{1 \leq l \leq p\}} \rho_{h,l} \right) \|\mathbf{u} - \mathbf{u}^{(k)}\|_\infty . \end{aligned}$$

We assumed that each $\chi_l(\cdot)$ has support in $\overline{\Omega}_l$. The desired result follows. □

Remark 15.34. In practice, the bounds for $\rho_{h,l}$ are uniformly independent of h (for sufficiently small h). These rates of convergence will typically be robust, and depend only mildly on the geometry and the coefficients, see [CH6]. The rate of convergence of Schwarz algorithms can be more rapid for singularly perturbed problems, such as the time stepped parabolic equation.

15.4 Analysis of Schwarz Nonmatching Grid Discretizations

In this section, we describe maximum norm estimates on the stability and convergence of discretizations of elliptic equations on nonmatching overlapping grids. We obtain a finite difference discretization of elliptic equation (15.1) by discretizing its Schwarz hybrid formulation (15.28) based on an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ of Ω. On each subdomain Ω_l^*, we let $\mathcal{T}_{h_l}(\Omega_l^*)$ denote a quasiuniform triangulation of size h_l. A global nonmatching grid discretization of (15.1) will require a local discretization of $L\, w_l(x) = f(x)$ on Ω_l^* using the grid $\mathcal{T}_{h_l}(\Omega_l^*)$, as well as constructing an intergrid *interpolation* stencil to discretize the boundary condition $w_l(x) = \sum_{j\neq l} \chi_j(x)\, w_j(x)$. We shall denote the resulting local algebraic equations as:

$$\begin{cases} A_{II}^{(l)}\mathbf{w}_I^{(l)} + A_{IB}^{(l)}\mathbf{w}_B^{(l)} = \mathbf{f}_I^{(l)} \\ \qquad\qquad\qquad\ \mathbf{w}_B^{(l)} = \mathcal{I}_{h_l}\mathbf{w} \end{cases} \quad \text{for } 1 \le l \le p. \tag{15.34}$$

Here $\mathbf{w}_I^{(l)}$ and $\mathbf{w}_I^{(l)}$ denote local nodal vectors corresponding to unknowns in the *interior* of Ω_l^* and on $B^{(l)} = \partial\Omega_l^* \cap \Omega$, while $\mathbf{w} = \left(\mathbf{w}^{(1)^T}, \ldots, \mathbf{w}^{(p)^T}\right)^T$, with $\mathbf{w}^{(l)} = \left(\mathbf{w}_I^{(l)^T}, \mathbf{w}_B^{(l)^T}\right)^T$. Each $\mathcal{I}_{h_l}$ will be an intergrid interpolation map.

Assumptions. In our analysis, we shall assume that the following hold for each local discretization and each intergrid interpolation stencil.

Assumption (A.1). Let $\{\Omega_l\}_{l=1}^p$ denote a non-overlapping decomposition of Ω of diameter h_0, and for some $0 < \beta < 1$ choose:

$$\Omega_l^* = \Omega_l^{\beta h_0} \equiv \{x \in \Omega \,:\, \mathrm{dist}(x, \Omega_l) < \beta h_0\}, \quad \text{for} \quad 1 \le l \le p.$$

Let $\{\chi_l(\cdot)\}_{l=1}^p$ be a partition of unity subordinate to $\{\Omega_l^{\epsilon h_0}\}_{l=1}^p$ for $0 \le \epsilon \ll \beta$. When $\epsilon = 0$, the partition will be *discontinuous* and subordinate to $\{\overline{\Omega}_l\}_{l=1}^p$:

$$\chi_l(x_i) = \begin{cases} 1 & \text{if } x_i \in \Omega_l \\ \dfrac{1}{d(x_i)} & \text{if } x_i \in B^{(l)} \\ 0 & \text{if } x_i \in \Omega \setminus \overline{\Omega}_l \end{cases} \qquad \text{for} \quad 1 \le l \le p,$$

where $d(x_i)$ denotes the number of subdomains Ω_j such that $x_i \in \overline{\Omega}_j$.

Assumption (A.2). On each grid $\mathcal{T}_{h_l}(\Omega_l^*)$ discretize $L\,w_l(x) = f(x)$ in Ω_l^* to be accurate to $O(h_l^{q_l})$. So if $v(x)$ is a smooth solution of $L\,v(x) = f(x)$ and $\boldsymbol{\pi}^{(l)}v = \left((\boldsymbol{\pi}_I^{(l)}v)^T, (\boldsymbol{\pi}_B^{(l)}v)^T\right)^T$ is the restriction of $v(\cdot)$ to nodes in $\overline{\Omega}_l^*$, then:

$$\left\|\left(A_{II}^{(l)}\,\mathbf{v}_I^{(l)} + A_{IB}^{(l)}\mathbf{v}_B^{(l)} - \mathbf{f}_I^{(l)}\right)\right\|_\infty \le c_l(v)\,h_l^{q_l}. \tag{15.35}$$

Here $c_l(v)$ will depend on higher order derivatives of $v(.)$.

Assumption (A.3). Let $A_{II}^{(l)}$ be a strictly diagonally dominant M-matrix with $A_{IB}^{(l)} \le 0$ entrywise.

Assumption (A.4). Let the inter-grid interpolation stencil $\mathcal{I}_{h_l}$ be chosen to discretize $w_l(y_{\tilde{i}}^{(l)}) = \sum_{j\ne l}\chi_j(y_{\tilde{i}}^{(l)})w_j(y_{\tilde{i}}^{(l)})$ using only nodal values of $w_j(x)$ in $\overline{\Omega}_j$ for $j \ne l$. In matrix terms, $(\mathcal{I}_{h_l}\,\mathbf{w})_{\tilde{i}}$ should employ only the nodal vectors $\mathbf{w}^{(j)}$ for $j \ne l$. Furthermore, for each $j \ne l$, only the nodal values of $\mathbf{w}^{(j)}$ for nodes in $\overline{\Omega}_j$ or $\Omega_j^{\epsilon h_0}$ must be used. Given a smooth function $w(x)$ on Ω, with $w_k(x) = w(x)$ on $\overline{\Omega}_k^*$ for $1 \le k \le p$, define $\mathbf{w}^{(k)} \equiv \boldsymbol{\pi}^{(k)}\,w$ and $\mathbf{w} = \left(\mathbf{w}^{(1)^T}, \ldots, \mathbf{w}^{(p)^T}\right)^T$. We shall assume that the stencil has accuracy:

$$\left\|w(y_{\tilde{i}}^{(l)}) - (\mathcal{I}_{h_l}\mathbf{w})_{\tilde{i}}\right\| \le c_l(w)\,h_l^{r_l}.$$

Assumption (A.5). Let each inter-grid interpolation stencil $\mathcal{I}_{h_l}$ have *non-negative* entries with unit row sum, yielding:

$$\|\mathcal{I}_{h_l}\mathbf{w}\|_\infty \le \|\mathbf{w}\|_\infty.$$

Such an interpolation stencil employs *convex* weights.

Truncation Errors. When the above properties hold, the *consistency* and *stability* of (15.34) can be analyzed in the maximum norm, as described below [ST, CA17]. Let $\boldsymbol{\pi}_I^{(l)}v$ and $\boldsymbol{\pi}_B^{(l)}v$ denote the interpolation of a smooth function $v(x)$ onto the interior grid points of Ω_l^* and the grid points on $B^{(l)}$, respectively. Let $\boldsymbol{\pi}\,v \equiv \left((\boldsymbol{\pi}_I^{(l)}v)^T, (\boldsymbol{\pi}_B^{(l)}v)^T, \ldots, (\boldsymbol{\pi}_I^{(p)}v)^T, (\boldsymbol{\pi}_B^{(p)}v)^T\right)^T$. Then, if $u(x)$ denotes the true solution of (15.1), we define the *local* discretization errors as $\boldsymbol{\mathcal{E}}_I^{(l)}$ and $\boldsymbol{\mathcal{E}}_B^{(l)}$ for $1 \le l \le p$:

$$\begin{cases} A_{II}^{(l)}\boldsymbol{\pi}_I^{(l)}u + A_{IB}^{(l)}\boldsymbol{\pi}_B^{(l)}u = \mathbf{f}_I^{(l)} + \boldsymbol{\mathcal{E}}_I^{(l)} \\ \qquad\qquad\qquad\quad \boldsymbol{\pi}_B^{(l)}u = \mathcal{I}_{h_l}\boldsymbol{\pi}u + \boldsymbol{\mathcal{E}}_B^{(l)} \end{cases} \quad \text{for } 1 \le l \le p. \tag{15.36}$$

Subtracting (15.36) from (15.34) will yield the following coupled system for the error $\mathbf{e}_I^{(l)} = \boldsymbol{\pi}_I^{(l)} u - \mathbf{w}_I^{(l)}$ and $\mathbf{e}_B^{(l)} = \boldsymbol{\pi}_B^{(l)} u - \mathbf{w}_B^{(l)}$ for $1 \le l \le p$:

$$\begin{cases} A_{II}^{(l)} \mathbf{e}_I^{(l)} + A_{IB}^{(l)} \mathbf{e}_B^{(l)} u = \boldsymbol{\mathcal{E}}_I^{(l)} \\ \qquad\qquad\qquad \mathbf{e}_B^{(l)} = \mathcal{I}_{h_l} \mathbf{e} + \boldsymbol{\mathcal{E}}_B^{(l)}, \end{cases} \quad \text{for } \; 1 \le l \le p. \tag{15.37}$$

In the remainder of this section, we shall analyze the solvability of (15.34) and obtain maximum norm bounds for $\|\mathbf{e}_I^{(l)}\|_\infty$ and $\|\mathbf{e}_B^{(l)}\|_\infty$. We shall employ the following result on the convergence of Picard iterates [AR3].

Theorem 15.35. *Suppose the following conditions hold.*

1. *Let $\mathcal{H}$ be a complete metric space with metric $d(.,.)$.*
2. *Let $T : \mathcal{H} \to \mathcal{H}$ be a contraction mapping, i.e., there is $0 \le \delta < 1$:*

$$d\,(T\,u, T\,v) \le \delta\; d\,(u, v)\,, \quad \forall\, u,\, v\, \in \mathcal{H}.$$

Then the following results will hold.

1. *There exists a unique $u_* \in \mathcal{H}$ which is a fixed point of T:*

$$u_* = T\,u_*.$$

2. *Given any $u^{(0)} \in \mathcal{H}$, the iterates $u^{(k)} \equiv T^k u^{(0)} \to u_*$ geometrically:*

$$\begin{aligned} d\left(u^{(k+1)}, u_*\right) &\le \delta\; d\left(u^{(n)}, u_*\right) \\ &\le \delta^n\, d\left(u^{(0)}, u_*\right). \end{aligned}$$

3. *For any $u^{(0)} \in \mathcal{H}$, the following will hold:*

$$d\left(u^{(0)}, u_*\right) \le \left(\frac{1}{1-\delta}\right) d\left(T\,u^{(0)}, u^{(0)}\right).$$

Proof. See [AR3]. □

To study the solvability of (15.34) and to obtain maximum norm estimates for the solution to (15.37), we study the following more general system:

$$\begin{cases} A_{II}^{(l)} \mathbf{v}_I^{(l)} + A_{IB}^{(l)} \mathbf{v}_B^{(l)} = \mathbf{g}_I^{(l)} \\ \qquad\qquad\qquad \mathbf{v}_B^{(l)} = \mathcal{I}_{h_l} \mathbf{v} + \mathbf{g}_B^{(l)} \end{cases} \quad \text{for } \; 1 \le l \le p. \tag{15.38}$$

Note that when $\mathbf{g}_I^{(l)} = \mathbf{f}_I^{(l)}$ and $\mathbf{g}_B^{(l)} = \mathbf{0}$, system (15.38) reduces to (15.34). In addition, when $\mathbf{g}_I^{(l)} = \boldsymbol{\mathcal{E}}_I^{(l)}$ and $\mathbf{g}_B^{(l)} = \boldsymbol{\mathcal{E}}_B^{(l)}$, system (15.38) reduces to (15.37).

Let n_l and m_l denote the number of nodal unknowns in Ω_l^* and $B^{(l)}$, and denote $n = \sum_{l=1}^p n_l$ and $m = \sum_{l=1}^p m_p$. Define $V = \mathbb{R}^{n+m}$ and equip it with the maximum norm $\|\cdot\|_\infty$. Given $\mathbf{g}_I^{(l)} \in \mathbb{R}^{n_l}$ and $\mathbf{g}_I^{(l)} \in \mathbb{R}^{m_l}$ for $1 \le l \le p$, we define the following metric space $\mathcal{H}_\mathbf{g} \subset V$ as an *affine* set:

$$\mathcal{H}_\mathbf{g} \equiv \left\{ (\mathbf{v}^{(1)^T}, \ldots, \mathbf{v}^{(p)^T})^T \subset V \;:\; A_{II}^{(l)} \mathbf{v}_I^{(l)} + A_{IB}^{(l)} \mathbf{v}_B^{(l)} \;=\; \mathbf{g}_I^{(l)}, \text{ for } \; 1 \le l \le p \right\}, \tag{15.39}$$

where $\mathbf{v}^{(l)} = \left(\mathbf{v}_I^{(1)^T}, \mathbf{v}_B^{(p)^T}\right)^T$. Equip $\mathcal{H}_\mathbf{g}$ with the metric $\mathrm{d}(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|_\infty$.

To study the solvability of (15.38), we employ a map $T_{\mathbf{g}} : \mathcal{H}_{\mathbf{g}} \rightarrow \mathcal{H}_{\mathbf{g}}$ which we associate with (15.38). Given $\mathbf{v} = (\mathbf{v}^{(1)^T}, \ldots, \mathbf{v}^{(p)^T})^T \in \mathcal{H}_{\mathbf{g}}$, define $\tilde{\mathbf{v}} = T_{\mathbf{g}}\,\mathbf{v} \in \mathcal{H}_{\mathbf{g}}$, where the components of $\tilde{\mathbf{v}} = (\tilde{\mathbf{v}}^{(1)^T}, \ldots, \tilde{\mathbf{v}}^{(p)^T})^T$ solve:

$$\begin{cases} A_{II}^{(l)}\tilde{\mathbf{v}}_I^{(l)} + A_{IB}^{(l)}\tilde{\mathbf{v}}_B^{(l)} = \mathbf{g}_I^{(l)} \\ \qquad\qquad\quad \tilde{\mathbf{v}}_B^{(l)} = \mathcal{I}_{h_l}\mathbf{v} + \mathbf{g}_B^{(l)} \end{cases} \quad \text{for } \; 1 \le l \le p. \tag{15.40}$$

By construction, $\mathbf{v}$ will solve (15.38) *iff* $T_{\mathbf{g}}\,\mathbf{v} = \mathbf{v}$. Under suitable assumptions, we shall show that $T_{\mathbf{g}} : \mathcal{H}_{\mathbf{g}} \rightarrow \mathcal{H}_{\mathbf{g}}$ is a *contraction.*

Lemma 15.36. *Let assumptions (A.1) through (A.5) hold and let* $\mathbf{v}$, $\mathbf{w} \in \mathcal{H}_{\mathbf{g}}$. *Then for node $y_{\tilde{i}} \in \overline{\Omega}_l$ with local index $\tilde{i}$, the following bound will hold:*

$$\max_{\{y_{\tilde{i}}^{(l)} \in \overline{\Omega}_l\}} \left| \left(\mathbf{v}_I^{(l)} - \mathbf{w}_I^{(l)}\right)_{\tilde{i}} \right| \le \rho_{h,l}\, \|\mathbf{v} - \mathbf{w}\|_\infty \qquad \text{for } \; 1 \le l \le p.$$

Proof. By construction, $\mathbf{v} - \mathbf{w}$ will be discrete harmonic in each subdomain:

$$A_{II}^{(l)}\left(\mathbf{v}_I^{(l)} - \mathbf{w}_I^{(l)}\right) + A_{IB}^{(l)}\left(\mathbf{v}_B^{(l)} - \mathbf{w}_B^{(l)}\right) = \mathbf{0}.$$

The result follows by the discrete maximum principle, see Lemma 15.20. □

The next result shows that $T_{\mathbf{g}}$ will be a contraction mapping.

Lemma 15.37. *Let assumptions (A.1) to (A.5) hold. Then, for* $\mathbf{v}$, $\mathbf{w} \in \mathcal{H}_{\mathbf{g}}$*:*

$$\mathrm{d}\left(T_{\mathbf{g}}\,\mathbf{v}, T_{\mathbf{g}}\,\mathbf{w}\right) = \|T_{\mathbf{g}}\,\mathbf{v} - T_{\mathbf{g}}\,\mathbf{w}\|_\infty \;\le\; \delta\,\|\mathbf{v} - \mathbf{w}\|_\infty = \delta\,\mathrm{d}\left(\mathbf{v}, \mathbf{w}\right),$$

where $\delta = \max\{\rho_{h,1}, \ldots, \rho_{h,p}\}$.

Proof. Let $\tilde{\mathbf{v}} = T_{\mathbf{g}}\,\mathbf{v}$ and $\tilde{\mathbf{w}} = T_{\mathbf{g}}\,\mathbf{w}$. By construction, $\tilde{\mathbf{v}} - \tilde{\mathbf{w}}$ will satisfy:

$$\begin{cases} A_{II}^{(l)}\left(\tilde{\mathbf{v}}_I^{(l)} - \tilde{\mathbf{w}}_I^{(l)}\right) + A_{IB}^{(l)}\left(\tilde{\mathbf{v}}_B^{(l)} - \tilde{\mathbf{w}}_B^{(l)}\right) = \mathbf{0} \\ \qquad\qquad\qquad\qquad\qquad\quad \left(\tilde{\mathbf{v}}_B^{(l)} - \tilde{\mathbf{w}}_B^{(l)}\right) = \mathcal{I}_{h_l}\left(\mathbf{v} - \mathbf{w}\right) \end{cases} \quad \text{for } \; 1 \le l \le p. \tag{15.41}$$

Thus, $\tilde{\mathbf{v}} - \tilde{\mathbf{w}}$ will be discrete harmonic, and the maximum principle will yield:

$$\|\tilde{\mathbf{v}}_I^{(l)} - \tilde{\mathbf{w}}_I^{(l)}\|_\infty \le \|\tilde{\mathbf{v}}_B^{(l)} - \tilde{\mathbf{w}}_B^{(l)}\|_\infty.$$

By construction, $\tilde{\mathbf{v}}_B^{(l)} - \tilde{\mathbf{w}}_B^{(l)} = \mathcal{I}_{h_l}\left(\mathbf{v} - \mathbf{w}\right)$ will only involve nodal values of $\mathbf{v} - \mathbf{w}$ within $\overline{\Omega}_j$ for $j \neq l$. Combining this with $\|\mathcal{I}_{h_l}\|_\infty \le 1$, yields:

$$\|\tilde{\mathbf{v}}_B^{(l)} - \tilde{\mathbf{w}}_B^{(l)}\|_\infty = \|\mathcal{I}_{h_l}\left(\mathbf{v} - \mathbf{w}\right)\|_\infty \le \max_{\{j \neq l\}} \|\mathbf{v}^{(j)} - \mathbf{w}^{(j)}\|_{\infty,\overline{\Omega}_j},$$

where $\|\mathbf{v}^{(j)} - \mathbf{w}^{(j)}\|_{\infty,\overline{\Omega}_j}$ denotes the maximum norm of $\left(\mathbf{v}^{(j)} - \mathbf{w}^{(j)}\right)$ for nodes restricted to $\overline{\Omega}_j$. Since $\mathbf{v}$, $\mathbf{w} \in \mathcal{H}$, an application of Lemma 15.36 yields $\max_{\{j \neq l\}} \|\mathbf{v}^{(j)} - \mathbf{w}^{(j)}\|_{\infty,\overline{\Omega}_j} \le \delta\,\|\mathbf{v} - \mathbf{w}\|_\infty$. Combining these two results yields the desired bound. □

Since $T_{\mathbf{g}}$ is a contraction, a unique fixed point will exist, and solve (15.38) for arbitrary $\{\mathbf{g}_I^{(l)}\}$ and $\{\mathbf{g}_B^{(l)}\}$. Maximum norm estimates of its solution can be obtained with the aid of Lemma 15.35. We shall choose $\mathbf{u}^{(0)} \in \mathcal{H}_{\mathbf{g}}$ and estimate $\|T_{\mathbf{g}}\,\mathbf{u}^{(0)} - \mathbf{u}^{(0)}\|_\infty$.

Lemma 15.38. *Suppose the solution to each local system:*

$$A_{II}^{(l)}\mathbf{w}_I^{(l)} + A_{IB}^{(l)}\mathbf{w}_B^{(l)} = \boldsymbol{\beta}_I^{(l)}, \tag{15.42}$$

satisfies the bound

$$\|\mathbf{w}_I^{(l)}\|_\infty \leq c_1\,\|\boldsymbol{\beta}_I^{(l)}\|_\infty + c_2\|\mathbf{w}_B^{(l)}\|_\infty. \tag{15.43}$$

Then, given $\{\mathbf{g}_I^{(l)}\}$ and $\{\mathbf{g}_B^{(l)}\}$, the following results will hold.

1. *There exists $\mathbf{u}^{(0)} \in \mathcal{H}_{\mathbf{g}}$ satisfying:*

$$\|\mathbf{u}^{(0)}\|_\infty \leq c_1 \left(\max_{1\leq l\leq p} \|\mathbf{g}_I^{(l)}\|_\infty \right).$$

2. *The Picard iterate $T_{\mathbf{g}}\,\mathbf{u}^{(0)} \in \mathcal{H}_{\mathbf{g}}$ will satisfy the bound:*

$$\|T_{\mathbf{g}}\,\mathbf{u}^{(0)}\|_\infty \leq (c_1 + c_2) \left(\max_{1\leq l\leq p} \|\mathbf{g}_I^{(l)}\|_\infty \right) + c_2 \left(\max_{1\leq l\leq p} \|\mathbf{g}_B^{(l)}\|_\infty \right).$$

3. *The distance $\mathrm{d}\left(\mathbf{u}^{(0)}, T_{\mathbf{g}}\,\mathbf{u}^{(0)}\right)$ will satisfy the bound:*

$$\mathrm{d}\left(\mathbf{u}^{(0)}, T_{\mathbf{g}}\,\mathbf{u}^{(0)}\right) \leq (2\,c_1 + c_2) \left(\max_{1\leq l\leq p} \|\mathbf{g}_I^{(l)}\|_\infty \right) + c_2 \left(\max_{1\leq l\leq p} \|\mathbf{g}_B^{(l)}\|_\infty \right).$$

4. *The solution $\mathbf{v}$ to (15.38) will satisfy:*

$$\|\mathbf{v}\|_\infty \leq \left(\frac{2\,c_1 + c_2}{1-\delta} \right) \|\mathbf{g}\|_\infty.$$

Proof. We construct $\mathbf{u}^{(0)} = \mathbf{w} \in \mathcal{H}_{\mathbf{g}}$ with *zero* boundary conditions on $B^{(l)}$:

$$\begin{cases} A_{II}^{(l)}\mathbf{w}_I^{(l)} + A_{IB}^{(l)}\mathbf{w}_B^{(l)} = \mathbf{g}_I^{(l)} \\ \hfill \mathbf{w}_B^{(l)} = \mathbf{0} \end{cases}$$

By (15.43), the following bound will hold for $\mathbf{w} \in \mathcal{H}_{\mathbf{g}}$:

$$\|\mathbf{w}_I^{(l)}\|_\infty \leq c_1\,\|\mathbf{g}_I^{(l)}\|_\infty \quad \text{for} \quad 1 \leq l \leq p.$$

Let $\mathbf{u}^{(0)} = \mathbf{w}$, then part 1 follows from the preceding bound.

To prove part 2, let $\tilde{\mathbf{w}} = T_{\mathbf{g}}\,\mathbf{w}$. By definition $\tilde{\mathbf{w}}$ will solve:

$$\begin{cases} A_{II}^{(l)}\tilde{\mathbf{w}}_I^{(l)} + A_{IB}^{(l)}\tilde{\mathbf{w}}_B^{(l)} = \mathbf{g}_I^{(l)} \\ \qquad\qquad\qquad\;\; \tilde{\mathbf{w}}_B^{(l)} = \mathcal{I}_{h_l}\mathbf{w} + \mathbf{g}_B^{(l)} \end{cases} \quad \text{for } 1 \le l \le p.$$

Since $\|\mathcal{I}_{h_l}\|_\infty \le 1$, we may estimate:

$$\|\tilde{\mathbf{w}}_B^{(l)}\|_\infty \le \|\mathbf{w}\|_\infty + \|\mathbf{g}_B^{(l)}\|_\infty.$$

Employing *a priori* estimate (15.43) we obtain:

$$\|\tilde{\mathbf{w}}_I^{(l)}\|_\infty \le c_1 \, \|\mathbf{g}_I^{(l)}\|_\infty + c_2 \left(\|\mathbf{w}\|_\infty + \|\mathbf{g}_B^{(l)}\|_\infty\right).$$

The desired result follows by maximizing over l.

Part 3 follows trivially from parts 1 and 2, and part 4 by an application of Picard's lemma. □

We may now apply the preceding result to estimate the errors:

$$\mathbf{e}_I^{(l)} = \boldsymbol{\pi}_I^{(l)} u - \mathbf{w}_I^{(l)} \quad \text{and} \quad \mathbf{e}_B^{(l)} = \boldsymbol{\pi}_B^{(l)} u - \mathbf{w}_B^{(l)},$$

of Schwarz nonmatching grid discretizations.

Theorem 15.39. *Schwarz discretization (15.34) will be solvable, and each subdomain solution* $\mathbf{w}_I^{(l)}$ *and* $\mathbf{w}_B^{(l)}$. *will satisfy the bound:*

$$\|\mathbf{w}^{(l)}\|_\infty \le \left(\frac{2\,c_1 + c_2}{1-\delta}\right) \|\mathbf{f}_I\|_\infty.$$

Given local truncation errors $\boldsymbol{\mathcal{E}}_I^{(l)}$ *and boundary interpolation errors* $\boldsymbol{\mathcal{E}}_I^{(l)}$*, the error* $\mathbf{e} = \boldsymbol{\pi} u - \mathbf{w}$ *in the Schwarz solution will satisfy (15.37) and the bound:*

$$\|\mathbf{e}\|_\infty \le \left(\frac{2\,c_1 + c_2}{1-\delta}\right) \max\{\|\boldsymbol{\mathcal{E}}_I\|_\infty, \|\boldsymbol{\mathcal{E}}_B\|_\infty\}.$$

Proof. Solvability of the Schwarz discretization follows by the preceding lemma, provided $\delta\,\|\mathcal{I}_{h_l}\|_\infty < 1$. Estimates for $\mathbf{w}_I^{(l)}$ and $\mathbf{w}_B^{(l)}$ can be obtained by using the preceding lemma, using $\mathbf{g}_I^{(l)} = \mathbf{f}_I^{(l)}$ and $\mathbf{g}_B^{(l)} = \mathbf{0}$.

An application of the preceding lemma will yield estimates for the subdomain and boundary errors in terms of $\mathcal{E}_I^{(l)}$, $\mathcal{E}_B^{(l)}$. Provided the intergrid interpolation errors $\|\mathcal{E}_B^{(l)}\|_\infty$ are smaller in magnitude than the local discretization errors, then the scheme will be accurate of optimal order:

$$\|\mathbf{e}\|_\infty \le \left(\frac{2\,c_1 + c_2}{1-\delta}\right) \max\{\|\boldsymbol{\mathcal{E}}_I\|_\infty, \|\boldsymbol{\mathcal{E}}_B\|_\infty\}. \qquad \square$$

Remark 15.40. Schwarz discretization (15.34) can be solved using Picard's contraction mapping $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$. The algorithm will be highly parallel, with a maximum norm convergence factor of $\delta = \max\{\rho_{h,1}, \ldots, \rho_{h,p}\}$.

Remark 15.41. When local grids are *matching*, the truncation error $\mathcal{E}_I^{(l)}$ due to intergrid interpolation will be zero, i.e., $\mathcal{E}_I^{(l)} = \mathbf{0}$ on each subdomain.

15.5 Analysis of Schwarz Heterogeneous Approximations

In this section, we shall estimate the maximum norm of the error introduced by a discretized *elliptic-hyperbolic* heterogeneous approximation of a discretized advection dominated elliptic equation. We consider the equation:

$$\begin{cases} L_\epsilon\, u = -\epsilon\, \nabla \cdot (a(x)\, \nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\, u = f, \text{ in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad u = 0, \text{ on } \partial\Omega. \end{cases} \tag{15.44}$$

Here $0 \le \epsilon \ll 1$, $a(x) \ge 0$ and $c(x) \ge c_0 > 0$. As $\epsilon \to 0^+$, we shall denote the formal limiting operator of $L_\epsilon\, u$ as $L_0 u = \mathbf{b}(x) \cdot \nabla u + c(x)\, u$. In order to construct a Schwarz elliptic-hyperbolic approximation of (15.44), we assume there are *two* overlapping subdomains Ω_1^* and Ω_2^* covering Ω, such that:

$$\epsilon\, |\nabla \cdot (a(x) \nabla u)| \ll |\mathbf{b}(x) \cdot \nabla u + c(x)\, u|\,, \quad \text{on } \Omega_1^*.$$

If this holds, we may formally drop the viscous term $-\epsilon\, \nabla \cdot (a(x)\, \nabla u)$ in Ω_1^*, and impose *inflow* conditions on its boundary, within a Schwarz hybrid formulation of (15.44). The latter will seek $w_1(x)$ on Ω_1^* and $w_2(x)$ on Ω_2^*:

$$\begin{cases} L_\epsilon\, w_1 = f, & \text{in } \Omega_1^* \\ \quad w_1 = w_2, & \text{on } B^{(1)} \\ \quad w_1 = 0, & \text{on } B_{[1]} \end{cases} \text{ and } \begin{cases} L_\epsilon\, w_2 = f, & \text{in } \Omega_2^* \\ \quad w_2 = w_1, & \text{on } B^{(2)} \\ \quad w_2 = 0, & \text{on } B_{[2]}. \end{cases} \tag{15.45}$$

The elliptic-hyperbolic approximation of (15.45) replaces $L_\epsilon\, w_1 = f$ on Ω_1^* by $L_0 v_1 = f$ on Ω_1^*, where $v_1(x) \approx w_1(x)$, and imposes *inflow* conditions on its boundary. For normal $\mathbf{n}(x)$, the *inflow* boundary $\Gamma_{1,in}$ of $L_0 v_1$ on $\partial\Omega_1^*$ is:

$$\Gamma_{1,in} \equiv \{x \in \partial\Omega_1^* \,:\, \mathbf{n}(x) \cdot \mathbf{b}(x) < 0\}\,.$$

The *elliptic-hyperbolic* approximation seeks $v_1(x) \approx w_1(x)$ and $v_2(x) \approx w_2(x)$:

$$\begin{cases} L_0\, v_1 = f, & \text{in } \Omega_1^* \\ \quad v_1 = v_2, & \text{on } B^{(1)} \cap \Gamma_{1,in} \\ \quad v_1 = 0, & \text{on } B_{[1]} \cap \Gamma_{1,in}. \end{cases} \text{ and } \begin{cases} L\, v_2 = f, & \text{in } \Omega_2^* \\ \quad v_2 = v_1, & \text{on } B^{(2)} \\ \quad v_2 = 0, & \text{on } B_{[2]}. \end{cases} \tag{15.46}$$

To obtain a stable discretization of (15.46), we shall assume that assumptions *(A.1)* to *(A.5)* from Chap. 15.4 holds. A discretization of (15.46) will be:

$$\begin{cases} C_{II}^{(1)} \mathbf{v}_I^{(1)} + C_{IB_{in}}^{(1)} \mathbf{v}_{B_{in}}^{(1)} = \mathbf{f}_I^{(1)} \\ \qquad\qquad\qquad\quad \mathbf{v}_{B_{in}}^{(1)} = \mathcal{I}_{1,in}\, \mathbf{v}_I^{(2)} \end{cases} \text{ and } \begin{cases} K_{II}^{(2)} \mathbf{v}_I^{(2)} + K_{IB}^{(2)} \mathbf{v}_B^{(2)} = \mathbf{f}_I^{(2)} \\ \qquad\qquad\qquad\quad \mathbf{v}_B^{(2)} = \mathcal{I}_2 \mathbf{v}_I^{(1)} \end{cases} \tag{15.47}$$

where $C^{(1)}$ is a first order *upwind* discretization of L_0 on Ω_1^*, whose stencil only involves nodes in Ω_1^* and $B_{in}^{(1)} = B^{(1)} \cap \Gamma_{1,in}$, while $K^{(2)} = \epsilon\, A^{(2)} + C^{(2)}$ is a discretization of L_ϵ on Ω_2^*. Both are strictly diagonally dominant M-matrices.

The equations $\mathbf{v}^{(1)}_{B_{in}} = \mathcal{I}_{1,in}\,\mathbf{v}^{(2)}_I$ and $\mathbf{v}^{(2)}_B = \mathcal{I}_2\mathbf{v}^{(1)}_I$ in (15.47) discretize $v_1 = v_2$ on $B^{(1)}_{in}$ and $v_2 = v_1$ on $B^{(2)}$, respectively. When $c(x) \ge c_0 > 0$ and assumptions *(A.1)* to *(A.5)* hold, the matrices $C^{(1)}_{II}$ and $K^{(2)}_{II}$ will be strictly diagonally dominant M-matrices, with $C^{(1)}_{IB_{in}} \le 0$, $K^{(2)}_{IB} = \epsilon\, A^{(2)}_{IB} + C^{(2)}_{IB} \le 0$. In particular, each row of the interpolation matrices $\mathcal{I}_{1,in}$ and $\mathcal{I}_2$ will have *non-negative* entries which sum to one. Since $-\mathcal{I}_{1,in} \le 0$ and $-\mathcal{I}_2 \le 0$ entrywise, the following block matrix form of (15.47) will be an M-matrix, and a discrete maximum principle will hold:

$$\begin{bmatrix} C^{(1)}_{II} & 0 & C^{(1)}_{IB} & 0 \\ 0 & K^{(2)}_{II} & 0 & K^{(2)}_{IB} \\ 0 & -\mathcal{I}_{1,in} & I & 0 \\ -\mathcal{I}_2 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{v}^{(1)}_I \\ \mathbf{v}^{(2)}_I \\ \mathbf{v}^{(1)}_{B_{in}} \\ \mathbf{v}^{(2)}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)}_I \\ \mathbf{f}^{(2)}_I \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{15.48}$$

When the preceding assumptions hold, system (15.48) will be *solvable*, as $T_{\mathbf{g}}$ described in Chap. 15.4 can be shown to be a *contraction*. Indeed, let $\mathcal{H}$:

$$\mathcal{H} = \left\{ (\mathbf{w}^{(1)^T}_I, \mathbf{w}^{(2)^T}_I, \mathbf{w}^{(1)^T}_{B_{in}}, \mathbf{w}^{(2)^T}_B)^T \, : \, \begin{matrix} C^{(1)}_{II}\mathbf{w}^{(1)}_I + C^{(1)}_{IB}\mathbf{w}^{(1)}_{B_{in}} = \mathbf{f}^{(1)}_I \\ K^{(2)}_{II}\mathbf{w}^{(2)}_I + K^{(1)}_{IB}\mathbf{w}^{(2)}_B = \mathbf{f}^{(2)}_I \end{matrix} \right\}.$$

Given $\mathbf{w} = (\mathbf{w}^{(1)^T}_I, \mathbf{w}^{(2)^T}_I, \mathbf{w}^{(1)^T}_{B_{in}}, \mathbf{w}^{(2)^T}_B)^T \in \mathcal{H}$, we define $\tilde{\mathbf{w}} = T\mathbf{w}$ as solving:

$$\begin{bmatrix} C^{(1)}_{II} & 0 & C^{(1)}_{IB} & 0 \\ 0 & K^{(2)}_{II} & 0 & K^{(2)}_{IB} \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{w}}^{(1)}_I \\ \tilde{\mathbf{w}}^{(2)}_I \\ \tilde{\mathbf{w}}^{(1)}_{B_{in}} \\ \tilde{\mathbf{w}}^{(2)}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(1)}_I \\ \mathbf{f}^{(2)}_I \\ \mathcal{I}_{1,in}\mathbf{w}^{(2)}_I \\ \mathcal{I}_2\mathbf{w}^{(1)}_I \end{bmatrix}. \tag{15.49}$$

By construction, if $\mathbf{u}, \mathbf{w} \in \mathcal{H}$, then $\tilde{\mathbf{u}} = T\mathbf{u}$, $\tilde{\mathbf{w}} = T\mathbf{w} \in \mathcal{H}$ and $\mathbf{u} - \mathbf{w}$, $\tilde{\mathbf{u}} - \tilde{\mathbf{w}}$ will each be discrete harmonic on the subdomains Ω^*_l for $1 \le l \le 2$:

$$\begin{cases} C^{(1)}_{II}(\mathbf{u}^{(1)}_I - \mathbf{w}^{(1)}_I) + C^{(1)}_{IB}(\mathbf{u}^{(1)}_{B_{in}} - \mathbf{w}^{(1)}_{B_{in}}) = \mathbf{0} \\ K^{(2)}_{II}(\mathbf{u}^{(2)}_I - \mathbf{w}^{(2)}_I) + K^{(1)}_{IB}(\mathbf{u}^{(2)}_B - \mathbf{w}^{(2)}_B) = \mathbf{0} \ \text{ and} \\ C^{(1)}_{II}(\tilde{\mathbf{u}}^{(1)}_I - \tilde{\mathbf{w}}^{(1)}_I) + C^{(1)}_{IB}(\tilde{\mathbf{u}}^{(1)}_{B_{in}} - \tilde{\mathbf{w}}^{(1)}_{B_{in}}) = \mathbf{0} \\ K^{(2)}_{II}(\tilde{\mathbf{u}}^{(2)}_I - \tilde{\mathbf{w}}^{(2)}_I) + K^{(1)}_{IB}(\tilde{\mathbf{u}}^{(2)}_B - \tilde{\mathbf{w}}^{(2)}_B) = \mathbf{0} \end{cases}$$

Thus, $\|(T\mathbf{u})^{(l)} - (T\mathbf{w})^{(l)}\|_\infty \le \max\{\|\mathbf{u}^{(1)} - \mathbf{w}^{(1)}\|_{\infty,\overline{\Omega}_1}, \|\mathbf{u}^{(2)} - \mathbf{w}^{(2)}\|_{\infty,\overline{\Omega}_2}\}$, since $\|\mathcal{I}_{1,in}\|_\infty \le 1$ and $\|\mathcal{I}_2\|_\infty \le 1$. Since $c(x) \ge c_0 > 0$ and assumptions *(A.1)* to *(A.5)* hold, $\rho_{h,l} \le e^{-\gamma\, d_l}$ independent of h and ϵ (for small h), using the comparison function $e^{-\gamma d_l(x)}$ on Ω^*_l. Since $\mathbf{u}^{(l)} - \mathbf{w}^{(l)}$ is discrete harmonic on each Ω^*_l, we obtain $\|\mathbf{u}^{(l)} - \mathbf{w}^{(l)}\|_{\infty,\overline{\Omega}_l} \le \rho_{h,l}\|\mathbf{u} - \mathbf{w}\|_\infty$ for $1 \le l \le 2$. Combining the preceding yields the contraction factor of $T_{\mathbf{g}}$ as $\delta = \max\{\rho_{h,1}, \rho_{h,2}\}$.

Truncation Errors. Since the heterogeneous system (15.46) omits the viscosity term $-\epsilon\,(\nabla \cdot a(x)\nabla v_1)$ in Ω_1^*, its discretization (15.47) or (15.48) will include a larger truncation error term due to the omitted viscosity term. Below, we estimate the global error. Accordingly, let $u(\cdot)$ denote the true solution of (15.44). We define the local discretization errors as:

$$\begin{cases} (\mathcal{E}_I^{(1)})_i \equiv \left(C_{II}^{(1)} \boldsymbol{\pi}_I^{(1)} u + C_{IB_{in}}^{(1)} \boldsymbol{\pi}_{B_{in}}^{(1)} u - \mathbf{f}_I^{(1)} \right) \\ \qquad\quad = (-\epsilon \nabla \cdot (a \nabla u) + L_0 u - f)(x_i) \qquad \text{for } x_i \in \Omega_1^* \\ (\mathcal{E}_I^{(2)})_i \equiv \left(K_{II}^{(2)} \boldsymbol{\pi}_I^{(2)} u + K_{IB}^{(2)} \boldsymbol{\pi}_B^{(2)} u - \mathbf{f}_I^{(2)} \right) \\ \qquad\quad = (L_\epsilon u - f)(x_i) \qquad\qquad\qquad\quad \text{for } x_i \in \Omega_2^* \\ \mathcal{E}_{B_{in}}^{(1)} \equiv \left(\boldsymbol{\pi}_{B_{in}}^{(1)} u - \mathcal{I}_{1,in} \boldsymbol{\pi}^{(2)} u \right) \\ \mathcal{E}_B^{(2)} \equiv \left(\boldsymbol{\pi}_B^{(2)} u - \mathcal{I}_2 \boldsymbol{\pi}^{(1)} u \right), \end{cases} \tag{15.50}$$

where $\boldsymbol{\pi}_I^{(1)} u$, $\boldsymbol{\pi}_{B_{in}}^{(1)} u$, $\boldsymbol{\pi}_I^{(2)} u$ and $\boldsymbol{\pi}_B^{(2)} u$ denote nodal interpolation of $u(.)$ onto the grid points of $\mathcal{T}_{h_1}(\Omega_1^*)$ in Ω_1^* and $B_{in}^{(1)}$ and onto the grid points of $\mathcal{T}_{h_2}(\Omega_2^*)$ in Ω_2^* and $B^{(2)}$. Importantly, $\mathcal{E}_I^{(1)} = O(h_1^{r_1}) + O(\epsilon\,|\nabla \cdot (a \nabla u)|)$ is a sum of the local truncation error $O(h_1^{r_1})$ for the discretization of $(L_0 u - f)$ on Ω_1^* and the omitted viscosity term. When $\epsilon\,|\nabla \cdot (a \nabla u)| = O(h_1^{r_1})$ in Ω_l^*, then omission of the viscosity term on Ω_1^* does not contribute significantly to the global error.

To ensure (maximum norm) stable local problems, we will require that:

$$\begin{cases} C_{II}^{(1)} \mathbf{w}_I^{(1)} + C_{IB_{in}}^{(1)} \mathbf{w}_{B_{in}}^{(1)} = \boldsymbol{\beta}_I^{(1)} \\ K_{II}^{(2)} \mathbf{w}_I^{(2)} + K_{IB}^{(2)} \mathbf{w}_B^{(2)} = \boldsymbol{\beta}_I^{(2)}, \end{cases} \tag{15.51}$$

satisfy the bounds:

$$\begin{cases} \|\mathbf{w}_I^{(1)}\|_\infty \le c_1\, \|\boldsymbol{\beta}_I^{(1)}\|_\infty + c_2 \|\mathbf{w}_{B_{in}}^{(1)}\|_\infty \\ \|\mathbf{w}_I^{(2)}\|_\infty \le c_1\, \|\boldsymbol{\beta}_I^{(2)}\|_\infty + c_2 \|\mathbf{w}_B^{(2)}\|_\infty. \end{cases} \tag{15.52}$$

Below, we state a result on the accuracy of (15.47).

Lemma 15.42. *Suppose the following conditions hold.*

1. *Let $c(x) \ge c_0 > 0$ and let assumptions (A.1) to (A.5) hold.*
2. *Define the local discretization errors as (15.50).*
3. *Suppose that the a priori estimates (15.52) hold.*

Then, the following error bounds will hold.

$$\|\mathbf{e}\|_\infty \le \left(\tfrac{2\,c_1 + c_2}{1-\delta} \right) \max\{ \|\mathcal{E}_I^{(1)}\|_\infty, \|\mathcal{E}_{B_{in}}^{(1)}\|_\infty, \|\mathcal{E}_I^{(2)}\|_\infty, \|\mathcal{E}_B^{(2)}\|_\infty \}.$$

Proof. Analogous to the proof in the preceding section. The map $T_{\mathbf{g}}$ can also be used as a parallel iterative solver. □

15.6 Applications to Parabolic Equations

The maximum norm theory described in the preceding sections can also be extended to discretizations of parabolic equations, provided the discretizations satisfy a discrete maximum principle and comparison theorem [MA35]. Below, we outline the salient points for the following parabolic equation:

$$\begin{cases} u_t + L\,u = f(x,t), \text{ in } \Omega \times (0,t_*) \\ \qquad\quad u = g(x,t), \text{ on } \partial\Omega \times (0,t_*) \\ \;\; u(x,0) = u_0(x), \;\text{ in } \Omega, \;\text{ for } \; t=0, \end{cases} \tag{15.53}$$

where $L\,u(x) \equiv -\nabla \cdot (a(x)\nabla u) + \mathbf{b}(x) \cdot \nabla u + c(x)\,u$. We shall assume that the coefficients satisfy $a(x) \geq 0$ and $c(x) \geq c_0 > 0$.

Let $A_{II}\mathbf{u}_I(t) + A_{IB}\mathbf{u}_B(t)$ denotes a finite difference discretization of $L\,u(.,t)$, where $\mathbf{u}_I(t)$ and $\mathbf{u}_B(t)$ denote nodal vectors associated with the discrete solution in the interior of Ω and on $\partial\Omega$, respectively. We assume that A_{II} is a strictly diagonally dominant M-matrix and that $A_{IB} \leq 0$ entrywise. Let $\mathbf{g}_B(t)$ denote the nodal vector associated with $g(.,t)$ on $\partial\Omega$. Then, a semi-discretization of (15.53) will be:

$$\begin{cases} \mathbf{u}_I'(t) + A_{II}\mathbf{u}_I(t) + A_{IB}\mathbf{u}_B(t) = \mathbf{f}_I(t), \quad \text{for } 0 < t < t_* \\ \qquad\qquad\qquad\qquad\quad\;\; \mathbf{u}_B(t) = \mathbf{g}_B(t), \quad \text{for } 0 < t < t_* \\ \qquad\qquad\qquad\qquad\quad\;\; \mathbf{u}_I(0) = \boldsymbol{\pi}_I u_0. \end{cases} \tag{15.54}$$

If $\tau = (t_*/m)$ denotes the time step and $\mathbf{u}_I^k$ the discrete solution at time $t_k = k\,\tau$, then a θ-scheme discretization of (15.53) will yield:

$$\begin{cases} (I + \tau\theta A_{II})\mathbf{u}_I^{(k+1)} - (I - \tau(1-\theta)A_{II})\mathbf{u}_I^{(k)} = \tilde{\mathbf{f}}_I^{(k+1)}, \text{ for } 0 \leq k \leq (m-1) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathbf{u}_I^{(0)} = \boldsymbol{\pi}_I u_0, \;\text{ for } k=0. \end{cases} \tag{15.55}$$

where $\tilde{\mathbf{f}}_I^{(k+1)} \equiv \tau\,\theta\,\mathbf{f}_I^{(k+1)} + \tau\,(1-\theta)\mathbf{f}_I^{(k)} - \tau\,(\theta\,A_{IB}\mathbf{g}_B^{(k+1)} + (1-\theta)A_{IB}\mathbf{g}_B^{(k)})$ and $\mathbf{g}_B^k = \mathbf{g}_B(k\tau)$ denotes the Dirichlet boundary data at time $k\tau$. The resulting system for determining $\mathbf{u}_I^{(1)}, \ldots, \mathbf{u}_I^{(m)}$ can be expressed in block matrix form:

$$\begin{bmatrix} D & & & \\ -E & D & & \\ & \ddots & \ddots & \\ & & -E & D \end{bmatrix} \begin{bmatrix} \mathbf{u}_I^{(1)} \\ \mathbf{u}_I^{(2)} \\ \vdots \\ \mathbf{u}_I^{(m)} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{f}}_I^{(1)} + E\mathbf{u}_I^{(0)} \\ \tilde{\mathbf{f}}_I^{(2)} \\ \vdots \\ \tilde{\mathbf{f}}_I^{(m)} \end{bmatrix}, \tag{15.56}$$

where $D = (I + \theta\tau A_{II})$ and $E = (I - (1-\theta)\tau A_{II})$. We impose the time step restriction $0 < \tau \leq \min_i \frac{1}{(1-\theta)\,A_{ii}}$ for maximum norm stability. It ensures that D is a strictly diagonally dominant M-matrix and $-E \leq 0$ entrywise.

Thus, the coefficient matrix in system (15.56) will be an M-matrix, yielding a maximum principle and comparison theorem [KA2, MA35].

When assumptions *(A.1)* to *(A.5)* hold and $c(x) \geq c_0 > 0$, and τ satisfies the constraint $0 < \tau \leq \min_i \frac{1}{(1-\theta)\, A_{ii}}$, the coefficient matrix in system (15.56) will be an M-matrix [KA2, MA35]. Thus, the inverse of the coefficient matrix in (15.56) will have *non-negative* entries. Since $E \geq 0$ and $-A_{IB} \geq 0$ entrywise, we will immediately obtain that the components $(\mathbf{u}_I^{(k)})_i$ of the solution to (15.56) are *monotone* increasing with respect to the entries $(\mathbf{f}_I^k)_j$ of the forcing data and the entries $(\mathbf{g}_B^k)_l$ of the boundary data. Importantly, if $\mathbf{w}_I^*$ is a comparison function for the associated discretized elliptic equation, it will be a *stationary* comparison function for the discretized parabolic equation:

$$\begin{bmatrix} D & & & \\ -E & D & & \\ & \ddots & \ddots & \\ & & -E & D \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^* \\ \mathbf{w}_I^* \\ \vdots \\ \mathbf{w}_I^* \end{bmatrix} \geq \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \tag{15.57}$$

As a result, the contraction factors will be the same as in the elliptic case, and this result also applies on any cylindrical space-time subregion.

Given an overlapping decomposition $\Omega_1^*, \ldots, \Omega_p^*$ of Ω, we obtain an overlapping decomposition of $\Omega \times (0, t_*)$ into space-time cylinders $\{\Omega_l^* \times (0, t_*)\}_{l=1}^p$. Most of the maximum norm results form preceding sections can be appropriately generalized to the time dependent case, including the well posedness of a Schwarz hybrid discretization on space-time domains and heterogeneous approximations. The salient point is that any *steady state* comparison grid function can be used as a comparison grid function in the parabolic case. This enables contraction factor estimates analogous to the elliptic case, and a contractive Picard mapping T. We omit the details [MA35].

16

Eigenvalue Problems

In this chapter, we describe domain decomposition and block matrix methods for large sparse *symmetric* eigenproblems [WI8, PA7, CH21, GO4, CI4, SA]. We focus on algorithms which *iteratively* approximate the *minimal* eigenvalue and corresponding eigenvector of a matrix, though most such methods can also be extended to simultaneously approximate several eigenvalues, and their associated eigenvectors, see [KR, KU2, BO10, BO11, BO12, BR10, MA9, LU5] and [LU6, KN2, BO13, KN3, CH16].

Our discussion will be organized as follows. In Chap. 16.1, we describe some background on the symmetric eigenvalue problem. Following this, Chap. 16.2 describes preconditioned *gradient methods* for eigenvalue problems. Chap. 16.3 describes block matrix methods for eigenvalue problems, involving a Schur complement matrix. Chap. 16.4 describes Schwarz subspace algorithms for eigenvalue problems. We conclude our discussion with an outline of the *modal synthesis* Rayleigh-Ritz approximation of eigenproblems in Chap. 16.5. We focus primarily of the matrix formulation of the underlying algorithms.

16.1 Background

We consider an eigenvalue problem associated with an elliptic operator L. It seeks $\lambda \in \mathbb{R}$ and sufficiently smooth $u(.) \neq 0$ such that:

$$\begin{cases} L\,u(x) \;\equiv\; -\nabla \cdot (a(x)\nabla u) + c(x)\,u = \lambda\,u(x) & \text{in } \Omega \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad u = 0, & \text{on } \partial\Omega. \end{cases}$$

A discretization of the eigenproblem, by a finite element or finite difference method, will seek $\lambda \in \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^n$ with $\mathbf{u} \neq \mathbf{0}$, satisfying:

$$A\,\mathbf{u} = \lambda\,M\,\mathbf{u}, \tag{16.1}$$

where $A^T = A$ is a *real* symmetric matrix of size n, obtained by the discretization of the self adjoint elliptic operator L, while $M = M^T > 0$ is a matrix of size n, corresponding to the mass matrix in finite element methods [ST14], or to the identity matrix $M = I$ in finite difference methods.

The *Rayleigh quotient* function associated with (16.1) is defined as:

$$\mathcal{R}\,(\mathbf{v}) \equiv \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T M \mathbf{v}}, \qquad \text{for } \mathbf{v} \neq \mathbf{0}. \tag{16.2}$$

Computing $\nabla\mathcal{R}(\mathbf{u})$ and solving $\nabla\mathcal{R}(\mathbf{u}) = \mathbf{0}$ yields:

$$\nabla\mathcal{R}(\mathbf{u}) = \left(\frac{2}{\mathbf{u}^T M \mathbf{u}}\right)(A\,\mathbf{u} - \mathcal{R}(\mathbf{u})\,M\,\mathbf{u}) = \mathbf{0} \;\;\Longrightarrow\;\; A\,\mathbf{u} = \mathcal{R}(\mathbf{u})\,M\,\mathbf{u}.$$

Thus, if $\mathbf{u}$ is a critical point of $\mathcal{R}(\cdot)$, then $\mathbf{u}$ will be an eigenvector of the generalized eigenvalue problem (16.1) corresponding to eigenvalue $\lambda = \mathcal{R}(\mathbf{u})$. Since A and M are Hermitian, applying inner products yields that the generalized eigenvalues of (16.1) are real, see [ST13], and can be ordered as:

$$\lambda_1 \;\leq\; \lambda_2 \;\leq\; \cdots \;\leq\; \lambda_n.$$

We let $\mathbf{u}_k \in \mathbb{R}^n$ denote the eigenvector corresponding to eigenvalue λ_k:

$$A\,\mathbf{u}_k \;=\; \lambda_k\,M\,\mathbf{u}_k.$$

By the preceding, the minimum value of $\mathcal{R}(\cdot)$ will be attained when:

$$\lambda_1\left(M^{-1}A\right) = \mathcal{R}(\mathbf{u}_1) = \min_{\{\mathbf{v}\neq\mathbf{0}\}} \mathcal{R}(\mathbf{v}).$$

Additionally, the following property will hold for the k'th eigenvalue:

$$\lambda_k \;=\; \mathcal{R}(\mathbf{u}_k) \;=\; \min_{\{\mathbf{v}\neq\mathbf{0}\,:\,\mathbf{v}^T M\mathbf{u}_i\,=\,0,\;1\leq i\leq k-1\}} \mathcal{R}(\mathbf{v}),$$

see for instance [GO4], where also the *min-max* characterization of the eigenvalues is described for the case $M = I$. In the following, we indicate several traditional methods for determining the minimal eigenvalue of a generalized eigenvalue problem [WI8, PA7, CH21, GO4, CI4, SA].

Shifted Inverse Power Method. The shifted inverse power method is motivated by the following observation. Suppose $M = I$ and $\mathbf{v}$ is a randomly chosen vector which has a nontrivial component in the direction of the unknown eigenvector $\mathbf{u}_1$. Then, if $\mu \approx \lambda_1 < \lambda_2$, the vector $(A - \mu I)^{-k}\,\mathbf{v}$ will approach a scalar multiple of $\mathbf{u}_1$ for large k, provided $\lambda_1 < \mu < (\lambda_1 + \lambda_2)/2$. Accordingly, the shifted inverse power method starts with a guess $\mathbf{v}^{(0)} \in \mathbb{R}^n$, defines $\mu^{(0)} = \mathcal{R}(\mathbf{v}^{(0)})$ and computes updates as follows (when $M = I$):

$$\begin{cases} \mathbf{v}^{(k+1)} \leftarrow \left(A - \mu^{(k)} I\right)^{-1} \mathbf{v}^{(k)} \\ \mu^{(k+1)} \leftarrow \mathcal{R}(\mathbf{v}^{(k+1)}). \end{cases}$$

For appropriately chosen starting iterates, the iterates $\mathbf{v}^{(k)}$ and $\mu^{(k)}$ will converge to the desired minimum eigenvector and corresponding eigenvalue. Efficient implementations can be found in [GO4, SA, CI4].

Lanczos Method. The Lanczos method is an iterative algorithm based on the computation of *Ritz vectors* and *Ritz values* associated with a Krylov space $\mathcal{K}_l \subset \mathbb{R}^n$. Given a subspace $\mathcal{V}_l \subset \mathbb{R}^n$, recall that a vector $\mathbf{v}_l \in \mathcal{V}_l$ is referred to as a Ritz vector provided it is a critical point of the Rayleigh quotient $\mathcal{R}(\cdot)$ *restricted to* $\mathcal{V}_l$. A Ritz vector $\mathbf{v}_l$ will approximate an eigenvector, while its associated Ritz value $\mathcal{R}(\mathbf{v}_l)$ will approximate an eigenvalue. Ritz vectors and values can be computed by solving a generalized eigenvalue problem of a smaller size. For instance, a Ritz value μ_l approximating the minimal eigenvalue, and its associated Ritz vector $\mathbf{v}_l$ will satisfy $\mathbf{v}_l \in \mathcal{V}_l$

$$\mu_l = \mathcal{R}(\mathbf{v}_l) = \min_{\{\mathbf{v} \in \mathcal{V}_l \backslash \mathbf{0}\}} \mathcal{R}(\mathbf{v}).$$

If $n_l = \dim(\mathcal{V}_l)$ and K_l is a matrix of size $n \times n_l$ whose columns span $\mathcal{V}_l$ with $\text{Range}(K_l) = \mathcal{V}_l$, then a Ritz vector $\mathbf{v}_l \in \mathcal{V}_l$ approximating the minimal eigenvector can be computed by determining $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}$ and $\mu_l \in \mathbb{R}$ satisfying:

$$\left(K_l^T A K_l\right) \mathbf{d}^{(l)} = \mu_l \left(K_l^T M K_l\right) \mathbf{d}^{(l)}.$$

This yields a eigenproblem of size n_l. Here μ_l denotes the minimal eigenvalue of $(K_l^T M K_l)^{-1}\,(K_l^T A K_l)$. Once $\mathbf{d}^{(l)}$ has been determined, the *Ritz vector* can be computed as $\mathbf{v}_l = K_l \mathbf{d}^{(l)} \in \mathcal{V}_l$.

Given a starting guess $\mathbf{v} \in \mathbb{R}^n$, the Lanczos iterates formally correspond to Ritz vectors and values based on a Krylov space $\mathcal{K}_l$ of dimension l:

$$\mathcal{K}_l(\mathbf{v}) = \text{span}\left\{\mathbf{v}, (M^{-1}A)\mathbf{v}, \ldots, (M^{-1}A)^{l-1}\mathbf{v}\right\}.$$

Efficient implementations of the Lanczos method can be found in [GO4, SA].

16.2 Gradient and Preconditioned Gradient Methods

Gradient methods determine approximate eigenvalues by seeking the *minima* of the *Rayleigh quotient* [SA6, KN6, BR10, KN6]. In *large* sparse symmetric eigenvalue computations arising from the discretization of elliptic eigenvalue problems, gradient methods have the advantage of low computational cost per iteration. With a careful choice of preconditioner, such methods can converge at rates independent of the mesh size h, and yield parallelizable algorithms [KN6, BR10, KN2]. In this section, we shall describe variants of the preconditioned gradient method for determining the lowest eigenvalue λ_1 and associated eigenvector $\mathbf{u}_1$ of (16.1).

Given an iterate $\mathbf{w}^{(k)} \approx \mathbf{u}_1$, the gradient update $\mathbf{w}^{(k+1)}$ is computed as:

$$\begin{cases} \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \tau\, \nabla \mathcal{R}(\mathbf{w}^{(k)}) \\ \qquad\quad\;\; = \mathbf{w}^{(k)} - \tilde{\tau}\, \left(A\,\mathbf{w}^{(k)} - \mathcal{R}(\mathbf{w}^{(k)})\, M\, \mathbf{w}^{(k)} \right), \end{cases}$$

where τ, $\tilde{\tau} > 0$ are step sizes. A preconditioner G of size n can be employed:

$$\begin{cases} \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} - \tau\, G^{-1}\, \nabla \mathcal{R}(\mathbf{v}^{(k)}) \\ \qquad\quad\; = \mathbf{v}^{(k)} - \tilde{\tau}\, G^{-1}\, \left(A\,\mathbf{v}^{(k)} - \mathcal{R}(\mathbf{v}^{(k)})\, M\, \mathbf{v}^{(k)} \right), \end{cases}$$

to speed up the gradient iteration. The idea of "preconditioning" an eigenvalue problem may be heuristically motivated by observing that a suitably chosen preconditioner G may increase the component of the update in the direction of the desired eigenvector $\mathbf{u}_1$. Ideally, G^{-1} must amplify the components in the direction of $\mathbf{u}_1$ and damp other components. Optimal preconditioners may be *indefinite* of the form $A - \mathcal{R}(\mathbf{w}^{(k)})M$ and nearly singular [CA24, KN2] (as in Davidson's method [DA3]). We focus only on preconditioners $G = G^T > 0$, since theoretical results are scant for indefinite preconditioners. We assume:

$$(1-\gamma)\, \mathbf{v}^T G \mathbf{v} \;\leq\; \mathbf{v}^T A \mathbf{v} \;\leq\; (1+\gamma)\, \mathbf{v}^T G \mathbf{v}, \quad \forall \mathbf{v} \in \mathbb{R}^n,$$

for some $0 < \gamma < 1$. In this case, the rate of convergence will depend only on γ, see [KN, KN6, BR10]. Note that if $\hat{G}$ is a preconditioner satisfying:

$$c_1 \left(\mathbf{v}^T \hat{G} \mathbf{v} \right) \;\leq\; \left(\mathbf{v}^T A \mathbf{v} \right) \;\leq\; c_2 \left(\mathbf{v}^T \hat{G} \mathbf{v} \right), \quad \forall \mathbf{v} \in \mathbb{R}^n,$$

we may define $G = \left(\frac{2}{c_1 + c_2} \right) \hat{G}$ by scaling, yielding $\gamma < 1$.

Remark 16.1. Given $\mathbf{v}^{(k)}$ and $\mathbf{d}^{(k)} \equiv G^{-1} \left(A\,\mathbf{v}^{(k)} - \mathcal{R}(\mathbf{v}^{(k)})\, M\, \mathbf{v}^{(k)} \right)$ as the *descent direction*, an *optimal* update $\mathbf{v}^{(k+1)}$ with choice of τ may be found by minimizing the Rayleigh quotient $\mathcal{R}(\cdot)$ in the two dimensional subspace $\text{span}\{\mathbf{v}^{(k)}, \mathbf{d}^{(k)}\}$ generated by $\mathbf{v}^{(k)}$ and $\mathbf{d}^{(k)}$, see [KN2]:

$$\mathcal{R}(\mathbf{v}^{(k+1)}) \;=\; \min_{\mathbf{v} \in \text{span}\{\mathbf{v}^{(k)}, \mathbf{d}^{(k)}\} \backslash \mathbf{0}} \mathcal{R}(\mathbf{v}).$$

This will require the solution of a generalized eigenvalue problem of size 2 (for which analytical expressions may be derived, if desired).

16.3 Schur Complement Methods

We next consider Schur complement based methods for eigenvalue problems. Consider the following block partition of the eigenvalue problem (16.1):

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \lambda \begin{bmatrix} M_{II} & M_{IB} \\ M_{IB}^T & M_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix}.$$

The above block partition may arise from a *non-overlapping* decomposition $\Omega_1, \ldots, \Omega_p$ with interface $B = \cup_{i=1}^p (\partial\Omega_i \cap \Omega)$, where $\mathbf{u}_I = (\mathbf{u}_I^{(1)^T}, \ldots, \mathbf{u}_p^{(p)^T})^T$ represents the unknowns in the interior of the subregions, and $\mathbf{u}_B$ the unknowns on the interface. The block matrix $A_{II} = \text{blockdiag}(A_{II}^{(1)}, \ldots, A_{II}^{(p)})$ will then be block diagonal, where $A_{II}^{(i)}$ is the stiffness matrix on Ω_i. Matrix M_{II} can be block partitioned similarly $M_{II} = \text{blockdiag}(M_{II}^{(1)}, \ldots, M_{II}^{(p)})$. The eigenvalue problem can be expressed as an equivalent *inhomogeneous* system which seeks a *non-zero* solution $(\mathbf{u}_I^T, \mathbf{u}_B^T)^T$ to:

$$\begin{bmatrix} A_{II} - \lambda M_{II} & A_{IB} - \lambda M_{IB} \\ A_{IB}^T - \lambda M_{IB}^T & A_{BB} - \lambda M_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_I \\ \mathbf{u}_B \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

Formally eliminating $\mathbf{u}_I$, we seek $\mathbf{u}_B \neq \mathbf{0}$ satisfying:

$$\begin{cases} S(\lambda)\,\mathbf{u}_B \; = \; \mathbf{0}, \qquad \text{where} \\ S(\lambda) \; = \; (A_{BB} - \lambda M_{BB}) - \left(A_{IB}^T - \lambda M_{IB}^T\right)(A_{II} - \lambda\, M_{II})^{-1}(A_{IB} - \lambda M_{IB}). \end{cases}$$

For the above reduction to be valid, it is sufficient that $\det{(A_{II} - \lambda M_{II})} \neq 0$, i.e., λ is not an eigenvalue of $(A_{II} - \lambda M_{II})$. If $A_{II} = \text{blockdiag}\left(A_{II}^{(1)}, \ldots, A_{II}^{(p)}\right)$, and $M_{II} = \text{blockdiag}\left(M_{II}^{(1)}, \ldots, M_{II}^{(p)}\right)$, then the preceding requirement will be equivalent to the condition that λ not be an eigenvalue of each of the subdomain eigenvalue problems $(A_{II}^{(l)} - \lambda\, M_{II}^{(i)})$. A *nontrivial* solution $\mathbf{u}_B$ of $S(\lambda)\,\mathbf{u}_B \; = \; 0$, will exist only when λ satisfies:

$$f(\lambda) \; \equiv \; \det(S(\lambda)) \; = \; 0. \tag{16.3}$$

In the following, we describe a *secant* method to determine λ (and hence $\mathbf{u}_B$).

Algorithm 16.3.1 *(Variant of Kron's Method)*
Choose starting guess $\mu^{(0)}$

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     Compute: $f(\mu^{(k)}) = \lambda_{min}(S(\mu^{(k)}))$
3.     Solve: $f(\mu) = 0$ using the secant method:
```

$$\mu^{(k+1)} = \mu^{(k)} - f(\mu^{(k)}) \left(\frac{\mu^{(k)} - \mu^{(k-1)}}{f(\mu^{(k)}) - f(\mu^{(k-1)})} \right)$$

```
4. Endfor
```

Remark 16.2. Since $S(\mu)$ is real symmetric, $\lambda_{min}(S(\mu^{(k)}))$ can be approximated using a Lanczos iteration, without assembly of $S(\mu^{(k)})$. Once $\lambda \in \mathbb{R}$ and $\mathbf{u}_B \neq \mathbf{0}$ have been determined using Lanczos, we may determine $\mathbf{u}_I$ as:

$$\mathbf{u}_I = -\left(A_{II} - \lambda\, M_{II}\right)^{-1} \left(A_{IB} - \lambda\, M_{IB}\right) \mathbf{u}_B,$$

provided λ is not an eigenvalue of $(A_{II} - \lambda M_{II})$, see [KR, SI4].

Remark 16.3. Each matrix vector product with $S(\mu)$ requires the solution of a linear system with an *indefinite* coefficient matrix of the form $(A_{II} - \lambda\, M_{II})$. If each $A_{II}^{(l)}$ is of sufficiently small size, then a direct method may be employed.

Below, we describe a block matrix gradient type algorithm [KN6, KN2] based on the Schur complement. We employ the notation $S(\lambda)$ as before. Let $G_B = G_B^T > 0$ be a multisubdomain Schur complement preconditioner for the standard Schur complement matrix $S(0) = S(0)^T > 0$:

$$G_B \asymp S(0) = (A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}).$$

Then, the algorithm below requires a careful choice of parameters μ_k and α_k.

Algorithm 16.3.2 *(Preconditioned Gradient Method in a Subspace)*
Input: G_B, α_k, μ_k, $\mathbf{u}_B^{(0)}$

```
1. For $k = 0, 1, \ldots$ until convergence do:
       $\mathbf{u}_B^{(k+1)} = \left(\alpha_k\, I - G_B^{-1} S(\mu_k)\right) \mathbf{u}_B^{(k)}$
2. Endfor
```

For choices of the parameters α_k and μ_k each iteration, see [KN6, KN2].

16.4 Schwarz Subspace Methods

In this section, we describe the sequential and parallel versions of Schwarz subspace methods for eigenvalue problems [MA9, LU5, LU6, CH16]. These methods seek to *minimize* the Rayleigh quotient associated with an eigenvalue problem, using a family of subspaces. Each minimization of a Rayleigh quotient within a subspace yields a lower dimensional eigenvalue problem, and as a result, Schwarz Subspace methods for eigenvalue problems require the solution of lower dimensional eigenvalue subproblems. Since the latter can still be computationally expensive when the number of unknowns in each subregion or subspace is large, multilevel or hierarchical approaches may be employed *recursively*, to introduce subproblems of a smaller size [CH16].

Consider the generalized eigenvalue problem (16.1). We will assume that its generalized eigenvalues are ordered as follows:

$$\lambda_1 < \lambda_2 \le \lambda_3 \le \cdots \le \lambda_n.$$

Let $V_1, \ldots, V_p$ denote subspaces of $V = \mathbb{R}^n$, satisfying:

$$\begin{cases} V_l = \text{Range}\left(R_l^T\right), & \text{for} \quad 1 \le l \le p \\ V = V_1 + \cdots + V_p \end{cases} \tag{16.4}$$

If n_l denotes the dimension of V_l, then each matrix R_l^T must be of size $n \times n_l$, with columns that span V_l. The requirement that the subspaces V_l sum to V imposes the constraint $\sum_{l=1}^p n_l \ge n$.

In Schwarz subspace methods, given an iterate $\mathbf{w} \in \mathbb{R}^n$ and a subspace $\text{Range}\left(R_l^T\right)$, the algorithm computes updates of the form $\alpha_* \mathbf{w} + R_l^T \mathbf{d}_*^{(l)}$ for $\alpha_* \in \mathbb{R}$ and $\mathbf{d}_*^{(l)} \in \mathbb{R}^{n_l}$, where $\alpha_* \mathbf{w} + R_l^T \mathbf{d}_*^{(l)}$ is chosen from the subspace generated by $\text{span}\{\mathbf{w}, R_l^T\}$ to *minimize* the Rayleigh quotient $\mathcal{R}(\cdot)$:

$$\mathcal{R}\left(\alpha_* \mathbf{w} + R_l^T \mathbf{d}_*^{(l)}\right) = \min_{\{\alpha \in \mathbb{R},\, \mathbf{d} \in \mathbb{R}^{n_l}\} \backslash \mathbf{0}} \mathcal{R}\left(\alpha \mathbf{w} + R_l^T \mathbf{d}\right).$$

Each such minimization requires the solution of a generalized eigenproblem involving (n_l+1) unknowns. Given $\mathbf{w}$, define $K_l = \left[\, R_l^T \quad \mathbf{w} \,\right]$ of size $n \times (n_l+1)$, whose columns consist of the n_l columns of R_l^T and $\mathbf{w}$. Then, solve:

$$\left(K_l^T A K_l\right) \begin{bmatrix} \mathbf{d}^{(l)} \\ \alpha \end{bmatrix} = \mu_{\min} \left(K_l^T M K_l\right) \begin{bmatrix} \mathbf{d}^{(l)} \\ \alpha \end{bmatrix}.$$

Provided $(n_l + 1)$ is sufficiently small, this may be solved by the QR method [GO4]. Otherwise, the Lanczos or gradient methods may be employed.

Algorithm 16.4.1 *(Sequential Schwarz Minimization Algorithm)*
Let $\mathbf{w}^{(0)} \in \mathbb{R}^n$ *be a starting iterate satisfying* $\lambda_1 < \mathcal{R}(\mathbf{w}^{(0)}) < \lambda_2$

```
1. For $k = 0, 1, \ldots,$ until convergence do:
2.     For $l = 1, \ldots, p$ determine $\mathbf{w}^{(k+\frac{l}{p})}$:
```

$$\mu^{(k+\frac{l}{p})} = \mathcal{R}(\mathbf{w}^{k+\frac{l}{p}}) = \min_{\{\alpha \in \mathbb{R},\, \mathbf{d}^{(l)} \in \mathbb{R}^{n_l}\} \backslash \mathbf{0}} \mathcal{R}\left(\alpha \mathbf{w}^{(k+\frac{l-1}{p})} + R_l^T \mathbf{d}^{(l)}\right)$$

```
3.     Endfor
4. Endfor
```

We next describe a parallel version of the preceding algorithm.

Algorithm 16.4.2 *(Parallel Schwarz Subspace Algorithm)*
Let $\mathbf{w}^{(0)} \in \mathbb{R}^n$ *be a starting iterate satisfying* $\lambda_1 < \mathcal{R}(\mathbf{u}^{(0)}) < \lambda_2$

```
1. For $k = 0, 1, \ldots$ do:
2.     For $l = 1, \ldots, p$ in parallel minimize:
           $\mathcal{R}\left(\mathbf{w}^{(k,l)}\right) \equiv \min_{\{\alpha \in \mathbb{R},\, \mathbf{d}^{l)} \in \mathbb{R}^{n_l}\} \backslash \mathbf{0}} \mathcal{R}\left(\alpha\, \mathbf{w}^{(k)} + R_l^T \mathbf{d}^{(l)}\right)$
3.     Endfor
4.     Minimize the Rayleigh quotient:
           $\mathcal{R}\left(\mathbf{w}^{(k+1)}\right) = \min_{\{(\alpha_1, \ldots, \alpha_p) \neq \mathbf{0}\}} \mathcal{R}\left(\alpha_1\, \mathbf{w}^{(k,1)} + \cdots + \alpha_p\, \mathbf{w}^{(k,p)}\right)$
5. Endfor
```

Both of the preceding algorithms converge, provided $\lambda_1 < \mathcal{R}(\mathbf{w}^{(0)}) < \lambda_2$.

Lemma 16.4. *Let* $\mathbf{w}^{(0)}$ *satisfy* $\mathcal{R}(\mathbf{w}^{(0)}) < \lambda_2$ *and* $V = V_1 + \cdots + V_p$. *Then,* $\mu^{(k+\frac{l}{N})} \to \lambda_1$ *as* $k \to \infty$.

Proof. See [LU5, LU6]. □

16.5 Modal Synthesis Method

We conclude our discussion of domain decomposition and block methods for eigenvalue problems by describing a *non-iterative* method for constructing global *approximations* of the minimal eigenvectors and associated eigenvalues of a generalized eigenvalue problem. The method, which is referred to as modal synthesis [BO10, BO11, BO12, BO13], has its origins in aeronautical and structural engineering applications. It is based on the decomposition of the domain of an elastic structure, into smaller structures, computing a few of the lowest eigenmodes (eigenvectors) associated with the substructures, and employing additional modes which couple the different structures. A *Rayleigh-Ritz* approximation is then employed to compute the lowest eigenmodes of the global structure based on the subspace of local and interface modes.

More specifically, let $A\,\mathbf{u} = \lambda\, M\,\mathbf{u}$ denote the matrix discretization of an elliptic eigenvalue problem on a domain Ω. Given a *non-overlapping* decomposition $\Omega_1, \ldots, \Omega_p$ of Ω, let $\mathbf{u}_I^{(1)}, \ldots, \mathbf{u}_I^{(p)}$ denote nodal vectors associated with each substructure. Let $\mathbf{u}_B$ denote the nodal vector associated with the interface $B = \cup_{l=1}^{p} (\partial\Omega_l \cap \Omega)$. The modal synthesis method constructs a low dimensional *subspace* $\mathcal{M} \subset \mathbb{R}^n$ with good approximation properties, and computes the Rayleigh-Ritz approximation of λ_1 and $\mathbf{u}_1$ based on $\mathcal{M}$.

Local Modes. On each substructure Ω_l, let $\mathcal{M}_l = \mathrm{span}(K_l^T)$ denote a subspace of low dimension, "based" on minimal eigenvectors of $A_{II}^{(l)}$ on Ω_l. More precisely, if n_l minimal eigenvectors of $A_{II}^{(l)}$ have been determined, then K_l^T will be a matrix of size $n \times n_l$, whose columns are extensions by zero to Ω, of the n_l minimal eigenvectors of $A_{II}^{(l)}$. These are the *local modes.*

Coupling Modes. On the interface B which separates the subdomains, define a space $\mathcal{M}_B$ of *coupling modes* as the subspace of *discrete harmonic* nodal vectors for arbitrary nodal values on B. Thus, if n_0 denotes the number of nodes on B, let I_0 be an identity matrix of size n_0, and define:

$$K_0^T = \begin{bmatrix} -A_{II}^{-1} A_{IB} I_0 \\ I_0 \end{bmatrix}.$$

Thus each column of K_0^T will correspond to a discrete harmonic extension of an elementary nodal vector on B into all the subdomains.

The space $\mathcal{M}$ generated by the local and coupling modes is defined as:

$$\mathcal{M} = \text{Range}\left(K^T\right), \quad \text{where} \quad K^T = \left[\, K_1^T \quad \cdots \quad K_p^T \quad K_0^T \,\right], \tag{16.5}$$

where K^T denotes a matrix of size $n \times (n_1 + \cdots + n_p + n_0)$. The modal synthesis method determines approximate eigenvectors and eigenvalues of (16.1) as the *Ritz vectors* and *Ritz values* based on the subspace $\mathcal{M}$:

$$\left(KAK^T\right) \mathbf{x} = \mu \left(KMK^T\right) \mathbf{x}, \quad \text{where} \quad \mathbf{x} \in \mathbb{R}^{n_1 + \cdots + n_p + n_0}.$$

Given a minimal Ritz vector $\mathbf{x}$ and Ritz value μ, the eigenvector approximation will be $K^T \mathbf{x} \approx \mathbf{u}_1$ corresponding to the approximate eigenvalue $\mu \approx \lambda_1$.

17

Optimization Problems

In this chapter, we describe extensions of the Schwarz subspace methods from Chap. 2 to iteratively solve minimization problems [TA4, TA5]. Such methods correspond to block generalizations of the Gauss-Seidel and Jacobi relaxation methods for minimization problems. In general terms, domain decomposition and multilevel methodology can be applied to minimization problems in two alternative ways. In the first approach, domain decomposition methods can be employed within an *inner* iteration, to solve the quadratic minimization problem occurring during each iteration of a traditional Newton or trust region method. Such an approach requires a *global* quadratic approximation of the underlying functional whose minimum is sought. In the second approach, the divide and conquer Schwarz subspace methodology seeks the global minimum using lower dimensional minimization problems on subspaces. This approach requires only *local* quadratic approximations.

Our discussion will focus on Schwarz subspace algorithms which employ lower dimensional minimization problems. In Chap. 17.1, we describe some background on traditional iterative methods for minimization (with selected theoretical results). In Chap. 17.2, we describe sequential and parallel variants of Schwarz subspace minimization algorithms. For a discussion of applications to nonlinear elliptic equations, see [CA6, TA4, LU7, TA5, LU9].

17.1 Traditional Algorithms

In this section, we describe traditional unconstrained minimization algorithms. We consider the problem of determining the minimum of a sufficiently smooth function $J : V \rightarrow \mathrm{I\!R}$, where $V = \mathrm{I\!R}^n$ is equipped with the Euclidean inner product $(.,.)$ and Euclidean norm $\|\cdot\|$. We shall seek $\mathbf{u}_* \in V$ satisfying:

$$J(\mathbf{u}_*) = \min_{\mathbf{v}\in V} J(\mathbf{v}). \tag{17.1}$$

Traditional unconstrained minimization algorithms include the gradient (or steepest descent) method, Gauss-Seidel and Jacobi relaxation methods, and Newton and trust region methods. Readers are referred to [OR, GI2, DE7, CI4] for detailed studies of such methods. In most applications, we shall assume that $J(\cdot)$ satisfies the properties described below.

Definition 17.1. *A function $J : V \rightarrow \mathrm{I\!R}$ is said to be convex, if for points $\mathbf{u}_1, \ldots, \mathbf{u}_k \in V$ and scalars $0 < \alpha_i < 1$ satisfying $\sum_{i=1}^k \alpha_i = 1$, it holds that:*

$$J\left(\sum_{i=1}^k \alpha_i\, \mathbf{u}_i\right) \le \sum_{i=1}^k \alpha_i\, J\left(\mathbf{u}_i\right).$$

If the inequality is strict when $\mathbf{u}_i \neq \mathbf{u}_j$, $J(\cdot)$ is said to be strictly convex.

Definition 17.2. *A functional $J : V \rightarrow \mathrm{I\!R}$ is said to be elliptic if there exists $\alpha > 0$ such that:*

$$\alpha\, \|\mathbf{u} - \mathbf{v}\|_V^2 \le \left(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v}), \mathbf{u} - \mathbf{v}\right), \quad \forall \mathbf{u}, \mathbf{v} \in V.$$

Definition 17.3. *A functional $J : V \rightarrow \mathrm{I\!R}$ is said to be Lipschitz if there exists $M > 0$ such that:*

$$\|\nabla J(\mathbf{u}) - \nabla J(\mathbf{v})\| \le M \|\mathbf{u} - \mathbf{v}\|_V, \quad \forall \mathbf{u}, \mathbf{v} \in V.$$

Remark 17.4. A convex *elliptic* functional can be shown to be strictly convex, see [CI4]. Such functionals will have a *unique* minimum. Below, we describe various traditional algorithms for the iterative solution of (17.1).

Gradient Method. The gradient (steepest descent) method seeks the minimum in short steps involving *line searches* in the direction of steepest descent. Given a current iterate $\mathbf{u}^{(k)}$, the gradient method either minimizes $J(\cdot)$ along the line $\mathbf{u}^{(k)} - \rho\, \nabla J(\mathbf{u}^{(k)})$ parameterized by $\rho \in \mathrm{I\!R}$, or moves a short step in the direction $-\nabla J(\mathbf{u}^{(k)})$ of steepest descent at $\mathbf{u}^{(k)}$. When the step size ρ_k is appropriately chosen, the gradient method can be robust, ensuring that $\{\mathbf{u}^{(k)}\}$ converges to a critical point of $J(\cdot)$ monotonically:

$$J(\mathbf{u}^{(0)}) \ge J(\mathbf{u}^{(1)}) \ge \cdots \ge J(\mathbf{u}^{(k)}) \ge J(\mathbf{u}^{(k+1)}).$$

The rate of convergence of the gradient method can be slow, and it requires computing $\nabla J(.)$, however, it does not require solving a linear system.

Algorithm 17.1.1 *(Gradient Method)*
Let $\mathbf{u}^{(0)}$ *denote a starting guess*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     Choose $\rho_k > 0$ and update:
           $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \rho_k \nabla J\left(\mathbf{u}^{(k)}\right)$
3. Endfor
```

The following result describes the convergence of the gradient method for a Lipschitz elliptic functional $J(\cdot)$ that is continuously differentiable.

Theorem 17.5. *Let* $J : V \to \mathrm{I\!R}$ *denote a sufficiently smooth convex functional which is elliptic and Lipschitz, with parameters* α *and* M *respectively. If the parameters* ρ_k *satisfy:*

$$0 < \delta_1 \le \rho_k \le \delta_2 < \frac{2\alpha}{M^2},$$

then the gradient iterates $\{\mathbf{u}^{(k)}\}$ *will converge geometrically to* $\mathbf{u}_*$*, with:*

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}_*\|_V \le \beta \, \|\mathbf{u}^{(k)} - \mathbf{u}_*\|_V,$$

where $\beta < 1$, *is a parameter that depends on* δ_1*,* δ_2*,* α *and* M*.*

Proof. We follow [CI4]. By definition, the iterates $\{\mathbf{u}^{(k)}\}$ will satisfy:

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \rho_k \nabla J(\mathbf{u}^{(k)}).$$

Since $\nabla J(\mathbf{u}_*) = \mathbf{0}$ at the minimum $\mathbf{u}_*$, the following will hold trivially:

$$\mathbf{u}_* = \mathbf{u}_* - \rho_k \nabla J(\mathbf{u}_*).$$

Subtracting the two, yields the following equation for the error $\mathbf{u}_* - \mathbf{u}^{(k+1)}$:

$$\left(\mathbf{u}_* - \mathbf{u}^{(k+1)}\right) = \left(\mathbf{u}_* - \mathbf{u}^{(k)}\right) - \rho_k \left(\nabla J(\mathbf{u}_*) - \nabla J(\mathbf{u}^{(k)})\right).$$

Equating the norms of both sides yields:

$$\begin{cases} \quad \|\mathbf{u}_* - \mathbf{u}^{(k+1)}\|^2 \\ = \|\mathbf{u}_* - \mathbf{u}^{(k)}\|_V^2 - 2\rho_k \left(\mathbf{u}_* - \mathbf{u}^{(k)}, \nabla J(\mathbf{u}_*) - \nabla J(\mathbf{u}^{(k)})\right) \\ \quad + \rho_k^2 \, \|\nabla J(\mathbf{u}_*) - \nabla J(\mathbf{u}^{(k)})\|^2 \\ \le \|\mathbf{u}_* - \mathbf{u}^{(k)}\|^2 - 2\rho_k \, \alpha \, \|\mathbf{u}_* - \mathbf{u}^{(k)}\|^2 + \rho_k^2 \|\nabla J(\mathbf{u}_*) - \nabla J(\mathbf{u}^{(k)})\|^2 \\ \le \left(1 - 2\,\alpha\,\rho_k + \rho_k^2 \, M^2\right) \|\mathbf{u}_* - \mathbf{u}^{(k)}\|^2. \end{cases}$$

The choice $\rho_k = \alpha / M^2$ will minimize the estimate for the contraction factor. It can be easily verified that $\left(1 - 2\,\alpha\,\rho_k + \rho_k^2 \, M^2\right) \ge 1$ for $\rho_k \le 0$ and $\rho_k \ge 1$. For $0 < \delta_1 \le \rho_k \le \delta_2 < 2\,\alpha / M^2$ the factor β will be less than 1, and geometric convergence would result. □

Remark 17.6. The choice of parameter ρ_k used in the estimate above is not optimal when $J(\mathbf{u})$ is quadratic with $J(\mathbf{u}) \equiv = \frac{1}{2}\,\mathbf{u}^T A\mathbf{u} - \mathbf{u}^T\mathbf{b}$, for A symmetric positive definite. In this case, the optimal choice of parameter ρ_k will be $\rho_{\text{opt}} = 2/\left(\lambda_{min}(A) + \lambda_{max}(A)\right)$, compared to $\rho_k = (\lambda_{min}(A)/\lambda_{max}(A)^2)$.

In applications to the discretizations of elliptic equations, the parameters α or M can depend on the mesh size h and the coefficients. In this case, the contraction factor β can deteriorate $h \rightarrow 0^+$ unless appropriate preconditioning is used. Below, we describe a "preconditioned" gradient method. We shall employ a matrix $H^T = H > 0$ of size n as the preconditioner.

Algorithm 17.1.2 *(Preconditioned Gradient Method)*
Let $\mathbf{u}^{(0)}$ denote a starting guess

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     Update:
           $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \rho_k\, H^{-1}\, \nabla\, J\left(\mathbf{u}^{(k)}\right)$
3. Endfor
```

Under appropriate assumptions, the preconditioned gradient method will converge geometrically. Below, we employ a norm:

$$(\mathbf{u}, \mathbf{v})_H \;\equiv\; \mathbf{u}^T H\mathbf{v}, \quad \text{with} \quad \|\mathbf{u}\|_H \;\equiv\; \left(\mathbf{u}^T H\mathbf{u}\right)^{1/2},$$

induced by the positive definite symmetric matrix $H^T = H > 0$.

Proposition 17.7. *Let H be a symmetric positive definite matrix, and let $J(\cdot)$ be a continuously differentiable elliptic functional satisfying:*

1. *Ellipticity in the H-induced norm:*

$$\left(H^{-1}(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v})), \mathbf{u} - \mathbf{v}\right)_H \;\geq\; \alpha\; (\mathbf{u} - \mathbf{v}, \mathbf{u} - \mathbf{v})_H\,.$$

2. *Lipschitz in the H-induced norm:*

$$\|H^{-1}\left(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v})\right)\|_H \leq M\|\mathbf{u} - \mathbf{v}\|_H.$$

Then, if the parameters ρ_k are chosen so that $0 < \delta_1 \leq \rho_k \leq \delta_2 \leq \left(2\alpha/M^2\right)$, the gradient iterates $\{\mathbf{u}^{(k)}\}$ will converge geometrically to $\mathbf{u}_*$:

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}_*\|_H \;\leq\; \beta\|\mathbf{u}^k - \mathbf{u}_*\|_H,$$

where $\beta < 1$ depends only on δ_1, δ_2, α and M.

Proof. The proof is the analogous to the proof in the unpreconditioned case (except for the H-induced inner product). □

Gauss-Seidel and Jacobi Relaxation. Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ denote a basis for $V = \mathbb{R}^n$. Each sweep of a point relaxation method will involve parallel or sequential minimizations of $J(\cdot)$ in the direction of the basis vectors $\mathbf{e}_j$. The Jacobi algorithm [CI4, TA4, TA5] employs parameters $\alpha_1, \ldots, \alpha_n$ satisfying $\sum_{l=1}^{n} \alpha_l = 1$ with $0 < \alpha_l < 1$ for $1 \leq l \leq n$.

Algorithm 17.1.3 *(Jacobi Relaxation)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ *be a starting guess*

```
For $k = 0, 1, \ldots$ until convergence do:
    For $i = 1, \ldots, n$ in parallel find $d_i^k \in \mathbb{R}$:
        $J(\mathbf{u}^{(k)} + d_i^k \, \mathbf{e}_i) = \min_{d \in \mathbb{R}} J\left(\mathbf{u}^{(k)} + d \, \mathbf{e}_i\right).$
    End
    Update $\mathbf{u}^{(k+1)} = \sum_{i=1}^{n} \alpha_i \, \left(\mathbf{u}^{(k)} + d_i^k \, \mathbf{e}_i\right) = \mathbf{u}^{(k)} + \sum_{i=1}^{n} \alpha_i \, d_i^k \mathbf{e}_i$
Endfor
```

Remark 17.8. A simple choice of parameters would be $\alpha_l \equiv (1/n)$. Since by construction $J(\mathbf{u}^{(k)} + d_i^k \, \mathbf{e}_i) \leq J(\mathbf{u}^{(k)})$, for $1 \leq i \leq n$, the following will hold, provided $J(\cdot)$ is a convex function:

$$\begin{cases} J\left(\mathbf{u}^{(k+1)}\right) = J\left(\sum_{i=1}^{n} \alpha_i (\mathbf{u}^{(k)} + d_i^k \mathbf{e}_i)\right) \leq \sum_{i=1}^{n} \alpha_i \, J\left(\mathbf{u}_i + d_i^k \mathbf{e}_i\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad \leq \sum_{i=1}^{n} \alpha_i \, J\left(\mathbf{u}^{(k)}\right) = J\left(\mathbf{u}^{(k)}\right). \end{cases}$$

Thus, Jacobi iterates *decrease monotonically.* Below, we describe the Gauss-Seidel relaxation, which *sequentially* minimizes along each direction $\mathbf{e}_j$.

Algorithm 17.1.4 *(Gauss-Seidel Relaxation)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ *be a starting guess*

```
For $k = 0, 1, \ldots$ until convergence do:
    For $i = 1, \ldots, n$ in sequence determine $d_i^k \in \mathbb{R}$:
        $J(\mathbf{u}^{(k+\frac{i-1}{n})} + d_i^k \mathbf{e}_i) = \min_{d_i \in \mathbb{R}} J\left(\mathbf{u}^{(k+\frac{i-1}{n})} + d_i \mathbf{e}_i\right).$
        Define $\mathbf{u}^{(k+\frac{i}{n})} \equiv \mathbf{u}^{(k+\frac{i-1}{n})} + d_i^k \mathbf{e}_i$
    Endfor
Endfor
```

By construction, $J(\mathbf{u}^{(k+\frac{i-1}{n})}) \geq J(\mathbf{u}^{(k+\frac{i}{n})})$ for each k, i, yielding *monotone* iterates. Under suitable assumptions on $J(\cdot)$, both relaxation algorithms will be globally convergent. The Schwarz algorithms that we shall describe in the next section correspond to generalizations of these relaxation algorithms, involving blocks of unknowns.

Newton's Method. Gradient and relaxation methods have the advantage of being globally convergent for Lipschitz and elliptic convex functions $J(.)$.

However, these algorithms have at best a geometric rate of convergence, with $\|\mathbf{u}^{(k+1)} - \mathbf{u}_*\| \leq \beta \|\mathbf{u}^{(k)} - \mathbf{u}_*\|$, for some $\beta < 1$. By comparison, Newton's method, which is based on minimizing a quadratic Taylor approximation of $J(\cdot)$ at $\mathbf{u}^{(k)}$, can converge "quadratically" close to the true minimum, i.e.:

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}_*\| \leq C \|\mathbf{u}^{(k)} - \mathbf{u}_*\|^2,$$

for some constant $C > 0$. This yields more rapid convergence closer to the true minimum, however, Newton's method can diverge if $\mathbf{u}^{(k)}$ is not sufficiently close to $\mathbf{u}_*$. It also requires computing the Hessian matrix and solving a linear system, which can be significant expenses.

In our discussion, we let $\nabla^2 J(\cdot)$ denote the (symmetric) Hessian matrix size n consisting of 2nd order partial derivatives of $J(\cdot)$:

$$\left(\nabla^2 J\right)_{ij}(\mathbf{u}) = \left(\frac{\partial^2 J}{\partial u_i \partial u_j}\right)\Bigg|_{\mathbf{u}}, \quad \text{for} \quad 1 \leq i, j \leq n.$$

At each iterate $\mathbf{u}^{(k)}$, we denote by $H_k = \nabla^2 J(\mathbf{u}^{(k)})$ as the current Hessian. A quadratic Taylor approximation $Q_k(\mathbf{u}) \approx J(\mathbf{u})$ "close" to $\mathbf{u}^{(k)}$ will be:

$$Q_k(\mathbf{u}) \equiv J(\mathbf{u}^{(k)}) + \nabla J(\mathbf{u}^{(k)}) \cdot (\mathbf{u} - \mathbf{u}^{(k)}) + \frac{1}{2}(\mathbf{u} - \mathbf{u}^{(k)})^T H_k (\mathbf{u} - \mathbf{u}^{(k)}).$$

Newton updates are computed as the global minimum of $Q_k(.)$.

Algorithm 17.1.5 *(Newton's Method)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ *denote a starting guess*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     Compute $\nabla J(\mathbf{u}^{(k)})$
3.     Compute $H_k \equiv \nabla^2 J(\mathbf{u}^{(k)})$
4.     Update: $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - H_k^{-1} \nabla J(\mathbf{u}^{(k)})$.
5. Endfor
```

Under suitable assumptions, Newton's method converges quadratically.

Proposition 17.9. *Let* $J : V \to \mathbb{R}$ *satisfy the following.*

1. *Let* K *denote a closed region containing a minimum* $\mathbf{u}_*$ *of* $J(\cdot)$.
2. *Let* $J(\cdot) \in C^3(K)$ *and be elliptic:*

$$\alpha\, \mathbf{w}^T \mathbf{w} \leq \mathbf{w}^T \left(\nabla^2 J(\mathbf{u})\right) \mathbf{w}, \quad \forall \mathbf{u}, \mathbf{w} \in V.$$

3. *Let the Taylor expansion of* $\nabla J(\cdot)$ *satisfy:*

$$\begin{cases} \nabla J(\mathbf{u} + \mathbf{w}) = \nabla J(\mathbf{u}) + \nabla^2 J(\mathbf{u})\mathbf{w} + R(\mathbf{u}, \mathbf{w}), \text{ with} & \\ \|R(\mathbf{u}, \mathbf{w})\| \leq C \|\mathbf{w}\|^2, & \textit{for } \mathbf{u} \in K, \end{cases}$$

for $\mathbf{u}, \mathbf{u} + \mathbf{w} \in K$. *Here* $R(\mathbf{u}, \mathbf{w})$ *denotes the Taylor series third order remainder terms evaluated at a point* $\theta\, \mathbf{u} + (1 - \theta)\, \mathbf{w}$ *for some* $\theta \in (0, 1)$.

Then, if $\mathbf{u}^{(0)} \in K$ *with* $\|\mathbf{u}^{(0)} - \mathbf{u}_*\| \leq \epsilon$ *sufficiently small, all subsequent Newton iterates* $\mathbf{u}^{(k)}$ *will remain in* K *with* $\|\mathbf{u}^{(k)} - \mathbf{u}_*\| \leq \epsilon$ *and satisfying:*

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}_*\| \, \leq \, C \, \|\mathbf{u}^{(k)} - \mathbf{u}_*\|^2.$$

Proof. We follow [CI4, DE7]. The Newton iterates satisfy:

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - H_k^{-1} \, \nabla J(\mathbf{u}^{(k)}),$$

At the minimum, since $\nabla J(\mathbf{u}_*) = \mathbf{0}$ it will trivially hold that:

$$\mathbf{u}_* = \mathbf{u}_* - H_k^{-1} \nabla J(\mathbf{u}_*),$$

Subtracting these two equations yields:

$$(\mathbf{u}_* - \mathbf{u}^{k+1}) = (\mathbf{u}_* - \mathbf{u}^k) - H_k^{-1} \left(\nabla J(\mathbf{u}_*) - \nabla J(\mathbf{u}^{(k)}) \right).$$

Expanding $\nabla J(\mathbf{u}_*) = \nabla J(\mathbf{u}^{(k)}) + H_k \left(\mathbf{u}_* - \mathbf{u}^{(k)} \right) + R(\mathbf{u}^{(k)}, \mathbf{u}_* - \mathbf{u}^{(k)})$ yields:

$$\begin{cases} \left(\mathbf{u}_* - \mathbf{u}^{(k+1)} \right) = \left(\mathbf{u}_* - \mathbf{u}^{(k)} \right) - H_k^{-1} \left(H_k \mathbf{u}_* - H_k \mathbf{u}^{(k)} + R(\mathbf{u}^{(k)}, \mathbf{u}_* - \mathbf{u}_k) \right) \\ \qquad\qquad\qquad\quad = -H_k^{-1} R(\mathbf{u}^{(k)}, \mathbf{u}_* - \mathbf{u}^{(k)}). \end{cases}$$

Using bounds for the remainder term, yields quadratic convergence:

$$\|\mathbf{u}_* - \mathbf{u}^{(k+1)}\| \leq C \|H_k^{-1}\| \|\mathbf{u}_* - \mathbf{u}^{(k)}\|^2.$$

Note that if $0 < \epsilon < 1$ is chosen so that $C\|H_k^{-1}\| \|\mathbf{u}_* - \mathbf{u}^{(k)}\| \leq \epsilon \ll 1$, then the above iteration will be a contraction, and the iterates will stay in a region in which the preceding bounds hold. □

Trust Region Method. The trust region method attempts to combine the stability of the gradient method with the rapid convergence of Newton's method [DE7]. Given a current iterate $\mathbf{u}^{(k)}$, we assume that the following quadratic Taylor approximation $Q_k(\mathbf{u})$ of $J(\mathbf{u})$ is a "good" approximation of $J(\mathbf{u})$ in the disk D_k:

$$\begin{cases} Q_k(\mathbf{u}) = J(\mathbf{u}^{(k)}) + \nabla J(\mathbf{u}^k) \cdot (\mathbf{u} - \mathbf{u}^{(k)}) + \frac{1}{2} (\mathbf{u} - \mathbf{u}_k)^T H_k (\mathbf{u} - \mathbf{u}^{(k)}) \\ \quad D_k = \left\{ \mathbf{u} : \|\mathbf{u} - \mathbf{u}^{(k)}\| \leq \delta \right\} \end{cases} \tag{17.2}$$

where $H_k = \nabla^2 J(\mathbf{u}^{(k)})$, and the user chooses the radius δ of the disk. Each such region D_k is called a *trust region*. Given $\mathbf{u}^{(k)}$ and the disk D_k, the trust region method defines the new iterate $\mathbf{u}^{(k+1)}$ as:

$$Q_k(\mathbf{u}^{(k+1)}) = \min_{\{\mathbf{v} \in D_k\}} Q_k(\mathbf{v}).$$

Thus, $\mathbf{u}^{(k+1)}$ is the minimum of $Q_k(\mathbf{u})$ within the trust region D_k.

To computationally determine the constrained minimum $\mathbf{u}^{(k+1)}$ of $Q_k(.)$ within D_k, we first determine the *global* critical point $\tilde{\mathbf{u}}^{(k+1)}$ of $Q_k(.)$:

$$\tilde{\mathbf{u}}^{(k+1)} \equiv \mathbf{u}^{(k)} - H_k^{-1} \nabla J(\mathbf{u}^{(k)}).$$

If $\tilde{\mathbf{u}}^{(k+1)}$ lies within D_k, then the update will be $\mathbf{u}^{(k+1)} \equiv \tilde{\mathbf{u}}^{(k+1)}$. However, if $\|\tilde{\mathbf{u}}^{(k+1)} - \mathbf{u}^{(k)}\| > \delta$, then the minimum of $Q_k(.)$ within D_k must lie on ∂D_k. Since $\partial D_k = \left\{\mathbf{u} : \|\mathbf{u} - \mathbf{u}^{(k)}\|^2 = \delta^2\right\}$, we may seek the minimum of $Q_k(.)$ on ∂D_k using *constrained* minimization. It will be the minimization of the quadratic functional $Q_k(.)$ subject to the constraints $\|\mathbf{u} - \mathbf{u}^{(k)}\|^2 = \delta^2$. Let λ denote the Lagrange multiplier variable and define the Lagrangian functional:

$$\mathcal{L}(\mathbf{u}, \lambda) \equiv Q_k(\mathbf{u}) + \frac{\lambda}{2} \left((\mathbf{u} - \mathbf{u}^{(k)})^T (\mathbf{u} - \mathbf{u}^{(k)}) - \delta^2 \right),$$

see Chap. 10. Seeking the saddle point of $\mathcal{L}(.,.)$ yields the equations:

$$\begin{cases} \nabla J(\mathbf{u}^{(k)}) + H_k(\mathbf{u} - \mathbf{u}^{(k)}) + \lambda\,(\mathbf{u} - \mathbf{u}^{(k)}) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad \|\mathbf{u} - \mathbf{u}^{(k)}\|^2 = \delta^2. \end{cases}$$

This yields the equations:

$$\begin{cases} (H_k + \lambda\, I)\,(\mathbf{u} - \mathbf{u}^{(k)}) = -\nabla J(\mathbf{u}^{(k)}) \\ \qquad\qquad \|\mathbf{u} - \mathbf{u}^{(k)}\|^2 = \delta^2. \end{cases}$$

Thus, λ may be chosen by requiring the solution of the first block row above, parameterized by λ, satisfies the constraint given in the second block row above. We outline the resulting algorithm below, see [DE7].

Algorithm 17.1.6 *(Trust Region Method)*
Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ *denote a starting guess*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     Compute $\tilde{\mathbf{u}}^{(k+1)} = \mathbf{u}^{(k)} - H_k^{-1} \nabla J(\mathbf{u}^{(k)})$.
3.     If $\|\tilde{\mathbf{u}}^{(k+1)} - \mathbf{u}^{(k)}\| \leq \delta$ define $\mathbf{u}^{(k+1)} \equiv \tilde{\mathbf{u}}^{(k+1)}$
4.     else determine $\lambda_k > 0$ such that:
           $\delta = \|\,(\lambda_k I + H_k)^{-1} \nabla J(\mathbf{u}^{(k)})\|$, and define:
           $\mathbf{u}^{(k+1)} \equiv \mathbf{u}^{(k)} - (\lambda_k I + H_k)^{-1} \nabla J(\mathbf{u}^{(k)})$
5.     Endif
6. Endfor
```

The diameter δ of the trust region D_k can be chosen adaptively as $\delta = \delta_k$ provided $\lambda_k \geq 0$ is chosen such that $\|\,(\lambda_k I + H_k)^{-1}\, \nabla J(\mathbf{u}^{(k)})\| = \delta_k$, using a bisection method [DE7]. As $\lambda_k \to 0^+$, the iteration reduces to Newton iteration, while as $\lambda_k \to \infty$, the iteration reduces to gradient iteration.

17.2 Schwarz Minimization Algorithms

Schwarz subspace minimization algorithms [TA4, TA5, CH16] formally correspond to generalized block Gauss-Seidel and Jacobi minimization algorithms. These algorithms are based on the solution of various minimization problems on subspaces, corresponding to blocks of unknowns. Below, we describe the algebraic version of such algorithms for the iterative solution of (17.1).

Accordingly, let $V_l \subset V$ denote subspaces of $V = \mathbb{R}^n$ satisfying:

$$V = V_1 + \cdots + V_p.$$

We shall let n_l denote the dimension of V_l and assume that $V_l = \text{Range}(R_l^T)$, where each R_l^T is an $n \times n_l$ matrix whose columns form a basis for V_l. The parallel Schwarz algorithm employs parameters α_l for $1 \le l \le p$ satisfying:

$$\sum_{l=1}^{p} \alpha_l = 1 \quad \text{with} \quad 0 < \alpha_l < 1 \quad \text{for } 1 \le l \le p.$$

A default choice would be $\alpha_l \equiv \frac{1}{p}$ for $1 \le l \le p$. We summarize the algorithm.

Algorithm 17.2.1 *(Parallel Schwarz Minimization Algorithm)*
Let $\mathbf{u}^{(0)}$ denote a starting guess

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $l = 1, \ldots, p$ in parallel determine $\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}$:
           $J\left(\mathbf{u}^{(k)} + R_l^T \mathbf{d}_*^{(l)}\right) = \min_{\{\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}\}} J\left(\mathbf{u}^{(k)} + R_l^T \mathbf{d}^{(l)}\right).$
3.     Endfor
4.     Define $\mathbf{u}^{(k+1)} = \sum_{l=1}^{p} \alpha_l \left(\mathbf{u}^{(k)} + R_l^T \mathbf{d}_*^{(l)}\right) = \mathbf{u}^{(k)} + \sum_{l=1}^{p} \alpha_l R_l^T \mathbf{d}_*^{(l)}$
5. Endfor
```

Remark 17.10. If $J(\cdot)$ is *convex*, the parallel Schwarz iterates will satisfy:

$$\begin{cases} J\left(\mathbf{u}^{(k+1)}\right) = J\left(\sum_l \alpha_l (\mathbf{u}^{(k)} + R_l^T \mathbf{d}_*^{(l)})\right) \le \sum_l \alpha_l \, J\left(\mathbf{u}^{(k)} + R_l^T \mathbf{d}_*^{(l)}\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \le \sum_l \alpha_l \, J\left(\mathbf{u}^{(k)}\right) = J\left(\mathbf{u}^{(k)}\right). \end{cases}$$

Thus, $J(\mathbf{u}^{(0)}) \ge J(\mathbf{u}^{(1)}) \ge \cdots \ge J(\mathbf{u}^{(k)})$ decreases monotonically.

We next describe a sequential version of the same algorithm.

Algorithm 17.2.2 *(Sequential Schwarz Subspace Minimization Algorithm)* *Let* $\mathbf{u}^{(0)}$ *denote a staring guess*

```
1. For $k = 0, 1, \ldots$ until convergence do:
2.     For $l = 1, \ldots, p$ determine $\mathbf{d}^{(l)}$:
           $$J(\mathbf{u}^{(k+\frac{l-1}{p})} + R_l^T \mathbf{d}_*^{(l)}) = \min_{\mathbf{d}^{(l)} \in \mathbb{R}^{n_l}} J\left(\mathbf{u}^{(k+\frac{l-1}{p})} + R_l^T \mathbf{d}^{(l)}\right).$$
3.         Define $\mathbf{u}^{(k+\frac{l}{p})} \equiv \mathbf{u}^{(k+\frac{l-1}{p})} + R_l^T \mathbf{d}_*^{(l)}$
4.     Endfor
5. Endfor
```

Remark 17.11. We may solve the local minimization problems in the Schwarz algorithms using point relaxation, gradient or trust region methods, in an *inner* iteration. For the trust region and Newton methods, we need to compute submatrices of the Hessian matrix, without a global linearization. The inner iteration can be solved up to some chosen local tolerance.

The following result concerns the convergence of Schwarz algorithms.

Proposition 17.12. *Suppose the following assumptions hold.*

1. *Let* $J : V \to \mathbb{R}$ *be twice continuously differentiable and elliptic:*
$$(\nabla J(\mathbf{u}) - \nabla J(\mathbf{v}), \mathbf{u} - \mathbf{v}) \geq \alpha \, (\mathbf{u} - \mathbf{v}, \mathbf{u} - \mathbf{v}), \quad \forall \mathbf{u}, \mathbf{v} \in V,$$
for some $\alpha > 0$.
2. *Let* ∇J *be Lipschitz continuous:*
$$\|\nabla J(\mathbf{u}) - \nabla J(\mathbf{v})\| \leq M \, \|\mathbf{u} - \mathbf{v}\|, \quad \forall \mathbf{u}, \mathbf{v} \in V,$$
for some $M > 0$.
3. *Let* V_i *be subspaces of* $V = \mathbb{R}^n$ *satisfying:*
$$V_1 + \cdots + V_m = V.$$
4. *The local minimization problems are solved exactly.*

Then, the iterates of the Schwarz minimization algorithms will converge to the unique minimum $\mathbf{u}_*$ *of* $J(\cdot)$. □

Proof. See [TA4, TA5]. □

Remark 17.13. Under additional assumptions, geometric convergence:

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}_*\|_{H_*} \leq \rho \, \|\mathbf{u}^{(k)} - \mathbf{u}_*\|_{H_*},$$

is also proved in [TA4, TA5], for some $0 < \beta < 1$, where $H_* = \nabla^2 J(\mathbf{u}_*)$.

18

Helmholtz Scattering Problem

In this chapter, we describe several domain decomposition methods and also outline a shooting method for determining the solution to the reduced wave equation in wave scattering. The Helmholtz wave scattering problem is the mathematical problem which describes the "scattered wave" when an incident electromagnetic or acoustic wave impinges on an object D and is scattered by it [KE, BL, KE2, KE3, BL2]. Mathematically, this leads to the problem of computing a "standing wave" solution of the wave equation (i.e., a time periodic solution of the special form $e^{ikt}v(x)$) in a domain Ω exterior to the object D, with appropriate boundary conditions. This yields the following "reduced wave equation" for $v(x)$:

$$\begin{cases} -\Delta\, v - \kappa^2\, n^2(x)\, v = 0, & \text{in } \Omega \setminus D, \\ v = g, & \text{on } \partial D, \\ \dfrac{\partial v}{\partial n} - i\kappa v = 0, & \text{on } \partial\Omega. \end{cases} \tag{18.1}$$

Here, $\Omega \supset D$ is a computational domain, on whose boundary the Bohr-Sommerfeld radiation boundary condition is applied. A discretization of the stationary problem (18.1), by finite difference or finite element methods, yields a sparse, non-Hermitian but complex symmetric linear system.

Our discussion in this chapter will focus only on iterative methods for solving the complex symmetric, but non-Hermitian, linear system arising from the discretization of the reduced wave equation. In Chap. 18.1, we discuss background on the reduced wave equation. Chap. 18.2 describes variants of non-overlapping and overlapping domain decomposition iterative methods for the reduced wave equation. Chap. 18.3 outlines an iterative method based on fictitious domain or domain imbedding control formulations. We conclude our discussion in Chap. 18.4 by outlining a control formulation based shooting method for determining the standing wave solution to a wave equation. In some sections, we shall formulate some algorithms in their continuous form, omitting matrix implementation. For a discussion of applications to Maxwell's equations, readers are referred to [TO10].

18.1 Background

Let $D \subset \mathbb{R}^d$ denote an obstacle with boundary ∂D. Consider an incident plane wave $Ae^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$ impinging on the obstacle D and resulting in a scattered wave. The sum (superposition) of the two waves (incident and scattered waves) should be a solution of the wave equation in the exterior $\mathbb{R}^d \setminus D$:

$$u_{tt} - c^2(x)\,\Delta u = 0, \qquad \text{in } (\mathbb{R}^d \setminus D) \times (0,T).$$

Here $c(x)$ denotes the local speed of the wave in the given medium. If we seek a scattered standing wave solution of the form $v(x)e^{-i\omega t}$, then substituting:

$$u(x,t) = Ae^{i(\mathbf{k}\cdot\mathbf{x}-\omega\, t)} + v(x)\, e^{-i\omega t},$$

into the wave equation yields the following reduced wave equation for $v(x)$:

$$\begin{cases} -\Delta v - \kappa^2\, n^2(x)\, v = 0, & \text{in } R^n \setminus D \\ v = -Ae^{i\mathbf{k}\cdot\mathbf{x}}, & \text{on } \partial D, \\ \lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\frac{\partial v}{\partial r} - i\kappa v \right) = 0, & \text{on } B_r(0), \end{cases} \tag{18.2}$$

where $\kappa\, n(x) = (\omega/c(x))$. The latter boundary condition, which is called the Bohr-Sommerfeld radiation condition, is required to ensure the uniqueness of the solution on the unbounded domain $(\mathbb{R}^d \setminus D)$, see [BL]. If a computational domain Ω with $D \subset \Omega \subset \mathbb{R}^d$ is employed, then the radiation boundary condition can be applied on $\partial\Omega$, as an approximation.

A classical method for constructing approximate solutions of the reduced wave equation is the *ray method* [KE, BL]. Ray methods have their origins in Hamilton's formulation of the laws of optics in terms of partial differential equations (based on Fermat's *principle of least time* for the path taken by a ray in a refractive medium). This establishes a connection between geometrical optics and the reduced wave equation. Mathematically, ray methods are asymptotic techniques valid for large κ, see [KE, BL]. Consider the reduced wave equation in the following form for large κ:

$$-\Delta v - \kappa^2 n^2(x) v = 0, \quad \mathbb{R}^d \setminus D,$$

where $n^2(x) = 1/c^2(x)$. Ray methods employ an asymptotic expansion:

$$v(x,\kappa) \asymp e^{-\kappa S(x)} \sum_{j=0}^{\infty} \frac{1}{(i\kappa)^j} A_j(x), \qquad \text{as} \quad \kappa \to \infty,$$

where $S(x)$ denotes an *unknown* phase function, and $\{A_j(x)\}_{j=0}^{\infty}$ denote *unknown* amplitude functions. Formally substituting this expansion into the Helmholtz equation and equating powers of κ (for κ assumed to be large), yields a nonlinear 1st order hyperbolic partial differential equation (called the *Eiconal* equation) for $S(x)$:

$$|\nabla S(x)|^2 = n^2(x).$$

Linear transport equations determine the amplitudes $A_j(x)$ along the rays (characteristic curves) of the Eiconal equation:

$$\begin{cases} 2\nabla S \cdot \nabla A_0 + A_0 \Delta S = 0, & \text{along rays,} \\ 2\nabla S \cdot \nabla A_n + A_n \Delta S = -\Delta A_{n-1},\, n \geq 1, & \text{along rays.} \end{cases}$$

Numerical solution of the ray equations can be obtained by time stepping the transport equations and the ordinary differential equations determining the rays (characteristic curves). For details and boundary conditions, see [BL].

18.2 Non-Overlapping and Overlapping Subdomain Methods

Consider a finite element discretization of (18.1) for $n(x) = 1$. Let $\mathbf{v}_I$ denote nodal unknowns in the *interior* of $(\Omega \setminus D)$, let $\mathbf{v}_B$ denote nodal unknowns on $\partial\Omega$ and $\mathbf{v}_D$ nodal unknowns on ∂D. Then, this discretization will be:

$$\begin{bmatrix} A_{II} - \kappa^2 M_{II} & A_{IB} - \kappa^2 M_{IB} \\ A_{IB}^T - \kappa^2 M_{IB}^T & -i\,\kappa G_{BB} + A_{BB} - \kappa^2 M_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I - A_{ID}\mathbf{g}_{\partial D} \\ \mathbf{f}_B \end{bmatrix}, \tag{18.3}$$

where $\mathbf{g}_{\partial D}$ denotes the discrete Dirichlet data on ∂D, matrices A_{XY} and M_{XY} are blocks of the stiffness and mass matrices, while G_{BB} is a boundary mass matrix on $\partial\Omega$. This is a complex symmetric linear system. Due to the indefinite blocks $(A_{XYI} - \kappa^2\, M_{XY})$, the eigenvalues of the above matrix will generally have positive and negative real parts, depending on κ. Complex preconditioned Krylov space methods such as GMREZ can be employed [SA2]. However, constructing preconditioners which are robust with respect to large κ and small h is challenging. Below, we outline two iterative solvers, one based on non-overlapping subdomains and employed without acceleration [DE8] and another based on overlapping subdomains [CA5].

Non-Overlapping Subdomains Solver. Formally, given a non-overlapping decomposition of Ω, any traditional Schur complement preconditioner may be employed to precondition (18.3). However, the convergence rate will deteriorate for large κ, and the preconditioners will be indefinite. Here, we shall describe an unaccelerated method of [DE8] in continuous form. Let $\Omega_1, \ldots, \Omega_p$ denote a nonoverlapping decomposition of $(\Omega \setminus D)$. We shall use $L\,u = -\Delta\, u - \kappa^2\, u$ to denote the Helmholtz operator. Consider the problem:

$$\begin{cases} L\,v = f, & \text{in } \Omega \setminus D, \\ v = g, & \text{on } \partial D, \\ \dfrac{\partial v}{\partial n} - i\,\kappa\, v = 0, & \text{on } \partial\Omega. \end{cases}$$

and let $v_l(x) = v(x)$ on Ω_l. The algorithm of [DE8] updates approximations $v_l^{(k)}(x)$ of $v_l(x)$ by matching mixed boundary conditions with the approximate solutions on adjacent subdomains, as in Robin-Robin algorithms.

Algorithm 18.2.1 *(Nonoverlapping Algorithm of [DE8])*
Let $(v_l^{(0)})_{l=1}^{p}$ *denote starting iterates.*

```
1. For $k = 0, 1 \ldots,$ until convergence do:
2.     Solve in parallel ($l = 1, \ldots, p$):
           $\begin{cases} L\, v_l^{(k+1)} = f, & \text{in } \Omega_l \\ v_l^{(k+1)} = g, & \text{on } \partial D \cap \partial \Omega_l \\ \frac{\partial v_l^{(k+1)}}{\partial n_{lj}} - i\,\kappa v_l^{(k+1)} = \frac{\partial v_j^{(k)}}{\partial n_{lj}} - i\,\kappa v_j^{(k)}, & \text{on } \partial \Omega_l \cap \partial \Omega_j \\ \frac{\partial v_l^{(k+1)}}{\partial n_l} - i\,\kappa v_l^{(k+1)} = 0, & \text{on } \partial \Omega. \end{cases}$
3. Endfor
```

Remark 18.1. Here $\frac{\partial v}{\partial n_{lj}}$ denotes the directional derivative along the exterior unit normal $\mathbf{n}_{lj}$ to Ω_l on $\partial \Omega_l \cap \partial \Omega_j$. The algorithm is similar to [LI8, DO13].

Remark 18.2. As $\kappa \to \infty$, the mixed boundary conditions:

$$\frac{\partial v_l^{(k+1)}}{\partial n_{lj}} - i\,\kappa v_l^{(k+1)} = \frac{\partial v_j^{(k)}}{\partial n_{lj}} - i\,\kappa v_j^{(k)}, \quad \text{on} \quad \partial \Omega_l \cap \partial \Omega_j,$$

reduces to Dirichlet boundary conditions. Since Dirichlet data will be transferred between adjacent subdomains at the end of each iteration, we obtain:

$$v_l^{(k+1)} = v_j^{(k)} = v_l^{(k-1)} \quad \text{on} \quad \partial \Omega_l \cap \partial \Omega_j.$$

Consequently,the boundary values of the iterates will oscillate, and repeat with period *two*. Consequently, as $\kappa \to \infty$, this algorithm will not be convergent (except if the initial iterate is the solution). For large κ, its convergence rate can be expected to deteriorate. Similarly, if $\kappa \to 0$, then in the limit, Neumann boundary data will be exchanged between neighboring subdomains:

$$\frac{\partial v_l^{(k+1)}}{\partial n_{lj}} = \frac{\partial v_j^{(k)}}{\partial n_{lj}} = \frac{\partial v_l^{(k-1)}}{\partial n_{lj}} \quad \text{on} \quad \partial \Omega_l \cap \partial \Omega_j.$$

Again, the Neumann boundary conditions will repeat every alternate iteration, and the algorithm will not be convergent (except if the initial iterate is the solution). Thus, the convergence rate will deteriorate as $\kappa \to 0$.

Remark 18.3. The above algorithm requires computing $\frac{\partial v}{\partial n_{lj}}$ on $\partial \Omega_l \cap \partial \Omega_j$. This can be computed using the weak form, for two subdomain interfaces. However, if many subdomains are involved, then the continuous version of the algorithm becomes *ill-defined* at the cross-points (vertices common to three or more subdomain boundaries). However, an alternative constraint can be imposed (for the discretized problem), see [DE8].

The following convergence result is proved in [DE8].

Proposition 18.4. *Let the following assumptions hold.*

1. *Let $L = -\Delta - \kappa^2$ be formally invertible for the given κ.*
2. *Let $e_l^{(k)} \equiv v - v_l^{(k)}$ denote the error in the computed solution on Ω_l.*

Then, the energy on $\cup_l^p \partial\Omega_l$ will converge to zero:

$$\sum_{l=1}^{p} \int_{\partial\Omega_l} |e^{(k)}|^2 ds \rightarrow 0, \qquad \text{as } k \rightarrow \infty.$$

Proof. See [DE8]. □

Remark 18.5. The above result shows strong convergence of the iterates, in the $L^2(\cup_{l=1}^p \partial\Omega_l)$ norm on the boundary. Unfortunately, the result does not yield geometric convergence, and the rate will depend on the parameter κ.

Overlapping Subdomain Preconditioner. We next consider the matrix form of a restricted Schwarz type preconditioner of [CA5] for system (18.3). Let $H\,\mathbf{u} = \mathbf{f}$ denote this system, and let $\Omega_1^*, \cdots, \Omega_p^*$ form an overlapping decomposition of $(\Omega \setminus D)$. We shall employ the notation:

- Let R_0 denote a matrix whose columns form a basis for a coarse space, and let H_0 denote the following matrix:

$$H_0 \equiv R_0^T H R_0.$$

- Let $\Phi_1, \ldots, \Phi_p$ be diagonal matrices $\Phi_l \geq 0$ that form a discrete partition of the identity $\Phi_1 + \cdots + \Phi_p = I$ of size n, subordinate to $\Omega_1^*, \cdots, \Omega_p^*$.
- Let $R_{I,l}$ denote the standard restriction matrix onto *interior* nodes in Ω_l^* and let $R_{B,l}$ denote the restriction matrix onto nodes on $\partial\Omega_l^* \setminus \partial D$.
- For $1 \leq l \leq p$, define the following subdomain matrices:

$$H^{(l)} \equiv \begin{bmatrix} A_{II}^{(l)} - \kappa^2 M_{II}^{(l)} & A_{IB}^{(l)} - \kappa^2 M_{IB}^{(l)} \\ A_{IB}^{(l)^T} - \kappa^2 M_{IB}^{(l)^T} & -i\,\kappa\, G_{BB}^{(l)} + A_{BB}^{(l)} - \kappa^2\, M_{BB}^{(l)} \end{bmatrix}.$$

Then, the preconditioner $\tilde{H}$ of [CA5] is described next.

For $0 < \theta < 1$, the action $\tilde{H}^{-1}$ of the inverse of the preconditioner $\tilde{H}$ is:

$$\tilde{H}^{-1}\,\mathbf{f} \equiv \theta \sum_{l=1}^{p} \Phi_l \begin{bmatrix} R_{I,l} \\ R_{B,l} \end{bmatrix} H^{(l)^{-1}} \begin{bmatrix} R_{I,l}^T\,\mathbf{f} \\ \mathbf{0} \end{bmatrix} + (1-\theta)\, R_0\, H_0^{-1} R_0^T\, \mathbf{f}.$$

The preceding preconditioner differs from an additive Schwarz preconditioner in two ways. First, the local subproblems employ radiation boundary conditions except at the nodes on ∂D. Second, a partition of identity sum is used, as in a restricted additive Schwarz preconditioner. No rigorous convergence estimates have been established for the above preconditioner [CA5].

18.3 Fictitious Domain and Control Formulations

Least squares-control [AT, NE7, DI2, GL12] and Lagrange multiplier methods, see [GL4, GL11], can be formulated on fictitious domains to provide iterative algorithms for the solution of (18.3). The primary *requirement* is that a fast solver, such as FFT based, be available for the discretized Helmholtz equation on a fictitious rectangular domain Ω_*, with Dirichlet, Neumann, periodic or radiation boundary conditions. In this section, we briefly outline, using continuous formulations, one class of fictitious domain or domain imbedding methods for (18.3). Other control formulations are also possible, extending methods from Chap. 13, see [AT, NE7, DI2, GL12, GL4, GL11].

The use of fictitious domain methodology is justified for the Helmholtz problem when there is an efficient solver available for the discretization of:

$$\begin{cases} -\Delta w - \kappa^2 n^2(x) w = \tilde{f}(x), & \text{in } \Omega_* \\ \qquad\qquad\qquad w = g, & \text{on } \partial\Omega_* \end{cases} \tag{18.4}$$

on some rectangular domain $\Omega_* \supset \Omega$. The Dirichlet conditions may be replaced by Neumann, periodic or radiation conditions on $\partial\Omega_*$. To obtain a solution of (18.1) on Ω, by employing the solution of (18.4) on Ω_*, define a squares norm error term $J(.)$ which is to be minimized:

$$J(w) \equiv \frac{1}{2}\,\|w - g\|^2_{0,\partial D} + \frac{1}{2}\,\|\frac{\partial w}{\partial n} - i\kappa w\|^2_{0,\partial\Omega}. \tag{18.5}$$

The control formulation will seek to minimize $J(\cdot)$ subject to constraints of the form (18.4). To illustrate this in matrix terms, for simplicity, we shall assume that $\Omega_* = \Omega$, and assume that a fast solver is available on the entire domain Ω_* with radiation boundary conditions on $\partial\Omega_*$. When $\Omega_* = \Omega$ with the original radiation conditions, the term $\frac{1}{2}\,\|\frac{\partial w}{\partial n} - i\kappa w\|^2_{0,\partial\Omega}$ may be omitted. In the more general case, the method may still be employed, with due modifications.

Formally, the forcing term $\tilde{f}(x)$ in (18.4) should be of the form:

$$\tilde{f}(x) = \begin{cases} 0, & \text{in } D \\ \gamma_D, & \text{on } \partial D \\ f(x), & \text{in } \Omega \setminus D \\ \gamma_*, & \text{on } \partial\Omega. \end{cases}$$

Such a forcing term $\tilde{f}(x)$ will be "formal" only because in a strict sense, if $\tilde{f}(\cdot) \in L^2(\Omega_*)$, then its restrictions γ_D and γ_* on ∂D and $\partial\Omega_*$ will have zero mass. We shall require $\tilde{f}(\cdot) \notin L^2(\Omega)$. This will not be an issue in the discretized problem, since the forcing terms will be nodal values at grid points. The control formulation to solve (18.1) will seek nontrivial controls γ_D and γ_* such that the solution to (18.4) with the above forcing, formally minimizes $J(w)$. When $J(w) = 0$, the desired solution will be obtained.

Suppose $\Omega_* = \Omega$ and that a fast solver is available on Ω_* with radiation boundary conditions on $\partial\Omega_*$. In this case, define subdomains $\Omega_1 = D$ and $\Omega_2 = (\Omega_* \setminus D)$. Block partition the unknowns into $\mathbf{w}_I^{(1)}$, $\mathbf{w}_I^{(2)}$, $\mathbf{w}_D$ and $\mathbf{w}_B$ where each nodal vector corresponds to unknowns in the interior of Ω_1, interior of Ω_2, on ∂D and on $\partial\Omega$, respectively. Then, the block matrix representation of a discretization of (18.4) will have the following form:

$$\begin{bmatrix} L_{II}^{(1)} & 0 & L_{ID}^{(1)} & 0 \\ 0 & L_{II}^{(2)} & L_{ID}^{(2)} & L_{IB}^{(2)} \\ L_{ID}^{(1)^T} & L_{ID}^{(2)^T} & L_{DD} & 0 \\ 0 & L_{IB}^{(2)^T} & 0 & L_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^{(1)} \\ \mathbf{w}_I^{(2)} \\ \mathbf{w}_D \\ \mathbf{w}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I^{(1)} \\ \mathbf{f}_I^{(2)} \\ \mathbf{f}_D \\ \mathbf{f}_B \end{bmatrix}. \tag{18.6}$$

Here $L_{XY}^{(l)} = (A_{XY}^{(l)} - \kappa^2 M_{XY}^{(l)})$, for $X, Y = I, D$, while $L_{IB}^{(l)} = (A_{IB}^{(l)} - \kappa^2 M_{IB}^{(l)})$ and $L_{BB}^{(l)} = (A_{BB}^{(l)} - \kappa^2 M_{BB}^{(l)} - i\,\kappa\, G_{BB}^{(l)})$. Here $A_{XY}^{(l)}$ and $M_{XY}^{(l)}$ denote block submatrices of the stiffness matrix associated with $-\Delta$ and the mass matrix on Ω_*, respectively. Matrix G_{BB} denotes the lower dimensional boundary mass matrix. We may formally choose the forcing term $\mathbf{f}_I^{(1)} = \mathbf{0}$ in the interior of the obstacle, and $\mathbf{f}_B = \mathbf{0}$ on $\partial\Omega_* = \partial\Omega$. The forcing term $\mathbf{f}_I^{(2)}$ is given, and $\mathbf{f}_D$ can be regarded as an unknown *control* vector. Define the discrete functional:

$$J(\mathbf{w}) = \frac{1}{2}\,\|\mathbf{w}_D - \mathbf{g}_D\|^2, \tag{18.7}$$

where $\mathbf{g}_D$ denotes the discretized Dirichlet data on ∂D. The least squares-control formulation will then seek the control $\mathbf{f}_D$ so that the solution to (18.6) minimizes the function $J(\cdot)$ in (18.7). The methodology will be analogous to that described in Chap. 13. We omit further details.

Remark 18.6. Preconditioners have been proposed in [ER2] for obtaining fast iterative solvers for the Helmholtz problem with radiation boundary conditions on rectangular domains. Fast Helmholtz solvers will also be available for the Dirichlet or periodic problem on Ω_*.

18.4 Hilbert Uniqueness Method for Standing Waves

In this section, we describe a *control* theory based method [BR37] to solve the Helmholtz scattering problem (18.1). This control method seeks the solution to (18.1) as the control data, to an initial value problem for an associated wave equation with a time periodic forcing term of a specified frequency. The initial data is then sought to yield a *time-periodic* solution of the wave equation, and a square norm functional is formulated to measure the time periodicity of the solution to the wave equation. The minimum of this functional solves the Helmholtz scattering problem (18.1), and this control problem can be solved using a *shooting* method. Computational tests in [BR37] indicate that its solution can be obtained efficiently provided an associated elliptic equation can be solved in optimal order complexity. In the discrete case, the computed solution will satisfy an alternate discretization of the Helmholtz scattering problem (due to discretization error introduced by time stepping).

In classical scattering theory for the wave equation, given a *convex* body D, the time average of the scattered solution $u(x,t)$ to the wave equation, with arbitrary initial conditions, converges to the solution $v(x)$ of the reduced wave equation (Helmholtz scattering problem):

$$\frac{1}{T}\int_0^T u(x,t)\,dt \to v(x).$$

In particular, exponential convergence is observed. Such a result, is however not valid for scattering objects which are not convex. In this context, the Hilbert uniqueness method [BR37] accelerates the convergence of time averages to the solution of the Helmholtz problem.

Our heuristic discussion will focus primarily on the continuous and semi-discrete versions of the Helmholtz problem (18.1). We seek $v(x)$ such that:

$$\begin{cases} -\Delta v - \kappa^2\, v = f(x), & \text{in} \quad \Omega \setminus D \\ \qquad\qquad v = g(x), & \text{in} \quad \partial D \\ \dfrac{\partial v}{\partial n} - i\kappa\, v = 0, & \text{on} \quad \partial\Omega. \end{cases} \tag{18.8}$$

Given $v(x)$ in (18.8), let $u(x,t) = v(x)e^{-i\,\kappa\,t}$ denote an associated *standing wave*. It can then be verified that $u(x,t) = v(x)e^{-i\,\kappa\,t}$ solves the wave equation:

$$\begin{cases} u_{tt} - \Delta u = f(x)e^{-i\kappa t}, & \text{in} \quad (\Omega \setminus D) \times (0, \dfrac{2\pi}{\kappa}) \\ u(x,t) = e^{-i\kappa t} g(x), & \text{on} \quad \partial D \times (0, \dfrac{2\pi}{\kappa}) \\ \dfrac{\partial u}{\partial t} + \dfrac{\partial u}{\partial n} = 0, & \text{on} \quad \partial\Omega \times (0, \dfrac{2\pi}{\kappa}) \\ u(x,0) = u(x, \dfrac{2\pi}{\kappa}) & \text{in} \quad (\Omega \setminus D) \\ u_t(x,0) = u_t(x, \dfrac{2\pi}{\kappa}) & \text{in} \quad (\Omega \setminus D). \end{cases} \tag{18.9}$$

Since $v(x) = u(x,0)$ and $-i\kappa v(x) = u_t(x,0)$, this suggests that the solution to (18.8) can be obtained from a *time-periodic* solution to (18.9).

Remark 18.7. Thus, we may determine the solution $v(.)$ to the Helmholtz problem by seeking initial conditions to (18.9) which yields a solution which is time-periodic of period $\frac{2\pi}{\kappa}$. If the periodic solution $u(x,t)$ to the wave equation (18.9) is *unique*, then $u(x,t)$ will be a standing wave of the form $u(x,t) = e^{-i\kappa t}\, v(x)$, yielding with $u(0,x) = v(x)$ and $u_t(0,x) = -i\,\kappa\, v(x)$.

Definition 18.8. *Given initial data $w_0(.)$ and $w_1(.)$ for the wave equation (18.9), we define the evolution map E:*

$$E : \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \rightarrow \begin{bmatrix} u(., \frac{2\pi}{\kappa}) \\ u_t(., \frac{2\pi}{\kappa}) \end{bmatrix} \tag{18.10}$$

where $u(.,.)$ solves (18.9) with initial data $w_0(.)$ and $w_1(.)$:

$$\begin{cases} u_{tt}(x,t) - \Delta u(x,t) = f(x)e^{-i\kappa t}, & in \quad (\Omega \setminus D) \times (0, \dfrac{2\pi}{\kappa}) \\ u(x,t) = e^{-i\kappa t} g(x), & on \quad \partial D \times (0, \dfrac{2\pi}{\kappa}) \\ \dfrac{\partial u}{\partial t}(x,t) + \dfrac{\partial u}{\partial n}(x,t) = 0, & on \quad \partial \Omega \times (0, \dfrac{2\pi}{\kappa}) \\ u(x,0) = w_0(x), & on \quad (\Omega \setminus D) \\ u_t(x,0) = w_1(x) & on \quad (\Omega \setminus D). \end{cases} \tag{18.11}$$

The map E will be affine linear, and if (w_0, w_1) is a fixed point of E:

$$E \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

then $v(.) = w_0(.)$ will solve Helmholtz's problem.

Control Problem. A control problem may now be posed for determining the initial data $v(.) = w_0(.)$ and $w_1(.)$ leading to a periodic solution of (18.11). A nonnegative functional $J(.,.) \geq 0$ measuring a square norm of the difference $E(w_0, w_1) - (w_0, w_1)$ may be employed, and the *minimum* of $J(.,.)$ may be sought, where E is the evolution map defined by (18.10). We define:

$$J(w_0, w_1) \equiv \frac{1}{2} \int_{\Omega \setminus D} \left(\left| \nabla (u(x, \frac{2\pi}{\kappa}) - w_0(x)) \right|^2 + \left| u_t(x, \frac{2\pi}{\kappa}) - w_1(x) \right|^2 \right) dx,$$

where $u(t,x)$ solves the wave equation (18.11). The preceding functional will be convex and quadratic, and is motivated by the energy:

$$\mathcal{E}(t) \equiv \int_{\Omega \setminus D} \left(|\nabla u|^2 + |u_t|^2 \right) dx.$$

A minimization algorithm such as a gradient (steepest) descent or the CG method may be employed to seek the minimum of $J(.,.)$. However, evaluating $J(.)$ will require evaluating the evolution map E, which in turn requires solving the wave equation on the time interval $(0, \frac{2\pi}{\kappa})$. In applications, time stepping can be employed, and this will introduce truncation errors.

We next heuristically outline the structure of the evolution map and the resulting control problem for a semi-discretization of the wave equation. For simplicity, we consider a finite difference discretization of (18.1):

$$\begin{bmatrix} A_{II} - \kappa^2 I & A_{IB} \\ A_{IB}^T & -i\,\kappa\, I + A_{BB} \end{bmatrix} \begin{bmatrix} \mathbf{v}_I \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} \mathbf{f}_I - A_{ID}\mathbf{g}_{\partial D} \\ \mathbf{f}_B \end{bmatrix}, \tag{18.12}$$

where A_{XY} for $X,\, Y = I,\, B$ is a block partition of a finite difference discretization of $-\Delta$ on $(\Omega \setminus D)$ with Neumann boundary conditions on $\partial\Omega$ and Dirichlet boundary conditions on ∂D. Here, $\mathbf{v}_I$ denotes a nodal vector of unknowns in $(\Omega \setminus D)$, and $\mathbf{v}_B$ with unknowns on $\partial\Omega$ and $\mathbf{v}_D$ on ∂D.

If $\mathbf{v}_I$ and $\mathbf{v}_B$ are the solution components from (18.12), then the *standing waves* $\mathbf{x}_I(t) \equiv e^{-i\,\kappa\, t}\,\mathbf{v}_I$ and $\mathbf{x}_B(t) \equiv e^{-i\,\kappa\, t}\,\mathbf{v}_B$ can be verified to solve the following first order system of ordinary differential equations for $\mathbf{x}_I(t)$, $\mathbf{x}_B(t)$ and $\mathbf{y}_I(t) = (d\,\mathbf{x}_I/\,dt)$ defined at the interior, boundary and interior grids points of $(\Omega \setminus D)$, $\partial\Omega$ and $(\Omega \setminus D)$, respectively. The 2nd order derivatives in time have been reduced to first order derivatives in time by introducing the variables $\mathbf{y}_I(t)$. The resulting system of differential equations will be:

$$\frac{d}{dt}\begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_B \\ \mathbf{y}_I \end{bmatrix} = \begin{bmatrix} 0 & 0 & I \\ -A_{IB}^T & -A_{BB} & 0 \\ -A_{II} & -A_{IB} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_B \\ \mathbf{y}_I \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{g}_I \end{bmatrix},$$

where $\mathbf{g}_I(t) \equiv \mathbf{f}_I e^{-i\,\kappa\, t} - A_{ID}\,\mathbf{g}_{\partial D} e^{-i\kappa\, t}$ and with initial conditions:

$$\begin{bmatrix} \mathbf{x}_I(0) \\ \mathbf{x}_B(0) \\ \mathbf{y}_I(0) \end{bmatrix} = \begin{bmatrix} \mathbf{w}_0 \\ \mathbf{w}_B \\ -i\,\kappa\,\mathbf{w}_0 \end{bmatrix},$$

provided $\mathbf{f}_B = \mathbf{0}$, $\mathbf{w}_0 = \mathbf{v}_I$ and $\mathbf{w}_B = \mathbf{v}_B$. This system can be formally verified using the reduced system (18.12). To derive an expression for the evolution map E_h we express the preceding system of differential equations compactly:

$$\frac{d\mathbf{U}}{dt} = L\,\mathbf{U} + \mathbf{F}(t), \quad \text{for } \; 0 < t < \frac{2\pi}{\kappa}, \quad \text{with } \; \mathbf{U}(0) \,=\, \mathbf{U}_0, \tag{18.13}$$

where the block vectors $\mathbf{U}(t)$, $\mathbf{U}_0$, $\mathbf{F}(t)$ and matrix L are defined as:

$$L \equiv \begin{bmatrix} 0 & 0 & I \\ -A_{IB}^T & -A_{BB} & 0 \\ -A_{II} & -A_{IB} & 0 \end{bmatrix}, \; \mathbf{U} \equiv \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_B \\ \mathbf{y}_I \end{bmatrix}, \; \mathbf{F} \equiv \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{g}_I \end{bmatrix}, \quad \mathbf{U}_0 \equiv \begin{bmatrix} \mathbf{w}_0 \\ \mathbf{w}_B \\ -i\,\kappa\,\mathbf{w}_0 \end{bmatrix}.$$

The solution to the linear system of differential equations (18.13) can be represented using Duhamel's principle as:

$$\mathbf{U}(t) \,\equiv\, e^{L\,t}\,\mathbf{U}_0 \,+\, \int_0^t e^{L\,(t-s)}\,\mathbf{F}(s)\,ds.$$

Thus, the semi-discrete evolution map E_h will satisfy:

$$E_h\,\mathbf{U}_0 = e^{\frac{2\pi}{\kappa}\,L}\,\mathbf{U}_0\; - \tilde{\mathbf{F}} \;\text{ where }\; \tilde{\mathbf{F}} \equiv -\int_0^{\frac{2\pi}{\kappa}} e^{L\,(t-s)}\,\mathbf{F}(s)\,ds.$$

Using E_h, we define the discrete control functional $J_h(.,.)$ as:

$$J_h(\mathbf{U}_0) \equiv \frac{1}{2}\,\|E_h\mathbf{U}_0 - \mathbf{U}_0\|_X^2,$$

where $\|\cdot\|_X$ denotes an appropriately chosen norm. The control functional $J_h(.)$ measures the square norm of the difference between solution at time $t = \frac{2\pi}{\kappa}$ and at time $t = 0$. The semi-discrete control problem seeks to enforce time periodicity of the solution by minimizing the control functional $J_h(.)$.

The linear system resulting from the condition for minimizing $J_h(.)$ will be symmetric and positive definite:

$$\nabla J_h = \mathbf{0} \;\Leftrightarrow\; \left(e^{\frac{2\pi}{\kappa}\,L} - I\right)^H X \left(e^{\frac{2\pi}{\kappa}\,L} - I\right)\mathbf{U}_0 = \left(e^{\frac{2\pi}{\kappa}\,L} - I\right)^H X\tilde{\mathbf{F}}.$$

A *heuristic* block matrix preconditioner can be formally obtained for the above least squares system using the approximation $(e^{\frac{2\pi}{\kappa}\,L} - I) \approx (\frac{2\pi}{\kappa})\,L$. This yields the preconditioner $M = (\frac{2\pi}{\kappa})^2\,L^T XL$ for the above system. To avoid cumbersome notation, we shall define the block matrix A as follows:

$$A \equiv \begin{bmatrix} A_{IB}^T & A_{BB} \\ A_{II} & A_{IB} \end{bmatrix} \;\text{ so that }\; L = \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix}.$$

This will yield the following 2×2 block structure for $L^H XL$ when $X = I$:

$$M = (\frac{2\pi}{\kappa})^2 \begin{bmatrix} A^2 & 0 \\ 0 & I \end{bmatrix}.$$

The action of M^{-1} can be computed at a computational cost proportional to solving two systems with coefficient matrix A. We omit further details.

References

[AB] Abdoulaev, G., Achdou, Y., Hontand, J.-C., Kuznetsov, Y., Pironneau, O., Prud'homme, C.: Domain decomposition for Navier-Stokes equations. ICIAM 99. Oxford Univ. Press (2000) 191–204

[AC] Achdou, Y., Japhet, C., Le Tallec, P., Nataf, F., Rogier, F., Vidrascu, M.: Domain decomposition methods for nonsymmetric problems. (Eds.) C.-H. Lai, P. E. Bjørstad, M. Cross, O. B. Widlund. Domain decomposition methods in science and engineering: Eleventh international conference. www.ddm.org (1999) 3–17

[AC2] Achdou, Y., Kuznetsov, Y.: Substructuring preconditioners for finite element methods on nonmatching grids. East-West J. Numer. Math. **3** No. 1 (1995) 1–28

[AC3] Achdou, Y., Kuznetsov, Y., Pironneau, O.: Substructuring preconditioners for the Q_1 mortar element method. Numer. Math. **71** No. 4 (1995) 419–449

[AC4] Achdou, Y., Le Tallec, P., Nataf, F., Vidrascu, M.: A domain decomposition preconditioner for an advection-diffusion problem. Comp. Meth. Appl. Mech. Engng. **184** (2000) 145–170

[AC5] Achdou, Y., Maday, Y.: The mortar element method with overlapping subdomains. SIAM J. Numer. Anal. **40** (2002) 601–628

[AC6] Achdou, Y., Maday, Y., Widlund, O.: Iterative substructuring preconditioners for mortar element methods in two dimensions. SIAM J. Numer. Anal. **32** No. 2 (1999) 551–580

[AC7] Achdou, Y., Nataf, F.: A Robin-Robin preconditioner for an advection diffusion problem. C. R. Acad. Sci. Paris. **325** Série I. (1997) 1211–1216

[AD] Adams, R. A.: Sobolev spaces. Academic Press (1975)

[AG] Agoshkov, V. I.: Poincaré-Steklov operators and domain decomposition methods in finite dimensional spaces. (Eds.) R. Glowinski, G. Golub, G. Meurant, J. Périaux. First international symposium on domain decomposition methods for partial differential equations. SIAM (1988) 73–112

[AG2] Agoshkov, V. I., Lebedev, V. I.: The Poincaré-Steklov operators and domain decomposition methods in variational problems. In Computational Processes and Systems. Nauka, Moscow. In Russian (1985) 173–227

[AI] Ainsworth, M.: A preconditioner based on domain decomposition for *h*-*p* finite element approximation on quasi-uniform meshes. SIAM J. Num. Anal. **33** (1996) 1358–1376

[AI2] Ainsworth, M., Sherwin, S.: Domain decomposition preconditioners for p and h-p finite element approximations of Stokes equations. Comp. Meth. Appl. Mech. Engrg. **175** (1999) 243–266

[AL] Alart, P., Barboteu, M., Le Tallec, P., Vidrascu, M.: Additive Schwarz method for nonsymmetric problems: Application to frictional multicontact problems. (Eds.) N. Debit, M. Garbey, R. H. W. Hoppe, D. E. Keyes, Y. A. Kuznetsov, J. Périaux. Domain decomposition methods in science and engineering. Thirteenth international conference on domain decomposition methods. (2002) 3–13

[AL2] Allen, M. B., Ewing, R. E., Lu, P.: Well conditioned iterative schemes for mixed finite element models of porous media flows. SIAM J. Sci. Comp. **13** No. 3 (1992) 794–814

[AL3] Almasi, G. S., Gottlieb, A.: Highly parallel computing. Addison-Wesley (1994)

[AL4] Alonso, A. M., Trotta, L., Valli, A.: Coercive domain decomposition algorithms for advection diffusion equations and systems. J. Comput. Appl. Math. **96** No. 1 (1998) 51–76

[AL5] Alonso, A. M., Valli, A.: An optimal domain decomposition preconditioner for low frequency time-harmonic Maxwell equations. Math. Comp. **68** No. 226 (1999) 607–631

[AL6] Alonso, A. M., Valli, A.: Domain decomposition methods for time harmonic Maxwell equations: Numerical results. In (Eds.) L. Pavarino, A. Toselli. Recent developments in domain decomposition methods. Lecture notes in computational science and engineering **23** (2002) 157–171

[AN] Anderson, E., Bai, Z., Bischof, C., Blackford, L., S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorenson, D.: LAPACK user's guide (Software, environment and tools, 9). SIAM (2000)

[AN2] Anderson, W. K., Gropp, W. D., Kaushik, D. K., Keyes, D. E., Smith, B. F.: Achieving high sustained performance in an unstructured mesh CFD application. Proceedings of SC1999. Bell prize award paper. IEEE Computer Society (1999)

[AR] Arbogast, T., Cowsar, L. C., Wheeler, M. F., Yotov, I.: Mixed finite element methods on nonmatching multiblock grids. SIAM J. Numer. Anal. **37** (2000) 1295–1315

[AR2] Arbogast, T., Wheeler, M. F., Yotov, I.: Mixed finite elements for elliptic problems with tensor coefficients as cell centered finite differences. SIAM J. Numer. Anal. **34** No. 2 (1997) 828–852

[AR3] Arnold, V. I.: Ordinary differential equations. Springer-Verlag (1992)

[AR4] Arnold, D. N., Brezzi, F.: Mixed and nonconforming finite element methods: Implementation, post processing and error estimates. Math. Model. Numer. Anal. **19** (1985) 7–32

[AR5] Arnold, D. N., Falk, R. S., Winther, R.: Preconditioning in H(div) and applications. Math. Comp. **66** No. 219 (1997) 957–984

[AR6] Arnold, D. N., Falk, R. S., Winther, R.: Multigrid in H(div) and H(curl). Numer. Math. **85** No. 2 (2000) 197–217

[AR7] Arrow, K., Hurwicz, L., Uzawa, H.: Studies in nonlinear programming. Stanford University Press, CA (1958)

[AS] Ascher, U. M., Matheij, R. M., Russell, R. R.: Numerical solution of boundary value problems for ordinary differential equations. Prentice Hall (1988)

[AS2] Ashby, S. F., Kelley, C. T., Saylor, P. E., Scroggs, J.: Preconditioning via asymptotically defined domain decomposition. Seventh international conference on domain decomposition methods in scientific and engineering computing. (Eds.) D. Keyes, J. Xu. Contemporary Mathematics **180**, AMS (1994) 131–136

[AS3] Ashby, S. F., Manteuffel, T. A., Saylor, P. E.: A taxonomy for conjugate gradient methods. SIAM J. Numer. Anal. **27** (1990) 1542–1568

[AS4] Astrakhantsev, G. P.: Method of fictitious domains for a second order elliptic equation with natural boundary conditions. USSR Comput. Math. and Math. Phys. **18** (1978) 114–121

[AT] Atamian, C., Dinh, Q. V., Glowinski, R., He, J., Périaux, J.: Control approach to fictitious domain methods application to fluid dynamics and electromagnetics. Fourth international symposium on domain decomposition methods for partial differential equations. SIAM (1991)

[AU] Auge, A., Kapurkin, A., Lube, G., Otto, F.-C.: A note on domain decomposition of singularly perturbed elliptic problems. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997) 163–170

[AU2] Auge, A., Lube, G., Otto, F.-C.: A non-overlapping domain decomposition method with adaptive interface conditions for elliptic problems. In Numerical treatment of multiscale problems. Notes Numer. Fluid. Mech. **70** Vieweg, Braunschweig (1999) 12–23

[AX] Axelsson, O.: Iterative solution methods. Cambridge University Press. (1996)

[AX2] Axelsson, O., Barker, V. A.: Finite element solution of boundary value problems- Theory and computations. Academic Press. (1984)

[AX3] Axelsson, O., Vassilevski, P.: Algebraic multilevel preconditioning methods, I. Numer. Math. **56** (1989) 157–177

[AX4] Axelsson, O., Vassilevski, P.: Algebraic multilevel preconditioning methods, II. SIAM. J. Num. Anal. **27** (1990) 1569–1590

[AZ] Aziz, A. K., Kellogg, R. B., Stephens, A. B.: Least squares methods for elliptic systems. Math. Comp. **44** (1985) 53–70 ZAMM, **37** No. 7/8 (1957) 243–245

[BA] Babuška, I.: Über Schwarzsche algorithmen in partielle differentialgleichungen der mathematischen physik. ZAMM, **37** No. 7/8 (1957) 243–245

[BA2] Babuška, I.: The Schwarz algorithm in partial differential equations of mathematical physics. Czech. Math. J. **83** No. 8 (1958) 328–343 (in Russian)

[BA3] Babuška, I., Aziz, A.: Survey lectures on the mathematical foundations of the finite element method. In A. Aziz (Ed.), The mathematical foundations of the finite element method with applications to partial differential equations. Academic Press (1972)

[BA4] Babuška, I., Craig, A., Mandel, J., Pitkäranta, J.: Efficient preconditioning for the p-version finite element method in two dimensions. SIAM J. Num. Anal. **28** No. 3 (1991) 624–661

[BA5] Babuška, I., Flaherty, J. E., Chandra, J.: Adaptive computational methods for partial differential equations. SIAM (1983)

[BA6] Babuška, I., Melenk, J. M.: The partition of unity finite element method: Basic theory and applications. Comp. Meth. Appl. Mech. Engrg. **139** (1996) 289–314

[BA7] Babuška, I., Melenk, J. M.: The partition of unity finite element method. Inter. J. Numer. Meth. Engrg. **40** (1997) 727–758

[BA8] Babuška, I., Rheinboldt, W. C.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15** (1978) 736–754

[BA9] Badea, L.: On the Schwarz alternating method with more than two subdomains for nonlinear monotone problems. SIAM J. Numer. Anal. **28** (1991) 179–204

[BA10] Badea, L., Wang, J.: An additive Schwarz method for variational inequalities. Math. Comp. **69** (2000) 1341–1354

[BA11] Bagrinovskii, K. A., Godunov, S. K.: Difference schemes for multidimensional problems. Dokl. Acad. Nauk. USSR **115** (1957)

[BA12] Bai, Z., Golub, G. H., Ng, M.: Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems. SIAM J. Matr. Anal. **24** (2003) 603–626

[BA13] Balay, S., Buschelman, Eijhkout, V., K., Gropp, W. D., Kaushik, D., Knepley, M. G., McInnes, L. C., Smith, B. F., Zhang, H.: PETSc users manual: ANL-95/11-Revision 2.1.5 Argonne National Laboratory. (2004)

[BA14] Balay, S., Buschelman, K., Gropp, W. D., Kaushik, D., Knepley, M. G., McInnes, L. C., Smith, B. F., Zhang, H.: PETSc web page: http://www.mcs.anl.gov/petsc (2001)

[BA15] Balay, S., Eijhkout, V., K., Gropp, W. D., Knepley, M. G., McInnes, L. C., Smith, B. F.: Efficient management of parallelism in object oriented numerical software libraries. Modern Software Tools in Scientific Computing. (Eds.) E. Arge, A. M. Bruaset, H. P. Langtangen. Birkhauser Press (1997) 163–202

[BA16] Bank, R. E., Dupont, T. F., Yserentant, H.: The hierarchical basis multigrid method. Numer. Math. **52** (1988) 427–458

[BA17] Bank, R. E., Scott, L. R.: On the conditioning of finite element equations with highly refined meshes. SIAM J. Numer. Anal. **26** No. 6 (1989) 1383–1394

[BA18] Bank, R. E., Welfert, B., Yserentant, H.: A class of iterative methods for solving saddle point problems. Numer. Math. **56** (1990) 645–666

[BA19] Barboteu, M., Alart, P., Vidrascu, M.: A domain decomposition strategy for nonclassical frictional multicontact problems. Comp. Methds. Appl. Mech. Engrg. **190** No. 37-38 (2001) 4785–4803

[BA20] Barnard, S. T., Simon, H. D.: A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. Concurrency: Practice and experience. **6** (1994) 101–117

[BA21] Barnes, E. R.: An algorithm for partitioning the nodes of a graph. SIAM J. Alg. Disc. Meth. **3** (1982) 541–550

[BA22] Barth, T. J.: Aspects of unstructured grids and finite volume solvers for the Euler and Navier-Stokes equations. Lecture Series 1994–05. Von Karman Institute for Fluid Dynamics (1994)

[BA23] Barth, T. J.: Figure of an unstructured grid provided by Dr. T. Barth. (2003)

[BA24] Barth, T. J., Chan, T. F., Tang, W.-P.: A parallel non-overlapping domain decomposition algorithm for compressible fluid flow problems on triangulated domains. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. Tenth international conference on domain decomposition methods. AMS Contemporary Mathematics **218** (1998) 23–41

[BA25] Bauer, A. C., Patra, A. K.: Performance of parallel preconditioners for adaptive *h-p* FEM discretization of incompressible flows. Comm. Numer. Methds. Engrg. **18** No. 5 (2002) 305–313

[BE] Beauwens, R.: Factorization iterative methods, M-operators and H-operators. Numer. Math. **31** (1979) 335–357

[BE2] Beauwens, R.: Modified incomplete factorization strategies. Preconditioning conjugate gradient methods. (Eds.) O. Axelsson, L. Kolotilina. Lecture notes in mathematics **1457**. Springer-Verlag (1990) 1–16

[BE3] Bell, J. B., Colella, P., Glaz, H. M.: A second-order projection method for the incompressible Navier-Stokes equations. J. Comp. Phys. **85** (1989) 257–283

[BE4] Ben Belgacem, F.: The mortar finite element method with Lagrange multipliers. Numer. Math. **84** No. 2 (1999) 173–197

[BE5] Ben Belgacem, F., Bernardi, C., Chorfi, N., Maday, Y.: Inf-sup conditions for the mortar spectral element discretization of the Stoke's problem. Numer. Math. **85** (2000) 257–281

[BE6] Ben Belgacem, F., Maday, Y.: The mortar element method for three dimensional finite elements. RAIRO Math. Modell. Numer. Anal. **31** No. 2 (1997) 289–302

[BE7] Benamou, J.-D.: A domain decomposition method for the optimal control of a system governed by the Helmholtz equation. In (Ed.) G. S. Cohen. Third international conference on mathematical and numerical wave propagation phenomena. SIAM (1995)

[BE8] Benamou, J.-D.: Domain decomposition methods with coupled transmission conditions for the optimal control of systems governed by elliptic partial differential equations. SIAM J. Numer. Anal. **33** No. 6 (1996) 2401–2416

[BE9] Benamou, J.-D., Després, B.: A domain decomposition method for the Helmholtz equation and related optimal control problems. J. Comp. Phys. **136** (1997) 68–82

[BE10] Benninghof, J. K., Lehoucq, R. B.: An automated multilevel substructuring method for eigenspace computation in linear elastodynamics. SIAM J. Sci. Comput. **25** No. 6 (2004) 2084–2106

[BE11] Benzi, M., Frommer, A., Nabben, R., Szyld, D.: Algebraic theory of multiplicative Schwarz methods. Numer. Math. **89** (2001) 605–639

[BE12] Benzi, M., Golub, G. H., Liesen, J.: Numerical solution of saddle point problems. Acta Numerica **14** (2005) 1–137

[BE13] Benzi, M., Meyer, C. D., Tuma, M.: A sparse approximate inverse preconditioner for the conjugate gradient method. SIAM J. Sci. Comp. **17** No. 5 (1996) 1135–1149

[BE14] Berger, M., Bokhari, S.: A partitioning strategy for nonuniform problems on multiprocessors. IEEE Trans. Comput. **36** (1987) 570–580

[BE15] Berger, M. J., Oliger, J.: Adaptive mesh refinement for hyperbolic partial differential equations. J. Comp. Phys. **53** (1984) 484–512

[BE16] Bergh, J., Lofstrom, J.: Interpolation spaces: An introduction. Springer-Verlag (1976)

[BE17] Berman, A., Plemmons, R. J.: Nonnegative matrices in the mathematical sciences. Academic Press (1979)

[BE18] Bernardi, C., Debit, N., Maday, Y.: Coupling finite element and spectral methods. Math. Comp. **54** (1990) 21–39

[BE19] Bernardi, C., Maday, Y.: Approximations spectrales de problèmes aux limites elliptiques. Mathèmatiques & applications **10**. Springer-Verlag (1992)

[BE20] Bernardi, C., Maday, Y.: Polynomial interpolation results in Sobolev spaces. J. Comput. Appl. Math. **43** (1992) 53–80

[BE21] Bernardi, C., Maday, Y.: Mesh adaptivity in finite elements using the mortar method. Revue Européenne des éléments finis. **9** (2000) 451–465

[BE22] Bernardi, C., Maday, Y., Patera, A.: A new nonconforming approach to domain decomposition: The mortar element method. College de France Seminar. Technical Report in 1989. (Eds.) H. Brezis, J. L. Lions. Pitman (1994)

[BE23] Bernardi, C., Maday, Y., Patera, A.: Domain decomposition by the mortar element method. Asymptotic and numerical methods for partial differential equations with critical parameters. NATO ASI (Eds.) H. G. Kaper, M. Garbey (1993) 269–286

[BE24] Bernardi, C., Maday, Y., Rapetti, F.: Discretisations variationelles de problemes aux limites elliptiques. Mathematiques et applications **45** Springer-Verlag (2004)

[BH] Bhardwaj, M., Day, D., Farhat, C., Lesoinne, M., Pierson, K., Rixen, D.: Application of the FETI method to ASCI problems: Scalability results on one thousand processors and discussion of highly heterogeneous problems. Internat. J. Numer. Methds. Engrg. **47** (2000) 513–535

[BI] Bica, I.: Iterative substructuring algorithms for the p-version finite element method for elliptic problems. PhD thesis. Computer Science TR 743, CIMS, New York University (1997)

[BI2] Bica, I.: Nonoverlapping domain decomposition algorithms for the p-version finite element method for elliptic problems. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. Tenth international conference on domain decomposition methods. AMS Contemporary Mathematics **218** (1998)

[BI3] Biegler, L. T., Ghattas, O., Heinkenschloss, M., van Bloeman Waanders, B.: (Eds.) Large scale PDE constrained optimization. Springer (2003)

[BI4] Biros, G., Ghattas, O.: Parallel Lagrange-Newton-Krylov-Schur methods for PDE constrained optimization, Part I: Krylov-Schur solver. SIAM J. Sci. Comp. **27** No. 2 (2005) 687–713

[BI5] Biros, G., Ghattas, O.: Parallel Lagrange-Newton-Krylov-Schur methods for PDE constrained optimization, Part II: The Lagrange-Newton solver and its application to optimal control of steady viscous flows. SIAM J. Sci. Comp. **27** No. 2 (2005) 714–739

[BJ] Bjørstad, P.: Multiplicative and additive Schwarz methods: Convergence in the two subdomain case. In (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. Second international symposium on domain decomposition methods. SIAM (1989) 147–159

[BJ2] Bjørstad, P., Dryja, M., Rahman, T.: Additive Schwarz methods for elliptic mortar finite element problems. Numer. Math. **95** No. 3 (2003) 427–457

[BJ3] Bjørstad, P., Espedal, M., Keyes, D. E.: (Eds.) Domain decomposition methods in science and engineering. Ninth international conference on domain decomposition methods. www.ddm.org (1997)

[BJ4] Bjørstad, P., Hvidsten, A.: Iterative methods for substructured elasticity problems in structural analysis. In (Eds.) R. Glowinski, G. Golub, G. Meurant, J. Périaux. First international symposium on domain decomposition methods for partial differential equations. SIAM (1988) 301–312

[BJ5] Bjørstad, P., Mandel, J.: On the spectra of sums of orthogonal projections with applications to parallel computing. BIT. **31** (1991) 76–88

[BJ6] Bjørstad, P., Moe, R., Skogen, M.: Parallel domain decomposition and iterative refinement algorithms. In (Ed.) W. Hackbusch. Parallel algorithms for partial differential equations. Proceedings of the 6th GAMM seminar. Vieweg-Verlag (1990)

[BJ7] Bjørstad, P., Skogen, M.: Domain decomposition algorithms of Schwarz type designed for massively parallel computers. Fifth international symposium on domain decomposition methods for partial differential equations. SIAM (1992)

[BJ8] Bjørstad, P., Widlund, O. B.: Solving elliptic problems on regions partitioned into substructures. Elliptic problem solvers II. (Eds.) G. Birkhoff, A. Schoenstadt. Academic Press (1984) 245–256

[BJ9] Bjørstad, P., Widlund, O. B.: Iterative methods for the solution of elliptic problems on regions partitioned into substructures. SIAM J. Numer. Anal. **23** No. 6 (1986) 1093–1120

[BJ10] Bjørstad, P., Widlund, O. B.: To overlap or not to overlap: A note on a domain decomposition method for elliptic problems. SIAM J. Sci. Stat. Comput. **10** No. 5 (1989) 1053–1061

[BL] Bleistein, N.: Mathematical methods for wave phenomena. Academic Press (1984)

[BL2] Bleistein, N., Cohen, J. K., Stockwell Jr., J. W. : Mathematics of multidimensional seismic imaging, migration, and inversion. Springer-Verlag (2001)

[BL3] Blum. H., Lisky, S., Rannacher, R.: A domain splitting algorithm for parabolic problems. Computing: Archiv fűr Informatik und Numerik. **49** No. 1. (1992) 11–23

[BO] Bochev, P. B., Gunzburger, M. D.: Accuracy of least squares methods for the Navier-Stokes equations. Comput. Fluids. **22** (1993) 549–563

[BO2] Bollobas, B.: Graph theory: An introductory course. Springer-Verlag (1979)

[BO3] Boppana, R. B.: Eigenvalues and graph bisection: An average case analysis. 28th Annual Symp. Found. Comp. Sci. (1987) 280–285

[BO4] Borgers, C.: The Neumann-Dirichlet domain decomposition method with inexact solvers on the subdomains. Numer. Math. **55** (1989) 123–136

[BO5] Borgers, C., Widlund, O. B.: On finite element domain imbedding methods. SIAM J. Numer. Anal. **27** No. 4 (1990) 963–978

[BO6] Bornemann, F., Yserentant, H.: A basic norm equivalence for the theory of multilevel methods. Numer. Math. **64** (1993) 455–476

[BO7] Bourgat, J.-F., Glowinski, R., Le Tallec, P., Vidrascu, M.: Variational formulation and algorithm for trace operator in domain decomposition calculations. In: Second international symposium on domain decomposition methods for partial differential equations. SIAM (1989) 3–16

[BO8] Bourgat, J.-F., Le Tallec, P., Tidriri, M. D.: Coupling Boltzmann and Navier-Stokes by friction. J. Comput. Phys. **127** (1996) 227–245

[BO9] Bourquin, F.: Analysis and comparison of several component mode synthesis methods on one dimensional domains. Numer. Math. **58** No. 1 (1990) 11–34

[BO10] Bourquin, F.: Component mode synthesis and eigenvalues of second order operators: Discretization and algorithms. RAIRO Model. Math. et Anal. Numer. **26** No. 3 (1992) 385–423

[BO11] Bourquin, F., d'Hennezel, F.: Application of domain decomposition techniques to modal synthesis for eigenvalue problems. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[BO12] Bourquin, F., d'Hennezel, F.: Numerical study of an intrinsic component mode synthesis method. Comp. Meth. Appl. Mech. Engrg. **97** (1992) 49–76

[BO13] Bourquin, F., Namar, R.: Decoupling and modal synthesis of vibrating continuous systems. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[BR] Braess, D.: Finite elements: Theory, fast solvers and applications to solid mechanics. Cambridge University Press. (1997)

[BR2] Braess, D., Dahmen, W.: Stability estimates of the mortar element method for three dimensional problems. East-West J. Numer. Math. **6** (1998) 249–263

[BR3] Braess, D., Dahmen, W., Wieners, C.: A multigrid algorithm for the mortar finite element method. SIAM J. Numer. Anal. **37** (2000) 48–69

[BR4] Braess, D., Verfurth, R.: Multigrid methods for nonconforming finite element methods. SIAM J. Numer. Anal. **27** No. 4 (1990) 979–986

[BR5] Bramble, J.: Multigrid methods. Chapman and Hall (1993)

[BR6] Bramble, J., Ewing, R. E., Pareshkevov, R., Pasciak, J.: Domain decomposition methods for problems with partial refinement. SIAM J. Sci. Comp. **13** No. 1 (1992) 397–410

[BR7] Bramble, J., Ewing, R. E., Pasciak, J., Schatz, A.: A preconditioning technique for the efficient solution of problems with local grid refinement. Comput. Meth. Appl. Mech. Engg. **67** (1988) 149–159

[BR8] Bramble, J., Pasciak, J.: A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. Math. Comp. **50** (1988) 1–18

[BR9] Bramble, J., Pasciak, J.: Iterative techniques for the time dependent Stokes equation. Math. Applic. **33** (1997) 13–30

[BR10] Bramble, J., Pasciak, J., Knyazev, A.: A subspace preconditioning algorithm for eigenvector/eigenvalue computation. Advances in computational mathematics. **6** No. 2 (1996) 159–189

[BR11] Bramble, J., Pasciak, J., Schatz, A.: An iterative method for elliptic problems on regions partitioned into substructures. Math. Comp. **46** No. 173, (1986) 361–369

[BR12] Bramble, J., Pasciak, J., Schatz, A.: The construction of preconditioners for elliptic problems by substructuring, I. Math. Comp. **47** (1986) 103–134

[BR13] Bramble, J., Pasciak, J., Schatz, A.: The construction of preconditioners for elliptic problems by substructuring, II. Math. Comp. **49** (1987) 1–16

[BR14] Bramble, J., Pasciak, J., Schatz, A.: The construction of preconditioners for elliptic problems by substructuring, III. Math. Comp. **51** (1988) 415–430

[BR15] Bramble, J., Pasciak, J., Schatz, A.: The construction of preconditioners for elliptic problems by substructuring, IV. Math. Comp. **53** (1989) 1–24

[BR16] Bramble, J., Pasciak, J., Vassilev, A.: Analysis of the inexact Uzawa algorithm for saddle point problems. SIAM J. Numer. Anal. **34** (1997) 1072–1092

[BR17] Bramble, J., Pasciak, J., Vassilevski, P.: Computational scales of Sobolev norms with application to preconditioning. Math. Comp. **69** (2000) 463–480

[BR18] Bramble, J., Pasciak, J., Wang, J., Xu, J.: Convergence estimates for product iterative methods with applications to domain decomposition. Math. Comp. **57** No. 195 (1991) 1–21

[BR19] Bramble, J., Pasciak, J., Wang, J., Xu, J.: Convergence estimates for multigrid algorithms without regularity assumptions. Math. Comp. **57** No. 195 (1991) 23–45

[BR20] Bramble, J., Pasciak, J., Xu, J.: Parallel multilevel preconditioners. Math. Comp. **55** (1990) 1–22

[BR21] Bramble, J., Xu, J.: Some estimates for a weighted L^2 projection. Math. Comp. **56** (1991) 163–176

[BR22] Brandt, A.: Multilevel adaptive solutions to boundary value problems. Math. Comp. **31** (1977) 333–390

[BR23] Brenner, S. C.: Two level additive Schwarz preconditioners for nonconforming finite elements. Domain decomposition methods in scientific and engineering computing. Contemporary Mathematics **180** AMS (1994) 9–14

[BR24] Brenner, S. C.: The condition number of the Schur complement in domain decomposition. Numer. Math. **83** (1999) 187–203

[BR25] Brenner, S. C.: Lower bounds of two level additive Schwarz preconditioners with small overlap. SIAM J. Sci. Comp. **21** No. 5 (2000) 1657–1669

[BR26] Brenner, S. C.: An additive Schwarz preconditioner for the FETI method. Numer. Math. **94** (2003) 1–31

[BR27] Brenner, S. C., He, Q.: Lower bounds for three dimensional nonoverlapping domain decomposition algorithms. Numer. Math. **93** No. 3 (2003) 445–470

[BR28] Brenner, S. C., Scott, L. R.: Mathematical theory of finite element methods. Springer-Verlag (1994)

[BR29] Brenner, S. C., Sung, L.-Y.: Discrete Sobolev and Poincaré inequalities via Fourier series. East-West J. Num. Math. **8** (2000) 83–92

[BR30] Brezina, M., Vanek, P.: A black box iterative solver based on a two level Schwarz method. Computing **63** No. 3 (1999) 233–263

[BR31] Brezzi, F.: On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers. RAIRO **8** (1974) 129–151

[BR32] Brezzi, F., Canuto, C., Russo, A.: A self adaptive formulation for the Euler-Navier-Stokes coupling. Comp. Meth. in Appl. Mech. Engrg. **73** (1989) 317–330

[BR33] Brezzi, F., Fortin, M.: Mixed and hybrid finite element methods. Springer-Verlag (1991)

[BR34] Brezzi, F., Franca, L. P., Marini, L. D., Russo, A.: Stabilization techniques for domain decomposition methods with nonmatching grids. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[BR35] Brezzi, F., Marini, L. D.: A three fields domain decomposition method. Sixth international conference on domain decomposition. AMS Contemp. Math. **157** (1994) 27–34

[BR36] Briggs, W. L., Henson, V. E., McCormick, S. F.: A multigrid tutorial. SIAM (2000)

[BR37] Bristeau, M. O., Glowinski, R., Périaux, J.: Controllability methods for the computation of time periodic solutions: Applications to scattering. J. Comput. Phys. **147** No. 2 (1998) 265–292

[BR38] Brown, P. N., Walker, H. F.: GMRES on nearly singular systems. SIAM J. Matr. Anal. Appl. **18** (1997) 37–51

[BR39] Bruaset, A. M., Tveito, A.: Numerical solution of partial differential equations on parallel computers. (Eds). Lecture notes in computational sciences and engineering **51**. Springer (2006)

[BU] Buzbee, B. L., Dorr, F., George, J., Golub., G.: The direct solution of the discrete Poisson equation on irregular regions. SIAM J. Num. Anal. **11** (1971) 722–736

[CA] Cai, X.-C.: Additive Schwarz algorithms for parabolic convection diffusion equations. Num. Math. **60** No. 1 (1991) 41–61

[CA2] Cai, X.-C.: An optimal two level overlapping domain decomposition method for elliptic problems in two and three dimensions. SIAM J. Sci. Comp. **14** (1993) 239–247

[CA3] Cai, X.-C.: Multiplicative Schwarz methods for parabolic problems. SIAM J. Sci. Comp. **15** No. 3 (1994) 587–603

[CA4] Cai, X.-C.: The use of pointwise interpolation in domain decomposition methods with non-nested meshes. SIAM J. Sci. Comput. **16** No. 1 (1995) 250–256

[CA5] Cai, X.-C., Casarin, M., Elliott Jr., F. W., Widlund, O. B. : Overlapping Schwarz algorithms for solving Helmholtz's equation. Tenth international conference on domain decomposition. (Eds.) J. Mandel, C. Farhat, X.-C. Cai. AMS Contemporary Mathematics **218** (1998) 391–399

[CA6] Cai, X.-C., Dryja, M.: Domain decomposition methods for monotone nonlinear elliptic problems. Domain decomposition methods in scientific and engineering computing. (Eds.) D. Keyes, J. Xu. AMS Contemporary Mathematics **180** (1994) 21–27

[CA7] Cai, X.-C., Dryja, M., Sarkis, M.: Overlapping nonmatching grid mortar element methods for elliptic problems. SIAM J. Numer. Anal. **36** (1999) 581–606

[CA8] Cai, X.-C., Dryja, M., Sarkis, M.: A restricted additive Schwarz preconditioner with harmonic overlap for symmetric positive definite linear systems. SIAM J. Numer. Anal. **41** No. 4 (2003) 1209–1231

[CA9] Cai, X.-C., Gropp, W. D., Keyes, D. E.: Convergence rate estimate for a domain decomposition method. Numer. Math. **61** (1992) 153–169

[CA10] Cai, X.-C., Gropp, W. D., Keyes, D. E.: A comparison of some domain decomposition and ILU preconditioned iterative methods for nonsymmetric elliptic problems. Numer. Lin. Alg. Applic. **1** No. 5 (1994) 477–504

[CA11] Cai, X.-C., Gropp, W. D., Keyes, D. E., Melvin, R. G., Young, D. P.: Parallel Newton-Krylov-Schwarz algorithms for the transonic full potential equations. SIAM J. Sci. Comp. **19** No. 1 (1998) 246–265

[CA12] Cai, X.-C., Gropp, W. D., Keyes, D. E., Tidriri, M. D.: Newton-Krylov-Schwarz methods in CFD. In (Ed.) R. Rannacher. Proceedings of the international workshop on the Navier-Stokes equations, Notes in numerical fluid mechanics. Vieweg-Verlag (1994) 17–30

[CA13] Cai, X.-C., Keyes, D. E.: Nonlinearly preconditioned inexact Newton algorithms. SIAM J. Sci. Comp. **24** No. 1 (2002) 183–200

[CA14] Cai, X.-C., Keyes, D. E., Marcinkowski, L.: Nonlinear additive Schwarz preconditioners and applications in computational fluid dynamics. Internat. J. Numer. Methds. Fluids. **40** No. 12 (2002) 1463–1470

[CA15] Cai, X.-C., Keyes, D. E., Venkatakrishnan, V.: Newton-Krylov-Schwarz: An implicit solver for CFD. In (Eds.) R. Glowinski, J. Périaux, Z.-C. Shi, O. B. Widlund. Domain decomposition methods in science and engineering. Eight international conference. John-Wiley (1996) 387–402

[CA16] Cai, X.-C., Keyes, D. E., Young, D. P.: A nonlinearly additive Schwarz preconditioned inexact Newton method for shocked duct flow. In (Eds.) N. Debit, M. Garbey, R. H. W. Hoppe, D. E. Keyes, Y. A. Kuznetsov, J. Périaux. Domain decomposition methods in science and engineering. Thirteenth international conference. CIMNE (2002) 345–352

[CA17] Cai, X.-C., Mathew, T. P., Sarkis, M. V.: Maximum norm analysis of overlapping nonmatching grid discretizations of elliptic equations. SIAM J. Numer. Anal. **37** (2000) 1709–1728

[CA18] Cai, X.-C., Mathew, T. P., Sarkis, M. V.: A polynomial coarse space on unstructured grids. Unpublished Work. (2003)

[CA19] Cai, X.-C., Sarkis, M. V.: A restricted additive Schwarz preconditioner for general sparse linear systems. SIAM J. Sci. Comput. **21** (1999) 239–247

[CA20] Cai, X.-C., Widlund, O. B.: Domain decomposition algorithms for indefinite elliptic problems. SIAM J. Sci. Comp. **13** No. 1 (1992) 243–258

[CA21] Cai, X.-C., Widlund, O. B.: Multiplicative Schwarz algorithms for some nonsymmetric and indefinite problems. SIAM J. Numer. Anal. **30** No. 4 (1993) 936–952

[CA22] Cai, Z., Goldstein, C. I., Pasciak, J.: Multilevel iteration for mixed finite element systems with penalty. SIAM J. Sci. Comp. **14** (1993) 1072–1088

[CA23] Cai, Z., Lazarov, R., Manteuffel, T., McCormick, S.: First order system least squares for 2nd order partial differential equations. Part I. SIAM J. Num. Anal. **31** No. 6 (1994) 1785–1802

[CA24] Cai, Z., Mandel, J., McCormick, S.: Multigrid methods for nearly singular linear equations and eigenvalue problems. SIAM J. Numer. Anal. **34** No. 1 (1997) 178–200

[CA25] Cai, Z., Pareshkevov, R., Russell, T. F., Ye, X.: Overlapping domain decomposition for a mixed finite element method in three dimensions. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[CA26] Calgaro, C., Laminie, J.: On the domain decomposition method for the generalized Stokes problem with continuous pressure. Numer. Methds. Part. Diff. Eqns. **16** No. 1 (2000) 84–106

[CA27] Canuto, C., Funaro, D.: The Schwarz algorithm for spectral methods. SIAM J. Numer. Anal. **25** No. 1 (1988) 24–40

[CA28] Canuto, C., Hussaini, M. Y., Quarteroni, A., Zang, T. A.: Spectral methods in fluid dynamics. Springer-Verlag (1988)

[CA29] Canuto, C., Russo, A.: Self adaptive coupling of mathematical models and/or numerical methods. Sixth international conference on domain decomposition. AMS, Contemporary Mathematics **157** (1994)

[CA30] Cao, W., Guo, B.: Preconditioners on element interfaces for the p-version spectral element method. SIAM J. Sci. Comp. **21** (1999) 522–551

[CA31] Carlenzoli, C., Quarteroni, A.: Adaptive domain decomposition methods for advection-diffusion problems. Modeling, mesh generation, and adaptive numerical methods for partial differential equations (Eds. I. Babuska, et al) IMA Volumes in Mathematics and its Applications **75**. Springer-Verlag (1995) 165–199

[CA32] Carlson, D., Markham, T.: Schur complements of diagonally dominant matrices. Czech Math. J. **29** No. 104 (1979) 246–251

[CA33] Carvalho, L. M., Girard, L., Meurant, G.: Local preconditioners for two-level nonoverlapping domain decomposition methods. Numer. Lin. Alg. Appl. **8** No. 4 (2001) 207–227

[CA34] Casarin, M.: Schwarz preconditioners for spectral and mortar finite element methods with applications to incompressible fluids. PhD thesis. Technical Report 717. Department of Computer Science. Courant Institute of Mathematical Sciences (1996)

[CA35] Casarin, M.: Quasioptimal Schwarz methods for conforming spectral element discretization. SIAM J. Numer. Anal. **34** No. 6 (1997) 2482–2502

[CA36] Casarin, M.: Diagonal edge preconditioners in p-version and spectral element methods. SIAM J. Sci. Comp. **31** No. 2 (1997) 610–620

[CA37] Casarin, M., Widlund, O. B.: A hierarchical preconditioner for the mortar finite element method. ETNA **4** (1996) 75–88

[CA38] Cazabeau, L., Lacour, C., Maday, Y.: Numerical quadratures and mortar methods. Computational science for the 21st century. John Wiley (1997) 119–128

[CH] Chan, R. H., Chan, T. F.: Circulant preconditioners for elliptic problems. J. Numer. Lin. Alg. Appl. **1** (1992) 77–101

[CH2] Chan, T. F.: Analysis of preconditioners for domain decomposition. SIAM J. Numer. Anal. **24** No. 2 (1987) 382–390

[CH3] Chan, T. F., Go, S., Zikatanov, L.: Lecture notes on multilevel methods for elliptic problems on unstructured grids. Technical Report. CAM 97-11. UCLA (1997)

[CH4] Chan, T. F., Goovaerts, D.: On the relationship between overlapping and nonoverlapping domain decomposition methods. SIAM J. Matr. Anal. Appl. **13** No. 2 (1992) 663–670

[CH5] Chan, T. F., Hou, T. Y.: Eigendecompositions of domain decomposition interface operators for constant coefficient elliptic problems. SIAM J. Sci. Comp. **12** No. 6 (1991) 1471–1479

[CH6] Chan, T. F., Hou, T. Y., Lions, P. L.: Geometry related convergence results for domain decomposition algorithms. SIAM J. Numer. Anal. **28** No. 2 (1991) 378–391

[CH7] Chan, T. F., Kako, T., Kawarada, H., Pironneau, O.: (Eds.) Domain decomposition methods in science and engineering. Twelveth international conference. www.ddm.org (2001)

[CH8] Chan, T. F., Keyes, D. E.: Interface preconditioning for domain decomposition convection diffusion operators. Third international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. b. Widlund. SIAM (1990)

[CH9] Chan, T. F., Mathew, T. P.: The interface probing technique in domain decomposition. SIAM J. Matr. Anal. Appl. **13** No. 1 (1992) 212–238

[CH10] Chan, T. F., Mathew, T. P.: Domain decomposition preconditioners for convection diffusion problems. Domain decomposition methods in science and engineering. (Eds.) A. Quarteroni, J. Périaux, Y. Kuznetsov, O. B. Widlund. AMS (1994) 157–175

[CH11] Chan, T. F., Mathew, T. P.: Domain decomposition algorithms. Acta Numerica. Cambridge University Press. (1994) 61–143

[CH12] Chan, T. F., Mathew, T. P., Shao, J. P.: Efficient variants of the vertex space domain decomposition algorithm. SIAM J. Sci. Comp. **15** No. 6 (1994) 1349–1374

[CH13] Chan, T. F., Resasco, D. C.: A survey of preconditioners for domain decomposition. Technical Report/DCS/RR-414. Yale University (1985)

[CH14] Chan, T. F., Resasco, D. C.: Analysis of domain decomposition preconditioners on irregular regions. In Advances in computer methods for partial differential equations-VI. (Eds.) R. Vichnevetsky, R. Stapleman. IMACS (1987) 317–322

[CH15] Chan, T. F., Shao, J.-P.: Optimal coarse grid size in domain decomposition. Technical Report 93-24. CAM report. UCLA (1993)

[CH16] Chan, T. F., Sharapov, I.: Subspace correction multilevel methods for elliptic eigenvalue problems. Num. Lin. Alg. Applic. **9** No. 1 (2002) 1–20

[CH17] Chan, T. F., Smith, B., Zou, J.: Overlapping schwarz methods on unstructured meshes using nonmatching coarse grids. Numer. Math. **73** No. 2 (1996) 149–167

[CH18] Chan, T. F., Weinan, E., Sun, J.: Domain decomposition interface preconditioners for fourth order elliptic problems. Appl. Numer. Math. **8** (1991) 317–331

[CH19] Chandra, R.: Conjugate gradient methods for partial differential equations. Report 129, Computer Science Department, Yale University (1978)

[CH20] Chartier, P., Philippe, B.: A parallel shooting technique for solving dissipative ODEs. Computing **51** (1993) 209–236

[CH21] Chatelin, F.: Spectral approximation of linear operators. Academic Press (1983)

[CH22] Chen, H., Lazarov, R. D.: Domain splitting algorithm for mixed finite element approximations to parabolic problems. East-West J. Numer. Math. **4**, No. 2 (1996) 121–135

[CH23] Cheng, H.: Iterative solution of elliptic finite element problems on partially refined meshes and the effect of using inexact solvers. Technical Report 649, Dept. of Comp. Sci., Courant Institute, New York University. (1993)

[CH24] Cheng, H.: Multilevel Schwarz methods with partial refinement. PhD thesis. Technical Report 654, Dept. of Comp. Sci., Courant Institute, New York University. (1994)

[CH25] Chevalier, P., Nataf, F.: Symmetrized method with optimized second order conditions for the Helmholtz equation. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. AMS Contemporary Mathematics **218** (1998)

[CH26] Chin, R. C. Y., Hedstrom, G. W., McGraw, J. R., Howes, F. A.: Parallel computation of multiple scale problems. In New computing environments: Parallel, Vector and Systolic. (Eds.) A. Wouk. SIAM (1986)

[CH27] Chorin, A. J.: Numerical solution of the Navier-Stokes equations. Math. Comp. **22** (1968) 745–762

[CH28] Chorin, A. J.: On the convergence of discrete approximations of the Navier-Stokes equations. Math. Comp. **23** (1969) 341–353

[CH29] Chorin, A. J., Marsden, J. E.: A mathematical introduction to fluid mechanics. Springer-Verlag (1990)

[CI] Ciarlet, P. G.: Discrete maximum principle for finite difference operators. Aequationes Math. **4** (1970) 338–353

[CI2] Ciarlet, P. G.: The finite element method for elliptic problems. North-Holland (1978)

[CI3] Ciarlet, P. G.: Mathematical elasticity. North-Holland (1988)

[CI4] Ciarlet, P. G.: Introduction to numerical linear algebra and optimization. Cambridge University Press (1989)

[CI5] Ciarlet, P. G., Raviart, P.-A.: Maximum principle and uniform convergence for the finite element method. Comp. Methods Appl. Mech. Engrg. **2** (1973) 17–31

[CI6] Ciarlet Jr., P., Chan, T., Szeto, W. K.: On the optimality of the median cut spectral bisection graph partitioning method. Technical Report. UCLA. CAM 93-14 (1993)

[CI7] Ciarlet Jr., P., Lamour, F.: Spectral partitioning methods and greedy partitioning methods: A comparison on finite element graphs. Technical Report. CAM 94-9. UCLA (1994)

[CI8] Ciarlet Jr., P., Lamour, F., Smith, B. F.: On the influence of the partitioning scheme on the efficiency of overlapping domain decomposition methods. Fifth symposium on the frontiers of massively parallel computation. (1995)

[CI9] Ciccoli, M. C.: Adaptive domain decomposition algorithms and finite volume-finite element approximation for advection-diffusion equations. **11** No. 4 (1996) 299–341

[CO] Coddington, E., Levinson, N.: Theory of ordinary differential equations. McGraw Hill (1984)

[CO2] Collino, F., Delbue, G., Joly, P., Piacentini, A.: A new interface condition in nonoverlapping domain decomposition. Comp. Methds. Appl. Mech. Engrg. **148** (1997) 195–207

[CO3] Collino, F., Ghanemi, S., Joly, P.: Domain decomposition method for harmonic wave propagation: A general presentation. Comp. Methds. Appl. Mech. Engrg. **184** (2000) 171–211

[CO4] Conceição, D.: Balancing domain decomposition preconditioners for nonsymmetric problems. Serie C 46, Instituto de matemática pura e aplicada, Brazil. (2006)

[CO5] Conceição, D., Goldfeld, P., Sarkis, M. V.: Robust two level lower order preconditioners for a higher-order Stokes discretization with highly discontinuous viscosities. 7th international conference on high performance computing in computational sciences. Lecture notes in computer sciences and engineering. Springer (2007)

[CO6] Cottle, R. W.: Manifestations of the Schur complement. Lin. Alg. and its Applic. **8** (1974) 189–211

[CO7] Courant, R., Hilbert, D.: Methods of mathematical physics, Vol. 2. Wiley Interscience (1962)

[CO8] Cowsar, L. C.: Domain decomposition methods for nonconforming finite elements of Lagrange type. Sixth copper mountain conference on multigrid methods. **2** No. 3224 (1993) 93–109

[CO9] Cowsar, L. C.: Dual variable Schwarz methods for mixed finite elements. Technical Report TR93-09, Dept. Math. Sciences. Rice University (1993)

[CO10] Cowsar, L. C., Mandel, J., Wheeler, M. F.: Balancing domain decomposition for mixed finite elements. Math. Comp. **64** No. 211 (1995) 989–1015

[CO11] Cowsar, L. C., Weiser, A., Wheeler, M. F.: Parallel multigrid and domain Decomposition algorithms for elliptic Equations. Fifth conference on domain decomposition methods for partial differential equations. (Eds.) T. Chan, D. E. Keyes, G. Meurant, J. Scroggs, R. Voigt. SIAM (1992)

[CO12] Cowsar, L. C., Wheeler, M. F.: Parallel domain decomposition method for mixed finite elements for elliptic partial differential equations. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[CR] Crabtree, D. E., Haynsworth, E. V.: An identity for the Schur complement of a matrix. Proc. Amer. Math. Soc. **22** (1969) 364–366

[CR2] Crank, J.: Free and moving boundary problems. Oxford Univ. Press (1984)

[CR3] Crank, J.: Mathematics of diffusion. Oxford Univ. Press (1989)

[CR4] Cryer, C. W.: The solution of a quadratic programming problem using systematic over-relaxation. SIAM J. Control **9** (1971) 385–392

[DA] Dagan, G.: Flow and transport in porous formations. Springer-Verlag (1989)

[DA2] Dautray, R., Lions, J. L.: Mathematical analysis and numerical methods for science and technology: Functional and variational methods. Springer-Verlag (1990)

[DA3] Davidson, E. R.: The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices. J. Comp. Phys. **17** (1975) 817–825

[DA4] Dawson C. N., Du, Q.: A domain decomposition method for parabolic equations based on finite elements. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) R. Glowinski, Y. Kuznetsov, G. Meurant, O. B. Widlund. SIAM (1991)

[DA5] Dawson C. N., Du, Q., Dupont, T. F.: A finite difference domain decomposition algorithm for numerical solution of the heat equation. Math. Comp. **57** No. 195 (1991)

[DA6] Dawson C. N., Dupont, T. F.: Explicit-implicit conservative Galerkin domain decomposition procedures for parabolic problems. Math. Comp. **58** No. 197 (1992) 21–34

[DE] De la Bourdonnaye, A., Farhat, C., Macedo Puppin, A., Magoulés, F., Roux, F.-X.: A nonoverlapping domain decomposition method for exterior Helmholtz problems. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. AMS Contemporary Masthematics **218** (1998) 42–66

[DE2] De Roeck, Y.-H.: A local preconditioner in a domain decomposed method. Technical Report TR89/10, Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique. Toulouse, France (1989)

[DE3] De Roeck, Y.-H., Le Tallec, P.: Analysis and test of a local domain decomposition preconditioner. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[DE4] Dean, E. J., Dinh, Q. V., Glowinski, R., He, J., Pan, T. W., Périaux, J.: Least squares-domain imbedding methods for Neumann problems: Applications to fluid dynamics. Fifth conference on domain decomposition methods for partial differential equations. (Eds.) T. Chan, D. E. Keyes, G. Meurant, J. Scroggs, R. Voigt. SIAM (1992) 451–475

[DE5] Debit, N., Garbey, M., Hoppe, R. H. W., Keyes, D. E., Kuznetsov, Y. A., Périaux, J.: (Eda.) Domain decomposition methods in science and engineering. Thirteenth international conference. CIMNE (2002)

[DE6] Demmel, J.: Applied numerical linear algebra. SIAM (1997)

[DE7] Dennis, J. E., Schnabel, R. B.: Numerical methods for unconstrained optimization and nonlinear equations. Prentice Hall. (1983)

[DE8] Despres, B.: Methodes de decomposition de domaine pour les problemes de propagation d'ondes en regime harmoniques. Ph.d. Thesis, University of Paris, IX, Dauphine (1991)

[DE9] Despres, B.: Domain decomposition method and the Helmholtz problem II. Second international conference on mathematical and numerical aspects of wave propagation. SIAM (1993) 197–206

[DE10] Despres, B., Joly, P., Roberts, J. E.:: A domain decomposition method for the harmonic Maxwell equations. Iterative methods in linear algebra. North-Holland (1992) 475–484

[DH] D'Hennezel, F.: Domain decomposition method with nonsymmetric interface operator. Fifth conference on domain decomposition methods for partial differential equations. (Eds.) T. Chan, D. E. Keyes, G. Meurant, J. Scroggs, R. Voigt. SIAM (1992)

[DI] Dinh, Q. V., Glowinski, R., Périaux, J.: Solving elliptic problems by domain decomposition methods with applications. In Elliptic problem solvers II. (Eds.) G. Birkhoff, A. Schoenstadt, Academic Press (1984) 395–426

[DI2] Dinh, Q. V., Glowinski, R., He, J., Kwock, V., Pan, T. W., Périaux, J.: Lagrange multiplier approach to fictitious domain methods: Applications to fluid dynamics and electromagnetics Fifth conference on domain decomposition methods for partial differential equations. (Eds.) T. Chan, D. E. Keyes, G. Meurant, J. Scroggs, R. Voigt. SIAM (1992) 151–194

[DO] Dohrmann, C. R.: A preconditioner for substructuring based on constrained energy minimization. SIAM J. Sci. Comp. **25** No. 1 (2003) 246–258

[DO2] Dohrmann, C., Lehoucq, R. B.: A primal based penalty preconditioner for elliptic saddle point systems. SIAM J. Num. Anal. **44** No. 1 (2006) 270–282

[DO3] Dolean, V., Lanteri, S., Nataf, F.: Optimized interface conditions for domain decomposition methods in fluid dynamics. Internat. J. Numer. Methds. Fluids. **40** (2002) 1539–1550

[DO4] Dorr, M.: On the discretization of interdomain coupling in elliptic boundary value problems. Second international symposium on domain decomposition methods for partial differential equations. SIAM (1989)

[DO5] Dostál, Z., Friedlander, A., Santos, S. A.: Solution of coercive and semicoercive contact problems by FETI domain decomposition. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. AMS Contemporary Mathematics **218** (1998) 82–93

[DO6] Dostál, Z., Gomes Neto, F. A. M., Santos, S. A.: Duality based domain decomposition with natural coarse space for variational inequalities. J. Comput. Appl. Math. **126** No. 1–2 (2000) 397–415

[DO7] Dostál, Z., Horák, D.: Scalability and FETI based algorithm for large discretized variational inequalities. Math. Comput. Simulation. (Modelling 2001) **61** No. 3–6 (2003) 347–357

[DO8] Douglas, C. C., Mandel, J.: Abstract theory for the domain reduction method. Computing. **48** (1992) 73–96

[DO9] Douglas Jr., J.: On the numerical solution of $U_{xx} + U_{yy} = U_t$ by implicit methods. J. Soc. Ind. Appl. Math. **3** (1955) 42–65

[DO10] Douglas Jr., J.: Alternating direction methods for three space variables. Numer. Math. **4** (1961) 41–63

[DO11] Douglas Jr., J., Ewing, R. E., Wheeler, M. F.: The approximation of the pressure by a mixed method in the simulation of miscible displacement. R.A.I.R.O Num. Anal. **17** No. 1 (1983) 17–33

[DO12] Douglas Jr., J., Gunn, J. E.: A general formulation of alternating direction method: Part I. Parabolic and hyperbolic problems. Numer. Math. **6** (1964) 428–453

[DO13] Douglas Jr., J., Paes Leme, P. J., Roberts, J. E., Wang, J.: A parallel iterative procedure applicable to the approximate solution of second order partial differential equations by mixed finite element methods. Numer. Math. **65** (1993) 95–108

[DO14] Douglas Jr., J., Roberts, J. E.: Numerical methods for a model for compressible miscible displacement in porous media. Math. Comp. **41** No. 164 (1983) 441–459

[DO15] Douglas Jr., J., Roberts, J. E.: Global estimates for mixed methods for 2nd order elliptic equations. Math. Comp. **44** (1985) 39–52

[DO16] Douglas Jr., J., Russel, T. F.: Numerical methods for convection dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedure. SIAM J. Numer. Anal. **19** (1982) 871–885

[DO17] Douglas Jr., J., Wang, J.: An absolutely stabilized finite element method for the Stokes problem. Math. Comp. **52** (1989) 495–508

[DO18] Douglas Jr., J., Yang, D. Q.: Numerical experiments of a nonoverlapping domain decomposition method for partial differential equations. Numerical Analysis: A. R. Mitchell 75th birthday volume. (Eds.) D. Griffiths and G. A. Watson. World Scientific Press. (1996)

[DR] Dryja, M.: A capacitance matrix method for Dirichlet problem on polygonal region. Numer. Math. **39** (1982) 51–64

[DR2] Dryja, M.: A finite element capacitance method for elliptic problems on regions partitioned into subregions. Numer. Math. **44** (1984) 153–168

[DR3] Dryja, M.: A method of domain decomposition for 3D finite element problems. First international symposium on domain decomposition methods for partial differential equations. (Eds.) R. Glowinski, G. Golub, G. Meurant, J. Périaux. SIAM (1988)

[DR4] Dryja, M.: An additive Schwarz algorithm for two and three dimensional finite element problems. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989)

[DR5] Dryja, M.: Substructuring methods for parabolic problems. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[DR6] Dryja, M.: Additive Schwarz methods for elliptic mortar finite element problems with a new coarse space. East-West J. Numer. Math. **5** No. 2 (1997) 79–98

[DR7] Dryja, M., Proskurowski, W.: On preconditioning mortar discretizations of elliptic problems. Num. Lin. Alg. with Appl. **10** (2003) 65–82

[DR8] Dryja, M., Proskurowski, W.: A generalized FETI-DP method for the mortar discretization of elliptic problems. Fourteenth conference on domain decomposition methods in science and engineering. I. Herrera et al (Eds.) UNAM Publ. Mexico (2003) 257–264

[DR9] Dryja, M., Sarkis, M. V., Widlund, O. B.: Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. Numer. Math. **72** (1996) 313–348

[DR10] Dryja, M., Smith, B., Widlund, O.: Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions. SIAM J. Numer. Anal. **31**, No. 6. (1994) 1662–1694

[DR11] Dryja, M., Widlund, O. B.: An additive variant of the Schwarz alternating method for the case of many subregions. Technical Report 339, Department of Computer Science, Courant Institute, New York University (1987)

[DR12] Dryja, M., Widlund, O. B.: On the optimality of an additive iterative refinement method. In Proc. Fourth copper mountain conference on multigrid methods. SIAM (1989)

[DR13] Dryja, M., Widlund, O. B.: Some domain decomposition algorithms for elliptic problems. In Iterative methods for large linear systems. (Eds.) L. Hayes, D. Kincaid. Academic Press (1989) 273–291

[DR14] Dryja, M., Widlund, O. B.: Towards a unified theory of domain decomposition algorithms for elliptic problems. Third international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1990)

[DR15] Dryja, M., Widlund, O. B.: Additive Schwarz methods for elliptic finite element problems in three dimensions. Fifth conference on domain decomposition methods for partial differential equations. (Eds.) T. Chan, D. E. Keyes, G. Meurant, J. Scroggs, R. Voigt. SIAM (1992)

[DR16] Dryja, M., Widlund, O. B.: Some recent results on Schwarz type domain decomposition algorithms. Sixth conference on domain decomposition methods for partial differential equations. (Ed.) A. Quarteroni. AMS (1993)

[DR17] Dryja, M., Widlund, O. B.: Domain decomposition algorithms with small overlap. SIAM J. Sci. Comput. **15**, No. 3. (1994) 604–620

[DR18] Dryja, M., Widlund, O. B.: Schwarz methods of Neumann-Neumann type for three dimensional elliptic finite element problems. Comm. Pure Appl. Math. **48** No. 2 (1995) 121–155

[DU] Duff, I. S., Erisman, A. M., Reid, J. K.: Direct methods for sparse matrices. Oxford science publications. (1986)

[E] E, W., Liu, J.-G.: Projection method I: Convergence and numerical boundary layers. SIAM J. Num. Anal. **32** (1995) 1017–1057

[EI] Eijkhout, V., Vassilevski, P. S.: The role of the strengthened Cauchy-Buniakowskii-Schwarz inequality in multilevel methods. SIAM Review **33** No. 3 (1991) 405–419

[EI2] Eisenstat, S. C., Elman, H. C., Schultz, M. H.: Variational iterative methods for nonsymmetric systems of linear equations. SIAM J. Numer. Anal. **20** No. 2 (1983) 345–357

[EI3] Eisenstat, S. C., Walker, H. F.: Globally convergent inexact Newton methods. SIAM J. Opt. **4** (1994) 393–422

[EL] Elliott, C. M., Ockendon, J. R.: Weak and variational methods for free and moving boundary problems. Pitman (1982)

[EL2] Elman, H.: Perturbation of eigenvalues of preconditioned Navier-Stokes operators. SIAM J. Matr. Anal. and Appl. **18** (1997) 733–751

[EL3] Elman, H.: Preconditioning for the steady state Navier-Stokes equations with low viscosity. SIAM J. Sci. Comput. **20** (1999) 1299–1316

[EL4] Elman, H.: Preconditioners for saddle point problems arising in computational fluid dynamics. Appl. Numer. Math. **43** (2002) 333–344

[EL5] Elman, H., Golub, G.: Inexact and preconditioned Uzawa algorithms for saddle point problems. SIAM J. Numer. Anal. **31** (1994) 1645–1661

[EL6] Elman, H., O'Leary, D. P.: Efficient iterative solution of the three dimensional Helmholtz equation. J. Comp. Phys. **142** (1998) 163–181

[EL7] Elman, H., O'Leary, D. P.: Eigenanalysis of some preconditioned Helmholtz problems. Numer. Math. **83** (1999) 231–257

[EL8] Elman, H., Silvester, D.: Fast nonsymmetric iterations and preconditioning for Navier-Stokes equations. SIAM J. Sci. Comp. **17** (1996) 33–46

[EL9] Elman, H., Silvester, D., Wathen, A.: Performance and analysis of saddle point preconditioners for the discrete steady state Navier-Stokes equations. Numer. Math. **90** (2002) 641–664

[EN] Enquist, B., Zhao, H.-K.: Absorbing boundary conditions for domain decomposition. Appl. Numer. Math. **27** No. 4 (1998) 341–365

[ER] Ernst, O.: A finite element capacitance matrix method for exterior Helmholtz problems. Numer. Math. **75** No. 2 (1996) 175–204

[ER2] Ernst, O., Golub, G.: A domain decomposition approach to solving the Helmholtz equation with a radiation boundary condition. Technical Report 92-08, Numerical analysis project, Computer Science Dept., Stanford University. (1992)

[EV] Evans, L. C.: Partial differential equations. Graduate studies in mathematics **19**. AMS (1998)

[EW] Ewing, R. E.: Domain decomposition techniques for efficient adaptive local grid refinement. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989)

[EW2] Ewing, R. E.: Mathematical modeling and simulation for applications of fluid flow in porous media. Current and future directions in applied mathematics. (Eds.) M. Alber, B. Hu, J. Rosenthal. Birkhauser (1997) 161–182

[EW3] Ewing, R. E., Jacobs, P., Parashkevov, R., Shen, J.: Applications of adaptive grid refinement methods. Advances in numerical partial differential equations and optimization. (Eds.) S. Gomez, J. P. Hennart, R. A. Tapia. SIAM (1991) 76–100

[EW4] Ewing, R. E., Koebbe, J. V., Gonzalez, R., Wheeler, M. F.: Mixed finite element methods for accurate fluid velocities. Finite elements in fluids **4**. Wiley (1985) 233–249

[EW5] Ewing, R. E., Lazarov, R. D., Pasciak, J. E., Vassilevski, P. S.: Domain decomposition type iterative techniques for parabolic problems on locally refined grids. SIAM J. Numer. Anal. **30** (1993) 1537–1557 Math. Comp. **56** (1991) 437–461

[EW6] Ewing, R. E., Lazarov, R. D., Vassilevski, P. S.: Local refinement techniques for elliptic problems on cell centered grids, I: Error analysis. Math. Comp. **56** (1991) 437–461

[EW7] Ewing, R. E., Lazarov, R. D., Vassilevski, P.: Local refinement techniques for elliptic problems on cell centered grids, II: Optimal order two grid methods. Numer. Lin. Alg. with Appl. **1** (1994) 337–368

[EW8] Ewing, R. E., Wang, J.: Analysis of the Schwarz algorithm for mixed finite element methods. RAIRO, Math. Modell. Numer. Anal. **26** No. 6 (1991) 739–756

[EW9] Ewing, R. E., Wang, J.: Analysis of multilevel decomposition iterative methods for mixed finite element methods. RAIRO, Math. Modell. Numer. Anal. **28** No. 4 (1994) 377–398

[FA] Faille, I., Flauraud, E., Nataf, F., Schneider, F., Willien, F.: Optimized interface conditions for sedimentary basin modelling. In (Eds.) N. Debit, M. Garbey, R. H. W. Hoppe, D. E. Keyes, Y. A. Kuznetsov, J. Périaux. Domain decomposition methods in science and engineering. Thirteenth international conference. CIMNE (2002) 463–470

[FA2] Farhat, C.: A Lagrange multiplier based divide and conquer finite element algorithm. J. Comput. System Engg. **2** (1991) 149–156

[FA3] Farhat, C.: A saddle point principle domain decomposition method for the solution of solid mechanics problems. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992) 271–292

[FA4] Farhat, C., Chandesris, M.: Time-decomposed parallel time integrators: Theory and feasibility studies for fluid, structure and fluid-structure applications. Int. J. Num. Meth. Engg. **58** (2003) 1397–1434

[FA5] Farhat, C., Chen, P.-S., Mandel, J.: A scalable Lagrange multiplier based domain decomposition method for time dependent problems. Internat. J. Numer. Methds. Engrg. **38** (1995) 3831–3853

[FA6] Farhat, C., Chen, P.-S., Mandel, J., Roux, F.-X.: The two level FETI method part II: Extension to shell problems, parallel implementation and performance results. Comput. Methds. Appl. Mech. Engrg. **155** (1998) 153–179

[FA7] Farhat, C., Chen, P.-S., Risler, F., Roux, F.-X.: A unified framework for accelerating the convergence of iterative substructuring methods with Lagrange multipliers. Internat. J. Numer. Methds. Engrg. **42** (1998) 257–288

[FA8] Farhat, C., Geradin, M.: On a component mode synthesis method and its application to incompatible structures. Computers and structures **51** (1994) 459–473

[FA9] Farhat, C., Lesoinne, M.: Automatic partitioning of unstructured meshes for parallel solution of problems in computational mechanics. Int. J. Num. Meth. Eng. **36** No. 5 (1993) 745–764

[FA10] Farhat, C., Lesoinne, M., Le Tallec, P., Pierson, K., Rixen, D.: FETI-DP: A dual primal unified FETI method- Part I: A faster alternative to the two level FETI method. Int. J. Numer. Meth. Engrg. **50** (2001) 1523–1544

[FA11] Farhat, C., Lesoinne, M., Pierson, K.: A scalable dual-primal domain decomposition method. Num. Lin. Alg. Appl. **7** (2000) 687–714

[FA12] Farhat, C., Macedo Puppin, A., Lesoinne, M.: A two level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems. Numer. Math. **85** No. 2 (2000) 283–303

[FA13] Farhat, C., Macedo Puppin, A., Tezaur, R.: FETI-H: A scalable domain decomposition method for high frequency exterior Helmholtz problems. In (Eds.) C.-H. Lai, P. E. Bjørstad, M. Cross, O. B. Widlund. Domain decomposition methods in science and engineering: Eleventh international conference. www.ddm.org (1999) 231–241

[FA14] Farhat, C., Mandel, J.: Scalable substructuring by Lagrange multipliers in theory and practice. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[FA15] Farhat, C., Mandel, J., Roux, F. X.: Optimal convergence properties of the FETI domain decomposition method. Comp. Meth. Appl. Mech. Engg. **115** (1994) 365–385

[FA16] Farhat, C., Roux, F. X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. Int. J. Num. Meth. Eng. **91** (1991)

[FA17] Farhat, C., Roux, F. X.: An unconventional domain decomposition method for an efficient parallel solution of large scale finite element systems. SIAM J. Sci. Comp. **13** (1992) 379–396

[FA18] Farhat, C., Roux, F. X.: Implicit parallel processing in structural mechanics. Comput. Mech. Adv. **2** (1994) 1–124

[FE] Feng, X.: Interface conditions and nonoverlapping domain decomposition methods for a fluid-solid interface problem. Tenth international conference on domain decomposition. (Eds.) J. Mandel, C. Farhat, X.-C. Cai. AMS Contemporary Mathematics **218**. (1998) 417–424

[FE2] Feng, X., Karakashian, O.: Two level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems. SIAM J. Numer. Anal. **39** No. 4 (2001) 1343–1365

[FE3] Ferket, P. J. J., Reusken, A. A.: A finite difference discretization method for elliptic problems on composite grids. Computing **56** (1996) 343–369

[FI] Fiedler, M.: Algebraic connectivity of graphs. Czech. Math. J. **23** (1973) 298–305

[FI2] Fiedler, M.: A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory. Czech. Math. J. **25** (1975) 619–633

[FI3] Finogenov, S. A., Kuznetsov, Y. A.: Two stage fictitious components method for solving the Dirichlet boundary value problem. Sov. J. Numer. Anal. Math. Modell. **3** No. 4 (1988) 301–323

[FI4] Fischer, P.: An overlapping Schwarz method for spectral element solution of the incompressible Navier-Stokes equations. J. Comput. Phys. **133** (1997) 84–101

[FI5] Fischer, P., Ronquist, E. M.: Spectral element methods for large scale parallel Navier-Stokes calculations. Comput. Methods Appl. Mech. Engrg. **116** (1994) 69–76

[FO] Fortin, M., Aboulaich, R.: Schwarz's decomposition method for incompressible flow problems. First international symposium on domain decomposition methods for partial differential equations. (Eds.) R. Glowinski, G. Golub, G. Meurant, J. Périaux. SIAM (1988)

[FO2] Fox, G.: A review of automatic load balancing and decomposition methods for the hypercube. Numerical algorithms for modern parallel computers. (Eds.) M. Schultz. Springer-Verlag (1988)

[FR] Fragrakis, Y., Papadrakakis, M.: The mosaic of high performance domain decomposition methods for structural mechanics: Formulation, interrelation and numerical efficiency of primal and dual methods. Computer Methods in Applied Mechanics and Engineering **192** (2003) 3799–3830

[FR2] Franca, L. P., Frey . S. L.: Stabilized finite element methods II: The incompressible Navier-Stokes equations. Comp. Meth. Appl. Mech. Engrg. **99** (1992) 209–233

[FR3] Franca, L. P., Frey, S. L., Hughes, T. J. R.: Stabilized finite element methods I: Application to the advective diffusive model. Comp. Meth. Appl. Mech. Engrg. **95** No. 2 (1992) 253–276

[FR4] Franca, L. P., Hughes, T. J. R.: Convergence analysis of Galerkin least squares method for nonsymmetric advective-diffusive forms of the Stokes and incompressible Navier-Stokes equations. Comp. Math. **105** (1993) 285–298

[FR5] Freund, R. H., Golub, G. H., Nachtigal, N.: Iterative solution of linear systems. Acta Numerica. Cambridge University Press (1992) 57–100

[FR6] Friedman, A.: Variational principles and free boundary problems. Krieger Publ. Co. (1988)

[FR7] Frommer, A., Szyld, D.: Weighted maximum norms, splittings, and overlapping additive Schwarz iterations. Num. Math. **83** (1999) 259–278

[FR8] Frommer, A., Szyld, D.: An algebraic convergence theory for restricted additive Schwarz methods using weighted maximum norms. SIAM J. Num. Anal. **39** (2001) 463–479

[FU] Funaro, D., Quarteroni, A., Zanolli, P.: An iterative procedure with interface relaxation for domain decomposition methods. SIAM J. Numer. Anal. **25** (1988) 1213–1236

[GA] Galvis, J., Sarkis, M. V.: Balancing domain decomposition methods for mortar coupling Stokes-Darcy systems. In (Eds.) O. B. Widlund, D. E. Keyes. Domain decomposition methods in science and engineering XVI. Lecture notes in computational science and engineering **55** Springer (2006) 373–380

[GA2] Gander, M. J.: Overlapping Schwarz for parabolic problems. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997) 97–104

[GA3] Gander, M. J., Golub, G. H.: A nonoverlapping optimized Schwarz method which converges with an arbitrarily weak dependence on h. In (Eds.) E. Herrera, D. E. Keyes, O. B. Widlund, R. Yates. Domain decomposition methods in science and engineering. Fourteenth international conference. Mexico (2003) 281–288

[GA4] Gander, M. J., Halpern, L., Nataf, F.: Optimal Schwarz waveform relaxation for the one dimensional wave equation, SIAM J. Numer. Anal. **41** No. 5 (2003) 1643–1681

[GA5] Gander, M. J., Magoulés, F., Nataf, F.: Optimized Schwarz methods without overlap for the Helmholtz equation. SIAM J. Sci. Comp. **24** No. 1 (2002) 38–60

[GA6] Gander, M. J., Stuart, A. M.: Space-time continuous analysis of waveform relaxation for the heat equation. SIAM J. Sci. Comput. **19** No. 6 (1998) 2014–2031

[GA7] Gander, M. J., Vandewalle, S.: On the superlinear and linear Convergence of the parareal algorithm. Proceedings of the 16th International Conference on Domain Decomposition Methods. (2005)

[GA8] Garbey, M.: Domain decomposition to solve layers and singular perturbation problems. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[GA9] Garbey, M.: Domain decomposition to solve layers and asymptotics. SIAM J. Sci. Comp. **15** No. 4 (1994) 866–891

[GA10] Garbey, M.: A Schwarz alternating procedure for singular perturbation problems. SIAM J. Sci. Comput. **17** (1996) 1175–1201

[GA11] Garbey, M., Kaper, H. G.: Heterogeneous domain decomposition for singularly perturbed elliptic boundary value problems. SIAM J. Numer. Anal. **34** No. 4 (1997) 1513–1544

[GA12] Garbey, M., Kuznetsov, Y. A., Vassilevski, Y. V.: A parallel Schwarz method for a convection diffusion problem. SIAM J. Sci. Comp. **22** No. 3 (2000) 891–916

[GA13] Garbey, M., Tromeur-Dervout, D.: Operator splitting and domain decomposition for multiclusters. Proc. Parallel CFD99. (Eds.) D. Keyes. (1999)

[GA14] Gastaldi, F., Gastaldi, L., Quarteroni, A.: Adaptive domain decomposition methods for advection dominated equations. East-West J. Numer. Math. **4** (1996) 165–206

[GA15] Gastaldi, F., Quarteroni, A., Sacchi-Landriani, G.: On the coupling of two dimensional hyperbolic and elliptic equations: analytical and numerical approach. Domain decomposition methods for partial differential equations. SIAM (1990) 22–63

[GE] Gear, W. C.: Numerical initial value problems in ordinary differential equations. Prentice Hall (1971)

[GE2] Geist, A., Beguelin, A., Dongarra, J., Jiang, W., Manchek, R., Sunderam, V. S.: PVM: Parallel Virtual Machine: A user's guide and tutorial for network parallel computing. MIT press (1994)

[GE3] Genseberger, M.: Domain decomposition in the Jacobi-Davidson method for eigenproblems. PhD thesis. Utrecht University (2001)

[GE4] George, A.: Nested dissection of a regular finite element mesh. SIAM J. Numer. Anal. **10** (1973) 345–363

[GE5] George, A., Liu, J.: Computer solution of large sparse positive definite systems. Prentice Hall (1981)

[GE6] George, P. L.: Automatic mesh generation. Wiley (1991)

[GI] Gilbarg, D., Trudinger, N. S.: Elliptic partial differential equations of second order. Springer-Verlag (1983)

[GI2] Gill, P. E., Murray, W., Wright, M. H.: Practical optimization. Academic press (1982)

[GI3] Girault, V., Raviart, P.-A.: Finite element methods for Navier-Stokes equations. Springer-Verlag (1986)

[GL] Glowinski, R.: Numerical methods for nonlinear variational problems. Springer-Verlag (1984)

[GL2] Glowinski, R., Dinh, Q. V., Periaux, J.: Domain decomposition methods for nonlinear problems in fluid dynamics. Comp. Meth. Appl. Mech. Engng. **40** (1983) 27–109

[GL3] Glowinski, R., Hesla, T. I., Joseph, D. D., Pan, T. W.: Distributed Lagrange multiplier methods for particulate flows. Computational science for the 21st century. (Eds.) M. O. Bristeau, G. J. Etgen, W. Fitzgibbon, J. L. Lions, J. Périaux, M. F. Wheeler. Wiley (1997) 270–279

[GL4] Glowinski, R., Hesla, T. I., Joseph, D. D., Pan, T. W.: A fictitious domain method with distributed Lagrange multipliers for the numerical simulation of particulate flows. Domain decomposition 10. (Eds.) J. Mandel, C. Farhat, X.-C. Cai. AMS (1998) 121–137

[GL5] Glowinski, R., Kinton, W., Wheeler, M. F.: Acceleration of domain decomposition methods for mixed finite elements by multilevel methods. Third international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. b. Widlund. SIAM (1990)

[GL6] Glowinski, R., Kuznetsov, Y. A.: On the solution of the Dirichlet problem for linear elliptic operators by a distributed Lagrange multiplier method. C. R. Acad. Sci. Paris Sér i Math. **327** No. 7 (1998) 693–698

[GL7] Glowinski, R., Le Tallec, P.: Augmented Lagrangian and operator splitting methods in nonlinear mechanics. SIAM (1989)

[GL8] Glowinski, R., Le Tallec, P.: Augmented Lagrangian interpretation of the nonoverlapping Schwarz alternating method. Third international symposium on domain decomposition methods for partial differential equations. SIAM (1990) 224–231

[GL9] Glowinski, R., Lions, J. L., Pironneau, O.: Decomposition of energy spaces and applications. C. R. A. S. Paris (1999)

[GL10] Glowinski, R., Lions, J. L., Tremolieres, R.: Numerical analysis of variational inequalities. North-Holland (1981)

[GL11] Glowinski, R., Pan, T. W., Hesla, T. I., Joseph, D. D., Périaux, J.: A distributed Lagrange multiplier/fictitious domain method for flows around moving rigid bodies: Applications to particulate flow. Int. J. Num. Meth. Fluids. **30** No. 8 (1999) 1043–1066

[GL12] Glowinski, R., Pan, T. W., Périaux, J.: A fictitious domain method for Dirichlet problem and applications. Comp. Meth. Appl. Mech. Engg. **111** (1994) 283–303

[GL13] Glowinski, R., Périaux, J., Terrasson, G.: On the coupling of viscous and inviscid models for compressible fluid flows via domain decomposition. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989)

[GL14] Glowinski, R., Wheeler, M. F.: Domain decomposition and mixed finite elements for elliptic problems. First international symposium on domain decomposition methods for partial differential equations. (Eds.) R. Glowinski, G. Golub, G. Meurant, J. Périaux. SIAM (1988)

[GM] Gmati, N., Farhat, C., Hetmaniuk, U.: An efficient substructuring method for analyzing acoustics in a concentric hole-cavity resonator. Mathematical and numerical aspects of wave propagation (Santiago de Compostela). SIAM (2000) 817–821

[GO] Goldfeld, P.: Balancing Neumann-Neumann preconditioners for mixed formulations of almost incompressible linear elasticity. PhD thesis. Courant Institute of Mathematical Sciences. TR-847, Dept. of Comp. Science. (2003)

[GO2] Goldfeld, P., Pavarino, L. F., Widlund, O. B.: Balancing Neumann-Neumann preconditioners for mixed approximations of heterogeneous problems in linear elasticity. Numer. Math. **95** No. 2 (2003) 283–324

[GO3] Golub, G., Mayers, D.: The use of preconditioning over irregular regions. Computational Methods in Applied Sciences and Engineering, VI. (Eds.) R. Glowinski, J. L. Lions. North Holland (1984)

[GO4] Golub, G., Van Loan, C. F.: Matrix computations. Johns Hopkins University Press, 2nd edition (1989)

[GO5] Gonçalves, E., Mathew, T. P., Sarkis, M., Schaerer, C.: A robust preconditioner for the Hessian system in elliptic optimal control problems. Proceedings of 17th conference on domain decomposition methods. Springer-Verlag (2007)

[GO6] Gonçalves, E., Mathew, T. P., Sarkis, M., Schaerer, C.: A robust preconditioner for an augmented Lagrangian based elliptic optimal control problem. Preprint (2007)

[GO7] Goossens, S., Cai, X.-C., Roose, D.: Overlapping nonmatching grid method: Some preliminary studies. Domain decomposition methods 10. (Eds.) J. Mandel, C. Farhat, X.-C. Cai. AMS (1998) 254–261

[GO8] Goovaerts, D.: Domain decomposition methods for elliptic partial differential equations. Ph.D. Thesis. Dept. Comp. Sci., Catholic University of Leuven (1989)

[GO9] Goovaerts, D., Chan, T., Piessens, R.: The eigenvalue spectrum of domain decomposed preconditioners. Appl. Numer. Math. **8** (1991) 389–410

[GO10] Gopalakrishnan, J.: On the mortar finite element method. Ph. D Thesis. Texas A&M University (1999)

[GO11] Gopalakrishnan, J., Pasciak, J.: Multigrid for the mortar finite element method. SIAM J. Numer. Anal. **37** (2000) 1029–1052

[GO12] Gopalakrishnan, J., Pasciak, J.: Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations. Math. Comp. **72** No. 241 (2003) 1–15

[GR] Grama, A., Gupta, A., Karypis, G., Kumar, V.: Introduction to parallel computing. Addison-Wesley (2003)

[GR2] Greenbaum, A.: Iterative methods for solving linear systems. SIAM Frontiers in applied mathematics. (1997)

[GR3] Griebel, M.: Multilevel algorithms considered as iterative methods on semidefinite systems. SIAM J. Sci. Stat. Comp. **15** No. 3 (1994) 547–565

[GR4] Griebel, M., Oswald, P.: Remarks on the abstract theory of additive and multiplicative Schwarz algorithms. SFB 342/29/91A. Technische universität München (1993)

[GR5] Griebel, M., Oswald, P.: On additive Schwarz preconditioners for sparse grid discretizations. Numer. Math. **66** (1994) 449–464

[GR6] Griebel, M., Oswald, P.: On the abstract theory of additive and multiplicative Schwarz algorithms. Numer. Math. **70** No.2 (1995) 163–180

[GR7] Griebel, M., Starke, G.: Multilevel preconditioning based on discrete symmetrization for convection diffusion equations. J. Comp. and Appl. Math. **83** (1997) 165–183

[GR8] Grisvard, P.: Elliptic problems on nonsmooth domains. Pitman Publishing. Boston (1985)

[GR9] Gropp, W. D.: Local uniform mesh refinement on loosely coupled parallel processors. Comput. Math. Appl. **15** (1988) 375–387

[GR10] Gropp, W. D.: Parallel computing and domain decomposition. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[GR11] Gropp, W. D., Kaushik, D., Keyes, D. E., Smith, B. F.: Performance modelling and tuning of an unstructured mesh CFD application. Proceedings of SC2000. IEEE Computer Society (2000)

[GR12] Gropp, W. D., Keyes, D. E.: Domain decomposition on parallel computers. Impact Comput. Sci. Eng. **1** (1989) 421–439

[GR13] Gropp, W. D., Keyes, D. E.: Domain decomposition as a mechanism for using asymptotic methods. Asymptotic and numerical methods for partial differential equations with critical parameters. (Eds.) H. Kaper, M. Garbey. **384** NATO ASI Series C. 93–106 (1993)

[GR14] Gropp, W. D., Keyes, D. E., McInnes, L. C., Tidriri, M. D.: Parallel implicit PDE computations: Algorithms and software. In (Eds.) A. Ecer et al. Parallel CFD (1997)

[GR15] Gropp, W., Lusk, E., Skjellum, A.: Using MPI: Portable parallel programming with Message Passing Interface. MIT Press, 2nd edition (1999)

[GR16] Gropp, W. D., Smith, B. F.: Experiences with domain decomposition in three dimensions: Overlapping Schwarz methods. Sixth international conference on domain decomposition. AMS, Contemporary Mathematics **157** (1994)

[GR17] Gropp, W. D., Smith, B. F.: Scalable, Extensible, and Portable Numerical Libraries. Proceedings of scalable parallel libraries conference. IEEE (1994) 87–93

[GU] Gundolf, H., Langer, U., Meyer, A.: A new approach to the Dirichlet domain decomposition method. In (Ed.) S. Hengst. Fifth multigrid seminar, Eberswalde. Karl-Weierstrass Institute Report R-MATH-09/90 (1990)

[GU2] Gunzburger, M., Heinkenschloss, M., Lee, H.-K.: Solution of elliptic partial differential equations by an optimization based domain decomposition method. Appl. Math. and Comp. **113** (2000) 111–139

[GU3] Gunzburger, M., Lee, H.-K.: A domain decomposition method for the Navier-Stokes equations. Proceedings of 17th workshop in pure mathematics. Korean Academic Council. Seoul. (1998) 13–33

[GU4] Guo, B., Cao, W.: A preconditioner for the *h*-*p* version of the finite element method in two dimensions. Numer. Math. **75** No. 1 (1996) 59–77

[GU5] Guo, B., Cao, W.: Additive Schwarz methods for the *h*-*p* version of the finite element method in two dimensions. SIAM J. Sci. Comp. **18** No. 5 (1997) 1267–1288

[GU6] Guo, B., Cao, W.: An iterative and parallel solver based on domain decomposition of the *h*-*p* version of the finite element method. J. Comp. Appl. Math. **83** (1997) 71–85

[GU7] Guo, B., Cao, W.: An additive Schwarz method for the *h-p* version of the finite element method in three dimensions. SIAM J. Num. Anal. **35** No.2 (1998) 632–654

[GU8] Guo, B., Cao, W.: Domain decomposition method for *h-p* version finite element method. Comp. Methds. Appl. Mech. Engrg. **157** (1998) 425–440

[GU9] Guo, B., Cao, W.: A preconditioner with inexact element face solvers for the three dimensional *p*-version finite element methods. J. Comp. Appl. Math. **144** (2002) 131–144

[GU10] Gustafsson, B., Hemmingsson-Franden, L.: A fast domain decomposition high order Poisson solver. SIAM J. Sci. Comp. **14** (1999) 223–243

[GU11] Gustafsson, B., Hemmingsson-Franden, L.: Implicit higher order difference methods and domain decomposition. Appl. Numer. Meth. **33** (2000) 493–500

[HA] Haber, E., Ascher, U. M.: Preconditioned all at once methods for large sparse parameter estimation problems. Inverse Problems **17** (2001) 1647–1684

[HA2] Hackbusch, W.: Multigrid methods and applications. Springer-Verlag (1985)

[HA3] Hackbusch, W.: Iterative solution of large sparse linear systems of equations. Springer-Verlag (1994)

[HA4] Hackbusch, W., Trottenberg, U.: Multigrid methods. Springer-Verlag (1982)

[HA5] Hagstrom, T., Tewarson, R., Jazcilevich, A.: Numerical experiments on a domain decomposition algorithm for nonlinear elliptic boundary value problems. Appl. Math. Lett. **1** No. 3 (1988) 299–302

[HA6] Hahn, W., Mittelmann, H.-D.: Some remarks on the discrete maximum principle for finite elements of higher order. Computing **27** (1981) 145–154

[HA7] Hairer, E., Norsett, S. P., Wanner, G.: Solving ordinary differential equations I: Nonstiff problems. Springer-Verlag (1987)

[HA8] Hairer, E., Wanner, G.: Solving ordinary differential equations II: Stiff and differential-algebraic problems. Springer-Verlag (1991)

[HA9] Hart, L., McCormick, S.: Asynchronous multilevel adaptive methods for solving partial differential equations on multiprocessors: Computational analysis. Parallel Computing. **12** (1989) 131–144

[HE] Hebeker, F. K., Kuznetsov, Y. A.: Unsteady convection and convection-diffusion problems via direct overlapping domain decomposition methods. Numer. Methds. part. Diff. Eqns. **14** No. 3 (1998) 387–406

[HE2] Hedstrom, G. W., Howes, F. A.: Domain decomposition for a boundary value problem with a shock layer. Third international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. B. Widlund. SIAM (1990)

[HE3] Heikkola, E., Kuznetsov, Y., Neittaanmaki, P., Toivanen, J.: Fictitious domain methods for the numerical solution of two dimensional scattering problems. J. Comput. Phys. **145** (1998) 89–109

[HE4] Heinkenschloss, M., Nguyen, H.: Balancing Neumann-Neumann methods for elliptic optimal control problems. Proceedings of 15th international conference on domain decomposition. (Eds.) R. Kornhuber, R. W. Hoppe, J. Periaux, O. Pironneau, O. B. Widlund, J. Xu. Lecture notes in computational science and engineering. **40**. Springer-Verlag (2004) 589–596

[HE5] Heinkenschloss, M., Nguyen, H.: Neumann-Neumann domain decomposition preconditioners for linear-quadratic elliptic optimal control problems. SIAM J. Sci. Comp. **28** No. 3 (2006) 1001–1028

[HE6] Hemmingsson, L.: A domain decomposition method for almost incompressible flow. Comput. and Fluids. **25** (1996) 771–789

[HE7] Hendrickson, B., Leland., R.: A multilevel algorithm for partitioning graphs. Technical Report. 93-1301. Sandia National Labs. Albuquerque, NM (1993)

[HE8] Hendrickson, B., Leland., R.: The CHACO user's guide. Technical Report. 93-2339. Sandia National Labs. Albuquerque, NM (1993)

[HE9] Henshaw, W. D.: Automatic grid generation. Acta Numerica (1996) 121–148

[HE10] Henshaw, W., Brislawn, K., Brown, D., Chesshire, G., Pao, K., Quinlan, D., Saltzman, J.: Overture: An object-oriented framework for solving PDEs on overlapping grids. LA-UR-97-4033. Third symposium on composite overset grid and solution technology. Los Alamos, New Mexico (1996)

[HE11] Herrera, I., Keyes, D. E., Widlund, O. B., Yates, R. A.: (Eds.) Domain decomposition methods in science and engineering. Fourteenth international conference. UNAM Publ. Mexico (2003)

[HE12] Hetmaniuk, U., Farhat, C.: A fictitious domain decomposition method for the solution of partially axisymmetric acoustic scattering problems II: Neumann boundary conditions. Internat. J. Numer. Methds. Engrg. **58** No. 1 (2003) 63–81

[HI] Hientzsch, B.: Fast solvers and domain decomposition preconditioners for spectral element discretizations of problems in $H(curl)$. PhD thesis. Courant Institute of Mathematical Sciences, TR-823 Comp. Sci. Dept. (2003)

[HI2] Hiptmair, R., Toselli, A.: Overlapping and multilevel Schwarz methods for vector valued elliptic problems in three dimensions. In (Eds.) P. Bjørstad, M. Luskin. Parallel solution of partial differential equations. IMA Vol. Math. Appl. **120** Springer (2000) 181–208

[HO] Hockney, R. W., Jesshope, C. R.: Parallel computers: Architecture, programming and algorithms. Adam Hilger (1988)

[HO2] Hoffman, C. M.: Geometric approaches to mesh generation. Modeling, mesh generation, and adaptive numerical methods for partial differential equations. IMA **75**. Springer-Verlag (1995)

[HO3] Hoffman, K. H., Zou, J.: Parallel algorithms of Schwarz variant for variational inequalities. Numer. Funct. Anal. Optim. **13** (1992) 449–462

[HO4] Hoffman, K. H., Zou, J.: Parallel efficiency of domain decomposition methods. Parallel Computing **19** (1993) 1375–1391

[HO5] Hoppe, R., Iliash, Y., Kuznetsov, Y., Vassilevski, P. S., Wohlmuth, B.: Analysis and parallel implementation of adaptive mortar element methods. East-West J. Numer. Anal. **6** No. 3 (1998) 223–248

[HO6] Hoppe, R., Kuznetsov, Y.: Overlapping domain decomposition methods with distributed Lagrange multipliers. East-West J. Numer. Anal. **9** No. 4 (2001) 285–293

[HO7] Houston, P., Schwab, C., Suli, E.: SIAM J. Numer. Anal. **39** No. 6 (2002) 2133–2163

[HU] Hu, Q., Zou, J.: A nonoverlapping domain decomposition method for Maxwell's equations in three dimensions. SIAM J. Numer. Anal. **41** (2003) 1682–1708

[HU2] Hu, Q., Zou, J.: Substructuring preconditioners for saddle point problems arising from Maxwell's equations in three dimensions. Math. Comp. **73** No. 245 (2004) 35–61

[HU3] Huang, Y., Xu, J.: A conforming finite element method for overlapping and nonmatching grids. Math. Comp. **72** (2003) 1057–1066

[IH] Ihlenburg, F., Babuška, I.: Finite element solution of the Helmholtz equation with high wave number, Part I: The h version of the FEM. Comp. Math. Engrg. **30** No. 9 (1995) 9–37

[IH2] Ihlenburg, F., Babuška, I.: Finite element solution of the Helmholtz equation with high wave number, Part II: The *h-p* version of the FEM. SIAM J. Num. Anal. **34** No. 1 (1997) 315–358

[IK] Ikeda, T.: Maximum principle in finite element models for convection diffusion phenomena. Lecture notes in numerical and applied analysis **4** North-Holland (1983)

[IS] Isaacson, E., Keller, H.: Analysis of numerical methods. Dover Publications (1994)

[JA] Jacobs, P. G.: Numerical methods for parabolic problems using local time stepping. Ph.D. Thesis. Dept. of Mathematics, University of Wyoming, Laramie (1995)

[JA2] Japhet, C., Nataf, F., Rogier, F.: The optimized order two method. Application to convection diffusion problems. Future generation computer systems. Future **18** (2001)

[JA3] Japhet, C., Nataf, F., Roux, F.-X.: Extension of a coarse grid preconditioner to non-symmetric problems. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. Tenth international conference. AMS Contemporary Mathematics **218** (1998) 279–286

[JE] Jenkins, E. W., Kelley, C. T., Miller, C. T., Kees, C. E.: An aggregation based domain decomposition preconditioner for groundwater flow. Technical Report TR00-13. Dept. of Maths. North Carolina State University (2000)

[JO] John, F.: Partial differential equations. Springer-Verlag (1978)

[JO2] Johnson, C.: Numerical solution of partial differential equations by the finite element method. Cambridge University Press (1987)

[KA] Kang, L. S.: Parallel algorithms and domain decomposition. In Chinese. Wuhan University Press. China (1987)

[KA2] Karafiat, A.: Discrete maximum principle in parabolic boundary value problems. Ann. Polon. Math. **53** (1991) 253–265

[KA3] Karypis, G., Kumar., V.: METIS, Unstructured graph partitioning and sparse matrix ordering system. Version 2.0 Computer Science Dept., Univ. of Minnesota (1995)

[KA4] Kaushik, D. K., Keyes, D. E., Smith, B. F.: NKS methods for compressible and incompressible flows on unstructured grids. Proceedings of the elevent international conference on domain decomposition methods. (Eds.) C.-H. Lai and P. E. Bjorstad and M. Cross and O. Widlund. (1999) 513–520

[KE] Keller, J. B.: Rays, waves and asymptotics. Bull. Amer. Math. Soc. **84** (1978) 727–750

[KE2] Keller, J. B.: One hundred years of diffraction theory. IEEE Trans. Antennas and Prop. **AP 33** (1985) 200-214

[KE3] Keller, J. B.: Semiclassical mechanics. SIAM Rev. **27** (1985) 485–504

[KE4] Kernighan, B. W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal. **49** No. 1 (1970) 291–307

[KE5] Kevorkian, J., Cole, J. D.: Perturbation methods in applied mathematics. Springer-Verlag (1981)

[KE6] Keyes, D. E.: Domain decomposition methods for the parallel computation of reacting flows. Comput. Phys. Commun. (1989)

[KE7] Keyes, D. E., Gropp, W. D.: A comparison of domain decomposition techniques for elliptic partial differential equations and their parallel implementation. SIAM J. Sci. Comput. **8** No. 2 (1987) 166–202

[KE8] Keyes, D. E., Gropp, W. D.: Domain decomposition techniques for the parallel solution of nonsymmetric systems of elliptic boundary value problems. Appl. Num. Math. **6** (1990) 281–301

[KE9] Keyes, D. E., Gropp, W. D.: Domain decomposition with mesh refinement. SIAM J. Sci. Comput. **13** (1992) 967–993

[KE10] Keyes, D. E., Kaushik, D. K., Smith, B. F.: Prospects for CFD on petaflops systems. CFD Review (1997)

[KI] Kim, C., Lazarov, R. D., Pasciak, J., Vassilevski, P.: Multiplier spaces for the mortar finite element method in three dimensions. SIAM J. Numer. Anal. **39** (2001) 519–538

[KI2] Kim, M.-Y., Park, E.-J., Park, J.: Mixed finite element domain decomposition for nonlinear parabolic problems. Comp. Math. Appl. **40** No. 9 (2000) 1061–1070

[KI3] Kimn, J.-H., Sarkis, M. V.: OBDD: Overlapping balancing domain decomposition methods and generalizations to Helmholtz equations. In (Eds.) O. B. Widlund, D. E. Keyes. Domain decomposition methods in science and engineering XVI. Lecture notes in computational science and engineeering **55** (2006) 317–324

[KI4] Kinderlehrer, D., Stampacchia, G.: An introduction to variational inequalities and their applications. SIAM. Classics in applied mathematics.(2000)

[KL] Klawonn, A.: Block triangular preconditioners for saddle point problems with a penalty term. SIAM J. Sci. Comput. **19** No. 1 (1998) 172–184

[KL2] Klawonn, A.: An optimal preconditioner for a class of saddle point problems with a penalty term. SIAM J. Sci. Comput. **19** No. 2 (1998) 540–552

[KL3] Klawonn, A., Pavarino, L.: Overlapping Schwarz methods for elasticity and Stokes problems. Comp. Methds. Appl. Mech. Engrg. **165** No. 1–4 (1998) 233–245

[KL4] Klawonn, A., Pavarino, L.: A comparison of overlapping Schwarz methods and block preconditioners for saddle point problems. Numer. Lin. Alg. Appl. **7** (2000) 1–25

[KL5] Klawonn, A., Rheinbach, O., Widlund, O. B.: Some computational results for dual-primal FETI methods for three dimensional elliptic problems. In (Eds.) R. Kornhuber, R. W. Hoppe, J. Periaux, O. Pironneau, O. B. Widlund, J. Xu. Proceedings of the fifteenth international conference on domain decomposition methods. Lecture notes in computational science and engineering. **40** Springer-Verlag (2004)

[KL6] Klawonn, A., Starke, G.: Block triangular preconditioners for nonsymmetric saddle point problems: Field of values analysis. Numer. Math. **81** (1999) 577–594

[KL7] Klawonn, A., Widlund, O. B.: A domain decomposition method wtih Lagrange multipliers and inexact solvers for linear elasticity. SIAM J. Sci. Comp. **22** No. 4 (2000) 1199–1219

[KL8] Klawonn, A., Widlund, O. B.: FETI and Neumann-Neumann iterative substructuring methods: Connections and new results. Comm. Pure Appl. Math. **54** (2001) 57–90

[KL9] Klawonn, A., Widlund, O. B.: Selecting constraints in dual-primal FETI methods for elasticity in three dimensions. In (Eds.) R. Kornhuber, R. W. Hoppe, J. Periaux, O. Pironneau, O. B. Widlund, J. Xu. Proceedings of the fifteenth international conference on domain decomposition methods. Lecture notes in computational science and engineering. **40** Springer-Verlag (2004)

[KL10] Klawonn, A., Widlund, O. B., Dryja, M.: Dual primal FETI methods for three dimensional elliptic problems with heterogeneous coefficients. SIAM J. Numer. Anal. **40** (2002) 159–179

[KL11] Klawonn, A., Widlund, O. B., Dryja, M.: Dual-primal FETI methods with face constraints. In (Eds.) L. Pavarino, A. Toselli. Recent developments in domain decomposition methods. Lecture notes in computational science and engineering. **23** Springer-Verlag (2002) 27–40

[KN] Knyazev, A.: Convergence rate estimates for iterative methods for mesh symmetric eigenvalue problems Sov. J. Num. Anal. and Math. Modell. **2** No. 5 (1987) 371–396

[KN2] Knyazev, A.: Preconditioned eigensolvers - an oxymoron? ETNA **7** (1998) 104–123

[KN3] Knyazev, A.: Preconditioned eigensolvers. Templates for the solution of algebraic eigenvalue problems. (Eds.) Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. Van der vorst. SIAM (2000) 337–368

[KN4] Knyazev, A.: Towards the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. SIAM. J. Sci. Comp. **23** N0. 2 (2001) 517–541

[KN5] Knyazev, A., Sharapov, I. A.: Variational Rayleigh quotient iteration methods for symmetric eigenvalue problems. East-West J. Num. Math. **1** No. 2 (1993) 121–128

[KN6] Knyazev, A., Skorokhodev, L.: Preconditioned gradient type iterative methods in a subspace for partial generalized symmetric eigenvalue problems. SIAM J. Numer. Anal. **31** No. 4 (1994) 1226–1239

[KO] Konovalov, A. N.: The method of fractional steps for solving the Cauchy problem for the multidimensional wave equation. Dokl. Akad. Nauk. **147** (1962) 25–27

[KO2] Korneev, V. G., Flaherty, J. E., Oden, J. T., Fish, J.: Additive Schwarz algorithms for solving *hp*-version finite element systems on triangular meshes. Appl. Numer. Math. **43** No. 4 (2002) 399–421

[KO3] Korneev, V. G., Jenson, S.: Preconditioning of the *p*-version of the finite element method. Comp. Methds. Appl. Mech. Engrg. **150** No. 1–4 (1997) 215–238

[KO4] Korneev, V. G., Langer, U., Xanthis, L.: On fast domain decomposition solving procedures for *hp*-discretizations of 3D elliptic problems. Comp. Methds. Appl. Math. **3** No. 4 (2003) 536–559

[KO5] Kornhuber, R.: Monotone multigrid methods for elliptic variational inequalities I. Numer. Math. **69** (1994) 167–184

[KO6] Kornhuber, R., Hoppe, R. H. W., Périaux, J., Pironneau, O., Widlund, O. B., Xu, J.: Proceedings of the fifteenth international conference on domain decomposition methods. Lecture notes in computational science and engineering. **40** Springer-Verlag (2004)

[KR] Kron, G.: A set of principles to interconnect the solutions of physical systems. J. Appl. Phys. **268** No. 8 (1953)

[KU] Kumfert, G., Pothen, A.: A multilevel nested dissection algorithm. Unpublished Work. (1995)

[KU2] Kuznetsov, Y. A.: Fictitious component and domain decomposition methods for the solution of eigenvalue problems. (Eds.) R. Glowinski, J. L. Lions. Computing methods in applied sciences and engineering VII. Elsevier Science Publ. (1986) 113–216

[KU3] Kuznetsov, Y. A.: New algorithms for approximate realization of implicit difference schemes. Sov. J. Numer. Anal. Math. Modell. **3** (1988) 99–114

[KU4] Kuznetsov, Y. A.: Multilevel domain decomposition methods. Appl. Numer. Math. **5** (1989)

[KU5] Kuznetsov, Y. A.: Domain decomposition methods for unsteady convection diffusion problems. IXth international conference on computing methods in applied science and engineering. INRIA (Eds.) R. Glowinski, J. L. Lions. Paris (1990) 327–344

[KU6] Kuznetsov, Y. A.: Overlapping domain decomposition methods for finite element problems with elliptic singular perturbed operators. Fourth international symposium on domain decomposition methods for partial differential equations. SIAM (1991)

[KU7] Kuznetsov, Y. A.: Efficient iterative solvers for elliptic problems on nonmatching grids. Russ. J. Numer. Anal. Math. Modeling. **10** No. 3 (1995) 187–211

[KU8] Kuznetsov, Y. A.: Overlapping domain decomposition with nonmatching grids. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[KU9] Kuznetsov, Y. A.: New iterative methods for singular perturbed positive definite matrices. Russ. J. Numer. Anal. Math. Modeling. **15** (2000) 65–71

[KU10] Kuznetsov, Y. A.: Domain decomposition and fictitious domain methods with distributed Lagrange multipliers. In (Eds.) N. Debit, M. Garbey, R. H. W. Hoppe, D. E. Keyes, Y. A. Kuznetsov, J. Périaux. Domain decomposition methods in science and engineering. Thirteenth international conference. CIMNE (2002) 67–75

[KU11] Kuznetsov, Y. A., Manninen, P., Vassilevski, Y.: On numerical experiments with Neumann-Neumann and Neumann-Dirichlet domain decomposition preconditioners. Technical Report. Univ. of Jyvaskyla, Finland (1993)

[KU12] Kuznetsov, Y. A., Neittaanmaki, P., Tarvainen, P.: Overlapping domain decomposition methods for the obstacle problem. Sixth international conference on domain decomposition. AMS, Contemporary Mathematics **157** (1994) 271–277

[KU13] Kuznetsov, Y. A., Neittaanmaki, P., Tarvainen, P.: Block relaxation methods for algebraic obstacle problems with M-matrices. East-West J. Numer. Math. **4** (1996) 69–82

[KU14] Kuznetsov, Y. A., Vassilevski, Y.: A Dirichlet-Dirichlet preconditioner for the mortar element method. Technical Report B12/1999. Univ. of Jyvaskyla, Finland. (1999)

[KU15] Kuznetsov, Y. A., Wheeler, M. F.: Optimal order substructuring preconditioners for mixed finite element methods on nonmatching grids. East-West J. Numer. Math. **3** No. 2 (1995) 127–143

[LA] Lacour, C.: Iterative substructuring preconditioners for the mortar finite element method. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[LA2] Lacour, C., Maday, Y.: Two different approaches for matching on nonconforming grids: The mortar element method and the FETI method. BIT **37** (1997) 720–738

[LA3] Laevsky, Y. M.: Direct domain decomposition method for solving parabolic equations. Preprint 940, in Russian. Novosibirsk Comput. Center., Siberian Branch of Russian Academy of Sciences. (1992)

[LA4] Laevsky, Y. M.: On the domain decomposition method for parabolic problems. Bull. Novosibirsk Comput. Center. **1** (1993) 41–62

[LA5] Lagerstrom, P. A.: Matched asymptotic expansions: Ideas and techniques. Springer-Verlag (1988)

[LA6] Lai, C. H., Bjørstad, P., Cross, M., Widlund, O. B.: Proceedings of the eleventh international conference on domain decomposition methods. (1999)

[LA7] Lambert, J. D.: Numerical methods for ordinary differential systems: The initial value problem. Wiley (1991)

[LA8] Lasser, C., Toselli, A.: Convergence of some two level overlapping domain decomposition preconditioners with smoothed aggregation coarse spaces. In (Eds.) L. Pavarino, A. Toselli. Recent developments in domain decomposition methods. Lecture notes in computational science and engineering. **23** Springer-Verlag (2002) 95–117

[LA9] Lasser, C., Toselli, A.: An overlapping domain decomposition preconditioner for a class of discontinuous Galerkin approximations of advection-diffusion problems. Math. Comp. **72** No. 243 (2003) 1215–1238

[LA10] Lax, P.: Linear algebra. John Wiley (1996)

[LA11] Lazarov, R. D., Pasciak, J. E., Vassilevski, P. S.: Iterative solution of a coupled mixed and standard galerkin discretization method for elliptic problems. Numer. Lin. Alg. Applic. **8** (2001) 13–31

[LE] Le Tallec, P.: Neumann-Neumann domain decomposition algorithms for solving 2D elliptic problems with nonmatching grids. East-West J. Numer. Math. **1** No. 2 (1993) 129–146

[LE2] Le Tallec, P.: Domain decomposition methods in computational mechanics. (Ed.) J. T. Oden. Computational Mechanics Advances **1** No. 2 North-Holland (1994) 121–220

[LE3] Le Tallec, P.: Neumann-Neumann domain decomposition algorithms for solving 2D elliptic problems with nonmatching grids. East-West J. Numer. Math. **1** No. 2 (1993) 129–146

[LE4] Le Tallec, P., De Roeck, Y.-H., Vidrascu, M.: Domain decomposition methods for large linear elliptic three dimensional problems. J. Comp. Appl. Math. **34** No. 1 (1991) 93–117

[LE5] Le Tallec, P., Mandel, J., Vidrascu, M.: A Neumann-Neumann domain decomposition algorithm for solving plate and shell problems. SIAM J. Numer. Anal. **35** No. 2 (1998) 836–867

[LE6] Le Tallec, P., Patra, A. K.: Non-overlapping domain decomposition methods for $h\,p$ approximations of the Stokes problem with discontinuous pressure fields. Comp. Methds. Appl. Mech. Engrg. **145** No. 3–4 (1997) 361–379

[LE7] Le Tallec, P., Perlat, J. P.: Coupling kinetic models with Navier-Stokes equations. CFD Review **II** (1998) 833–855

[LE8] Le Tallec, P., Sassi, T.: Domain decomposition with nonmatching grids: Augmented Lagrangian approach. Math. Comp. **64** No. 212 (1995) 1367–1396

[LE9] Le Tallec, P., Sassi, T., Vidrascu, M.: Three dimensional domain decomposition methods with nonmatching grids and unconstrained grid solvers. (Eds.) D. Keyes, J. Xu. Seventh international conference on domain decomposition methods methods in scientific and engineering computing. AMS Contemporary Mathematics **180** (1994)

[LE10] Le Tallec, P., Tidriri, M. D.: Convergence analysis of domain decomposition algorithms with full overlapping for advection diffusion problems. Math. Comp. **68** No. 226 (1999) 585–606

[LE11] Le Tallec, P., Tidriri, M. D.: Application of maximum principle to the analysis of the time marching algorithm. JMAA **229** (1999) 158–169

[LE12] Lebedev, V. I.: Decomposition methods (in Russian). USSR Academy of Sciences. Moscow (1986)

[LE13] Lee, D., Yu, M.: A study on parallel flow computation by Newton-Schwarz method.A Proceedings of the sixth national conference on computational fluid dynamics. Taiwan (1999)

[LE14] Lees, M.: Alternating direction and semi-explicit difference methods for parabolic equations. Numer. Math. **3** (1961) 398–412

[LE15] Lesoinne, M.: A FETI-DP corner selection algorithm for three dimensional problems. n (Eds.) I. Herrera, D. E. Keyes, O. B. Widlund, R. A. Yates. Domain decomposition methods in science and engineering. Fourteenth international conference. UNAM Publ. Mexico (2003) 217–223

[LE16] Lewis, T. G., El-Rewini, H.: Introduction to parallel computing. Prentice Hall (1992)

[LI] Li, J.: Dual-primal FETI methods stationary Stokes and Navier-Stokes equations. PhD thesis. Courant Institute of Mathematical Sciences. TR-830. Dept. of Comp. Sci. (2002)

[LI2] Lions, J. L.: Some methods in the mathematical analysis of systems and their control. Taylor and Francis (1981)

[LI3] Lions, J. L., Maday, Y., Turinici, G.: Resolution d'EDP par un schema en temps parareel. C. R. Acad. Sci. Paris Ser. I Math. **332** (2001) 661–668

[LI4] Lions, J. L., Magenes, E.: Nonhomogeneous boundary value problems and applications: Vol. I. Springer-Verlag (1972)

[LI5] Lions, P. L.: Interpreétation stochastique de la méthode alternée de Schwarz. C. R. Acad. Sci. Paris **268** (1978) 325–328

[LI6] Lions, P. L.: On the Schwarz alternating method. I: First international symposium on domain dcecomposition methods for partial differential equations. SIAM (1988)

[LI7] Lions, P. L.: On the Schwarz alternating method. II: Second international symposium on domain dcecomposition methods for partial differential equations. SIAM (1989)

[LI8] Lions, P. L.: On the Schwarz alternating method. III: A variant for nonoverlapping subdomains. Domain dcecomposition methods for partial differential equations. SIAM (1990)

[LI9] Lions, P. L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16** (1979) 964–979

[LI10] Liu, M., Wang, J.: Pricing american options by domain decomposition methods. Iterative methods in scientific computation. (Eds.) J. Wang, M. Allen, B. Chen, T. Mathew. IMACS publications (1998)

[LO] Lorentz, R., Oswald, P.: Constructing economical Riesz basis for Sobolev spaces. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[LU] Lu, T., Shih, T., Liem, C.: Domain decomposition methods: New numerical techniques for solving PDE (in Chinese). Science Publishers, Beijing (1992)

[LU2] Lube, G., Muller, L., Muller, H.: A new non-overlapping domain decomposition method for stabilized finite element methods appled to the non-stationary Navier-Stokes equations. Numerical linear algebra methods for computational fluid flow problems. Num. Lin. Alg. Appl. **7** No. 6 (2000) 449–472

[LU3] Luenberger, D. G.: Optimization by vector space methods. Wiley-Interscience (1997)

[LU4] Luenberger, D. G.: Introduction to dynamic systems: Theory, models, and applications. Wiley (2001)

[LU5] Lui, S. H.: Some recent results on domain decomposition methods for eigenvalue problems. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997)

[LU6] Lui, S. H.: Domain decomposition methods for eigenvalue problems. J. Comp. Appl. Math. **117** No. 1 (2000) 17–34

[LU7] Lui, S. H.: On Schwarz alternating methods for nonlinear elliptic partial differential equations. SIAM J. Sci. Comp. **21** (2000) 1506–1523

[LU8] Lui, S. H.: On monotone and Schwarz alternating methods for nonlinear elliptic partial differential equations. M2AN Math. Model. Num. Anal. **35** No. 1 (2001) 1–15

[LU9] Lui, S. H.: On linear monotone and Schwarz alternating methods for nonlinear elliptic partial differential equations. Numer. Math. **93** (2002) 109–129

[LU10] Lui, S. H., Golub, G.: Homotopy method for the numerical solution of the eigenvalue problem of self adjoint partial differential equations. **10** No. 3–4 (1995) 363–378

[LU11] Luo, J. C.: Solving eigenvalue problems by implicit decomposition. Numer. Meth. part. Diff. Eqns. **7** No. 2 (1991) 113–145

[LU12] Luo, J. C.: A domain decomposition method for eigenvalue problems. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992) 306–321

[LY] Lynch, R. E., Rice, J. R., Thomas, D. H.: Direct solution of partial differential equations by tensor product methods. Numer. Math. **6** (1964) 185–199

[MA] Macedo Puppin, A.: On the choice of Krylov methods and preconditioners for a domain decomposed iterative solution of the exterior Helmholtz problem. Proceedings of the elevent international conference on domain decomposition methods. (Eds.) C.-H. Lai and P. E. Bjorstad and M. Cross and O. Widlund. (1999) 531–538

[MA2] Maday, Y., Meiron, D., Patera, A., Ronquist, E.: Analysis of iterative methods for the steady and unsteady Stokes problem: Application to spectral element discretizations. SIAM J. Sci. Comp. **14** No. 2 (1993) 310–337

[MA3] Maday, Y., Patera, A.: Spectral element methods for the Navier-Stokes equations. In State of the art surveys in computational mechanics (Eds. A. K. Noor and J. T. Oden), ASME, New York (1989)

[MA4] Maday, Y., Patera, A., Mavriplis, C.: Nonconforming mortar element methods: Application to spectral discretizations. Second international symposium on domain decomposition methods for partial differential equations. SIAM (1989)

[MA5] Maday, Y., Rapetti, F., Wohlmuth, B.: The influence of quadrature formulas in 2D and 3D mortar element methods. Lecture notes in computational science and engineering. **23** (2001) 203–221

[MA6] Maday, Y., Ronquist, E. M., Staff, G. A.: The parareal-in-time algorithm: Basics, stability analysis and more. Preprint. (2007)

[MA7] Maday, Y., Turinici, G.: A parareal in time procedure for the control of partial differential equations. C. R. Acad. Sci. Paris Ser. I Math. **335** (2002) 387–392

[MA8] Magolu, M. M., Notay, Y.: Dynamically relaxed block incomplete factorizations for solving two and three dimensional problems. SIAM J. Sci. Comp. **21** (2000) 2008–2028

[MA9] Maliassov, S. Y.: On the Schwarz alternating method for eigenvalue problems. (In Russian) Russ. J. Numer. Anal. Modell. **13** No. 1 (1998) 45–56

[MA10] Mandel, J.: A multilevel iterative method for symmetric positive definite linear complementarity problems. Appl. Math. Optim. **11** (1984) 77–95

[MA11] Mandel, J.: On block diagonal and Schur complement preconditioning. Numer. Math. **58** (1990) 79–93

[MA12] Mandel, J.: Iterative solvers by substructuring for the p-version finite element method. Comp. Methods Appl. Mech. Engrg. **80** (1990) 117–128

[MA13] Mandel, J.: Two level domain decomposition preconditioning for the p-version finite element version in three dimensions. Int. J. Numer. Meth. Engrg. **29** (1990) 1095–1108

[MA14] Mandel, J.: Balancing domain decomposition. Comm. Appl. Numer. Meth. **9** (1993) 233–241

[MA15] Mandel, J.: Hybrid domain decomposition with unstructured subdomains. Sixth international conference on domain decomposition. AMS, Contemporary Mathematics **157** (1994) 103–112

[MA16] Mandel, J.: Iterative substructuring with Lagrange multipliers for coupled fluid-solid scattering. Fourteenth international conference on domain decomposition. Preprint (2002) 107–117

[MA17] Mandel, J., Brezina, M.: Balancing domain decomposition for problems with large jumps in coefficients. Math. Comp. **65** (1996) 1387–1401

[MA18] Mandel, J., Dohrmann, C. R.: Convergence of a balancing domain decomposition by constraints and energy minimization. Numer. Lin. Alg. Appl. **10** (2003) 639–659

[MA19] Mandel, J., Dohrmann, C. R., Tezaur, R.: An algebraic theory for primal and dual substructuring methods by constraints. Sixth IMACS International Symposium on Iterative Methods in Scientific Computing, Denver, CO, 2003. Appl. Numer. Math. **54** No. 2 (2005) 167–193

[MA20] Mandel, J., Farhat, C., Cai, X.-C.: Domain decomposition methods 10. Tenth international conference. AMS Contemporary Mathematics **218** (1998)

[MA21] Mandel, J., McCormick, S.: Iterative solution of elliptic equations with refinement: The two level case. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989)

[MA22] Mandel, J., McCormick, S.: Iterative solution of elliptic equations with refinement: The model multilevel case. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989)

[MA23] Mandel, J., Scott Lett, G.: Domain decomposition preconditioning of p-version finite elements with high aspect ratios. Appl. Numer. Math. **8** (1991) 411–425

[MA24] Mandel, J., Sekerka, B.: A local convergence proof for the iterative aggregation method. Lin. Alg. Appl. **51** (1983) 163–172

[MA25] Mandel, J., Tezaur, R.: On the convergence of a substructuring method with Lagrange multipliers. Num. Math. **73** (1996) 473–487

[MA26] Mandel, J., Tezaur, R.: On the convergence of a dual-primal substructuring method. Num. Math. **88** No. 3 (2001) 543–558

[MA27] Mandel, J., Tezaur, R., Farhat, C.: A scalable substructuring method by Lagrange multipliers for plate bending problems. SIAM J. Num. Anal. **36** No. 5 (1999) 1370–1391

[MA28] Marchuk, G. I., Kuznetsov, Y. A., Matsokin, A. M.: Fictitious domain and domain decomposition methods. Soviet J. Num. Anal. Math. Modeling **1** (1986) 3–61

[MA29] Marini, L. D., Quarteroni, A.: A relaxation procedure for domain decomposition methods using finite elements. Numer. Math. **56** (1989) 575–598

[MA30] Martin, O. C., Otto, S. W.: Partitioning of unstructured meshes for load balancing. Concurrency: Practice and Experience. **7** No. 4 (1995) 303–314

[MA31] Mathew, T. P.: Schwarz alternating and iterative refinement methods for mixed formulations of elliptic problems, Part I: Algorithms and numerical results. Numer. Math. **65** No. 4 (1993) 445–468

[MA32] Mathew, T. P.: Schwarz alternating and iterative refinement methods for mixed formulations of elliptic problems, Part II: Theory. Numer. Math. **65** No. 4 (1993) 469–492

[MA33] Mathew, T. P.: Uniform convergence of the Schwarz alternating method for solving singularly perturbed advection diffusion equations. SIAM J. Numer. Anal. **35** No. 4 (1998) 1663–1683

[MA34] Mathew, T. P., Polyakov, P. L., Russo, G., Wang, J.: Domain decomposition operator splittings for the solution of parabolic equations. SIAM J. Sci. Comp. **19** No. 3 (1998) 912–932

[MA35] Mathew, T. P., Russo, G.: Maximum norm stability of difference schemes for parabolic equations on overset nonmatching space-time grids. Math. Comp. **72** No. 242 (2003) 619–656

[MA36] Mathew, T. P., Sarkis, M. V., Schaerer, C. E.: Analysis of block matrix preconditioners for elliptic optimal control problems. Num. Lin. Alg. Appls. **14** No. 3 (2007)

[MA37] Matsokin, A. M., Nepomnyaschikh, S. V.: A Schwarz alternating method in a subspace, Soviet Mathematics **29** No. 10 (1985) 78–84

[MA38] Matsokin, A. M., Nepomnyaschikh, S. V.: Norms on the space of traces of mesh functions. Sov. J. Numer. Anal. Math. Modell. **3** (1988) 199–216

[MA39] Matsokin, A. M., Nepomnyaschikh, S. V.: On using the bordering method for solving systems of mesh equations. Sov. J. Numer. Anal. Math. Modell. **4** (1989) 487–492

[MA40] Mavriplis, D.: Unstructured mesh generation and adaptivity. VKI Lecture series. VKI-LS 1995-02 Von Karman Institute for Fluid Dynamics. Belgium (1995)

[MA41] Mavriplis, D.: Unstructured grid techniques. Ann. Rev. Fluid. Mach. **29** (1997) 473–514

[MC] McCormick, S. F.: Fast adaptive composite grid (FAC) methods. Defect correction methods: Theory and applications. (Eds.) K. Böhmer, H. Stetter. Computing Supplementum 5. Springer-Verlag (1984) 115–121

[MC2] McCormick, S. F.: Multigrid methods. Frontiers in Appl. Math. Vol. 6, SIAM (1987)

[MC3] McCormick, S. F.: Multilevel adaptive methods for partial differential equations. SIAM (1989)

[MC4] McCormick, S. F., Thomas, J.: The fast adaptive composite grid (FAC) method for elliptic equations. Math. Comp. **46** No. 174 (1986) 439–456

[MC5] McHugh, P. R., Knoll, D., Keyes, D. E.: Application of a Schwarz preconditioned Newton-Krylov algorithm to a low speed reacting flow problem. AIAA Journal **36** (1998) 290–292

[MC6] McInnes, L. C., Susan-Resigna, R. F., Keyes, D. E., Atassi, H. M.: Additive Schwarz methods with nonreflecting boundary conditions for the parallel computation of Helmholtz problems. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. Tenth international conference. AMS Contemporary Mathematics **218** (1998) 325–333

[ME] Meddahi, S.: The Schwarz algorithm for multidomain spectral approximation of elliptic problems. Calcolo **31** No. 3 (1993) 241–253

[ME2] Meddahi, S.: Schwarz algorithms for the Raviart-Thomas mixed method. Calcolo **31** No. 1–2 (1994) 95–114

[ME3] Meijerink, J. A., van der Vorst, H. A.: An iterative solution method for linear systems for which the coefficient matrix is a symmetric M-matrix. Math. Comp. **31** No. 137 (1977) 148–162

[ME4] Melenk, J. M.: On condition numbers in $h\,p$-FEM with Gauss-Lobatto based shape functions. J. Comp. Appl. Math. **139** No. 1 (2002) 21–48

[ME5] Meurant, G.: Domain decomposition preconditioners for the conjugate gradient method. Calcolo **25** No. 1–2 (1988) 103–119

[ME6] Meurant, G.: Numerical experiments with a domain decomposition method for parabolic problems on parallel computers. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[ME7] Meurant, G.: A domain decomposition method for parabolic problems. Appl. Numer. Math. **8** (1991) 427–441

[ME8] Meurant, G.: Computer solution of large linear systems. Studies in Mathematics and Its Applications **28**. North-Holland (1999)

[MI] Miller, K.: Numerical analogs of the Schwarz alternating procedure. Numer. Math. **7** (1965) 91–103

[MI2] Mitra, S., Parashar, M., Browne, J. C.: DAGH: User's guide. Dept. Comp. Sci. Univ. of Texas, Austin. www.cs.utexas.edu/users/dagh (2001)

[MI3] Mizukami, A., Hughes, T.: A Petrov-Galerkin finite element method for convection dominated flows: An accurate upwinding technique for satisfying the maximum principle. Comput. Methods. Appl. Mech. Engrg. **50** (1985) 181–193

[MO] Morchoisne, Y.: Inhomogeneous flow calculations by spectral methods: Monodomain and multidomain techniques. Spectral methods for partial differential equations. (Eds.) R. Voigt, D. Gottlieb, M. Y. Hussaini. SIAM-CBMS (1984) 181–208

[MO2] Morgenstern, D.: Begründung des alternierenden verfahrens durch orthogonalprojektion. ZAMM **36** (1956) 7–8

[MU] Munoz-Sola, R.: Polynomial liftings on a tetrahedron and applications to the *h-p* version of the finite element method in three dimensions. SIAM J. Num. Anal. **34** No. 1 (1997) 282–314

[MU2] Murphy, M. F., Golub, G. H., Wathen, A.: A note on preconditioning for indefinite linear systems. SIAM J. Sci. Comp. **21** No. 6 (2000) 196–197

[NA] Nabben, R.: A characterization of M-matrices. Unpublished work (1995)

[NA2] Nabben, R.: Comparisons between additive and multiplicative Schwarz iterations in domain decomposition methods. Numer. Math. **95** (2003) 145–162

[NA3] Nabben, R., Szyld, D.: Convergence theory of restricted multiplicative Schwarz methods. SIAM J. Num. Anal. **40** (2003) 2318–2336

[NA4] Nataf, F., Nier, F.: Convergence rate of some domain decomposition methods for overlapping and nonoverlapping subdomains. Numer. Math. **75** No. 3 (1997) 357–377

[NA5] Nataf, F., Rogier, F.: Factorization of the convection diffusion operator and the Schwarz algorithm. M^3AS **5** No. 1 (1995) 67–93

[NA6] Natarajan, R.: Domain decomposition using spectral expansions of Steklov-Poincaré operators. SIAM J. Sci. Comp. **16** No. 2 (1995) 470–495

[NA7] Natarajan, R.: Domain decomposition using spectral expansions of Steklov-Poincaré operators II: A matrix formulation. SIAM J. Sci. Comp. **18** No. 4 (1997) 1187–1199

[NE] Nečas, J.: Les méthodes directes en théorie des équations elliptiques. Academia Prague (1967)

[NE2] Nedelec, J.-C.: Mixed finite elements in R^3. Numer. Math. **35** (1980) 315–341

[NE3] Nepomnyaschikh, S.: Domain decomposition and Schwarz methods in a subspace for the approximate solution of elliptic boundary value problems. PhD thesis. Computing Center, USSR Academy of Sciences. Novosibirsk. (1986)

[NE4] Nepomnyaschikh, S.: On the application of the method of bordering for elliptic mixed boundary value problems and on difference norms of $W_2^{1/2}(S)$. Sov. J. Numer. Anal. Math. Modell. **4** (1989) 493–506

[NE5] Nepomnyaschikh, S.: Application of domain decomposition to elliptic problems with discontinuous coefficients. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[NE6] Nepomnyaschikh, S.: Mesh theorems of traces, normalizations of functions, traces and their inversions. Sov. J. Numer. Anal. Math. Modell. **6** (1991) 1–25

[NE7] Nepomnyaschikh, S.: Domain decomposition and fictitious domains for elliptic boundary value problems. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[NO] Nocedal, J., Wright, S. J.: Numerical optimization. Springer (1999)

[NO2] Notay, Y.: On the robustness of modified incomplete factorization methods. Int. J. Comput. Math. **40** (1992) 121–141

[OD] Oden, J. T., Patra, A. K., Feng, Y.: Parallel domain decomposition solver for adaptive hp finite element methods. SIAM J. Num. Anal. **34** (1997) 2090–2118

[OD2] Oden, J. T., Reddy, J. N.: An introduction to the mathematical theory of finite elements. John Wiley. New York (1982)

[OH] Ohmori, K.: The discrete maximum principle for nonconforming finite element approximation to stationary convection diffusion equations, Math. Rep. Toyama Univ. **2** (1979) 33–52

[OL] O'Leary, D. P., Widlund, O. B.: Capacitance matrix methods for the Helmholtz equation on general three dimensional regions. Math. Comput. **33** (1979) 849–879

[OM] O'Malley, R. E.: Singular Perturbation Methods for Ordinary Differential Equations. Springer (1991)

[ON] Ong, M. E.: Hierarchical basis preconditioners for second order elliptic problems in three dimensions. Technical Report 89-3. Dept. Appl. Math., Univ. of Washington. Seattle (1989)

[OR] Ortega, J. M., Rheinboldt, W. C.: Iterative solution of nonlinear equations in several variables. Academic Press (1970)

[OS] Oswald, P.: On function spaces related to finite element approximation theory. Z. Anal. Anwendungen. **9** No. 1 (1990) 43–64

[OS2] Oswald, P.: On discrete norm estimates related to multilevel preconditioners in the finite element method. Proc. Int. Conf. Constructive theory of functions, Varna. Bulg. Acad. Sci., Sofia (1992) 203–214

[OS3] Oswald, P.: Multilevel finite element approximation: Theory and Applications. B. G. Teubner, MR **95k**:651110, Stuttgart, Germany (1994)

[OS4] Oswald, P.: Multilevel norms for $H^{-1/2}$. Computing **61** (1998) 235–255

[OW] Owen, S. J.: A survey of unstructured mesh generation technology. Proceedings of 7th international meshing roundtable. Michigan. http://www.andrew.cmu.edu/user/sowen/survey (1998)

[PA] Pacheco, P.: Parallel programming with MPI. Morgan-Kaufman (1997)

[PA2] Pahl, S.: Schwarz type domain decomposition methods for spectral element discretizations. Master's thesis. University of Witwatersrand. Johannesburg, S. Africa. (1993)

[PA3] Paige, C. C., Saunders, M.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12** (1975) 617–629

[PA4] Palansuriya, C. J., Lai, C. H., Ierotheou, C. S., Pericleous, K. A.: A domain decomposition based algorithm for non-linear 2D inverse heat conduction problems. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. Tenth international conference. AMS Contemporary Mathematics **218** (1998) 515–522

[PA5] Park, J., Park, E.-J.: A primal mixed domain decomposition method for elliptic problems. Proceedings of workshop on pure and applied mathematics (DaeWoo, Korea) **19** No. 2 (1999)

[PA6] Park, K.-C., Justino, M. R., Felippa, C. A.: An algebraically partitioned FETI method for parallel structural analysis: Algorithm description. Int. J. Num. Methds. Engrg. **40** (1997) 2717–2737

[PA7] Parlett, B. N.: The symmetric eigenvalue problem. Prentice Hall (1980)

[PA8] Pasciak, J. E.: Two domain decomposition techniques for Stokes problems. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989) 419–430

[PA9] Pavarino, L. F.: Additive Schwarz methods for the p-version finite element method. Numer. Math. **66** No. 4 (1994) 493–515

[PA10] Pavarino, L. F.: Schwarz methods with local refinement for the p-version finite element method. Numer. Math. **69** No. 2 (1994) 185–211

[PA11] Pavarino, L. F.: Neumann-Neumann algorithms for spectral elements in three dimensions. RAIRO Math. Modell. Numer. Anal. **31** (1997) 471–493

[PA12] Pavarino, L. F.: Preconditioned mixed spectral element methods for elasticity and Stokes problems. SIAM J. Sci. Comp. **19** No. 6 (1997) 1941–1957

[PA13] Pavarino, L. F.: Indefinite overlapping Schwarz methods for time dependent Stokes problems. Comp. Methds. Appl. Mech. Engrg. **187** No. 1–2 (2000) 35–51

[PA14] Pavarino, L. F., Warburton, T.: Overlapping Schwarz methods for unstructured spectral elements. J. Comp. Phys. **160** No. 1 (2000) 298–317

[PA15] Pavarino, L., Widlund, O. B.: A polylogarithmic bound for an iterative substructuring method for spectral elements in three dimensions. SIAM J. Num. Anal. **33** No. 4 (1996) 1303–1335

[PA16] Pavarino, L., Widlund, O. B.: Iterative substructuring methods for spectral elements: Problems in three dimensions based on numerical quadrature. Comp. Math. Appl. **33** No. 1/2 (1997) 193-209

[PA17] Pavarino, L., Widlund, O. B.: Balancing Neumann-Neumann methods for incompressible Stokes equations. Comm. Pure Appl. Math. **55** No. 3 (2002) 302–335

[PE] Peaceman, D. W., Rachford Jr., H. H.: The numerical solution of parabolic and elliptic differential equations. J. Soc. Ind. Appl. Math. **3** (1955) 28–41

[PE2] Peisker, P.: On the numerical solution of the first biharmonic equation. RAIRO Math. Mod. Num. Anal. **22** (1998) 655–676

[PE3] Pencheva, G., Yotov, I.: Balancing domain decomposition for mortar mixed finite element methods on nonmatching grids. Numer. Lin. Alg. Appl. **10** No. 1–2 (2003) 159–180

[PE4] Peyret, R., Taylor, T. D.: Computational methods for fluid flow. Springer-Verlag (1983)

[PH] Phillips, T. N.: Pseudospectral domain decomposition techniques for the Navier-Stokes equations. Fifth international symposium on domain decomposition methods for partial differential equations. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[PI] Picard, E.: Sur l'application des méthodes d'approximations successives á l'étude de certain équations différentielles ordinaires. J. de Math. Pures et Appl. **9** (1893) 217–271

[PI2] Pierson, K. H.: A family of domain decomposition methods for the massively parallel solution of computational mechanics problems. PhD thesis. Dept. of Aerospace engineering. Univ. of Colorado (2000)

[PO] Poincaré, H.: La méthode de Neumann et le probléme de Dirichlet. Acta Mathematica **20** (1896)

[PO2] Pothen, A., Simon, H. D., Liou, K. P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. **11** (1990) 430–452 Kluwer Academic Press (1996)

[PO3] Pothen, A.: Graph partitioning algorithms with applications to scientific computing. Parallel numerical algorithms, D. E. Keyes, A. H. Sameh, V. Venkatakrishnan (eds.). Kluwer Academic Press (1996)

[PR] Proskurowksi, W., Vassilevski, P. S.: Preconditioning capacitance matrix problems in domain imbedding. SIAM J. Sci. Comput. **15** (1994) 77–88

[PR2] Proskurowksi, W., Widlund, O. B.: On the numerical solution of Helmholtz's equation by the capacitance matrix method. Math. Comp. **30** (1976) 433–468

[PR3] Prudencio, E., Byrd, R., Cai, X.-C.: Parallel full space SQP Lagrange-Newton-Krylov-Schwarz algorithms for PDE constrained problems. SIAM J. Sci. Comp. a**27** (2006) 1305–1328

[PR4] Przemieniecki, J. S.: Matrix structural analysis of substructures. Am. Inst. Aero. Astro. J. **1** (1963) 138–147

[PR5] Przemieniecki, J. S.: Theory of matrix structural analysis. McGraw Hill (1968)

[QU] Quarteroni, A.: Domain decomposition algorithms for the Stokes equations. problems. Second international conference on domain decomposition methods. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. Widlund. SIAM (1989)

[QU2] Quarteroni, A.: Domain decomposition methods for systems of conservation laws: Spectral collocation approximations. SIAM J. Sci. Comp. **11** (1990) 1029–1052

[QU3] Quarteroni, A., Pasquarelli, F., Valli, A.: Heterogeneous domain decomposition: Principles, algorithms, applications. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992) 129–150

[QU4] Quarteroni, A., Stolcis, L.: Homogeneous and heterogeneous domain decomposition for compressible fluid flows at high Reynolds numbers. Numerical methods in fluid dynamics **5** (1995) 113–128

[QU5] Quarteroni, A., Valli, A.: Theory and applications of Steklov-Poincaré operators for boundary value problems: The heterogeneous operator case. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[QU6] Quarteroni, A., Valli, A.: Domain decomposition methods for partial differential equations. Oxford Science Publications (1999)

[QU7] Queck, W.: The convergence factor of preconditioned algorithms of the Arrow-Hurwicz type. SIAM J. Numer. Anal. **26** (1989) 1016–1030

[QU8] Quinn, M. J.: Parallel computing, theory and practice. McGraw-Hill (1994)

[RA] Rannacher, R.: On Chorin's projection method for Navier-Stokes equations. Lecture notes in mathematics. Springer-Verlag (1992)

[RA2] Rapetti, F., Toselli, A.: A FETI preconditioner for two dimensional edge element approximations of Maxwell's equations on non-matching grids. SIAM J. Sci. Comp. **23** No. 1 (2001) 92–108

[RA3] Rapin, G., Lube, G.: Comparison of two iterative substructuring methods for advection diffusion problems. Thirteenth international conference on domain decomposition methods. (Eds.) N. Debit, M. Garbey, R. Hoppe, J. Periaux, D. Keyes, Y. Kuznetsov (2001)

[RA4] Raviart, P. A., Thomas, J. M.: A mixed finite element method for 2nd order elliptic problems. Mathematical aspects of finite element methods. Lecture notes in mathematics **606** Springer-Verlag (1975)

[RE] Resasco, D. C.: Domain decomposition algorithms for elliptic partial differential equations. Ph.D. Thesis, Dept. of Comp. Sci., Yale University. (1990)

[RH] Rheinbach, O.: FETI- a dual iterative substructuring method for elliptic partial differential equations. Master's thesis. Mathe. Inst., Univ. zu Koln (2002)

[RH2] Rheinboldt, W. C.: On a theory of mesh refinement processes. SIAM J. Numer. Anal. **17** (1980) 766–778

[RI] Richtmyer, R., Morton, K.: Difference methods for initial value problems. Wiley-Interscience. (1995)

[RI2] Rivlin, T. J.: The Chebyshev polynomials. Wiley Interscience (1990)

[RI3] Rixen, D., Farhat, C.: Preconditioning the FETI and balancing domain decomposition methods for problems with intra- and inter-subdomain coefficient jumps. In (Eds.) P. Bjorstad, M. Espedal, D. Keyes. Ninth international conference: Domain decomposition methods in science and engineering. www.ddm.org (1997) 472–479

[RI4] Rixen, D., Farhat, C.: A simple and efficient extension of a class of substructure based preconditioners to heterogeneous structural mechanics problems. Int. J. Num. Methds. Engrg. **44** (1999) 489–516

[RI5] Rixen, D. J., Farhat, C., Tezaur, R., Mandel, J.: Theoretical comparison of the FETI and algebraically partitioned FETI methods, and performance comparisons with a direct sparse solver. Int. J. Numer. Methds Engrg. **46** (1999) 501–534

[RO] Rønquist, E. M.: A domain decomposition solver for the steady Navier-Stokes equations. In (Eds.) A. V. Ilin, L. R. Scott. Proc. of ICOSAHAM 95. Houston J. Math. (1996) 469–485

[RO2] Rønquist, E. M.: Domain decomposition methods for the steady Stokes equations. In (Eds.) C.-H. Lai, P. E. Bjørstad, M. Cross, O. B. Widlund. Domain decomposition methods in science and engineering: Eleventh international conference. www.ddm.org (1999) 330–340

[RU] Russell, T. F.: Time stepping along characteristics with incomplete iteration for a Galerkin approximation of miscible displacement in porous media. SIAM J. Numer. Anal. **22** (1985) 970–1013

[RU2] Russell, T. F.: Local refinement via domain decomposition techniques for mixed finite element methods with rectangular Raviart-Thomas elements. Third international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. B. Widlund. SIAM (1990)

[RU3] Rusten, T.: Iterative methods for mixed finite element systems. Ph.D. Thesis. Univ. of Oslo, Norway (1991)

[RU4] Rusten, T., Winther, R.: Mixed finite element methods and domain decomposition. Computational methods in water resources IX No. 1. Comput. Mech., Southampton (1992) 597–604

[RU5] Rusten, T., Winther, R.: A preconditioned iterative method for saddle point problems. SIAM J. Matrix Anal. **13** No. 3 (1991) 887–904

[SA] Saad, Y.: Numerical methods for large eigenvalue problems. Halstead Press, John Wiley (1992)

[SA2] Saad, Y.: Iterative methods for sparse linear systems. PWS publishing company (1996)

[SA3] Saad, Y., Schultz, M. H.: GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comp. **7** (1986) 856–869

[SA4] Sala, M: An algebraic two level domain decomposition preconditioner with applications to the compressible Euler equations. Int. J. Num. Methds. Fluids. **40** No. 12 (2002) 1551–1560

[SA5] Samarskii, A. A.: Locally one dimensional difference schemes on nonuniform grids. Z. Vycisl. Mat. Mat. Fiz. **3** (1963) 431–466

[SA6] Samokish, B. A.: The steepest descent method for an eigenvalue problem with semibounded operators. Izv. Vuzov. Math. (In Russian) **5** (1958) 105–114

[SA7] Sarkis, M. V.: Two level Schwarz methods for nonconforming finite elements and discontinuous coefficients. Proceedings of sixth Copper Mountain conference on multigrid methods. **2** No. 3224 (1993) 543–566

[SA8] Sarkis, M. V.: Schwarz preconditioners for elliptic problems with discontinuous coefficients using conforming and nonconforming elements. Ph.D. thesis. Technical Report 671, Comp. Sci. Dept., Courant Institute, New York University (1994)

[SA9] Sarkis, M. V.: Multilevel methods for P_1 nonconforming finite elements and discontinuous coefficients in three dimensions. Seventh international conference on domain decomposition methods methods in scientific and engineering computing. AMS Contemporary Mathematics **180** (1994) 119–124

[SA10] Sarkis, M. V.: Nonstandard coarse spaces and Schwarz methods for elliptic problems with discontinuous coefficients using nonconforming elements. Numer. Math. **77** No. 3 (1997) 383–406

[SA11] Sarkis, M. V.: Partition of unity coarse spaces and Schwarz methods with harmonic overlap. Lecture Notes in Computational Science and Engineering. (Eds.) L. Pavarino and A. Toselli. Springer-Verlag **23** (2002) 75–92

[SA12] Sarkis, M. V.: Partition of unity coarse spaces. Fluid flows and transport in porous media. Mathematical and numerical treatment. (Eds.) Z. Chen and R. Ewing. AMS Contemporary Mathematics **295** (2002) 445–456

[SA13] Sarkis, M. V.: Partition of unity coarse spaces: Enhanced versions, discontinuous coefficients and applications to elasticity. Proceedings of the 14th international conference on domain decomposition methods, Cocoyoc, Mexico (2002)

[SA14] Sarkis, M. V.: A coarse space for elasticity. Applied mathematics and scientific computing. (Eds.) Z. Drmac et al. Kluwer Academic/Plenum Publishers (2003) 261–273

[SA15] Sarkis, M. V., Szyld, D. B.: A proposal for a dynamically adapted inexact additive Schwarz preconditioner. In (Eds.) O. B. Widlund, D. E. Keyes. Domain decomposition methods in science and engineering XVI. Lecture notes in computational science and engineeering **55** (2006) 333–337

[SA16] Sarkis, M. V., Szyld, D. B.: Optimal left and right additive Schwarz preconditioning for minimal residual methods with Euclidean and energy norms. Comp. Methds. Appl. Mech. Engrg. **196** (2007) 1507–1514

[SA17] Sarkis, M. V., Tu, X.: Singular function mortar finite element methods. Comp. Meth. Appl. Math. **3** No. 1 (2003)

[SC] Scapini, F.: The alternating Schwarz method applied to some biharmonic variational inequalities. Calcolo **27** (1990) 57–72

[SC2] Schaerer, C. E., Mathew, T. P., Sarkis, M. V.: Block diagonal parareal preconditioner for parabolic optimal control problems. Proceedings of the seventeenth international conference on domain decomposition methods. Lecture notes in computational science and engineering. Springer (2007)

[SC3] Schaerer, C. E., Mathew, T. P., Sarkis, M. V.: Temporal domain decomposition for a linear quadratic optimal control problem. Seventh international conference on high performance computing in the computational sciences. Lecture notes in computer science. Springer (2007)

[SC4] Schatz, A. H.: An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. Math. Comp. **28** No. 128 (1974) 959–962

[SC5] Schwarz, H. A.: Gesammelte mathematische abhandlungen. Vierteljahrsschrift der naturforschenden gesselschaft in Zürich **15** (1870) 272–286 (republished by Springer-Verlag)

[SC6] Scott, L. R., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comp. **54** (1990) 483–493

[SC7] Scroggs, J. S.: A parallel algorithm for nonlinear convection diffusion equations. Third international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, J. Périaux, O. B. Widlund. SIAM (1990)

[SE] Sehmi, N. S.: Large order structural eigenanalysis techniques. Mathematics and its applications. Ellis-Horwood Ltd. Chichester (1989)

[SE2] Seshaiyer, P., Suri, M.: Uniform $h-p$ convergence results for the mortar finite element method. Math. Comp. **69** (1999) 521–546

[SH] Shampine, L. F., Gordon, M. K.: Computer solution of ordinary differential equations: The initial value problem. Freeman (1975)

[SH2] Shampine, L. F., Gear, C. W.: A user's view of solving stiff ordinary differential equations. SIAM Rev. **21** (1979) 1–17

[SH3] Sherwin, S. J., Casarin, M.: Low energy bases preconditioning for elliptic substructured solvers based on spectral/hp element discretizations. J. Comp. Phys. **171** (2001) 1–24

[SI] Silvester, D., Elman, H., Kay, D., Wathen, A.: Efficient preconditioning of the linearized Navier-Stokes equations. J. Comp. Appl. Math. **128** (2001) 261–279

[SI2] Simon, H. D.: Partitioning of unstructured problems for parallel processing. Comp. Sys. Engrg. **2** No. 2/3 (1991) 135–148

[SI3] Simoncini, V., Szyld, D. B.: On the occurrence of super-linear convergence of exact and inexact Krylov subspace methods. SIAM Rev. **47** No. 2 (2005) 247–272

[SI4] Simpson, A., Tabarrok, B.: On Kron's eigenvalue procedure and related methods of frequency analysis.Quart. J. Mech. Appl. Math. **21** (1968) 1–39

[SK] Skogen, M.: Schwarz methods and parallelism. Ph.D. Thesis, Dept. of Informatics, Univ. of Bergen, Norway (1992)

[SM] Smith, B.: Domain decomposition algorithms for the partial differential equations of linear elasticity. Ph.D. Thesis, Courant Institute, New York University (1990)

[SM2] Smith, B.: A domain decomposition algorithm for elliptic problems in three dimensions. Numer. Math. **60** No. 2 (1991) 219–234

[SM3] Smith, B.: An optimal domain decomposition preconditioner for the finite element solution of linear elasticity problems. SIAM J. Sci. Stat. Comp. **13** No. 1 (1992) 364–378

[SM4] Smith, B.: A parallel implementation of an iterative substructuring algorithm for problems in three dimensions. SIAM J. Sci. Comput. **14** No. 2 (1993) 406–423

[SM5] Smith, B., Bjørstad, P. E., Gropp, W. D.: Domain decomposition: Parallel multilevel methods for elliptic partial differential equations. Cambridge University Press (1996)

[SM6] Smith, B., Widlund, O. B.: A domain decomposition algorithm using a hierarchical basis. SIAM J. Sci. Stat. Comp. **11** No. 6 (1990) 1212–1220

[SM7] Smoller, J.: Shock waves and reaction-diffusion systems. Springer-Verlag (1994)

[SO] Sobolev, S. L.: L'algorithme de Schwarz dans la théorie de l'elasticité. Comptes rendus doklady de l'académie des sciences de l'URSS **IV** No. XIII **6** (1936) 243–246

[SO2] Sod, G.: Numerical methods in fluid dynamics: Initial and initial-boundary value problems. Cambridge University Press (1985)

[ST] Starius, G.: Composite mesh difference methods for elliptic boundary value problems. Numer. Math. **28** (1977) 243–258

[ST2] Starke, G.: Alternating direction preconditioning for nonsymmetric systems of linear equations. SIAM J. Sci. Comp. **15** (1994) 369–384

[ST3] Starke, G.: Field of values analysis of preconditioned iterative methods for nonsymmetric elliptic problems. Numer. Math. **78** (1997) 103–117

[ST4] Stefanica, D.: A numerical study of FETI algorithms for mortar finite elements. SIAM J. Sci. Comp. **23** No. 4 (2001) 1135–1160

[ST5] Stefanica, D., Klawonn, A.: The FETI method for mortar finite elements. Proceedings of eleventh international conference on domain decomposition methods. (Eds.) C.-H. Lai, P. E. Bjørstad, M. Cross, O. Widlund. www.ddm.org (1999) 121–129

[ST6] Steger, J., Benek, J.: On the use of composite grid schemes in computational aerodynamics. Comp. Meth. Appl. Mech. Engin. **64** (1987) 301–320

[ST7] Stein, E. M.: Singular integrals and differentiability properties of functions. Princeton university press (1970)

[ST8] Steklov, V. A.: General methods for solving basic problems of mathematical physics. Mathematical Society of Charkov (1901)

[ST9] Sternberg, S.: Lectures on differential geometry. Chelsea, New York (1964)

[ST10] Stoer, J., Bulirsch, R.: Introduction to numerical analysis. Springer-Verlag (1980)

[ST11] Strang, G.: On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5** (1968) 506–517

[ST12] Strang, G.: An introduction to applied mathematics. Wellesley Cambridge Press (1986)

[ST13] Strang, G.: Linear algebra and its applications. International Thomson Publishing (1988)

[ST14] Strang, G., Fix, G. J.: An analysis of the finite element method. Prentice-Hall (1973)

[SU] Sun, H., Tang, W.-P.: An overdetermined Schwarz alternating method. SIAM J. Sci. Comp. **17** No. 4 (1996) 884–905

[SU2] Sun, J., Zou, J.: Domain decomposition preconditioner for 4th order problems using B-spline finite element method. Fourth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. Chan, R. Glowinski, Y. Kuznetsov, O. B. Widlund. SIAM (1991)

[SZ] Szyld, D., Widlund, O.: Applications of conjugate gradient type methods to eigenvalue calculations. Advances in computer methods for partial differential equations III. Proc. Thirds. IMAC Int. Symp. IMACS (1979) 167–173

[TA] Tai, X.-C.: Domain decomposition for linear and nonlinear elliptic problems via function or space decomposition. Domain decomposition methods in scientific and engineering computing. (Eds.) D. Keyes, J. Xu. AMS (1994)

[TA2] Tai, X.-C.: Convergence rate analysis of domain decomposition methods for obstacle problems. East-West J. Numer. Anal. **9** No. 3 (2001) 233–252

[TA3] Tai, X.-C.: Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities. Numer. Math. **93** (2003) 755–786

[TA4] Tai, X.-C., Espedal, M. S.: Rate of convergence of some space decomposition method for linear and nonlinear elliptic problems. SIAM J. Numer. Anal. **35** (1998) 1559–1570

[TA5] Tai, X. C., Xu, J.: Global convergence of subspace correction methods for convex optimization problems. Math. Comput. **71** No. 237 (2001) 105–124

[TA6] Tan, K. H., Borsboom, M. J. A.: On generalized Schwarz coupling applied to advection dominated problems. In (Eds.) D. Keyes, J. Xu. Seventh international conference on domain decomposition methods methods in scientific and engineering computing. AMS Contemporary Mathematics **180** (1994) 125–130

[TA7] Tang, W.-P.: Schwarz splitting and template operators. Ph.D. Thesis, Dept. of Comp. Sci. Stanford University (1988)

[TA8] Tang, W.-P.: Generalized Schwarz splittings. SIAM J. Sci. Stat. Comp. **13** No. 2 (1992) 573–595

[TE] Temam, R.: Navier-Stokes equations: Theory and numerical analysis. AMS (1985)

[TE2] Teng, S.-H., Wong, C. W.: Unstructured mesh generation: Theory, practice and perspectives. Int. J. Comp. Geom. and Appl. **10** No. 3 (2000) 227–266

[TE3] Tezaur, R.: Analysis of Lagrange multiplier based domain decomposition. Ph.D. thesis. Univ. of Colorado, Denver (1998)

[TH] Thomas, J. M.: Finite element matching methods. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[TH2] Thomée, V.: Galerkin finite element methods for parabolic problems. Lecture notes in mathematics **1054** Springer (1984)

[TH3] Thompson, J. F., Soni, B. K., Weatherill, N. P., Weatherill, N. P.: Handbook of grid generation. CRC Press (1998)

[TO] Tong, C. H., Chan, T. F., Kuo, C. C. J.: A domain decomposition preconditioner based on a change to a multilevel nodal basis. SIAM J. Sci. Comput. **12** (1991) 1486–1495

[TO2] Toselli, A.: Some results on overlapping Schwarz methods for the Helmholtz equation employing perfectly matched layers. In (Eds.) C.-H. Lai, P. E. Bjørstad, M. Cross, O. B. Widlund. Domain decomposition methods in science and engineering: Eleventh international conference. www.ddm.org (1999) 539–545

[TO3] Toselli, A.: Neumann-Neumann methods for vector field problems. ETNA **11** (2000) 1–24

[TO4] Toselli, A.: Overlapping Schwarz methods for Maxwell's equations in three dimensions. Numer. Math. **86** No. 4 (2000) 733–752

[TO5] Toselli, A.: FETI domain decomposition methods for scalar advection diffusion problems. Comp. Methds. Appl. Mech. Engrg. **190** N0. 43–44 (2001) 5759–5776

[TO6] Toselli, A.: Two iterative substructuring methods for Maxwell's equations with discontinuous coefficients in two dimensions. In (Eds.) T. F. Chan, T. Kako, H. Kawarada, O. Pironneau. Domain decomposition methods in science and engineering. Twelveth international conference. www.ddm.org (2001) 215–222

[TO7] Toselli, A., Klawonn, A.: A FETI domain decomposition method for edge element approximations in two dimensions with discontinuous coefficients. SIAM J. Num. Anal. **39** (2001) 932–956

[TO8] Toselli, A., Vasseur, X.: A numerical study on Neumann-Neumann and FETI methods for $h\,p$ approximations on geometrically refined boundary layer meshes in two dimensions. Comp. Methds. Appl. Mech. Engrg. **192** (2003) 4551–4579

[TO9] Toselli, A., Vasseur, X.: Domain decomposition preconditioners of Neumann-Neumann type for $h\,p$ approximations on boundary layer meshes in three dimensions. IMA J. Num. Anal. **24** No. 1 (2004) 123–156

[TO10] Toselli, A., Widlund, O. B.: Domain decomposition methods: Algorithms and theory. Springer (2004)

[TO11] Toselli, A., Wohlmuth, B., Widlund, O. B.: An iterative substructuring method for Maxwell's equation in two dimensions. Math. Comp. **70** No. 235 (2001) 935–949

[TR] Trefethen, L. N., Bau III, D.: Numerical linear algebra. SIAM (1997)

[TR2] Trotta, R. L.: Multidomain finite elements for advection-diffusion equations. Appl. Numer. Math. **21** No. 1 (1996) 91–118

[TS] Tsui, W.: Domain decomposition of biharmonic and Navier-Stokes equations. Ph.D. Thesis. Dept. of Mathematics, UCLA, Los Angeles (1991)

[TU] Tuminaro, R. S., Shadid, J. N., Walker, H. F.: On backtracking failure in Newton-GMRES methods. J. Comput. Phys. **180** (2002) 549–558

[VA] Vabischevich, P. N.: Parallel domain decomposition algorithms for time dependent problems of mathematical physics. Advances in numerical methods and applications $O(h^3)$. (Eds.) I. T. Dimov, B. Sendov, P. S. Vassilevski. World Scientific, Singapore (1994) 293–299

[VA2] Vabischevich, P. N., Matus, P.: Difference schemes on grids locally refined in space as well as in time. Advances in numerical methods and applications $O(h^3)$. (Eds.) I. T. Dimov, B. Sendov, P. S. Vassilevski. World Scientific, Singapore. (1994) 146–153

[VA3] Van Driessche, R., Roose, D.: A graph contraction algorithm for the fast calculation of the Fiedler vector of a graph. Proceedings of the seventh SIAM conference on parallel computing for scientific computing. (1995) 621–626

[VA4] Van Loan, C.: Computational frameworks for the fast Fourier transform. Frontiers in applied mathematics 10. SIAM (1992)

[VA5] Vanek, P., Brezina, M., Mandel, J.: Convergence of algebraic multigrid based on smoothed aggregation. Numer. Math. **88** (2001) 559–679

[VA6] Vanek, P., Brezina, M., Tezaur, R.: Two grid method for linear elasticity on unstructured meshes. SIAM J. Sci. Comp. **21** (1999) 900–923

[VA7] Vanek, P., Mandel, J., Brezina, M.: Algebraic multigrid by smooth aggregation for second and fourth order elliptic problems. Computing **56** (1996) 179–196

[VA8] Vanek, P., Mandel, J., Brezina, M.: Solving a two dimensional Helmholtz problem using algebraic multigrid. Technical Report 110. Center for Computational Mathematics, Univ. of Colorado, Denver (1997)

[VA9] Varga, R. S.: Matrix iterative analysis. Prentice Hall (1962)

[VA10] Vassilevski, P. S.: Preconditioning nonsymmetric and indefinite finite element matrices. Num. Lin. Alg. with Appl. **1** No. 1 (1992) 59–76

[VA11] Vassilevski, P., Wang, J.: Multilevel iterative methods for mixed finite element discretizations of elliptic problems. Numer. Math. **63** (1992) 503–520

[VA12] Vassilevski, P., Wang, J.: An application of the abstract multilevel theory to nonconforming finite element methods. SIAM J. Num. Anal. **32** No. 1 (1995) 235–248

[VE] Verfurth, R.: A combined conjugate gradient method-multigrid algorithm for the numerical solution of the Stokes problem. IMA J. Numer. Anal. **4** (1984) 441–455

[VE2] Versieux, H., Sarkis, M. V.: A three-scale finite element method for elliptic equations with rapidly oscillating periodic coefficients. In (Eds.) O. B. Widlund, D. E. Keyes. Domain decomposition methods in science and engineering XVI. Lecture notes in computational science and engineeering **55** (2006) 763–770

[WA] Walker, H. F.: An adaptation of Krylov subspace methods to path following problems. SIAM J. Sci. Comput. **21** (2000) 1191–1198

[WA2] Wang, J.: Convergence analysis of the Schwarz algorithm and multilevel decomposition iterative methods I: Self adjoint and positive definite elliptic problems. Iterative methods in linear algebra. (Eds.) R. Beauwens, P. de Groen. North-Holland (1992) 93–110

[WA3] Wang, J.: Convergence analysis of the Schwarz algorithm and multilevel decomposition iterative methods II: Non-self adjoint and indefinite elliptic problems. SIAM J. Numer. Anal. **30** (1993) 953–970

[WA4] Wang, J.: Convergence estimates for multilevel techniques for finite element approximations. J. Comp. Appl. Math. **50** (1994) 593–604

[WA5] Wang, J., Mathew, T. P.: Mixed finite element methods over quadrilaterals. Proceedings of the third international conference on advances in numerical methods and applications. (Eds.) I. T. Dimov, B. Sendov, P. Vassilevski. World Scientific (1994) 203–214

[WA6] Wang, J., Xie, R.: Domain decomposition for elliptic problems with large jumps in coefficients. Proceedings of conference on scientific and engineering computing, National Defense Industry Press, Beijing, China. (1994) 74–86

[WA7] Wang, J., Yan, N.: A parallel domain decomposition procedure for convection diffusion problems. Eight international conference on domain decomposition methods. (Eds.) R. Glowinski, J. Périaux, Z. Shi. (1997)

[WA8] Wathen, A., Fischer, B., Silvester, D.: The convergence rate of the minimal residual method for the Stokes problem. Num. Math. **71** (1995) 121–134

[WA9] Wathen, A., Silvester, D.: Fast iterative solution of stabilised Stokes systems I: Using simple diagonal preconditioners. SIAM J. Num. Anal. **30** No. 3 (1993) 630–649

[WE] Weiser, A., Wheeler, M. F.: On convergence of block centered finite differences for elliptic problems. SIAM J. Numer. Anal. **25** No. 2 (1988) 351–375

[WE2] Wetton, B. R.: Error analysis for Chorin's original fully discrete projection method with regularization in space and time. SIAM J. Num. Anal. **34** (1997) 1683–1697

[WH] Wheeler, M. F., Gonzalez, R.: Mixed finite element methods for petroleum reservoir engineering problems. Computing methods in applied sciences and engineering VI. (Eds.) R. Glowinski, J. L. Lions. North-Holland (1984) 639–658

[WH2] Wheeler, M. F., Yotov, I.: Multigrid on the interface for mortar mixed finite element methods for elliptic problems. Vistas in domain decomposition and parallel processing in computational mechanics. Special issue of: Comp. Meth. in Appl. Mech. Engrg. **184** (2000) 287–302

[WH3] Wheeler, J., Wheeler, M. F., Yotov, I.: Enhanced velocity mixed finite element methods for flow in multiblock domains. Comp. Geosciences. **6** No. 3–4 (2002) 315–332

[WI] Widlund, O. B.: An extension theorem for finite element spaces with three applications. In Numerical Techniques in Continuum Mechanics, (Eds.) W. Hackbusch and K. Witsch, Notes on Numerical Fluid Mechanics, Vol. **16** Friedr. Vieweg und Sohn. (1987) 110–122

[WI2] Widlund, O. B.: A comparison of some domain decomposition and iterative refinement algorithms for elliptic finite element problems. Technical Report BSC 88/15. IBM Bergen Scientific Centre, Bergen, Norway (1988)

[WI3] Widlund, O. B.: Iterative substructuring methods: Algorithms and theory for elliptic problems in the plane. In (Eds.) R. Glowinski, G. Golub, G. Meurant, J. Périaux. First international symposium on domain decomposition methods for partial differential equations SIAM (1988) 113–128

[WI4] Widlund, O. B.: On the rate of convergence of the classical Schwarz alternating method in the case of more than two subregions. Technical Report. Dept. of Comp. Sci., Courant Institute (1989)

[WI5] Widlund, O. B.: Some Schwarz methods for symmetric and nonsymmetric elliptic problems. In (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. Fifth international symposium on domain decomposition methods for partial differential equations SIAM (1992) 19–36

[WI6] Widlund, O. B.: Exotic coarse spaces for Schwarz methods for lower order and spectral finite elements. Seventh international conference on domain decomposition methods methods in scientific and engineering computing. AMS Contemporary Mathematics **180** (1994) 131–136

[WI7] Widlund, O. B.: Preconditioners for spectral and mortar finite element methods. In (Eds.) R. Glowinski, J. Périaux, Z.-C. Shi, O. B. Widlund. Domain decomposition methods in science and engineering. Eight international conference. John-Wiley (1996)

[WI8] Wilkinson, J. H.: The algebraic eigenvalue problem. Oxford University Press (1980)

[WI9] Willien, F., Faille, I., Nataf, F., Schneider, F.: Domain decomposition methods for fluid flow in porous medium. In 6th European conference on the mathematics of oil recovery (1998)

[WI10] Wilmott, P., Howison, S., Dewynne, J.: Option pricing: Mathematical models and computation. Oxford Financial Press (1993)

[WI11] Wilmott, P., Howison, S., Dewynne, J.: The mathematics of financial derivatives. Cambridge University Press (1995)

[WO] Wohlmuth, B.: Mortar finite element methods for discontinuous coefficients. ZAMM **79** S I (1999) 151–154

[WO2] Wohlmuth, B.: Hierarchical a posteriori error estimators for mortar finite element methods with Lagrange multipliers. SIAM J. Numer. Anal. **36** (1999) 1636–1658

[WO3] Wohlmuth, B.: A residual based error estimator for mortar finite element discretizations. Numer. Math. **84** (1999) 143–171

[WO4] Wohlmuth, B.: A mortar finite element method using dual spaces for the Lagrange multipliers. SIAM J. Numer. Anal. **38** (2000) 989–1012

[WO5] Wohlmuth, B.: Discretization techniques and iterative solvers based on domain decomposition. Lecture notes in computational science and engineering. **17** Springer (2001)

[WO6] Wohlmuth, B., Toselli, A., Widlund, O. B.: Iterative substructuring for Raviart-Thomas vector fields in three dimensions. SIAM J. Numer. Anal. **37** No. 5 (2000) 1657–1676

[WU] Wu, L., Allen, M. B., Park, E.-J.: Mixed finite element solution of reaction-diffusion equations using a two grid method. Comp. Meth. Water Res. **12** (1998) 217–224

[WU2] Wu, Y., Cai, X. C., Keyes, D. E.: Additive Schwarz methods for hyperbolic equations. In (Eds.) J. Mandel, C. Farhat, X.-C. Cai. Domain decomposition methods 10. Tenth international conference. AMS Contemporary Mathematics **218** (1998) 513–521

[XU] Xu, J.: Theory of multilevel methods. Ph.D. thesis, Cornell University. Technical Report AM-48, Penn State University (1989)

[XU2] Xu, J.: Counter examples concerning a weighted L^2 projection. Math. Comp. **57** (1991) 563–568

[XU3] Xu, J.: Iterative methods by space decomposition and subspace correction. SIAM Review **34** No. 4 (1992) 581–613

[XU4] Xu, J.: A new class of iterative methods for non-selfadjoint or indefinite problems. SIAM J. Numer. Anal. **29** No. 2 (1992) 303–319

[XU5] Xu, J.: Iterative methods by SPD and small subspace solvers for nonsymmetric or indefinite problems. Fifth international symposium on domain decomposition methods for partial differential equations. (Eds.) T. F. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, R. G. Voigt. SIAM (1992)

[XU6] Xu, J.: Two grid discretization techniques for linear and nonlinear PDES. SIAM J. Num. Anal. **33** (1996) 1759–1777

[XU7] Xu, J.: The method of subspace corrections. J. Comp. Appl. Math. **128** No. 1–2 (2001) 335–362

[XU8] Xu, J., Cai, X.-C.: A preconditioned GMRES method for nonsymmetric or indefinite problems. Math. Comput. **59** (1992) 311–319

[XU9] Xu, J., Zikatanov, L.: The method of alternating projections and the method of subspace corrections in a Hilbert space. J. Amer. Math. Soc. **15** No. 3 (2002) 573–597

[XU10] Xu, J., Zou, J.: Some nonoverlapping domain decomposition methods. SIAM Review **40** (1998) 857–914

[YA] Yanenko, N. N.: On weak approximation of systems of differential equations. Sibirsk. Mat. Zh. **5** (1964)

[YA2] Yang, D. Q.: A parallel iterative nonoverlapping domain decomposition procedure for elliptic problems. IMA J. of Numer. Anal. **16** (1996) 75–91

[YA3] Yanik, E. G.: Sufficient conditions for a discrete maximum principle for high order collocation methods. Comput. Math. Appl. **17** (1989) 1431–1434

[YO] Yotov, I.: Interface solvers and preconditioners of domain decomposition type for multiphase flow in multiblock porous media. Advances in computation: Theory and practice **7** (2001) 157–167

[YS] Yserentant, H.: Hierarchical bases of finite element spaces in the discretization of nonsymmetric elliptic boundary value problems. Computing **35** (1985) 39–49

[YS2] Yserentant, H.: On the multilevel splitting of finite element spaces. Numer. Math. **49** (1986) 379–412

[YS3] Yserentant, H.: On the multilevel splitting of finite element spaces for indefinite elliptic boundary value problems. SIAM J. Numer. Anal. **23** (1986) 581–595

[ZE] Zeng, J., Zhou, S.: On monotone and geometric convergence of Schwarz methods for two sided obstacle problems. SIAM J. Numer. Math. **35** No. 2 (1998) 600–616

[ZH] Zhang, X.: Studies in domain decomposition: Multilevel methods and the biharmonic Dirichlet problem. Ph.D. Thesis, Courant Institute, New York University (1991)

[ZH2] Zhang, X.: Multilevel Schwarz methods. Numer. Math. **63** No. 4 (1992) 521–539

[ZH3] Zhang, X.: Multilevel Schwarz methods for the biharmonic Dirichlet problem. SIAM J. Sci. Comput. **15** No. 3 (1994) 621–644

[ZH4] Zhang, X.: Two level Schwarz methods for the biharmonic problem discretized by C^1 elements. SIAM J. Numer. Anal. **33** No. 2 (1996) 555–570

[ZH5] Zhuang, Y., Sun, X.-H.: Stabilized explicit-implicit domain decomposition methods for the numerical solution of parabolic equations. SIAM J. Sci. Comp. **24** No. 1 (2002) 335–358

[ZO] Zou, J., Huang, H. C.: Algebraic subproblem decomposition methods and parallel algorithms with monotone convergence. J. Comp. Math. **10** (1992) 47–59

[ZU] Zulehner, W.: Analysis of iterative methods for saddle point problems: A unified approach. Math. Comp. **71** (2002) 479–505

Index

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Lecture and seminar notes
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged**. The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgment on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

- at least 100 pages of text;
- a table of contents;
- an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
- a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact Lecture Notes in Computational Science and Engineering at the planning stage.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Format. Only works in English are considered. They should be submitted in camera-ready form according to Springer-Verlag's specifications.
Electronic material can be included if appropriate. Please contact the publisher.
Technical instructions and/or LaTeX macros are available via http://www.springer.com/authors/book+authors?SGWID=0-154102-12-417900-0. The macros can also be sent on request.

General Remarks

Lecture Notes are printed by photo-offset from the master-copy delivered in camera-ready form by the authors. For this purpose Springer-Verlag provides technical instructions for the preparation of manuscripts. See also *Editorial Policy.*

Careful preparation of manuscripts will help keep production time short and ensure a satisfactory appearance of the finished book.

The following terms and conditions hold:

Categories i), ii), and iii):
Authors receive 50 free copies of their book. No royalty is paid. Commitment to publish is made by letter of intent rather than by signing a formal contract. Springer- Verlag secures the copyright for each volume.

For conference proceedings, editors receive a total of 50 free copies of their volume for distribution to the contributing authors.

All categories:
Authors are entitled to purchase further copies of their book and other Springer mathematics books for their personal use, at a discount of 33.3% directly from Springer-Verlag.

Addresses:

Timothy J. Barth
NASA Ames Research Center
NAS Division
Moffett Field, CA 94035, USA
e-mail: barth@nas.nasa.gov

Michael Griebel
Institut für Numerische Simulation
der Universität Bonn
Wegelerstr. 6
53115 Bonn, Germany
e-mail: griebel@ins.uni-bonn.de

David E. Keyes
Department of Applied Physics
and Applied Mathematics
Columbia University
200 S. W. Mudd Building
500 W. 120th Street
New York, NY 10027, USA
e-mail: david.keyes@columbia.edu

Risto M. Nieminen
Laboratory of Physics
Helsinki University of Technology
02150 Espoo, Finland
e-mail: rni@fyslab.hut.fi

Dirk Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
e-mail: dirk.roose@cs.kuleuven.ac.be

Tamar Schlick
Department of Chemistry
Courant Institute of Mathematical
Sciences
New York University
and Howard Hughes Medical Institute
251 Mercer Street
New York, NY 10012, USA
e-mail: schlick@nyu.edu

Mathematics Editor at Springer:
Martin Peters
Springer-Verlag,
Mathematics Editorial IV
Tiergartenstrasse 17
D-69121 Heidelberg, Germany
Tel.: *49 (6221) 487-8185
Fax: *49 (6221) 487-8355
e-mail: martin.peters@springer.com

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*

2. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*

4. P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, R. D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*

5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*

6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.

7. R. von Schwerin, ***Multi B****ody* ***S****ystem* ***SIM****ulation.* Numerical Methods, Algorithms, and Software.

8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*

9. T. J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*

10. H. P. Langtangen, A. M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing.*

11. B. Cockburn, G. E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.

12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.

13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*

14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*

15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*

16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.

17. B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*

18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*

19. I. Babuška, P. G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*

20. T. J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.

21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*

22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.

23. L. F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H. H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T. J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M. A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R. H. W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M. A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H. P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B. N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

Monographs in Computational Science and Engineering

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition

2. A. Quarteroni, F. Saleri, *Scientific Computing with MATLAB and Octave.* 2nd Edition

3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition

4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.

5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.