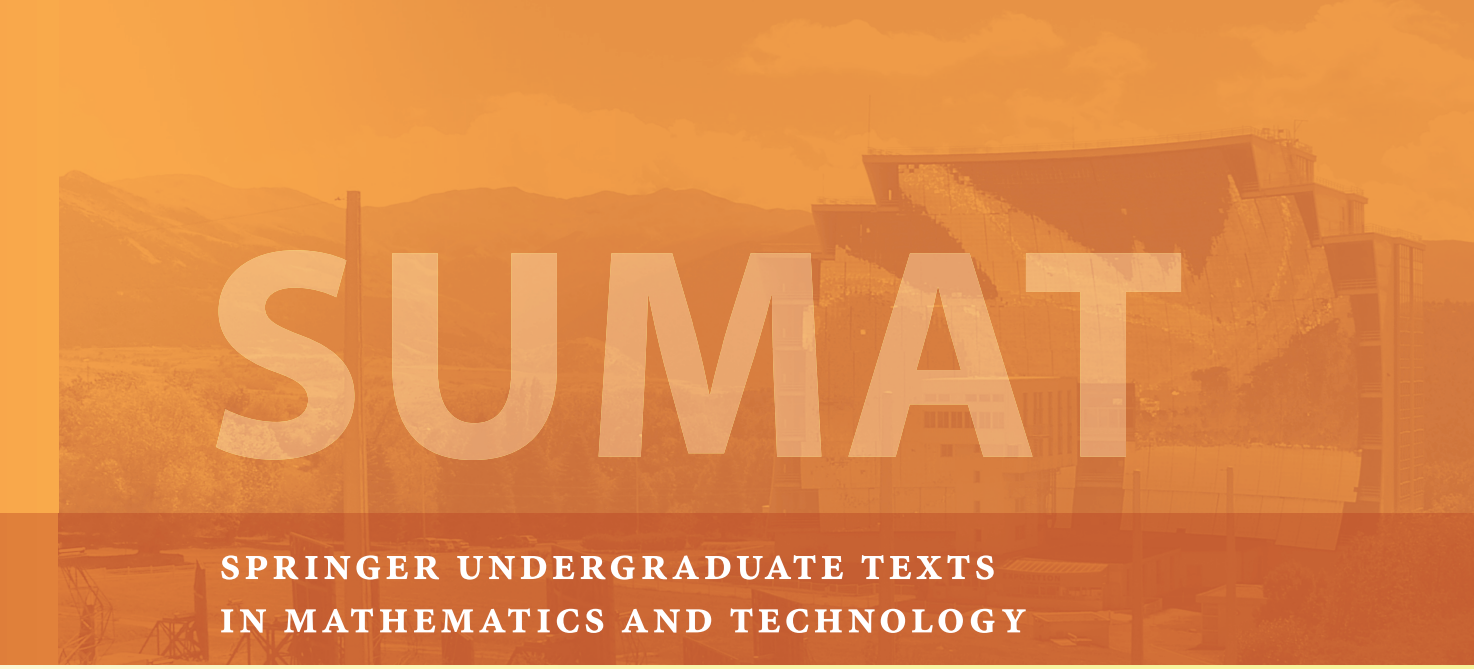


SPRINGER UNDERGRADUATE TEXTS
IN MATHEMATICS AND TECHNOLOGY

Christiane Rousseau
Yvan Saint-Aubin

Mathematics and Technology

Chris Hamilton, Translator

Springer

Springer Undergraduate Texts
in Mathematics and Technology

Christiane Rousseau
Yvan Saint-Aubin

Mathematics and Technology

With the participation of Hélène Antaya and Isabelle Ascah-Coallier

Chris Hamilton, Translator

Springer

Christiane Rousseau
Département de mathématiques
et de statistique
Université de Montréal
C.P. 6128, Succursale Centre-ville
Montréal, Québec H3C 3J7
Canada
rousseac@dms.umontreal.ca

Yvan Saint-Aubin
Département de mathématiques
et de statistique
Université de Montréal
C.P. 6128, Succursale Centre-ville
Montréal, Québec H3C 3J7
Canada
saint@dms.umontreal.ca

ISBN: 978-0-387-69215-9 e-ISBN: 978-0-387-69216-6
DOI: 10.1007/978-0-387-69216-6

Library of Congress Control Number: 2008926885

Mathematics Subject Classification (2000): 00-01,03-01,42-01,49-01,94-01,97-01

Preface

Of what use is mathematics? Hasn't everything in mathematics already been discovered? These are natural questions often asked by undergraduates. The answers provided by their professors are often quite brief. Most university courses, pressed for time and rigidly structured, offer little opportunity to present and study actual applications and real-world examples.

Even more high-school students ask the same questions with more insistence. Teachers in these schools generally work under even tighter constraints than university professors. If they are able to competently respond to these questions it is probably because they received good answers from their teachers and professors. And if they do not have the answers, then whose fault is it?

The genesis of this text

It is impossible to introduce this text without first discussing the course in which it originated. The course "Mathematics and Technology" was created at the Université de Montréal and taught for the first time in the winter semester of 2001. It was created after observing that most courses in the department neglect to present real applications. Since its creation the course has been open to both undergraduate mathematics students and future high-school teachers.

Since no appropriate text or manual for the course we envisioned existed, we were led to write our own course notes, from which we taught. We got so caught up in writing these notes that they quickly grew to the size of a textbook, containing much more material than could possibly be taught in one semester. Despite the two of us being career mathematicians, we must admit that we both knew little or nothing about most of the applications presented in the following chapters.

The goal of the "Mathematics and Technology" course

The primary goals of the course are to demonstrate the active and evolving character of mathematics, its omnipresence in the development of technologies, and to initiate students into the process of modeling as a path to the development of various mathematical applications.

Although a few of the included subjects fall outside the strict domain of technology, we hope to make it clear that, yes, mathematics is useful, and it plays a major role in everyday technologies. Several of the subjects treated in this text are still being actively developed, and this allows students to see, often for the first time, that the field of mathematics remains open and dynamic.

Since the students taking our course include future high school-teachers, it is important to stress that the point is not simply to provide them with examples and applications that they can repeat to their future students, but rather to give them the tools to formulate and develop real-world examples appropriate to their students. They should be instilled with the feeling that they are teaching a subject that is intrinsically elegant, of course, but whose applications have helped shape our physical environment and our understanding of it.

The choice of subjects

In choosing subjects we have paid particular attention to the following points:

- The applications should be recent or affect the students' day-to-day life. Moreover, contrary to the mature mathematics typically taught in other undergraduate courses, some of the mathematics used should be modern or even still in development.
- The mathematics should be relatively elementary and if it exceeds the typical first-year undergraduate curriculum (calculus, linear algebra, probability theory), the missing pieces must be covered within the chapter. A special effort is made to make extensive use of high-school-level mathematics, particularly Euclidean geometry. Basic high-school and undergraduate mathematics form a remarkable toolkit, provided they are well understood and mastered, allowing students to readily explore their wide applications and, often for the first time, to discover their power when used together.
- The level of mathematical sophistication required should remain at a minimum: ideas are a scientist's most precious commodity, and behind most technological successes there lies a brilliant yet sometimes elementary observation.

As a result, the mathematics used in the book covers a very wide spectrum:

- Lines and planes appear in all of their forms (regular equations, parametric equations, subspaces), often in unexpected ways (using the intersection of several planes to decode a Reed–Solomon encoded message).
- A large number of subjects make use of basic geometric objects: circles, spheres, and conics. The concept of *locus of points* in Euclidean geometry is often repeated, for example in problems where we calculate the position of an object through triangulation (Chapter 1 on GPS, and Chapter 15 on *Science Flashes*).
- The different types of affine transformations in the plane or in space (in particular rotation and symmetries) appear several times: in Chapter 11 on image compression using fractals, in Chapter 2 on mosaics and friezes, and in Chapter 3 on robot motion.
- Finite groups appear as symmetry groups (Chapter 2 on mosaics and friezes) and also in the development of primality tests in cryptography (Chapter 7).

- Finite fields make an appearance in Chapter 6 on error-correcting codes, in Chapter 1 on GPS and in Chapter 8 on random-number generation.
- Chapter 7 on cryptography and Chapter 8 on random-number generation both make use of arithmetic modulo n, while Chapter 6 on error-correcting codes makes use of arithmetic modulo 2.
- Probability theory appears in several unexpected places: in Chapter 9 on *Google*'s *PageRank* algorithm, and in the construction of large prime numbers in Chapter 7. It is also used more classically in Chapter 8 on random-number generation.
- Linear algebra is omnipresent: in Chapter 6 on Hamming and Reed–Solomon codes, in Chapter 9 on the *PageRank* algorithm, in Chapter 3 on robot motion, in Chapter 2 on mosaics and friezes, in Chapter 1 on GPS, in Chapter 12 on the JPEG standard, etc.

Using this book as a course text

The text is written for students who have a familiarity with Euclidean geometry and have mastered multivariable calculus, linear algebra, and elementary probability theory. We hope that we have not implicitly assumed any other background knowledge. Working through the text nonetheless requires a certain scientific maturity: it involves integrating a variety of mathematical tools in a setting different from the one in which they were originally taught. For that reason, undergraduates in their junior or senior years are the ideal audience for the course.

The text presents applications in two forms: the main chapters (all except Chapter 15) are long and detailed, while the *Science Flashes* (sections of Chapter 15) are short and narrow in scope. Readers will notice a certain unity in the form of the longer chapters: the first sections describe the application and the underlying mathematical problem; this is followed by an exploration of simple cases of the problem and, if necessary, a development of the required mathematics. We call these parts the *basic* portion of the chapter. Afterward, one or more sections may explore more-complicated examples, provide more details to the mathematical tools discussed earlier, or simply discuss the fact that mathematics alone is not always sufficient! We refer to this latter part of a chapter as the *advanced* portion. Each application is typically covered in 5–6 hours of class: two hours for the basic theory, two hours for examples and exercises and, if time permits, one or two hours for advanced topics. Often we are able only to touch briefly on the advanced material, unless a second week is spent on the chapter. Each *Science Flash* can be treated in an hour of class or even assigned as an exercise without being preceded by any theory development. During a single semester we aim to cover a significant part of 8 to 12 chapters and a handful of *Science Flashes*. Another option is to significantly reduce the number of chapters being covered and to dig further into their advanced sections.

We are thus forced to select subjects as a function of their intrinsic interest or the students' mathematical knowledge. The chapters not selected or the advanced portions of those that were covered are natural points of departure for course projects. Self-guided students who are reading this text on their own may simply jump from chapter

to chapter as the mood strikes them. Each chapter is (mathematically) independent (or very nearly so), and any links among them are explicitly stated.

One last note for professors using this book as a course text. Teaching this course has forced us to revise our usual pedagogical methods: here no subject is prerequisite for further courses, the definitions and theorems are not the ultimate goals of the course, and the problems are not drill. These factors can cause some anxiety on the students' side. Moreover, we are not specialists in any of the technologies we discuss here. So we had to revise our teaching. We try to make as many links as possible to the technology. We encourage students to participate in the course. This allows us to check their background relative to the mathematical tools being used. As for exams, we choose to reassure them from the beginning by stating that the exams are open book, noncumulative, and limited to the basic material. Emphasis is put on simple mathematical modeling and problem solving. Our sets of exercises focus on these skills.

Using this book as a self-directed reader

During the writing of this text we have always been passionate about presenting the mathematics underlying technology and demonstrating both its intrinsic beauty and power. We believe that this text will be of interest to any reader, from young scientist to experienced mathematician, curious to understand the mathematics that drives technological innovation. Since the chapters are largely independent, the reader can hop from subject to subject at will. Hopefully, the reader will be equally interested in the many historical notes scattered throughout the text and, who knows, even find time to work through a few of the exercises.

The contributions of Hélène Antaya and Isabelle Ascah-Coallier

The first draft of Chapter 14 on the calculus of variations was written by Hélène Antaya during a summer internship at the end of her junior college. Chapter 13 on computing with DNA was written the following summer by Hélène Antaya and Isabelle Ascah-Coallier while they were supported by an Undergraduate Student Research Award from the National Sciences and Engineering Research Council (NSERC) of Canada.

How to use the chapters

For the most part, chapters are independent. The beginning of each chapter contains a brief "how-to," describing the required basic knowledge, the relationships between the sections, and, if necessary, their relative difficulty.

Christiane Rousseau
Yvan Saint-Aubin

Département de mathématiques et de statistique
Université de Montréal
June 2008

Acknowledgments

The genesis of the "Mathematics and Technology" course and accompanying course notes can be traced back to the winter of 2001. We had to learn a variety of subjects that we knew only incidentally or not at all, and also had to construct sets of exercises and student projects. Throughout the many years of evolution of this text we have asked numerous questions that required a great level of explanation. We would like to thank those who have supported us in this endeavor. Their assistance has helped us reduce the inevitable ambiguities and errors; we are responsible for any that remain, and we invite our readers to report any they may find.

We learned much from Jean-Claude Rizzi, Martin Vachon, and Annie Boily, all from Hydro-Québec, who helped us learn about storm tracking; from Stéphane Durand and Anne Bourlioux about the finer points of GPS; from Andrew Granville on recent integer factorization algorithms; from Mehran Sahami about the inner workings of Google; from Pierre L'Ecuyer about random-number generators; from Valérie Poulin and Isabelle Ascah-Coallier about how quantum computers function; from Serge Robert, Jean LeTourneux, and Anik Soulière on the relationship between math and music; from Paul Rousseau and Pierre Beaudry about basic computer architecture; from Mark Goresky about linear shift registers and the properties of the sequences they generate. David Austin, Robert Calderbank, Brigitte Jaumard, Jean LeTourneux, Robert Moody, Pierre Poulin, Robert Roussarie, Kaleem Siddiqi, and Loïc Teyssier provided us with references and precious commentary.

Many of our friends and colleagues read portions of the manuscript and provided us with feedback, notably Pierre Bouchard, Michel Boyer, Raymond Elmahdaoui, Alexandre Girouard, Martin Goldstein, Jean LeTourneux, Francis Loranger, Marie Luquette, Robert Owens, Serge Robert, and Olivier Rousseau. Nicolas Beauchemin and André Montpetit helped us on more than one occasion with graphics and the subtleties of LaTeX. We were lucky to have colleagues Richard Duncan, Martin Goldstein, and Robert Owens help us with the English terminology.

Since the first draft we have freely shared our manuscript. Many of our friends and colleagues have encouraged us throughout this adventure, including John Ball, Jonathan Borwein, Bill Casselman, Carmen Chicone, Karl Dilcher, Freddy Dumortier, Stéphane Durand, Ivar Ekeland, Bernard Hodgson, Nassif Ghoussoub, Frédéric Gourdeau, Jacques Hurtubise, Louis Marchildon, Odile Marcotte, and Pierre Mathieu.

We wish to thank Chris Hamilton, who worked for many months on the excellent English translation of our manuscript. Moreover, it was a great pleasure working with him. We appreciate his judicious commentary and suggestions. His clever adaptations, when needed, and his discovery of many errors helped to improve the original French version of the text.

We thank David Kramer, our copyeditor for his expert assistance and excellent suggestions. We are grateful to Ann Kostant and Springer, who showed great interest in our book, from the first version to the printed one.

We would also like to thank those nearest to us, Manuel Giménez, Serge Robert, Olivier Rousseau, Valérie Poulin, Anaïgue Robert, and Chi-Thanh Quach, who have always supported us, including listening to us talk about this project over the years.

Contents

1

Positioning on Earth and in Space

This chapter is the best example in the book of how diverse the applications of mathematics can be to a simple technical question: how can one locate people or events on Earth? This diversity is striking, and to spend more than one week on this chapter can be a good idea for that reason. Two hours are sufficient to cover the theory behind GPS (Section 1.2) and to briefly touch upon the application of GPS to storm tracking (Section 1.3). Afterward, there is a choice to be made. If you have already introduced finite fields in Chapter 6 on error-correcting codes or Chapter 8 on random-number generators, then the mechanics of the GPS signal can be covered in a little more than an hour, since you may skip the review of finite fields. If time is limited and finite fields have not yet been introduced, a reasonable compromise is simply to state Theorem 1.4 and to illustrate it using several examples such as Example 1.5. Section 1.5 on cartography will require a minimum of two hours, unless the students are already familiar with the notion of conformal maps. Section 1.2 requires only Euclidean geometry and basic linear algebra, while Section 1.3 uses elementary probability concepts. Section 1.4 is more difficult unless one has some knowledge of finite fields. Section 1.5 makes use of multivariable calculus.

1.1 Introduction

Since the beginning of time man has been interested in determining his position on the Earth. He started with primitive instruments, navigating through the use of the magnetic compass, the astrolabe, and later the sextant. Recent history has seen the development of significantly more complex and accurate navigational aids, such as the Global Positioning System (GPS). In this chapter we walk backward through time: we will start by discussing modern-day GPS, followed by a brief discussion of ancient techniques, mostly in the exercises.

Since such navigational techniques are really useful only if we have accurate maps of the world, we will dedicate a section to cartography. Since the Earth is a sphere,

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_1,

it is impossible to represent it on a sheet of paper in a manner that preserves angles, relative distances, and relative areas. The chosen compromise depends largely on the application. The Peters Atlas has chosen to use projections that preserve relative area [2]. Marine charts, on the other hand, have chosen projections that preserve angles.

1.2 Global Positioning System

1.2.1 Some Facts about GPS

The GPS constellation of satellites was completed in July 1995 by the Defense Department of the USA, and was authorized for use by the general public. When it was first deployed, the system consisted of 24 satellites designed such that at least 21 would be functioning 98% of the time. In 2005 the system consisted of 32 satellites, of which at least 24 are to be functioning while the others are ready to take over in case a satellite fails. The satellites are positioned 20,200 km from the surface of the Earth. They are distributed across 6 orbital planes, each tilted at an angle of 55 degrees to the equatorial plane (see Figure 1.1). There are at least 4 satellites per orbital plane, roughly equidistant from each other. Each satellite completes a circular orbit around the Earth in 11 hours and 58 minutes. The satellites are situated such that at any moment and at any location on Earth we may observe at least 4 satellites.

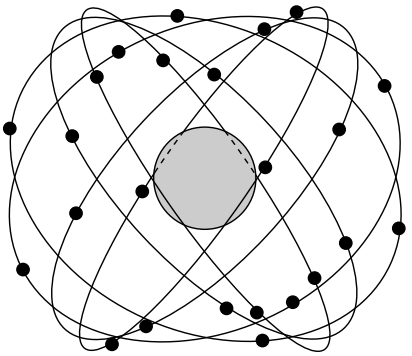


Fig. 1.1. The 24 satellites on 6 orbital planes.

The 24 satellites emit a signal that repeats periodically, and which is received with the aid of a special receiver. When we buy a GPS we are in fact buying a device (which we will call the receiver) that receives the GPS signals and uses the information in them to calculate its location. It contains an almanac with which it is able to calculate the absolute position of each satellite at any given moment of time. However, since slight errors in the orbit are inevitable, correcting information for each satellite is coded

directly within the emitted signal (this correcting information is updated every hour). Each satellite emits its signal continuously. The period of the signal is fixed, and the start time of the cycle may be determined through the use of the almanac. Additionally, each satellite is equipped with an extremely precise atomic clock allowing it to stay synchronized to the start times contained in the almanac. When a receiver records a signal from a satellite, it immediately starts comparing it with one that it generates and that is supposed to match perfectly the one received. In general, these signals will not immediately match. Thus, the receiver shifts the copy it generates until it is in phase with the received signal (which it determines through calculating the correlation between the two). In such a manner, the device is able to calculate the time it takes for the signal to arrive from the satellite. We will discuss this system in much more detail in Section 1.4.

The system described above is the standard precision GPS system. In absence of more sophisticated ground-based corrections, it permits the calculation of receiver position to about 20 meters. Prior to May 2000, the Department of Defense intentionally introduced inaccuracies to the satellite signals in order to reduce the precision of the system to 100 meters.

1.2.2 The Theory Behind GPS

How does the receiver calculate its position? We will start by assuming that the clocks of the receiver and all of the satellites are perfectly synchronized. The receiver calculates its position through triangulation. The basic principle of triangulation methods is to determine where a person (object) is located by using some knowledge relating the position of the person (object) with respect to reference objects whose positions are known. In the case of the receiver of the GPS, it calculates its distance to the satellites, whose positions are known.

- The receiver measures the time t_1 it takes for the signal emitted from satellite P_1 to reach it. Given that the signal travels at the speed of light c, the receiver can calculate its distance from the satellite as $r_1 = ct_1$. The set of points situated at a distance r_1 from the satellite P_1 forms a sphere S_1 centered at P_1 with radius r_1. So we know that the receiver is on S_1. Consider these points as defined in a Cartesian coordinate system. Let (x, y, z) be the unknown position of the receiver and let (a_1, b_1, c_1) be the known position of the satellite P_1. Then (x, y, z) must satisfy the equation describing points on the sphere S_1, namely

$$(x - a_1)^2 + (y - b_1)^2 + (z - c_1)^2 = r_1^2 = c^2 t_1^2. \tag{1.1}$$

- This piece of information is insufficient to determine the precise position of the receiver. The receiver therefore records the signal of a second satellite P_2, recording the time t_2 that the signal took to arrive and calculating the distance $r_2 = ct_2$ to the satellite. As before, it must be that the receiver lies on the sphere S_2 of radius r_2 centered at (a_2, b_2, c_2):

$$(x-a_2)^2+(y-b_2)^2+(z-c_2)^2=r_2^2=c^2t_2^2. \tag{1.2}$$

This narrows down our search, since the intersection of two overlapping spheres is a circle. Thus, we have now narrowed down the position of the receiver to a circle $C_{1,2}$ on which the receiver must lie. However, we again do not know precisely where the receiver is on this circle.

- In order for the receiver to calculate its final position, it needs to capture and process the signal received from a third satellite P_3. Once again, the receiver measures the time t_3 for the signal to arrive and calculates its distance $r_3=ct_3$ from it. As before, it follows that the receiver lies somewhere on the sphere S_3 of radius r_3 centered at (a_3,b_3,c_3):

$$(x-a_3)^2+(y-b_3)^2+(z-c_3)^2=r_3^2=c^2t_3^2. \tag{1.3}$$

The receiver is therefore at the intersection of the circle $C_{1,2}$ and the sphere S_3. Since a sphere and a circle intersect at two points, it may seem that we are not yet sure of the position of the receiver. Fortunately, this is not the case. In fact, the satellites have been positioned such that one of the two solutions will be completely unrealistic, being quite far away from the surface of the Earth. Thus, by finding the two solutions of the system $(*)$ of equations formed by equations (1.1), (1.2), and (1.3), and subsequently eliminating the spurious solution, the receiver may calculate its precise position.

Solving the system $(*)$. The equations of system $(*)$ are quadratic, not linear, which complicates the solution. You may have observed, however, that if we subtract one of the equations from another we obtain a linear equation, since the terms x^2, y^2, and z^2 cancel. Thus, we replace the system $(*)$ by an equivalent system obtained by replacing the first equation by (1.1)−(1.3) and the second equation by (1.2)−(1.3) and by keeping the third equation. This results in the system

$$2(a_3-a_1)x+2(b_3-b_1)y+2(c_3-c_1)z=A_1, \tag{1.4}$$
$$2(a_3-a_2)x+2(b_3-b_2)y+2(c_3-c_2)z=A_2, \tag{1.5}$$
$$(x-a_3)^2+(y-b_3)^2+(z-c_3)^2=r_3^2=c^2t_3^2, \tag{1.6}$$

where

$$\begin{aligned} A_1 &= c^2(t_1^2-t_3^2)+(a_3^2-a_1^2)+(b_3^2-b_1^2)+(c_3^2-c_1^2),\\ A_2 &= c^2(t_2^2-t_3^2)+(a_3^2-a_2^2)+(b_3^2-b_2^2)+(c_3^2-c_2^2). \end{aligned} \tag{1.7}$$

The satellites have been placed in such a manner that no three satellites will ever fall along a line. This property guarantees that at least one of the 2×2 determinants

$$\begin{vmatrix} a_3-a_1 & b_3-b_1 \\ a_3-a_2 & b_3-b_2 \end{vmatrix}, \quad \begin{vmatrix} a_3-a_1 & c_3-c_1 \\ a_3-a_2 & c_3-c_2 \end{vmatrix}, \quad \begin{vmatrix} b_3-b_1 & c_3-c_1 \\ b_3-b_2 & c_3-c_2 \end{vmatrix}$$

is nonzero. In fact, if all three determinants are zero, then the vectors $(a_3-a_1, b_3-b_1, c_3-c_1)$ and $(a_3-a_2, b_3-b_2, c_3-c_2)$ are collinear (their cross product is zero), implying that the three points P_1, P_2, and P_3 fall on a line.

Suppose that the first determinant is nonzero. Using Cramer's rule, the first two equations of (1.6) can give us solutions for x and y as a function of z:

$$\begin{aligned} x &= \frac{\begin{vmatrix} A_1 - 2(c_3 - c_1)z & 2(b_3 - b_1) \\ A_2 - 2(c_3 - c_2)z & 2(b_3 - b_2) \end{vmatrix}}{\begin{vmatrix} 2(a_3 - a_1) & 2(b_3 - b_1) \\ 2(a_3 - a_2) & 2(b_3 - b_2) \end{vmatrix}}, \\ y &= \frac{\begin{vmatrix} 2(a_3 - a_1) & A_1 - 2(c_3 - c_1)z \\ 2(a_3 - a_2) & A_2 - 2(c_3 - c_2)z \end{vmatrix}}{\begin{vmatrix} 2(a_3 - a_1) & 2(b_3 - b_1) \\ 2(a_3 - a_2) & 2(b_3 - b_2) \end{vmatrix}}. \end{aligned} \tag{1.8}$$

Substituting these values into the third equation of (1.6) yields a quadratic equation in z, which we may solve to find the two solutions z_1 and z_2. Back-substituting z for the values z_1 and z_2 into the two above equations yields the corresponding values x_1, x_2, y_1, and y_2. We could easily find closed forms to these solutions, but the formulas involved quickly become too large to offer any insight or convenience.

Choosing the axes of our coordinate system. Nowhere in the above discussion did we mention or were we forced to choose a set of axes for our coordinate system. However, to facilitate the translation from absolute coordinates to latitude, longitude, and altitude we make the following choice:

- the center of the coordinate system is the center of the Earth;
- the z axis passes through the two poles, and is oriented toward the North Pole;
- the x and y axes both lie in the equatorial plane;
- the positive x axis passes through the point of 0 degrees longitude;
- the positive y axis passes through the point of longitude 90 degrees west;

Since the radius R of the Earth is approximately 6365 km, a solution (x_i, y_i, z_i) is considered acceptable if $x_i^2 + y_i^2 + z_i^2 \approx (6365 \pm 50)^2$. The uncertainty of 50 km allows a window for the altitudes of mountains and airplanes. A more natural coordinate system for expressing points near the surface of the Earth is the longitude L, the latitude l, and the distance h from the center of the Earth (the altitude above sea level is therefore given by $h - R$). Longitude and latitude are angles that will be expressed in degrees. If a point (x, y, z) lies exactly on the sphere of radius R (in other words, if the point lies at altitude zero), then its longitude and latitude may be found by solving the following system of equations:

$$\begin{aligned} x &= R \cos L \cos l, \\ y &= R \sin L \cos l, \\ z &= R \sin l. \end{aligned} \tag{1.9}$$

Since $l \in [-90°, 90°]$, we obtain

$$l = \arcsin \frac{z}{R}, \tag{1.10}$$

allowing us to calculate $\cos l$. The longitude L is therefore uniquely determined by the two equations

$$\begin{cases} \cos L = \dfrac{x}{R \cos l}, \\ \\ \sin L = \dfrac{y}{R \cos l}. \end{cases} \tag{1.11}$$

Calculating the position of the receiver. Let (x, y, z) be the position of the receiver. We begin by calculating the distance h of the receiver from the center of the Earth, given by

$$h = \sqrt{x^2 + y^2 + z^2}.$$

We now have two choices for calculating the latitude and longitude: adapt the formulas (1.10) and (1.11) by replacing all occurrences of R with h, or project the position (x, y, z) to the surface of the sphere and use these values in the equations (1.10) and (1.11):

$$(x_0, y_0, z_0) = \left(x\frac{R}{h}, y\frac{R}{h}, z\frac{R}{h} \right).$$

The altitude of the receiver is given by $h - R$.

1.2.3 Dealing with Practical Difficulties.

We have just presented the theory behind calculating the position, which holds true in a perfect world. Unfortunately, real life is vastly more complicated, since the times being measured are extremely short and must be measured to high precision. The satellites are each equipped with an extremely precise (and expensive!) atomic clock allowing them to be (very nearly) perfectly in sync. Meanwhile, the average receiver is typically equipped with only a mediocre clock, allowing it to be within the budget of most everyone. Assuming that the clocks of the satellites are in sync, the receiver is easily capable of calculating precise transit times for the signals from the satellites. However, given that the receiver is not perfectly in sync, it will actually be calculating three fictitious transit times T_1, T_2, and T_3. How do we deal with these inaccurate measurements? When we had three unknowns, x, y, z, we had needed three measured times t_1, t_2, t_3, to find the unknowns. Now the fictitious time T_i measured by the receiver is given by

$$\begin{aligned} T_i \;=\; & (\text{arrival time of the signal on the receiver's clock}) \\ & - (\text{departure time of the signal on the satellite's clock}). \end{aligned}$$

The solution comes from the fact that the error between the fictitious time T_i calculated by the receiver and the actual time t_i is the same, regardless of the satellite from which the measurement was taken. That is, $T_i = \tau + t_i$, for $i = 1, 2, 3$, where

$$\begin{aligned} t_i &= \text{(arrival time of the signal on the satellite's clock)} \\ &\quad - \text{(departure time of the signal on the satellite's clock)} \end{aligned}$$

and τ is given by the equation

$$\begin{aligned} \tau &= \text{(arrival time of the signal on the receiver's clock)} \\ &\quad - \text{(arrival time of the signal on the satellite's clock).} \end{aligned} \tag{1.12}$$

The constant τ represents the clock offset between the clocks on the satellites and the receiver's clock. This introduces a fourth unknown, τ, to our original system of three unknowns x, y, z. In order to resolve the system of equations to a finite set of solutions we must obtain a fourth equation. This is simple to do in our context: the receiver simply measures the offset signal transit time T_4 between itself and a fourth satellite P_4. Since $t_i = T_i - \tau$ for $i = 1, \ldots, 4$, our system then becomes:

$$\begin{aligned} (x-a_1)^2 + (y-b_1)^2 + (z-c_1)^2 &= c^2(T_1-\tau)^2, \\ (x-a_2)^2 + (y-b_2)^2 + (z-c_2)^2 &= c^2(T_2-\tau)^2, \\ (x-a_3)^2 + (y-b_3)^2 + (z-c_3)^2 &= c^2(T_3-\tau)^2, \\ (x-a_4)^2 + (y-b_4)^2 + (z-c_4)^2 &= c^2(T_4-\tau)^2 \end{aligned} \tag{1.13}$$

where we have the four unknowns x, y, z, and τ. As before, we can use elementary operations to replace three of these quadratic equations by linear equations. To do this we subtract the fourth equation from each of the first three, resulting in:

$$\begin{aligned} 2(a_4-a_1)x + 2(b_4-b_1)y + 2(c_4-c_1)z &= 2c^2\tau(T_4-T_1) + B_1, \\ 2(a_4-a_2)x + 2(b_4-b_2)y + 2(c_4-c_2)z &= 2c^2\tau(T_4-T_2) + B_2, \\ 2(a_4-a_3)x + 2(b_4-b_3)y + 2(c_4-c_3)z &= 2c^2\tau(T_4-T_3) + B_3, \\ (x-a_4)^2 + (y-b_4)^2 + (z-c_4)^2 &= c^2(T_4-\tau)^2, \end{aligned} \tag{1.14}$$

where

$$\begin{aligned} B_1 &= c^2(T_1^2 - T_4^2) + (a_4^2 - a_1^2) + (b_4^2 - b_1^2) + (c_4^2 - c_1^2), \\ B_2 &= c^2(T_2^2 - T_4^2) + (a_4^2 - a_2^2) + (b_4^2 - b_2^2) + (c_4^2 - c_2^2), \\ B_3 &= c^2(T_3^2 - T_4^2) + (a_4^2 - a_3^2) + (b_4^2 - b_3^2) + (c_4^2 - c_3^2). \end{aligned} \tag{1.15}$$

In the system of equations (1.14), Cramer's rule applied to the first three equations allows us to determine values for x, y, and z as a function of τ:

$$
\begin{aligned}
x &= \frac{\begin{vmatrix} 2c^2\tau(T_4-T_1)+B_1 & 2(b_4-b_1) & 2(c_4-c_1) \\ 2c^2\tau(T_4-T_2)+B_2 & 2(b_4-b_2) & 2(c_4-c_2) \\ 2c^2\tau(T_4-T_3)+B_3 & 2(b_4-b_3) & 2(c_4-c_3) \end{vmatrix}}{\begin{vmatrix} 2(a_4-a_1) & 2(b_4-b_1) & 2(c_4-c_1) \\ 2(a_4-a_2) & 2(b_4-b_2) & 2(c_4-c_2) \\ 2(a_4-a_3) & 2(b_4-b_3) & 2(c_4-c_3) \end{vmatrix}}, \\
y &= \frac{\begin{vmatrix} 2(a_4-a_1) & 2c^2\tau(T_4-T_1)+B_1 & 2(c_4-c_1) \\ 2(a_4-a_2) & 2c^2\tau(T_4-T_2)+B_2 & 2(c_4-c_2) \\ 2(a_4-a_3) & 2c^2\tau(T_4-T_3)+B_3 & 2(c_4-c_3) \end{vmatrix}}{\begin{vmatrix} 2(a_4-a_1) & 2(b_4-b_1) & 2(c_4-c_1) \\ 2(a_4-a_2) & 2(b_4-b_2) & 2(c_4-c_2) \\ 2(a_4-a_3) & 2(b_4-b_3) & 2(c_4-c_3) \end{vmatrix}}, \qquad (1.16) \\
z &= \frac{\begin{vmatrix} 2(a_4-a_1) & 2(b_4-b_1) & 2c^2\tau(T_4-T_1)+B_1 \\ 2(a_4-a_2) & 2(b_4-b_2) & 2c^2\tau(T_4-T_2)+B_2 \\ 2(a_4-a_3) & 2(b_4-b_3) & 2c^2\tau(T_4-T_3)+B_3 \end{vmatrix}}{\begin{vmatrix} 2(a_4-a_1) & 2(b_4-b_1) & 2(c_4-c_1) \\ 2(a_4-a_2) & 2(b_4-b_2) & 2(c_4-c_2) \\ 2(a_4-a_3) & 2(b_4-b_3) & 2(c_4-c_3) \end{vmatrix}}.
\end{aligned}
$$

None of this makes sense unless the denominator is nonzero. However, the denominator is zero if and only if the four satellites are situated in the same plane (see Exercise 1). Once again, the satellites are laid out such that no four of them that are visible from a given point on the Earth will ever lie in the same plane. We forward-substitute the solutions to the first three equations into the fourth, resulting in a final quadratic equation in τ, which yields two solutions τ_1 and τ_2. Back-substituting these into (1.16) yields two possible positions for the receiver, and we use the same trick as before to eliminate the spurious solution.

Which satellites should the receiver choose if it can see more than four? In this case, the receiver has a choice for which data to use in the calculations. It makes sense to use the data that will introduce the minimal amount of error. In reality, the time measurements are all approximate. This implies that the calculated distances to the satellites are only approximate. Graphically, we could represent the area of incertitude by thickening the shell of each sphere. The intersection of the thick spheres then becomes a set, the size of which is related to the uncertainty of the solution. Thinking geometrically, it is easy to convince ourselves that the greater the angle between the surfaces of two intersecting thick spheres, the smaller the volume of space swept out by the intersection. Conversely, if the spheres intersect almost tangentially, then the volume of intersection (and hence uncertainty) is bigger. We thus want to choose the spheres S_i that intersect each other at as large an angle as possible (see Figure 1.2).

This is the geometric intuition behind our choice. Algebraically, we see that the values of x, y, and z in terms of τ are obtained by dividing by

Fig. 1.2. A small angle of intersection at the left (loss of precision) and a large angle at the right.

$$\begin{vmatrix} 2(a_4 - a_1) & 2(b_4 - b_1) & 2(c_4 - c_1) \\ 2(a_4 - a_2) & 2(b_4 - b_2) & 2(c_4 - c_2) \\ 2(a_4 - a_3) & 2(b_4 - b_3) & 2(c_4 - c_3) \end{vmatrix}.$$

The smaller the denominator, the larger the error. Thus, we want to choose the four satellites that maximize this denominator.

More advanced investigations into this topic would fit easily into a course project.

A few refinements:

- **Differential GPS (DGPS):** One source of imprecision in GPS comes from the fact that distances are calculated to the satellites using the constant c, which is the speed of light in a vacuum. In reality, the signal is traveling and refracting through the atmosphere, which both lengthens its trajectory and decreases its speed. To obtain a better approximation to the actual average speed of the signal on the path from the satellite to the receiver we can employ a differential GPS system. The idea is to refine the value of c to be used in calculating satellite distance. We do this by comparing the transit time measured at the receiver and the transit time measured at another nearby receiver at a precisely known position. This allows us to accurately calculate the average speed of light along the path from a given satellite to the receiver, which in turn results in more accurate distance calculations. When helped with such a fixed ground station, GPS precision is on the order of centimeters.
- The signal sent by each satellite is a random signal that repeats at regular known intervals. The period of the signal is relatively short, such that the distance covered by the signal in one period is on the order of a few hundred kilometers. When the receiver sees the beginning of a period of the signal it must determine at precisely which moment in time this period was emitted from the satellite. A priori we have an uncertainty of some integer number of periods.
- Rapidly moving GPS receiver: installing a GPS receiver on a fast-moving object (an airplane, for example) is a very natural application: if an airplane needs to land in

inclement weather, the pilot needs to know its precise position at every instant, and the time to calculate the position must be reduced to an absolute minimum.
- The Earth is not really round! In fact, the Earth is more an ellipsoid that is slightly flattened at the poles and bulging at the equator (an "oblate spheroid"). The radius of the Earth is roughly 6356 km at the poles and 6378 km at the equator. Thus, the calculations for translating Cartesian coordinates (x, y, z) into latitude, longitude, and altitude must be refined to accommodate this fact.
- **Relativistic corrections.** The speed of the satellites is sufficiently large that all of the calculations must be adapted to account for the effects of special relativity. In fact, the clocks on the satellites are traveling very fast compared to those on Earth. As such, the theory of special relativity predicts that these clocks will run slower than those on Earth. Furthermore, the satellites are in relatively close proximity to the Earth, which has significant mass. General relativity predicts a small increase in the speed of the clocks on board the satellites. As a first approximation, we may model the Earth as a large nonrotating spherical mass without any electrical charge. The effect is relatively easy to compute using the Schwarzschild metric, which describes the effects of general relativity under these simplified conditions. As it turns out, this simplification is sufficient to capture the actual effect to high precision. The two effects must both be considered because even though they are in opposite directions, they only partially cancel each other out. For more details see [4].

Applications of GPS. The applications of GPS are numerous, and we name only a few here:

- A GPS receiver allows a person to easily find his/her position when outdoors. As such, it is immediately useful to hikers, kayakers, hunters, sailors, boaters, etc. Most receivers allow the marking of waypoints, which can be saved either when one is physically present at the location (in which case the receiver has calculated its position) or by manually entering map coordinates into the receiver. By joining waypoints with line segments we can in turn represent a route. The receiver may then provide us with our position relative to a chosen waypoint or even give us instructions on how to follow our route. More sophisticated receivers may even store detailed map information. The receiver may then display our position on a portion of the map appearing on the screen, annotated with our waypoints and routes.
- More and more vehicles (especially taxis) are equipped with GPS navigation systems that allow their drivers to find their way to a particular address. In Western Europe and North America there exist several products that provide precise directions to nearly any address.
- Imagine you have an ancient map on which you wish to plot a route you have followed. The route can be saved in the GPS as it is taken, and later uploaded to a computer with the appropriate software. Such software can then superimpose the followed route on the digitized map. If you do not already have a digital version of the map, you may first scan it and (using appropriate software) overlay it with

a coordinate system by simply showing it the location of three known points (see Exercise 5).

- The ubiquitous use of GPS on airplanes allows the size of airways (imaginary corridors of air that airplanes follow between points) to be decreased while still ensuring that airplanes on different airways will stay a safe distance from each other.
- A fleet of delivery vehicles may be equipped with GPS receivers that permit the simultaneous tracking of all vehicles. Such a system is presently used to direct taxis in Paris. In this application the GPS system must be coupled with a communication system allowing the coordinates of each vehicle to be broadcast (an example of such a system is the Global System for Mobile Communications, or GSM). Similar systems are used for tracking wildlife in environmental studies. It is not hard to imagine the impact on our lives if a car rental company equipped its fleet with a GPS-GSM system, allowing it to ensure that clients respect the territorial limits imposed by the rental contract!
- GPS may be used to help blind people find their way.
- Geographers use GPS to measure the growth of Mount Everest: this mountain continues to grow slowly as its glacier, the Khumbu, descends. Similarly, every two years an expedition ascends Mont Blanc to update its official height at the peak. In the nineties, geographers once again asked whether K2 was in fact taller than Mount Everest. Since their 1998 expedition, where they used GPS, the matter is now definitely closed: Mount Everest is the tallest mountain on Earth, at 8830 m. In 1954, the height of Everest was estimated at 8848 m by B. L. Gulatee. At the time, the estimate was computed using theodolite measurements taken from six stations on the north Indian plains (a theodolite is an optical instrument for measuring angles, used in the field of geodesy).
- There are many military applications, considering that the system was originally developed for the use of the American military. One such use is the precise guidance of bombs.

The future: GPS and Galileo. Up until now the United States has had a monopoly in this market. Given that they maintain exclusive control over GPS, the American government can choose to scramble the GPS signal to block access to it or degrade its accuracy over a certain region for military reasons (under the NAVWAR program, for navigational warfare). In March 2002 the European Union and European Space Agency agreed to fund the development and deployment of Galileo, a positioning system designed as an alternative to GPS. Two test satellites were launched in 2005 with the remaining 28 satellites to be launched before the end of 2010. GPS satellites do not actually transmit information regarding the status of the satellite or the quality of the signal itself. It can thus take several hours before a malfunctioning satellite is detected and shut down, with the system accuracy being degraded severely during that time. This restricts the applications of GPS for guiding airplanes in inclement weather. The Galileo satellites are designed to constantly transmit signal quality information, allowing receivers to ignore the signal from malfunctioning satellites. This is done through a

system of ground stations that accurately measure the actual position of the satellite and compare it to the satellite's calculated position. This information is sent to the malfunctioning satellite, which in turn relays it back to receivers. The US government is planning a similar improvement to the GPS system.

1.3 How Hydro-Québec Manages Lightning Strikes

New solutions to existing problems often become apparent as new technology is made available. Hydro-Québec[1] uses GPS as part of its approach to managing lightning strikes. Mathematics is at work in several places in their lightning-strike-monitoring system. As such, this section focuses not only on the application of GPS to managing lightning strikes, but on the mathematics involved elsewhere in their approach.

1.3.1 Locating Lightning Strikes

In 1992, Hydro-Québec installed a lightning strike locating system throughout its network. The basic problem is to determine the boundaries of areas affected by storms, in order to reduce the power transmitted on affected power lines and to reroute it through power lines outside of the stormy area. In doing so, the potential impact of a lightning strike on a power line is minimized: damage caused by a lighting strike is kept localized, thereby minimizing the number of customers affected and increasing the overall reliability of the power grid.

To accomplish this goal, Hydro-Québec uses a system of 13 detectors distributed across the lower two-thirds of the province of Québec (the territory covered by power lines). Their positions are precisely known, but since the system relies on precise time measurements, the clocks in the detectors are required to be perfectly synchronized. To do this, they each use a GPS receiver.

Using a GPS receiver as a time reference. It may seem a little surprising that a GPS receiver can be used to tell time. We just observed that GPS receivers are typically equipped with cheap clocks of relatively low precision. However, we also observed that in calculating its position, the receiver calculates τ, the clock shift between its clock and those on board the GPS satellites. Thus, the receiver actually calculates the precise time as measured by the clocks on board the satellites. When great precision is desired and the receiver is stationary, it is better to replace the calculated values of x, y, z, and τ by the average of several calculated values $(x_i, y_i, z_i, \tau_i)_{i=1}^{N}$ at different times. Indeed, there is an error in each calculation (x_i, y_i, z_i, τ_i). The errors in space can be in any direction around the true receiver position, and they obey a nice statistical law (they are uniform and Gaussian). Similarly, the error in the calculation of the time shift can

[1] Hydro-Québec is the largest producer, transporter, and distributor of electricity in the province of Québec. Its name comes from the fact that 95% percent of its power generation is hydroelectric.

be positive or negative. So the position of the receiver and time shift are more precisely approximated by $(\frac{1}{N}\sum_{i=1}^{N} x_i, \frac{1}{N}\sum_{i=1}^{N} y_i, \frac{1}{N}\sum_{i=1}^{N} z_i, \frac{1}{N}\sum_{i=1}^{N} \tau_i)$.

In such a manner, a GPS receiver is capable of synchronizing its clock to the satellites with a precision of roughly 100 nanoseconds (a nanosecond is 1 billionth of a second). Such an approach is used in the Hydro-Québec detectors. Indeed, the GPS allows the 13 detectors to synchronize their clocks up to 100 nanoseconds. Once the receiver is synchronized with the satellites' clocks it can also "beat the second," i.e., send a pulse every second. This is used for other measurements.

Locating lightning strikes. In addition to maintaining a synchronized clock, the 13 detectors are also responsible for monitoring all abnormal electromagnetic activity and identifying such activity caused by lightning strikes. Hydro-Québec has positioned the detectors quite far from the actual power lines since the electromagnetic fields caused by the power lines would disrupt accurate signal detection. The detectors are typically placed on the roof of Hydro-Québec management buildings, distributed as uniformly as possible throughout the territory to be monitored. When lightning strikes within this territory with sufficient energy to threaten the grid, it is typically recorded by at least five detectors. In fact, the detectors are sufficiently sensitive to locate extremely large lightning strikes as far away as Mexico, but with less precision.

The lightning strike generates an electromagnetic wave, which travels through space at the speed of light. Each detector notes the precise time when the wave was perceived. For this, they use a fast oscillator (for example, a quartz crystal) that is synchronized to the GPS time source. The frequency of such oscillators typically varies from 4 to 16 megahertz (a megahertz is a frequency of one million oscillations per second, abbreviated MHz). The detectors relay this information to a central computer as soon as they have measured the wave. This system then calculates the position of the lightning strike through triangulation (in other words, by using the differences in the times at which the wave was observed by the individual detectors, as explored in Exercise 2).

Identifying lightning strikes. There exist three types of lightning strikes:

- Lightning strikes between clouds. This type forms the majority of lightning strikes. They are not detected, but they do not affect the grid, since they do not strike the ground.
- Negative lightning strikes. In this case the cloud is negatively charged, and the lightning strike consists of a flow of electrons traveling from the cloud to the ground.
- Positive lightning strikes. In this case the cloud is positively charged, and the lightning strike consists of a flow of electrons traveling from the ground to the cloud. As you may have guessed, the wave of a positive strike is thus the mirror image of that of a negative strike.

If we limit ourselves to lightning strikes between the ground and the clouds, 90% of such strikes are negative. However, during a strong storm this percentage is reversed, and 90% of the ground lightning strikes are positive. The detectors can differentiate between a negative and a positive lightning strike: one is the mirror image of the other.

If one detector were to register a wave for a positive lightning strike, and another were to register a negative lightning strike, it stands to reason that these two waves could not have been generated by the same strike. Unfortunately, it is a little more complicated than that. A wave that has traveled more than 300 km from its source may be reflected by the ionosphere, inverting the signal. Thus a detector situated far enough away may actually be measuring a reflected signal.

To differentiate between lightning strikes and other electromagnetic signals, the detector analyzes the shape of the wave by filtering the signal and looking for the specific signature of a lightning strike. In particular, the detector notes the beginning of the signal, the maximum amplitude, the number of peaks, and the slope of the rise, sending this information to the central computer. Signal processing is a beautiful subject of applied mathematics, but we will not discuss it here.

From theory to practice. There are several additional tricks that may be employed to correctly identify received signals.

- Let P and Q be the two detectors that are the furthest apart from each other, and let T be the time necessary for a signal to travel between P and Q at the speed of light. We can be sure that the time difference between the two detected signals for the same lightning strike can be no more than T. Thus, if two detectors registered a strike at times t_1 and t_2 such that $|t_1 - t_2| > T$, then these signals could not have come from the same lightning strike.
- The amplitude of the wave generated by the lightning strike is inversely proportional to the square of the distance to its source. Thus, in order for two detected signals to correspond to the same lightning bolt, the amplitudes of the measured signals must be compatible with the calculated location.
- If lightning strikes within 20 km of a detector, the readings from the detector are eliminated from the calculation. This is because the measured amplitude is too large, and the detector is not able to detect difference between a signal of a single lightning strike and a superposed signal from two lightning strikes.

With these methods Hydro-Québec is able to locate lightning strikes within 500 m of accuracy when they fall within the area covered by the detectors. The accuracy diminishes for lightning strikes outside of the covered territory.

Locating faults in the power lines. A similar method is used to locate faults in the transport network: for example, if a lightning strike has damaged a power line, technicians need to know where to go in order to repair it. On either end of each power line to be protected an oscilloperturbograph is installed, synchronized by GPS. This device measures the form of the 60 Hz signal traveling through the power line. Depending on the fault, there will be different types of perturbation observed. The perturbation travels along the power line at the speed of light. The two detectors measure the times t_1 and t_2 at which the perturbation is observed, and using the difference $t_1 - t_2$, the location of the fault can be calculated. These techniques are precise only within a few hundred meters, but this is generally sufficient. In Quebec the power lines are often

very long and traverse immense uninhabited areas; thus the system allows for a rapid deployment of a repair team to the area of the fault.

Redistributing power transmission. Lightning-strike detection can be used to determine the size and location of a storm. Since lightning strikes occur in a random manner within a territory, statistical models can be used. For this, the territory is divided into an even grid, and a spatiotemporal density of lightning strikes is calculated. For example, a storm with two lightning strikes per km^2 within 10 minutes is very strong. Using the information from the model, the heart of the storm is calculated (the storm centroid). The calculation is repeated every five minutes, and the displacement of the calculated centroid used to infer the speed and direction of the storm (which can be anywhere from 0 to 200 km/h). This information can in turn be used to predict which areas of the power grid will be affected next. One of the more difficult problems to be solved is that of two storms near each other: the system must decide whether there are in fact two separate storms, or a single larger storm. An interesting challenge for engineers!

Armed with this information, the distributor draws upon his experience to decide whether to lower the amount of power being transmitted over a potentially affected power line. Keeping a power grid in equilibrium is a very delicate operation. There must always be a balance between the amount of power being generated, the amount of power being transmitted, and the amount of power being used. In order to diminish the amount of power being transmitted on one line, there must be excess capacity on one or more other lines. Thus, in order to make such decisions the distributor must have a certain margin for maneuver. Each line has a maximum capacity, but as a rule, power grids are always operated slightly below capacity so that the system can absorb the loss of an entire line at any given moment.

1.3.2 Threshold and Quality of Lightning-Strike Detection

The detector equipment is tested to meet minimum standards for detection, but one can generally do better. It is thus worthwhile to accurately gauge their actual capabilities, a process that relies principally on statistical methods.

To this end we will draw upon an empirical law of probability for the random variable X, giving the intensity of a lightning strike. Rather than using the density function $f(I)$ of lightning strikes, we will use the distribution function

$$P(I) = \text{Prob}(X > I) = \frac{1}{1 + \left(\frac{I}{M}\right)^K}. \tag{1.17}$$

We have that $P(0) = \text{Prob}(X > 0) = 1$. The values of M and K to be used depend on the geographic zone and the particulars of its environment and are determined empirically. The value of I is given in kA (kiloamperes). Certain values are used sufficiently often to merit being given a name. Thus, the function P from (1.17) is called the Popolansky function when $M = 25$ and $K = 2$. It is called the Anderson–Erikson

function when $M = 31$ and $K = 2.6$. Figure 1.3 represents the Popolansky function and Figure 1.4 represents the density function $f(I)$ of the associated variable X. Recall that $P(I) = \int_I^\infty f(J)dJ$ and therefore that $f(I) = -P'(I)$.

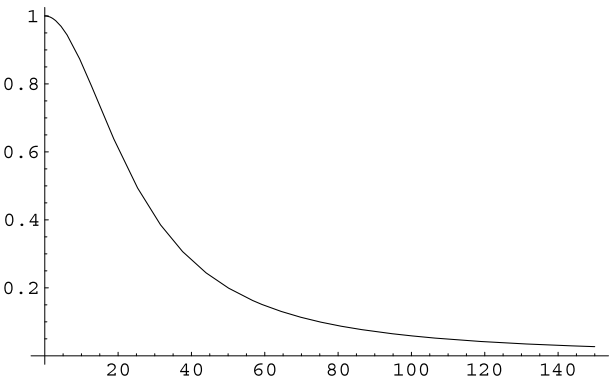


Fig. 1.3. The Popolansky function.

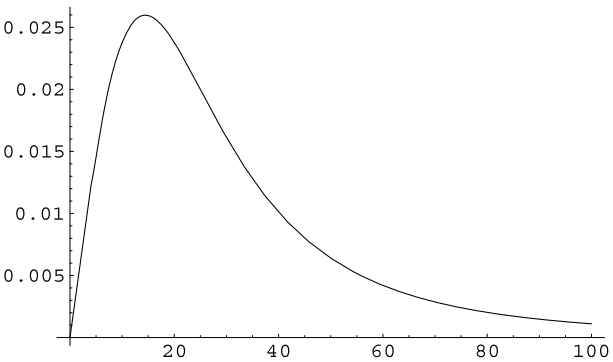


Fig. 1.4. The density function associated with the Popolansky function.

We demonstrate how this empirical law can be used in practice.

Example 1.1 The Popolansky function

1. *The probability of a random lightning bolt having an amperage greater than 50 kA is*

$$P(50) = \frac{1}{1 + \left(\frac{50}{25}\right)^2} = \frac{1}{5} = 0.2. \tag{1.18}$$

2. *The median of this distribution is the value I_m of I such that*

$$Prob(X > I_m) = P(I_m) = \frac{1}{2}. \tag{1.19}$$

This gives us the equation $\frac{1}{1+\left(\frac{I_m}{25}\right)^2} = \frac{1}{2}$. Thus $1 + \left(\frac{I_m}{25}\right)^2 = 2$, or in other words $\left(\frac{I_m}{25}\right)^2 = 1$. Hence $I_m = 25$.

Calculating the rate of lightning-strike detection. In practice we do not actually detect all lightning strikes, but only those with energy greater than a certain threshold. This threshold depends on the position of the lightning strike with respect to the detectors and on various sources of interference that may decrease the reception quality of the detectors at any given moment in time. We will explore how to determine the percentage of lightning strikes that are detected. In our example we determined that 50% of lightning strikes have an amperage higher than 25 kA. Suppose for now that in a sample of detected lightning strikes we observed that 60% had an amperage higher than 25 kA. Let E be the event "the lightning strike is detected." Then we wish to calculate Prob(E). We know the probability that a *detected* lightning strike (in other words, that event E took place) had an amperage higher than 25 kA. This is a conditional probability because we have assumed that the lightning strike was detected, and it may be written as

$$\text{Prob}(X > 25|E) = 0.6. \tag{1.20}$$

On the other hand, we know that the conditional probability of $X > 25$ knowing that E has occurred can be expressed as

$$\text{Prob}(X > 25|E) = \frac{\text{Prob}(X > 25 \text{ and } E)}{\text{Prob}(E)}. \tag{1.21}$$

As such we cannot do much with this expression, since both the numerator and the denominator are unknown. But suppose we can assume that all lightning strikes with an amperage higher than 25 kA are detected. Then the event "$X > 25$ and E" becomes simply $X > 25$, whose probability is known. Thus (1.21) provides

$$\text{Prob}(E) = \frac{\text{Prob}(X > 25)}{\text{Prob}(X > 25|E)} = \frac{0.5}{0.6} = \frac{5}{6} = 0.83. \tag{1.22}$$

Suppose now that for a given limited geographic region we can assume the hypothesis (with a reasonable margin of error) that the only lightning strikes not detected are those that have a weaker amperage. We may wish to determine the threshold amperage I_0 below which lightning strikes are not detected. For this calculation the event E becomes $X > I_0$. We have seen that Prob(E) = $\frac{5}{6}$ = 0.83. Since Prob(E) = Prob($X > I_0$) = $P(I_0)$, this gives the equation

$$P(I_0) = \frac{0.5}{0.6} = \frac{5}{6}, \tag{1.23}$$

while comes to $\frac{1}{1+\left(\frac{I_0}{25}\right)^2} = \frac{5}{6}$, or equivalently $1 + \left(\frac{I_0}{25}\right)^2 = \frac{6}{5}$. Hence I_0 is the value satisfying $\left(\frac{I_0}{25}\right)^2 = \frac{1}{5} = 0.2$, yielding

$$I_0 = 25\sqrt{0.2} = 11.18. \tag{1.24}$$

We can therefore conclude that for the given region, the threshold of detection is $I_0 = 11.18$ kA, and that lightning strikes with amperages below this value are not detected.

1.3.3 Long-Term Risk Management

Managing lightning strikes is not limited to the task of detecting and locating storms. Hydro-Québec keeps detailed long-term statistics that are used to construct isokeraunic maps giving the density of lightning strikes over a period of five years. Such a map can then be used to identify which zones are subject to more risk. In the case of power lines that have already been built, this information can be used to decide which sections should be better protected. Similarly, such maps allow for ready identification of routes to take during the construction of new power lines. These choices can be rewritten in a risk-management framework.

Risks due to violent storms are but one of many risks faced by a company that produces, transports, and distributes electricity. Hence, all of the tasks of locating lightning strikes, storm tracking and identifying risk zones may be used as part of a general risk-management framework. The problem is to make the distribution network as reliable as possible. Investment in the grid for this purpose represents the cost. Thus, individual investments in the grid have to be evaluated in terms of their profitability. The more dangerous a given event and the larger its financial impact, the more prepared we are to invest in protecting the system from the event or limiting its impact. Naturally, this is always subject to the condition that the cost of the protection not be too high! To formalize such a system we introduce three variables:

- the probability p of the event at risk;
- the projected cost C_i were the event to occur without precautions taken to mitigate it;
- the cost of attenuation C_a, which is to be paid to protect equipment and limit the impact of the event occurring.

We introduce the index

$$\frac{pC_i}{C_a}. \tag{1.25}$$

We see that the numerator represents the expected cost of repairs and that the denominator represents the cost of protection. We must have this ratio at least 1 in order for investing in protection to be profitable. However, there are several other factors

that come into play in practice. We are more likely to purchase protection if it is valid for multiple events. Similarly, the situation changes if the protection is only partial, meaning we incur some reduced repair cost if the event occurs, as opposed to being total.

1.4 Linear Shift Registers and the GPS Signal

Linear shift registers allow the generation of sequences that have excellent properties in terms of allowing a receiver to synchronize with them. These simple-to-build devices (one can build a linear shift register with a few basic electrical components) generate pseudorandom signals. That is, they generate signals that appear to be largely random even though they are generated by deterministic algorithms.

We will construct a linear shift register that generates a periodic signal with a period of $2^r - 1$. It will have the property that it is extremely poorly correlated with all translations of itself and with other signals generated by the same register using different coefficients. This property of having a signal that correlates poorly to its translations and other similar signals permits GPS receivers to easily identify the signals of individual GPS satellites and synchronize to them. The signal produced by a linear shift register can be imagined as a sequence of zeros and ones. The register itself may be imagined as a ribbon of r boxes containing the entries $a_{n-1}, \ldots, a_{n-r}$, each of which holds a value of 0 or 1 (see Figure 1.5). Each box is associated with a number $q_i \in \{0, 1\}$. The r values

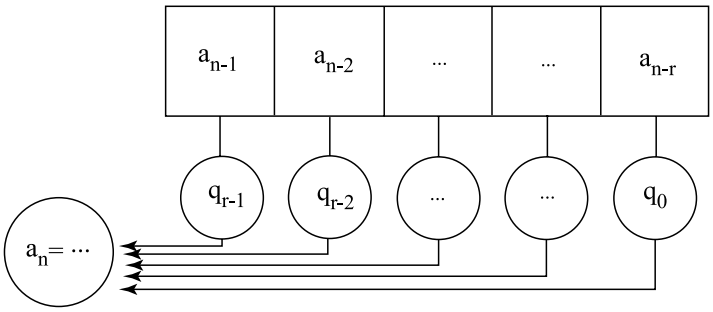


Fig. 1.5. A linear shift register.

q_i are fixed and distinct for all satellites. We generate a pseudorandom sequence in the following manner:

- We give ourselves a set of initial conditions $a_0, \ldots, a_{r-1} \in \{0, 1\}$, not all zero.
- Given $a_{n-r}, \ldots, a_{n-1}$, the register calculates the next element in the sequence as

$$a_n \equiv a_{n-r}q_0 + a_{n-r+1}q_1 + \cdots + a_{n-1}q_{r-1} = \sum_{i=0}^{r-1} a_{n-r+i}q_i \pmod 2. \tag{1.26}$$

(To calculate modulo 2 we perform the calculation as normal. The final result is 0 if the number is even, and 1 otherwise. As such we write $a \equiv 0 \pmod 2$ if a is even, and $a \equiv 1 \pmod 2$ if it is odd.)

- We shift each entry to the right, forgetting a_{n-r}. The calculated value a_n is inserted into the leftmost box.
- We iterate the above procedure.

Since the above procedure is perfectly deterministic and the number of initial conditions is finite, we will generate a sequence that must become periodic. Similarly we can see that the period of the sequence can be at most 2^r, since there are only 2^r distinct sequences of length r. In fact, we can convince ourselves that if at some moment $a_{n-r} = \cdots = a_{n-1} = 0$, then for all $m \geq n$ we will have $a_m = 0$. Thus an "interesting" periodic sequence must never contain a sequence of r zeros, and will therefore have a maximal period of $2^r - 1$. In order to generate a sequence with interesting properties we need only carefully choose the coefficients $q_0, \ldots, q_{r-1} \in \{0, 1\}$ and the initial conditions $a_0, \ldots, a_{r-1} \in \{0, 1\}$.

We never see the entire sequence, but rather we only observe a window of $M = 2^r - 1$ consecutive entries $\{a_n\}_{n=m}^{n=m+M-1}$, which we label $B = \{b_1, \ldots, b_M\}$. We wish to compare it with another window $C = \{c_1, \ldots, c_M\}$ of the form $\{a_n\}_{n=p}^{n=p+M-1}$. For example, sequence B is sent by the satellite, and sequence C is a cyclic shift of the same sequence generated by the GPS receiver. To determine the shift between the two, the receiver shifts repeatedly its sequence by one unit (by making $p \mapsto p + 1$) until it is identical with B.

Definition 1.2 *We call the correlation between two sequences B and C of length M the number of entries i where $b_i = c_i$ minus the number of entries i where $b_i \neq c_i$. We denote this by* $\mathrm{Cor}(B, C)$.

Remark: If the register consists of r entries, then the correlation between any pairs of sequences B and C must satisfy $-M \leq \mathrm{Cor}(B, C) \leq +M$, where $M = 2^r - 1$. We say that the sequences are poorly correlated if $\mathrm{Cor}(B, C)$ is close to zero.

Proposition 1.3 *The correlation between two sequences is given by*

$$\mathrm{Cor}(B, C) = \sum_{i=1}^{M} (-1)^{b_i}(-1)^{c_i}. \tag{1.27}$$

Proof. The number $\mathrm{Cor}(B, C)$ is calculated as follows: each time $b_i = c_i$ we must add 1. Similarly, each time $b_i \neq c_i$, we must subtract 1. Recall that b_i and c_i may take on only the value 0 or 1. Thus if $b_i = c_i$, then either $(-1)^{b_i} = (-1)^{c_i} = 1$ or $(-1)^{b_i} = (-1)^{c_i} = -1$. In either case we see that $(-1)^{b_i}(-1)^{c_i} = 1$. Similarly, if

$b_i \neq c_i$, exactly one of $(-1)^{b_i}$ and $(-1)^{c_i}$ is equal to 1 and the other to -1. Hence $(-1)^{b_i}(-1)^{c_i} = -1$. □

The following theorem shows that we may initialize a linear shift register in such a manner that it will generate a sequence that is poorly correlated to every translation of itself.

Theorem 1.4 *Given a linear shift register as shown in Figure 1.5, there exist coefficients* $q_0, \ldots, q_{r-1} \in \{0, 1\}$ *and initial conditions* $a_0, \ldots, a_{r-1} \in \{0, 1\}$ *such that the sequence generated by the register has a period of length* $2^r - 1$. *Consider two windows* B *and* C *of this sequence of length* $M = 2^r - 1$, *where* $B = \{a_n\}_{n=m}^{n=m+M-1}$ *and* $C = \{a_n\}_{n=p}^{n=p+M-1}$ *with* $p > m$. *If* M *does not divide* $p - m$, *then*

$$\mathrm{Cor}(B, C) = -1. \tag{1.28}$$

In other words, the number of bits in disagreement is always one more than the number of bits in agreement.

The proof of this theorem makes use of finite fields. We will begin by walking through an example that illustrates the theorem. The proof will follow in Section 1.4.2.

Example 1.5 *In this example we take* $r = 4$, $(q_0, q_1, q_2, q_3) = (1, 1, 0, 0)$, *and* $(a_0, a_1, a_2, a_3) = (0, 0, 0, 1)$. *We let the reader verify that these values generate a sequence with period* $2^4 - 1 = 15$, *repeating the following block of symbols:*

$$0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \ .$$

If we translate the sequence to the left by one symbol, we send the first 0 *to the end, yielding*

$$0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \ .$$

We see that the two blocks of symbols differ at positions 3, 4, 6, 8, 9, 10, 11, *and* 15. *Thus, they differ at eight positions and agree at seven positions, yielding a correlation of* -1.

In order to calculate the correlation with the other 14 translations of the sequence we explicitly write all possible translations. Inspection shows that any two of the following sequences agree at exactly seven places and differ in the remaining eight. We leave it to the reader to verify this:

0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
0	0	1	0	0	1	1	0	1	0	1	1	1	1	0
0	1	0	0	1	1	0	1	0	1	1	1	1	0	0
1	0	0	1	1	0	1	0	1	1	1	1	0	0	0
0	0	1	1	0	1	0	1	1	1	1	0	0	0	1
0	1	1	0	1	0	1	1	1	1	0	0	0	1	0
1	1	0	1	0	1	1	1	1	0	0	0	1	0	0
1	0	1	0	1	1	1	1	0	0	0	1	0	0	1
0	1	0	1	1	1	1	0	0	0	1	0	0	1	1
1	0	1	1	1	1	0	0	0	1	0	0	1	1	0
0	1	1	1	1	0	0	0	1	0	0	1	1	0	1
1	1	1	1	0	0	0	1	0	0	1	1	0	1	0
1	1	1	0	0	0	1	0	0	1	1	0	1	0	1
1	1	0	0	0	1	0	0	1	1	0	1	0	1	1
1	0	0	0	1	0	0	1	1	0	1	0	1	1	1

In the preceding example we did not explicitly show why we chose those specific values for $q_0, \ldots, q_3$ and $a_0, \ldots, a_3$. In order to show this and to prove Theorem 1.4 we will need to use the theory of finite fields. In particular, we will need to make use of the field $\mathbb{F}_{2^r}$ containing 2^r elements. For the case $r = 1$ the field $\mathbb{F}_2$ is the field of 2 elements $\{0, 1\}$ with addition and multiplication modulo 2.

1.4.1 The Structure of the Field $\mathbb{F}_2^r$

The structure and construction of finite fields of order p^n (for p prime) are explored in Sections 6.2 and 6.5 of Chapter 6. These sections are self-contained and may be read without reading the rest of Chapter 6. For the remainder of this chapter, we will assume that the reader has knowledge of the material covered in these sections.

The elements of $\mathbb{F}_2^r$ are the r-tuples $(b_0, \ldots, b_{r-1})$, where $b_i \in \{0, 1\}$. The addition of two such r-tuples is simply addition modulo 2, performed entry by entry,

$$(b_0, \ldots, b_{r-1}) + (c_0, \ldots, c_{r-1}) = (d_0, \ldots, d_{r-1}), \tag{1.29}$$

where $d_i \equiv b_i + c_i \pmod 2$. To define a multiplication operator we start by choosing an irreducible polynomial

$$P(x) = x^r + p_{r-1}x^{r-1} + \cdots + p_1 x + p_0 \tag{1.30}$$

over the field $\mathbb{F}_2$. We interpret each r-tuple $(b_0, \ldots, b_{r-1})$ as a polynomial of degree less than or equal to $r - 1$:

$$b_{r-1}x^{r-1} + \cdots + b_1 x + b_0. \tag{1.31}$$

In order to multiply the two r-tuples we multiply the two associated polynomials. The product is a polynomial of degree less than or equal to $2(r - 1)$, which is then reduced to a polynomial in x of degree $r - 1$ by taking its remainder when divided by

P (a process analogous to long division as applied to integers). This is equivalent to applying the rule $P(x) = 0$, i.e. $x^r = p_{r-1}x^{r-1} + \cdots + p_1 x + p_0$ (recall that $-p_i = p_i$ in $\mathbb{F}_2$) and iterating. We then interpret the coefficients of the resulting degree-$(r-1)$ polynomial as the entries of an r-tuple. The following is a classic theorem from the theory of finite fields. We will give only an overview of its proof without dwelling too much on the underlying algebra. If you are unfamiliar with the material covered in the following discussion you may safely skip it. The above discussion has explicitly shown that the vector elements $\mathbb{F}_2^r$ may be interpreted as polynomials.

Theorem 1.6 *1. The set $\mathbb{F}_{2^r}$ together with addition and multiplication as defined above is a field.*

2. There exists an element α such that the nonzero elements of $\mathbb{F}_{2^r}$ are precisely the elements α^i for $i = 0, \ldots, 2^r - 2$. In other words,

$$\mathbb{F}_{2^r} \setminus \{0\} = \{1, \alpha, \alpha^2, \ldots, \alpha^{2^r-2}\}. \tag{1.32}$$

An element α satisfying this property is called a primitive root, and satisfies $\alpha^{2^r-1} = 1$.

3. The elements $\{1, \alpha, \ldots, \alpha^{r-1}\}$ are linearly independent when interpreted as elements of the vector space $\mathbb{F}_2^r$ over $\mathbb{F}_2$ (which is isomorphic to the field $\mathbb{F}_{2^r}$).

4. If α is a primitive root of the field $\mathbb{F}_{2^r}$ constructed with an irreducible polynomial P over $\mathbb{F}_2$, then α is a root of a polynomial of degree r,

$$Q(x) = x^r + q_{r-1}x^{r-1} + \cdots + q_1 x + q_0,$$

irreducible over $\mathbb{F}_2$. The field constructed using the polynomial Q in the definition of multiplication is isomorphic to the field constructed using the polynomial P.

Definition 1.7 *A polynomial $Q(x)$ with coefficients in $\mathbb{F}_2$ is called primitive if it is irreducible and if the polynomial x is a primitive root of the field $\mathbb{F}_{2^r}$ constructed with respect to $Q(x)$.*

Outline of the Proof of Theorem 1.6

1. The proof is identical to the proof that $\mathbb{F}_p$ (also called $\mathbb{Z}_p$) is a field if p is prime (see Exercise 24 of Chapter 6). This proof makes use of Euclid's algorithm for polynomials, which finds the greatest common divisor of two given polynomials.
2. The nonzero elements of $\mathbb{F}_{2^r}$ form a multiplicative group G with $2^r - 1$ elements. Each nonzero element y generates a finite subgroup $H = \{y^i, i \in \mathbb{N}\}$. Lagrange's theorem (Theorem 7.18) states that the number of elements of H must divide the number of elements of G. Moreover, since H is finite, there must exist some minimum s such that $y^s = 1$. This s, called the order of the element y, is equal to the number of elements of H. Thus y is a root of the polynomial $x^s + 1 = 0$. Since $s \mid 2^r - 1$ then y is a root of $R(x) = x^{2^r-1} + 1$. (Exercise: why?) We have therefore shown that all elements of G are roots of the polynomial $R(x) = x^{2^r-1} + 1$. Suppose

now that there exists m, a strict divisor of 2^r-1, such that the order of all elements of G divides m. Then all elements of G must be roots of the polynomial x^m+1. This is a contradiction, since this polynomial has only $m < 2^r-1$ roots. Thus there exist elements y_i with orders m_i (for $i=1,\ldots,n$) such that the least common multiple of the m_i is 2^r-1. As such, the order of the product $y_1 \cdots y_n = \alpha$ is 2^r-1.

3. We will simply assume that the elements $\{1,\alpha,\ldots,\alpha^{r-1}\}$ are linearly independent when interpreted as vectors in the space $\mathbb{F}_2^r$.
4. The vectors $\{1,\alpha,\ldots,\alpha^r\}$ are linearly dependent because any $r+1$ vectors in a vector space of dimension r must be. Since the vectors $\{1,\alpha,\ldots,\alpha^{r-1}\}$ are linearly independent, there exist coefficients q_i such that $\alpha^r = q_0 + q_1\alpha + \cdots + q_{r-1}\alpha^{r-1}$. Thus α is a root of the polynomial $Q(x) = x^r + q_{r-1}x^{r-1} + \cdots + q_1 x + q_0$. This polynomial must be irreducible over $\mathbb{F}_2$, for otherwise, α would be the root of a polynomial with degree smaller than r, which would be in contradiction to the fact that $\{1,\alpha,\ldots,\alpha^{r-1}\}$ are linearly independent in $\mathbb{F}_2^r$. □

Remark: We could have chosen to write $\mathbb{F}_{2^r}$ with the polynomial $Q(x)$ rather than with the polynomial $P(x)$. The advantage of this last result is that it always allows us to ensure that $\alpha = x$ is a primitive root. One must be careful, however, since the progression α^i is not the same when computed modulo $Q(x)$ as it is when computed modulo $P(x)$!

Definition 1.8 *The trace function is the function* $T : \mathbb{F}_{2^r} \to \mathbb{F}_2$ *given by* $T(b_{r-1}x^{r-1} + \cdots + b_1 x + b_0) = b_{r-1}$.

Proposition 1.9 *The function* T *is linear and surjective. It has the value* 0 *on exactly half of the elements of* $\mathbb{F}_{2^r}$ *and* 1 *on the remaining half.*

Proof: Exercise!

1.4.2 Proof of Theorem 1.4

We choose a primitive polynomial $P(x)$ over $\mathbb{F}_2$,

$$P(x) = x^r + q_{r-1}x^{r-1} + \cdots + q_1 x + q_0,$$

permitting us to construct the field $\mathbb{F}_{2^r}$.

The q_i of the linear shift register are the coefficients of the polynomial $P(x)$. In order to construct good initial conditions we choose any nonzero polynomial $b = b_{r-1}x^{r-1} + \cdots + b_1 x + b_0$ from $\mathbb{F}_{2^r}$. We define the initial conditions as

$$\begin{array}{lcll} a_0 & = & T(b) & = \; b_{r-1}, \\ a_1 & = & T(xb), & \\ & \vdots & & \\ a_{r-1} & = & T(x^{r-1}b). & \end{array} \tag{1.33}$$

Consider how the value of a_1 is calculated:

$$\begin{aligned} a_1 = T(bx) &= T(b_{r-1}x^r + b_{r-2}x^{r-1} + \cdots + b_0 x) \\ &= T(b_{r-1}(q_{r-1}x^{r-1} + \cdots + q_1 x + q_0) + b_{r-2}x^{r-1} + \cdots + b_0 x) \\ &= T((b_{r-1}q_{r-1} + b_{r-2})x^{r-1} + \cdots) \\ &= b_{r-1}q_{r-1} + b_{r-2}. \end{aligned} \tag{1.34}$$

A similar calculation allows for the determination of the values $a_2, \ldots, a_{r-1}$. The formulas quickly become large, but the calculation can be performed very quickly in practice when the q_i and b_i are substituted by zeros and ones.

Example 1.10 *In Example 1.5 the polynomial used was $P(x) = x^4 + x + 1$. (Exercise: verify that the polynomial is irreducible and primitive.) The polynomial b that was chosen was simply $b = 1$. This creates the initial conditions $a_0 = T(1) = 0$, $a_1 = T(x) = 0$, $a_2 = T(x^2) = 0$, and $a_3 = T(x^3) = 1$.*

Proposition 1.11 *Let us choose the coefficients $q_0, \ldots, q_{r-1}$ of a shift register as those of a primitive polynomial*

$$P(x) = x^r + q_{r-1}x^{r-1} + \cdots + q_1 x + q_0.$$

Let $b = b_{r-1}x^{r-1} + \cdots + b_1 x + b_0$. We choose the initial elements $a_0, \ldots, a_{r-1}$ as in (1.33). *Then the sequence $\{a_n\}_{n \geq 0}$ generated by the shift register is given by $a_n = T(x^n b)$, and it repeats with a period that divides $2^r - 1$.*

Proof. We use the fact that $P(x) = 0$, which is to say $x^r = q_{r-1}x^{r-1} + \cdots + q_1 x + q_0$. Then

$$\begin{aligned} T(x^r b) &= T((q_{r-1}x^{r-1} + \cdots + q_1 x + q_0)b) \\ &= q_{r-1}T(x^{r-1}b) + \cdots + q_1 T(xb) + q_0 T(b) \\ &= q_{r-1}a_{r-1} + \cdots + q_1 a_1 + q_0 a_0 \\ &= a_r. \end{aligned} \tag{1.35}$$

We proceed by induction. Suppose now that the elements of the sequence satisfy $a_i = T(x^i b)$ for $i \leq n-1$. Then

$$\begin{aligned} T(x^n b) = T(x^r x^{n-r} b) &= T((q_{r-1}x^{r-1} + \cdots + q_1 x + q_0)x^{n-r}b) \\ &= q_{r-1}T(x^{n-1}b) + \cdots + q_1 T(x^{n-r+1}b) + q_0 T(x^{n-r}b) \\ &= q_{r-1}a_{n-1} + \cdots + q_1 a_{n-r+1} + q_0 a_{n-r} \\ &= a_n. \end{aligned} \tag{1.36}$$

Thus multiplication by x corresponds exactly to the calculation performed by the shift register, and therefore $a_n = T(x^n b)$ for all n. We see immediately that the minimal period has length at most $2^r - 1$, since $x^{2^r - 1} = 1$. □

We may ask ourselves, what is the minimal period of this sequence? To begin, it must be a divisor of $2^r - 1$ (see Exercise 11). In fact, we will show that the minimal period is exactly $2^r - 1$ when P is primitive. The proof will be indirect. If the period were given by $s \in \mathbb{N}$ such that $2^r - 1 = sm$ and $1 < s < 2^r - 1$, then the infinite sequence $\{a_n\}_{n\geq 0}$ and the sequence $\{a_{n+s}\}_{n\geq 0}$ would have to be identical. We will show that this cannot be true. Do not forget our original goal of creating sequences that are poorly correlated with translations of themselves. We will compute at the same time the correlation between any two windows B and C of length $M = 2^r - 1$, $B = \{a_n\}_{n=m}^{n=m+M-1}$ and $C = \{a_n\}_{n=p}^{n=p+M-1}$.

Proposition 1.12 *If $B = \{a_n\}_{n=m}^{n=m+M-1}$ and $C = \{a_n\}_{n=p}^{n=p+M-1}$, then $\mathrm{Cor}(B, C) = -1$ if M does not divide $p - m$.*

Proof. We can suppose $m \leq p$. Then

$$\begin{array}{rcl} \mathrm{Cor}(B,C) & = & \sum_{i=0}^{M-1}(-1)^{a_{m+i}}(-1)^{a_{p+i}} \\ & = & \sum_{i=0}^{M-1}(-1)^{T(x^{m+i}b)}(-1)^{T(x^{p+i}b)} \\ & = & \sum_{i=0}^{M-1}(-1)^{T(x^{m+i}b)+T(x^{p+i}b)} \\ & = & \sum_{i=0}^{M-1}(-1)^{T(x^{m+i}b+x^{p+i}b)} \\ & = & \sum_{i=0}^{M-1}(-1)^{T(bx^{i+m}(1+x^{p-m}))} \\ & = & \sum_{i=0}^{M-1}(-1)^{T(x^{i+m}\beta)}, \end{array} \tag{1.37}$$

where $\beta = b(1+x^{p-m})$. By our choice of P we know that x is a primitive root of our field and therefore that $x^M = 1$ and $x^N \neq 1$ if $1 \leq N < M$. We deduce that $x^N = 1$ if and only if M divides N. If M divides $p - m$ then $x^{p-m} = 1$ and $\beta = b(1+1) = b \cdot 0 = 0$, in which case $\mathrm{Cor}(B, C) = M$. If M does not divide $p-m$ then the polynomial $(1+x^{p-m})$ is not the zero polynomial; hence $\beta = b(1 + x^{p-m})$ is nonzero as well, since it is the product of two nonzero elements. Thus β is of the form x^k, where $k \in \{0, \ldots, 2^r - 2\}$, which implies that the set $\{\beta x^{i+m}, 0 \leq i \leq M-1\}$ is a permutation of the elements of $\mathbb{F}_{2^r} \setminus \{0\} = \{1, x, \ldots, x^{2^r-2}\}$. The trace function T take a 1 value on half of the elements of $\mathbb{F}_{2^r}$ and a 0 value on the remaining elements. Since it takes a 0 value on the zero element, it takes a 1 value on 2^{r-1} of the elements of $\mathbb{F}_{2^r} \setminus \{0\}$ and a 0 value on the remaining $2^{r-1} - 1$. Hence $\mathrm{Cor}(B, C) = -1$. □

Corollary 1.13 *The period of the pseudorandom sequence generated by the linear shift register is exactly $M = 2^r - 1$.*

Proof. If the period were equal to $K < M$, then the sequence would coincide with its translation by K elements, and the two sequences would have a correlation equal to M. This is in contradiction to Proposition 1.12. □

If we now want to generate other pseudorandom sequences of the same length, we may use the same principle but change the polynomial $P(x)$. (We want a distinct

sequence for each satellite.) Galois theory lets us (in certain cases) calculate the correlation of this new sequence with the first one and its translations. Engineers, however, content themselves with looking up these correlation values in precalculated tables.

1.5 Cartography

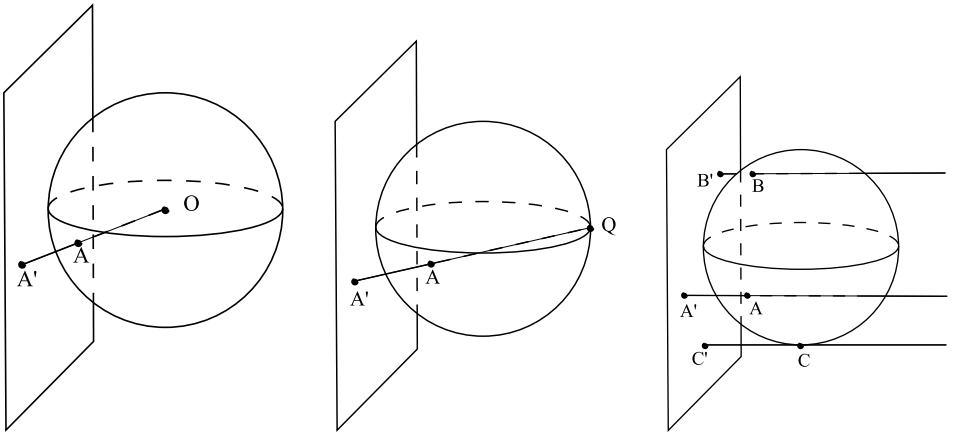
As mentioned in the introduction, the field of cartography encounters certain nontrivial problems in trying to faithfully represent the surface of the Earth. Maps are generally used to orient or guide us. Depending on the application it may be more important to us that the map preserve distances, for example if we desire the shortest path between two points on the map to correspond to the shortest path between two points in reality. This condition is generally not important on terrestrial maps because when traveling by car we are constrained to travel on highways, and when traveling on foot, the distances involved are sufficiently small that any deviation from the true shortest path is negligible. In contrast, in choosing the route to be flown by an airplane or taken by a boat, the problem becomes noticeable. Moreover, for someone navigating a sailboat or small airplane with relatively rudimentary equipment, it is not sufficient just to plot a course on a map. The course must also be able to be followed and held by the pilot. Prior to the invention of GPS it was very common to use a magnetic compass as a primary means of navigation. Using a magnetic compass we can assure ourselves that we are following a trajectory that maintains a constant angle with respect to the Earth's magnetic field. Such a trajectory is not necessarily the shortest path between two points, but since it is an easy path to follow, it would be convenient to have maps on which such paths are represented by straight lines. Marine and some aeronautical charts have this property. However, on these charts relative areas are not preserved: two regions of the globe that have the same surface area are not in general represented by domains with the same surface area on the map.

We begin by stating the rules of the game. A theorem in differential geometry states that it is impossible to map a portion of the surface of the sphere into the plane while preserving both distances and angles. (For those who are familiar with the terminology, such a transformation is called an "isometry" and preserves the "Gaussian curvature" of the surface. The Gaussian curvature of a sphere of radius R is $1/R^2$, while the Gaussian curvature of both a plane and a cylinder is zero.) Thus, we must make a compromise. The specific compromise to be made depends on the application.

Cartography is principally concerned with projections, and there are many different types.

Projection onto a plane tangent to the sphere. This is the most elementary type of projection. There exist several variations on this type of projection: where the projection goes through the center of the sphere (*gnomonic projection*); where the projection goes through the point antipodal to the tangent point (*stereographic projection*); and, where the projection is taken along lines that are orthogonal to the plane of projection

(*orthographic projection*). (See Figure 1.6.) This family of projections gives reasonable results if we are interested in mapping only a small portion of the sphere centered at the point of tangency. However, the distortions become very pronounced as we move away from the point of tangency. From a mathematical point of view these projections offer little interest (except the stereographic projection discussed in Exercise 24), and we will not discuss them further.



(a) Gnomonic projection (b) Stereographic projection (c) Orthographic projection

Fig. 1.6. Three types of projections onto a tangential plane.

For the remainder of this section we will limit our discussion to projections onto a cylinder. After the projection the cylinder may be unrolled, yielding a plane. Already we can see that progress has been made. Instead of there being just one point of tangency (where the map is most accurate), there is an entire circle of tangency around the sphere. However, there will still be severe distortions as we move toward the poles of the sphere. As before, there are several variations on this projection, and depending on the method chosen, the resulting map will have different properties. There is generally a strong desire to map lines of latitude (parallels) to horizontal lines and lines of longitude (meridians) to vertical lines. Such a projection means that there is an easy mapping between Cartesian coordinates on the map and longitude and latitude on the globe (but there will be distortion of distances along distinct parallels).

Projection onto the cylinder via the center of the sphere. Under this projection the sphere maps to an infinite cylinder, with the poles being mapped to the infinite extremes of the cylinder. This projection has little use or interest beyond the fact that its formula is simple.

Horizontal projection onto the cylinder. This projection is known to geographers as Lambert projection, but in fact it was studied in detail by Archimedes. Let S be a sphere of radius R whose surface points satisfy the equation $x^2 + y^2 + z^2 = R^2$. We want to project the sphere onto the cylinder C satisfying the equation $x^2 + y^2 = R^2$. The projection $P : S \to C$ is given by the formula

$$P(x, y, z) = \left(\frac{Rx}{\sqrt{x^2 + y^2}}, \frac{Ry}{\sqrt{x^2 + y^2}}, z \right) \tag{1.38}$$

(see Figure 1.7). The point $P(x_0, y_0, z_0)$ is therefore the point of intersection between the cylinder and the horizontal half-line starting at $(0, 0, z_0)$ (on the vertical axis) and passing through the point (x_0, y_0, z_0).

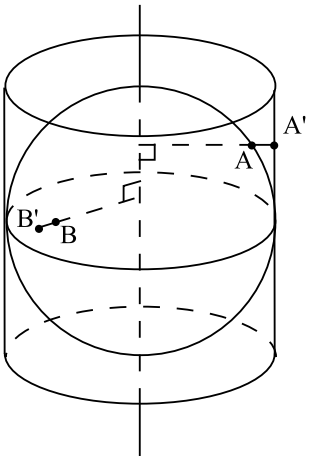


Fig. 1.7. Horizontal projection onto a cylinder.

Although it has less distortion than the cylindrical projection via the center of the sphere, this projection distorts distances as we move away from the equator. However, this projection has a rather remarkable property: it preserves area. This property was discovered for the first time by Archimedes. This projection was therefore chosen in producing the Peters atlas (see Figure 1.8). In other atlases using different projections, the Nordic countries have a greatly exaggerated size. In the Peters atlas [2], the relative sizes of these countries are precisely preserved, although they appear less tall and wider. We will now prove this remarkable property of the Lambert projection.

Theorem 1.14 *The projection $P : S \to C$ given by equation* (1.38) *preserves area. (In geographic and cartographic terms, we say that this projection is* equivalent.*)*

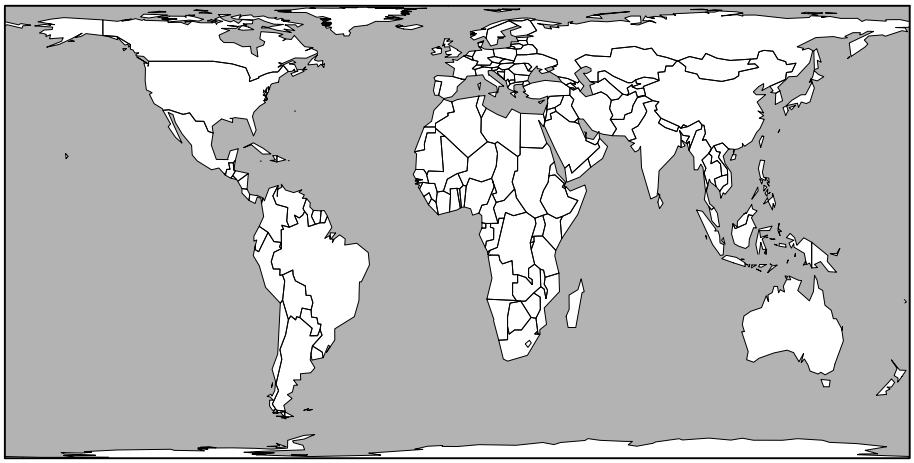


Fig. 1.8. The world map using the Lambert cylindrical projection.

PROOF. To make the proof simpler we will first change our coordinate system. We parameterize the sphere using two angular coordinates, θ and ϕ, which can be mapped back to Cartesian coordinates using the following mapping:

$$\begin{aligned} &F : (-\pi, \pi] \times [\tfrac{-\pi}{2}, \tfrac{\pi}{2}] \to S, \\ &(\theta, \phi) \mapsto F(\theta, \phi) = (x, y, z) = (R\cos\theta\cos\phi, R\sin\theta\cos\phi, R\sin\phi). \end{aligned} \tag{1.39}$$

These are the spherical coordinates. We can interpret θ as being the longitude, expressed in radians rather than degrees, with $\theta = 0$ corresponding to the Greenwich meridian, $\theta > 0$ corresponding to eastern longitudes, and $\theta < 0$ corresponding to western longitudes. In the same way, ϕ is the latitude, positive values of ϕ corresponding to northern latitudes. Similarly, using the same parameters we may parameterize the cylinder as

$$\begin{aligned} &G : (-\pi, \pi] \times [-\tfrac{\pi}{2}, \tfrac{\pi}{2}] \to C, \\ &(\theta, \phi) \mapsto G(\theta, \phi) = (x, y, z) = (R\cos\theta, R\sin\theta, R\sin\phi). \end{aligned} \tag{1.40}$$

Under these coordinate systems the projection P may be rewritten as $(\theta, \phi) \mapsto (\theta, \phi)$. Let A be a region of the sphere and let $P(A)$ be the corresponding projected region on the cylinder. Both of these regions are the images of the same set B with

$$B \subset (-\pi, \pi] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right].$$

The area on the surface of the sphere of A is given by (we will justify this formula a little later)

$$\text{Area}(A) = \iint_B \left| \frac{\partial F}{\partial \theta} \wedge \frac{\partial F}{\partial \phi} \right| \, d\theta \, d\phi \tag{1.41}$$

where $v \wedge w$ represents the cross product of v and w and $|v \wedge w|$ represents its length (see [1] or any multivariable calculus textbook). This yields

$$\begin{aligned}
\tfrac{\partial F}{\partial \theta} &= (-R\sin\theta\cos\phi, R\cos\theta\cos\phi, 0),\\
\tfrac{\partial F}{\partial \phi} &= (-R\cos\theta\sin\phi, -R\sin\theta\sin\phi, R\cos\phi),\\
\tfrac{\partial F}{\partial \theta} \wedge \tfrac{\partial F}{\partial \phi} &= (R^2\cos\theta\cos^2\phi, R^2\sin\theta\cos^2\phi, R^2\sin\phi\cos\phi),\\
\left|\tfrac{\partial F}{\partial \theta} \wedge \tfrac{\partial F}{\partial \phi}\right| &= R^2|\cos\phi|.
\end{aligned}$$

Similarly, for the cylinder the area of $P(A)$ is given by

$$\text{Area}(P(A)) = \iint_B \left|\frac{\partial G}{\partial \theta} \wedge \frac{\partial G}{\partial \phi}\right| \, d\theta \, d\phi. \tag{1.42}$$

Here we see that

$$\begin{aligned}
\tfrac{\partial G}{\partial \theta} &= (-R\sin\theta, R\cos\theta, 0),\\
\tfrac{\partial G}{\partial \phi} &= (0, 0, R\cos\phi),\\
\tfrac{\partial G}{\partial \theta} \wedge \tfrac{\partial G}{\partial \phi} &= (R^2\cos\theta\cos\phi, R^2\sin\theta\cos\phi, 0),\\
\left|\tfrac{\partial G}{\partial \theta} \wedge \tfrac{\partial G}{\partial \phi}\right| &= R^2|\cos\phi|.
\end{aligned}$$

It is easy to see that the integrals for the areas of A and $P(A)$ need to be calculated over the same domain B. Since the two integrands are identical, the above shows that these two areas are in fact equal. □

Justification of Equations (1.41) **and** (1.42)**.** This is a quick reminder (most likely from your multivariable calculus course) about how to calculate the area of a surface. We consider cutting B into infinitesimally small rectangular pieces with side lengths $d\theta$ and $d\phi$. The area of A (respectively $P(A)$) is given by the sum of the areas of the images of the pieces under the mapping F (respectively G). We will consider the area of A. We can think of $d\theta$ and $d\phi$ as being little segments that are tangential to the curves $\phi =$ constant and $\theta =$ constant. Thus their images are little segments that are tangential to the images of these two curves: the vectors $\frac{\partial F}{\partial \theta}\, d\theta$ and $\frac{\partial F}{\partial \phi}\, d\phi$. These vectors will in general inscribe a parallelogram whose area is precisely $\left|\frac{\partial F}{\partial \theta} \wedge \frac{\partial F}{\partial \phi}\right| \, d\theta \, d\phi$ (the product of the lengths of the vectors, multiplied by the sines of the angle between them).

In our proof, the image under F of this piece of B resembles a little rectangle with sides of lengths $R\, d\theta|\cos\phi|$ and $R\, d\phi$. Similarly, its image under G is a little rectangle with sides of length $R\, d\theta$ and $R|\cos\phi|\, d\phi$. In both of these cases the images have an area of $R^2|\cos\phi|\, d\theta\, d\phi$.

Mercator projection. The Lambert projection preserves areas but it does not preserve angles. In making marine charts, projections that preserve angles are preferred, since they allow for the easy plotting of courses that can be followed using a magnetic compass.

The Mercator projection $M : S \to C$ does exactly this. This projection covers the entire infinitely long cylinder. Here again we will use spherical coordinates (1.39) for representing a point Q on the sphere, given by $F(\theta, \phi)$. Its image under M is given by

$$M(Q) = M(F(\theta, \phi)) = \left(R\cos\theta, R\sin\theta, R\log\left(\tan\tfrac{1}{2}(\phi + \tfrac{\pi}{2})\right)\right). \tag{1.43}$$

As before, the final projection will be given by the unrolled cylinder. Let θ represent the horizontal coordinate (abscissa) on the unrolled cylinder and let z represent the vertical coordinate (ordinate). This gives us a mapping $N : S \to \mathbb{R}^2$ of the sphere onto the plane. If (θ, ϕ) are the spherical coordinates of a point Q, we will map this point to

$$N(F(\theta, \phi)) = \left(\theta, \log\left(\tan\tfrac{1}{2}(\phi + \tfrac{\pi}{2})\right)\right) \tag{1.44}$$

(see Figures 1.9 and 1.10).

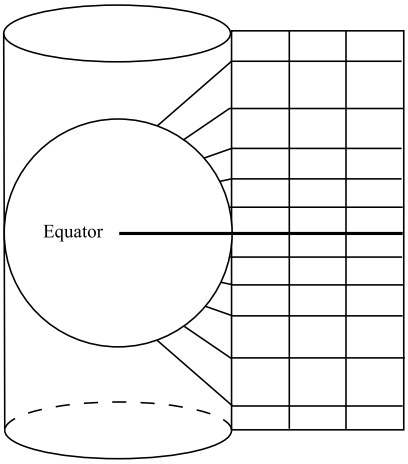


Fig. 1.9. Mercator projection: project onto a cylinder and unroll it. A given distance along a meridian appears longer the further away it is from the equator.

Definition 1.15 *A transformation $N : S_1 \to S_2$ from a surface S_1 to a surface S_2 is conformal if it preserves angles. That is, if two curves on S_1 intersect each other at point Q with an angle α, the images of these two curves on S_2 will intersect each other at point $N(Q)$ with the same angle α.*

Theorem 1.16 *The transformations M and N defined in equations* (1.43) *and* (1.44) *are conformal.*

Fig. 1.10. A map of the world using the Mercator projection. Since the entire map would have infinite height, only the portion between 85°S and 85°N is shown here.

PROOF. We will content ourselves with giving the proof for the mapping N. Then it will follow that M is conformal if we convince ourselves that rolling or unrolling a cylinder cannot change the angles of intersection between curves inscribed on it. Since two curves tangent to each other are mapped to two curves tangent to each other, it suffices to consider tiny line segments tangent to the original curves at the point of intersection. Consider a point (θ_0, ϕ_0) and two little line segments passing through this point, which may be written as

$$\begin{aligned} v(t) &= (\theta_0 + t\cos\alpha, \phi_0 + t\sin\alpha), \\ w(t) &= (\theta_0 + t\cos\beta, \phi_0 + t\sin\beta). \end{aligned}$$

We will consider the tangent vectors $F \circ v = v_1$ and $F \circ w = w_1$ in $Q = F(\theta_0, \phi_0)$, and show that they inscribe the same angle as the vectors $N \circ F \circ v = v_2$ and $N \circ F \circ w = w_2$ in $N(Q)$. The tangent vectors may be calculated using the chain rule and are given by

$$\begin{aligned} v_1'(0) &= R(-\sin\theta_0\cos\phi_0\cos\alpha - \cos\theta_0\sin\phi_0\sin\alpha, \\ &\qquad \cos\theta_0\cos\phi_0\cos\alpha - \sin\theta_0\sin\phi_0\sin\alpha, \cos\phi_0\sin\alpha), \\ w_1'(0) &= R(-\sin\theta_0\cos\phi_0\cos\beta - \cos\theta_0\sin\phi_0\sin\beta, \\ &\qquad \cos\theta_0\cos\phi_0\cos\beta - \sin\theta_0\sin\phi_0\sin\beta, \cos\phi_0\sin\beta), \\ v_2'(0) &= (\cos\alpha, \tfrac{\sin\alpha}{\cos\phi_0}), \\ w_2'(0) &= (\cos\beta, \tfrac{\sin\beta}{\cos\phi_0}). \end{aligned}$$

To show that the transformation is conformal we use the following criteria:

Lemma 1.17 *The transformation is conformal if for all θ_0, ϕ_0, there exists a positive constant $\lambda(\theta_0, \phi_0)$ such that for all α and β, the following relation holds for the scalar product of $v_i'(0)$ and $w_i'(0)$:*

$$\langle v_1'(0), w_1'(0)\rangle = \lambda(\theta_0, \phi_0)\langle v_2'(0), w_2'(0)\rangle. \tag{1.45}$$

PROOF. Let ψ_i be the angle between $v_i'(0)$ and $w_i'(0)$ for $i = 1, 2$. We want to show that $\cos\psi_1 = \cos\psi_2$. If (1.45) is satisfied, we see that

$$\begin{aligned}
\cos\psi_1 &= \frac{\langle v_1'(0), w_1'(0)\rangle}{|v_1'(0)|\,|w_1'(0)|} \\
&= \frac{\langle v_1'(0), w_1'(0)\rangle}{\langle v_1'(0), v_1'(0)\rangle^{1/2}\langle w_1'(0), w_1'(0)\rangle^{1/2}} \\
&= \frac{\lambda(\theta_0,\phi_0)\langle v_2'(0), w_2'(0)\rangle}{(\lambda(\theta_0,\phi_0)\langle v_2'(0), v_2'(0)\rangle)^{1/2}(\lambda(\theta_0,\phi_0)\langle w_2'(0), w_2'(0)\rangle)^{1/2}} \\
&= \frac{\langle v_2'(0), w_2'(0)\rangle}{\langle v_2'(0), v_2'(0)\rangle^{1/2}\langle w_2'(0), w_2'(0)\rangle^{1/2}} \\
&= \frac{\langle v_2'(0), w_2'(0)\rangle}{|v_2'(0)|\,|w_2'(0)|} \\
&= \cos\psi_2.
\end{aligned}$$

(The requirement that $\lambda(\theta_0, \phi_0)$ be positive ensures that there is no division by zero and that square roots are real.) □

Verifying (1.45) for the Mercator projection requires a bit of work but simplifies nicely. We obtain that

$$\begin{aligned}
\langle v_1'(0), w_1'(0)\rangle &= R^2(\cos^2\phi_0\cos\alpha\cos\beta + \sin\alpha\sin\beta), \\
\langle v_2'(0), w_2'(0)\rangle &= \cos\alpha\cos\beta + \tfrac{\sin\alpha\sin\beta}{\cos^2\phi_0}.
\end{aligned}$$

From this it follows that $\lambda(\theta_0, \phi_0) = R^2\cos^2\phi_0$. □

The shortest path between two points on a sphere. We consider two points Q_1 and Q_2 on the surface of a sphere. If they are not antipodal, the points cannot be in line with the center of the sphere; thus they form a plane with it. The intersection between the plane and the sphere traces out a great circle, with the points Q_1 and Q_2 both lying on it. The points cut the circle into two arcs, and the shorter of the two is the shortest path on the surface of the sphere between Q_1 and Q_2. Let O be the center of the sphere. Then the length of this path is $R\alpha$, where $\alpha \in [0, \pi)$ is the angle between OQ_1 and OQ_2, and R is the radius of the sphere. In maritime navigation the shortest path between two points is called an *orthodrome*. In mathematics the shortest path between two points on some surface is usually called a *geodesic*. The geodesics of a sphere are all great circles. If we consider a chart constructed using the Mercator projection, the orthodrome between two points Q_1 and Q_2 does not correspond to a straight line on the chart, unless the points lie along the same longitude. In the vocabulary of marine navigation the *loxodrome* (also called a *rhumb line*) between two

points is the route joining them that intersects all lines of meridians at the same angle. Under the Mercator projection, this route corresponds to a straight line joining the two points, and this in fact proves that such a route always exists. A loxodrome is usually longer and never shorter than an orthodrome. However, on a Mercator projection of the sphere this relationship is inverted (see Figure 1.11).

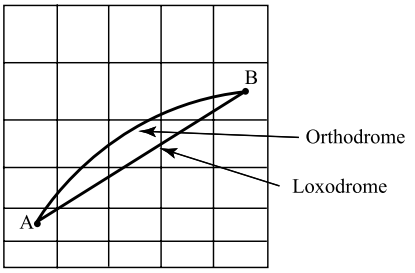


Fig. 1.11. Orthodromic and loxodromic routes between two points A and B.

Following a trajectory. If we want to proceed from point A to point B using only traditional navigation techniques (in other words, without using GPS), it is easier to follow the loxodromic route (which appears as a straight line in a Mercator projection). This trajectory intersects each line of meridian at a constant angle. The traditional tool of navigation is a simple magnetic compass, which indicates the direction to magnetic north. The magnetic field lines surrounding the Earth resemble lines of meridian, originating at the north magnetic pole and terminating at the south magnetic pole. However, the magnetic north and south poles do not perfectly coincide with the Earth's true poles. Moreover, the magnetic poles are not static, but rather wander slowly. Thus, in practice, the magnetic field lines intersect the lines of meridian at an angle, and this angle is not the same at every position on Earth, nor is it the same at one location from one year to the next. The exact value of the variation between true north and magnetic north can be quickly looked up in tables, and is usually included directly on marine and aeronautical charts. Alternatively, it can be calculated assuming that we know our location and those of one or more nearby landmarks. If we are navigating sufficiently far away from the poles we can assume that the variation is nearly constant. Thus, in order to follow a loxodromic route it suffices to keep a compass pointed at the desired angle, to be calculated in view of the angle between the magnetic field lines and the meridians at the current position.

Cartography in the vicinity of the poles. If we want to make charts of the areas around the poles, the projections discussed previously are not very convenient. Thus we instead consider projections onto oblique cylinders or cones. If we want a conformal projection, we can use the Mercator projection onto an oblique cylinder. However, in doing so we lose the property that lines of longitude and latitude map to straight lines.

We may also consider conformal projections onto the surface of a cone. Such projections are called *Lambert projections* (see Exercise 26 for an example).

The UTM coordinate system. When we want to enter a waypoint into a GPS receiver, we must calculate its coordinate on a chart. Many charts make use of the UTM (Universal Transverse Mercator) coordinate system, which comes from 60 projections of the same type as the Mercator projection: the difference is that the cylinder is no longer vertical, but horizontal, hence tangent to the Earth along a meridian. The corresponding projection is called a transverse Mercator projection. A longitude zone covers an interval of longitude of width 6 degrees. Each of the 60 longitude zones in the UTM system is based on a transverse Mercator projection. This allows us to map a region of large north–south extent with a low amount of distortion. This system was originally designed by the North Atlantic Treaty Organization (NATO) in 1947.

1.6 Exercises

GPS ("Global positioning system")

1. Show that the denominator of equation (1.16) is zero if and only if the four satellites lie in the same plane.

2. The *Loran* (for "LOng RANge") navigational system was widely used for marine navigation for many years, particularly just off the North American coasts. Since many boats are still equipped with Loran receivers, the system has not been decommissioned, even though GPS is becoming increasingly popular. Loran transmitters are organized into chains of three to five transmitters, one being designated as the master or principal station M and the others as the slave or secondary stations W, X, Y, and Z.

- The principal station transmits a signal.
- The secondary station W receives the signal, delays a predetermined amount of time, and retransmits the same signal.
- The secondary station X receives the signal, delays a predetermined amount of time, and retransmits the same signal.
- etc.

The delays used by each secondary station are chosen such that there will be no doubt as to the origin of a signal received anywhere within the designated service area of the chain of transmitters. The idea behind the system is that the Loran receiver (on the boat) measures the phase shift between the received signals. Since there are between three and five signals received, there are at least two phase shifts that will be independent.
(a) Explain how we can determine our position knowing two phase shifts.
(b) In practice, the phase shift between the first antenna and the second antenna allows the receiver to locate itself on a branch of a hyperbola. Why?

Comment: These hyperbolic positioning curves are drawn on marine charts. A position on a marine chart can therefore be identified as the point of intersection between two hyperbolic curves drawn on the chart.

3. In order to calculate its position a GPS receiver needs to know the signal transit time for four satellites. If we constrain the problem by saying that the receiver is at an altitude of zero (in other words, at sea level), show that only three satellites are required in order to calculate the receiver's position. Explain the details of the calculations to be performed.

4. Meteorites regularly enter the atmosphere, rapidly heat up, disintegrate, and finally explode before hitting the surface of the Earth. This explosion generates a shock wave that travels in all directions at the speed of sound v. The shock wave is detected by seismographs installed at various locations on the surface of the Earth.

If four stations (equipped with perfectly synchronized clocks) note the moment that the shock wave arrives, explain how to calculate both the position and time of the explosion.

5. Consider a map that does not explicitly show any lines of latitude or longitude, nor the direction of north. Explain how knowing the locations of any three nonaligned landmarks on the chart allows for the position of any point on the chart to be calculated. What hypothesis must be made in order for this to work?

Lightning strikes and storms

6. What is the minimum number of detectors that must observe a lightning strike in order for it to be located? Give the system of equations that the central computer must resolve in order to calculate this position.

7. Given the two times t_1 and t_2 measured by the oscilloperturbographs on either end of a power line of length L, calculate the location of the fault on the power line.

8. A nanosecond is one billionth of a second: 10^{-9} s. Calculate the distance traveled by light in 100 nanoseconds and from this deduce the accuracy of the position calculated by a system that measures light transit times within 100 nanoseconds.

9. Given that $P(I)$ is the Popolansky function, calculate the density function $f(I)$ of the variable X representing the amperage of lightning strikes. What is the mode of this distribution (the value of I where the density takes its maximum)?

10. In other regions the Anderson–Erikson function P given in (1.17) is typically used, where $M = 31$ and $K = 2.6$. In contrast to the Popolansky function, you will have to use numerical methods.

(a) Calculate the median of this distribution.

(b) Calculate the 90th percentile of this distribution. In other words, find the value I such that $\text{Prob}(X \leq I) = 0.9$.
(c) If 58% of detected lightning strikes have an amperage higher than the median, calculate the percentage of lightning strikes that are not detected. By making the further assumption that only the weakest lightning strikes avoid detection, calculate the threshold amperage I_0 below which lightning strikes will not be detected.
(d) Calculate the mode of this distribution.

Linear shift registers

11. Consider a sequence $\{a_n\}$ that is periodic with length N, that is, $a_{n+N} = a_n$ for all n. Show that the minimal period of this sequence, the least integer M such that $a_{n+M} = a_n$ for all n, must be a divisor of N.

12. **(a)** Show that the polynomial $x^4 + x^3 + 1$ is primitive over $\mathbb{F}_2$.
(b) Calculate the sequence generated by the linear shift register where $(q_0, q_1, q_2, q_3) = (1, 0, 0, 1)$ and the initial conditions are $(a_0, a_1, a_2, a_3) = (T(b), T(xb), T(x^2b), T(x^3b))$ with $b = 1$. Verify that this sequence has a minimal period of length 15.
(c) Verify that this sequence is not the same as that given in Example 1.5.
(d) Calculate the correlation between this sequence and the different translations of the sequence of Example 1.5.

13. Show that the polynomial $x^4 + x^3 + x^2 + x + 1$ is not primitive over $\mathbb{F}_2$. Calculate the sequence generated by the linear shift register where $(q_0, q_1, q_2, q_3) = (1, 1, 1, 1)$ and the initial conditions are $(a_0, a_1, a_2, a_3) = (T(b), T(xb), T(x^2b), T(x^3b))$ with $b = 1$. Verify that this sequence has a minimal period of length less than 15.

A few elementary ways of calculating position
Before the invention of GPS humankind used several other (mathematical!) methods and ingenious tools for calculating position: the position of the North Star, the position of the sun at noon, the sextant, etc. Some of these techniques are still in use today. In fact, even though GPS is much more precise and simple to use, we cannot guarantee that the system will never break down, or that we will always have a fresh set of batteries on hand. Hence the continued importance and use of these simpler techniques.

14. The North Star is situated very nearly on the axis of rotation of the Earth and is visible only from the Northern Hemisphere.
(a) If we are situated on the 45th parallel, with what angle over the horizon will we see the North Star? What about from the 60th parallel?
(b) Suppose that you see the North Star with an angle θ above the horizon. At what latitude are you?

15. The axis of rotation of the Earth is at an angle of 23.5 degrees with the normal to the ecliptic plane (the plane of the Earth's orbit around the sun).

(a) The Arctic Circle is situated at 66.5 degrees north latitude. If you are at the Arctic Circle, at what angle above the horizon will you see the sun at noon during the equinox? During the summer solstice? During the winter solstice? (It is this last property that led to the naming of this particular parallel.)
(b) Answer the same question assuming that you are at the equator.
(c) Answer the same question if you are at a latitude of 45 degrees north.
(d) The Tropic of Cancer is situated at a latitude of 23.5 degrees north. Show that the sun is vertically above the Tropic of Cancer at noon during the summer solstice.
(e) For which points on the surface of the Earth is the sun vertically above at noon on at least one day of the year?

16. We can also use the height of the sun at noon to calculate latitude. If the sun is at an angle θ above the horizon at noon during the summer solstice, calculate your latitude. Answer the same question during the equinoxes and the winter solstice.

17. In order to determine your approximate longitude you can use the following technique. Set your watch to the local time at the Greenwich meridian. Note the indicated time when the sun is at its zenith. Explain how you can use this information to calculate your longitude. This method is not terribly accurate, since it is rather difficult to tell when the sun is at its zenith. Instead, marine navigators typically interpolate the results of two measures, one taken before zenith and another after.

18. The workings of a sextant: as shown in Exercises 14 and 17 we can determine longitude and latitude by measuring the angle above the horizon of the sun or North Star. This is nice in theory, but in practice how do we get an accurate measurement while standing on a rocking boat? This is where the sextant is useful. Sextants use a system of two mirrors. The navigator adjusts the angle between the two mirrors until he sees the reflected image of the sun or North star at the same level of the horizon, as shown in Figure 1.12.
(a) Show that if the angle between the two mirrors is θ, then the angle above the horizon made by the sun or the North Star is 2θ.
(b) Explain why the measurement is not too strongly affected by the rocking of the boat.

Cartography

19. Consider two points $Q_1 = (x_1, y_1, z_1)$ and $Q_2 = (x_2, y_2, z_2)$ on the surface of an idealized spherical Earth of radius R. Let the longitudes of these two points be θ_1 and θ_2 and the latitudes be ϕ_1 and ϕ_2, respectively. Calculate the minimal distance along the surface of the Earth between these two points.

20. Consider a chart constructed using the standard Mercator projection. Calculate the equation of the orthodrome between the point at longitude 0° and latitude 0° and the point at longitude 90°W and latitude 60°N.

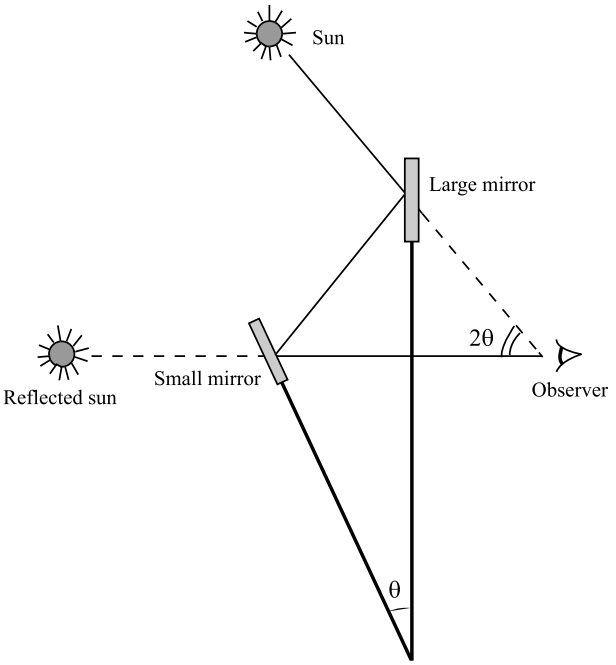


Fig. 1.12. The workings of a sextant (Exercise 18).

21. Consider a chart constructed using the horizontal cylindrical projection. Calculate the equation of the orthodrome between the point at longitude 0° and latitude 0° and the point at longitude 90°W and latitude 60°N.

22. Consider projecting the sphere onto a vertical cylinder via the center of the sphere.
(a) Give the formula describing the projection.
(b) What is the image of the meridians? What about the parallels?
(c) What is the image of a great circle?

23. Conic projections use cones that are tangent or secant to the sphere and project through the center of the sphere. Imagine a conic projection and draw the grid of meridians and parallels on the unwrapped cone.

24. Stereographic projection: Consider projecting the sphere onto a plane tangent to the sphere at a point P. Let P' be the point on the sphere diametrically opposed to P. The projection is performed as follows: if Q is a point on the sphere, then its projection is the intersection of the line $P'Q$ with the plane tangent to the sphere at P.

(a) Give the formula for this projection in the case that P is the South Pole and we consider the sphere to have radius 1. (In this case the point P' is the North Pole and the tangent plane is described by the equation $z = -1$.)
(b) Show that this projection is conformal.

25. In order to accurately represent the Earth we need to model it as an ellipsoid of revolution $\frac{x^2}{a^2} + \frac{y^2}{a^2} + \frac{z^2}{b^2} = 1$. In general, the spherical coordinates of an ellipsoid may be written as

$$(x, y, z) = (a\cos\theta\cos\phi, a\sin\theta\cos\phi, b\sin\phi).$$

The notion of longitude is the same as that of a sphere, but most geographers tend to use geodesic latitude, defined as follows: the geodesic latitude of a point P on an ellipsoid is the angle between the normal vector at the point P and the equatorial plane (the plane $z = 0$). Calculate the geodesic latitude as a function of ϕ.

26. Lambert conic conformal projection: Consider the sphere $x^2 + y^2 + z^2 = 1$ and a cone centered above the North Pole at a point z.
(a) What are the coordinates of the peak of the cone if the cone is tangent to the sphere along the parallel ϕ_0?
(b) If we cut the cone along the meridian $\theta = \pi$ and unroll it, we obtain a sector of a circle. Show that the angular width of this sector is $2\pi\sin\phi_0$.
(c) Show that the distance ρ_0 between the peak of the cone and all points of tangency between the cone and the sphere is $\rho_0 = \cot\phi_0$.
(d) **Harder!** Suppose that the sector is unrolled and aligned as shown in Figure 1.13.

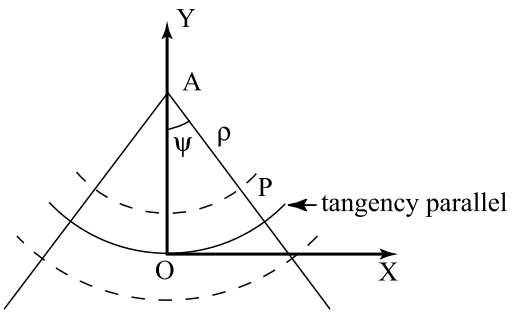


Fig. 1.13. The unrolling of the cone for Exercise 26: If P is a point, then $\rho = |AP|$ and ψ is the angle $\widehat{OAP}$.

The Lambert projection of the sphere onto this unrolled sector is defined as follows. Let $(x, y, z) = (\cos\theta\cos\phi, \sin\theta\cos\phi, \sin\phi)$ be a point on the sphere. Map it to the point

$$\begin{cases} X = \rho \sin \psi, \\ Y = \rho_0 - \rho \cos \psi, \end{cases}$$

where

$$\begin{cases} \rho = \rho_0 \left(\frac{\tan \frac{1}{2}(\frac{\pi}{2} - \phi)}{\tan \frac{1}{2}(\frac{\pi}{2} - \phi_0)} \right)^{\sin \phi_0}, \\ \psi = \theta \sin \phi_0. \end{cases}$$

Verify that the projection from the sphere to the cone given by $(x, y, z) \mapsto (X, Y)$ is conformal.

References

[1] M. Do Carmo. *Differential Geometry of Curves and Surfaces.* Prentice Hall, 1976.

[2] A. Peters, editor. *Peters World Atlas.* Turnaround Distribution, 2002.

[3] P. Richardus and R.K. Adler. *Map Projections.* North-Holland, 1972.

[4] E.F. Taylor and J.A. Wheeler. *Exploring Black Holes: Introduction to General Relativity.* Addison Wesley Longman, New York, 2000. (Chapters 1 and 2 and project on GPS.)

2

Friezes and Mosaics

This chapter discusses the classification of friezes and several concepts related to mosaics. The first section introduces the concept of operations that leave a frieze unchanged, using basic geometry and intuition. It also describes what will be the main steps of the classification theorem. Section 2.2 defines affine transformations and their matrix representation, and isometries. The highlight of this chapter is the classification theorem shown in Section 2.3. In less detail, the last section discusses mosaics. There is no advanced section to this chapter, the proof of the classification theorem being the most difficult element. Sections 2.1 and 2.4 can be covered in three hours of class. The tools are then purely geometric and the possibility of classification is made clear. If the classification theorem is the goal, four hours should be devoted to the first three sections. In all cases, the lecturer should bring copies of Figure 2.2 on transparencies to the classroom. Their use on a projector helps students to understand quickly the concept of symmetry. Only a basic knowledge of linear algebra and Euclidean geometry is required to understand this chapter. The proof of the classification theorem requires a familiarity with abstract reasoning.

This subject offers several interesting directions for further study: aperiodic tilings (end of Section 2.4) is one such direction, while Exercises 13, 14, 15, and 16 present several others.

Friezes and mosaics have been used in decoration for several millennia. The ancient world's Sumerian, Egyptian, and Mayan civilizations all used them to great effect. It would be a lie, however, to pretend that ancient mathematics developed the "technology" behind the art. The formal mathematical study of tilings is relatively recent, having started no more than two centuries ago. The memoir of Bravais [1], a French physicist, is among the first scientific studies of the subject.

Mathematics is able to provide a way to systematically classify the friezes and mosaics commonly seen in architecture and art. These classifications have allowed mathematicians to better understand the rules behind them and to create truly new patterns by breaking some of these rules.

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_2,

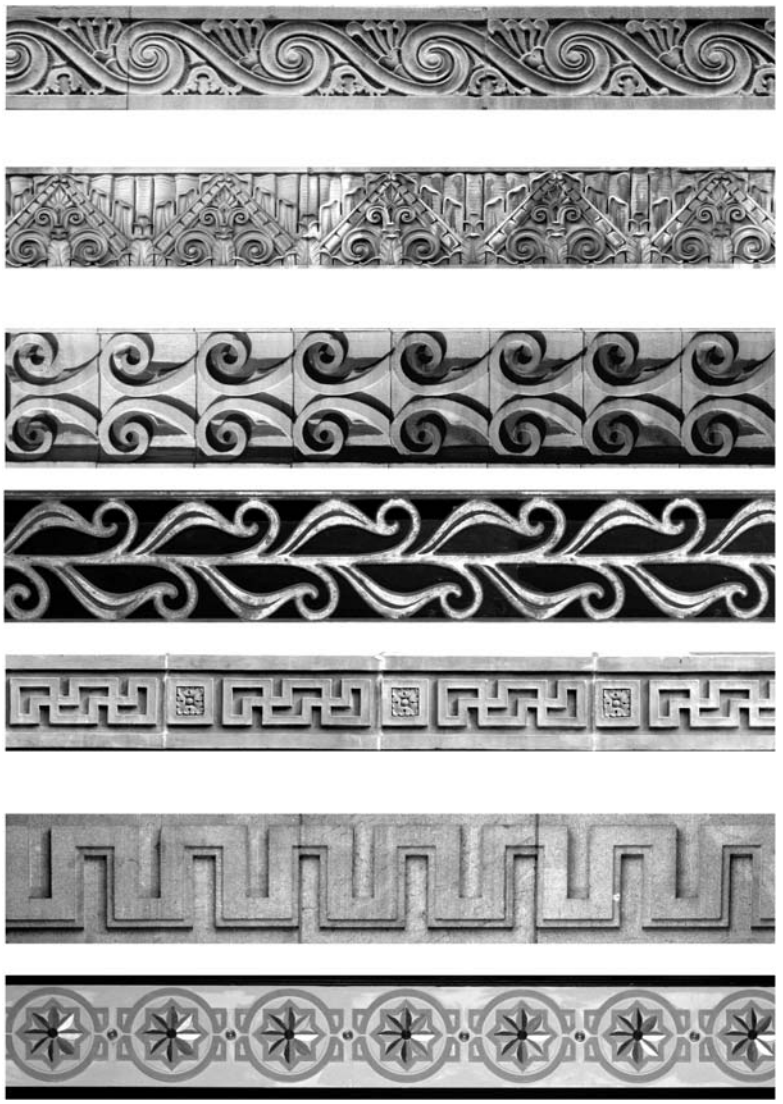


Fig. 2.1. Seven friezes. (Each of the above friezes has its pattern displayed in simplified form in Figure 2.2.)

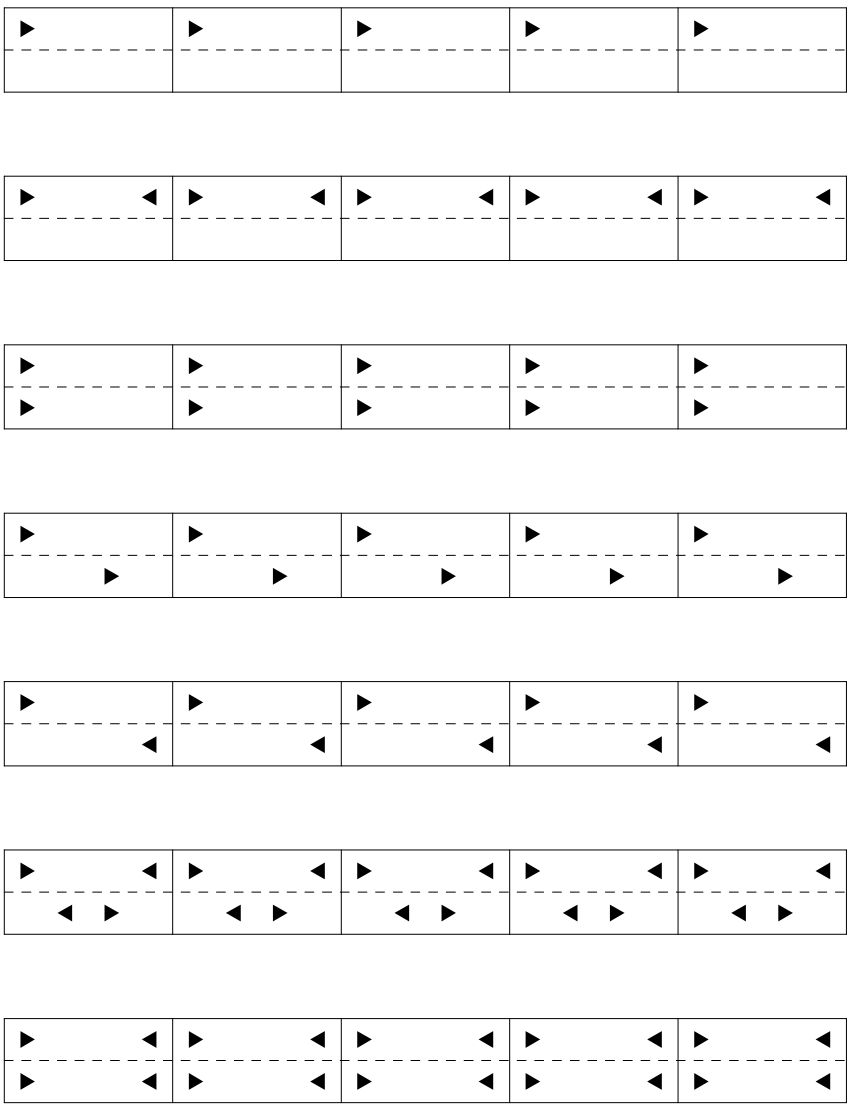


Fig. 2.2. Seven simplified friezes. (Each of the above friezes is a simplified form of the corresponding frieze in Figure 2.1.)

Classification of objects is a fairly common mathematical activity. The reader who has followed a course on multivariable calculus will remember the classification of extrema of a function of two variables using the second partial derivative test. If the matrix of second derivatives (the Hessian matrix) is nonsingular, the extremum can be classified as either a local minimum or a local maximum or a saddle point. The reader might also have encountered the classification of conics, either in an advanced linear algebra class or in Euclidean geometry. And for those having read Chapter 6 on error-correcting codes, Theorems 6.17 and 6.18 classify finite fields. These are examples of classifications of abstract objects. It may be surprizing to learn that mathematics can classify objects as concrete as architectural patterns. Here is how it is done.

2.1 Friezes and Symmetries

The *Oxford English Dictionary* defines *frieze* as *a band of painted or sculptured decoration.* It is also defined as *that member in the entablature of an order which comes between the architrave and cornice*, referring to the architectural location where such patterns are commonly used. Figure 2.1 shows seven friezes taken from architecture. To discuss these objects from a mathematical point of view, we will modify the definition to include the following elements: *(i)* a frieze has a constant and finite width (the height of the friezes in Figure 2.1) and is infinitely long in the perpendicular direction (the horizontal one in our examples); and, *(ii)* it is periodic, meaning that there exists some minimal distance $L > 0$ such that a translation of the frieze by a distance L along the direction in which it is infinite will leave the frieze unchanged. The length L is called the *period* of the frieze. This definition does not fit perfectly with real-world friezes (specifically those in Figure 2.1) because they are not infinitely long. However, we can easily imagine extending them infinitely in both directions by simply continuing the pattern.

Figure 2.2 presents seven more friezes. They are much less detailed but much simpler to study. Each of these seven friezes has the same period L, equal to the distance between two neighboring vertical bars. In the remaining discussion we will imagine that these vertical bars *do not appear* in the frieze pattern, since they have been drawn simply to make the period explicit. Some of these friezes are invariant under various geometric transformations other than translations. For example, the third and seventh friezes remain the same even if we flip them so as to exchange their top and bottom. In this case we say that they are invariant under *reflection by a horizontal mirror.* The second, sixth, and seventh friezes remain unchanged if flipped from left to right; we say that they are invariant under *reflection by a vertical mirror.* These distinctions between various friezes raise a natural question: *is it possible to classify all friezes by considering the set of operations under which they are invariant?* For example, the set of operations leaving the first frieze unchanged includes neither the horizontal nor the vertical reflection just discussed. This set of operations is distinct from that characterizing the third frieze, which may be reflected horizontally. Note that the

friezes in Figures 2.1 and 2.2 have been ordered such that they each display the same respective symmetries. Thus, corresponding pairs will be left unchanged by the same operations. For example, the third frieze in both figures is invariant under translations and horizontal reflection.

When a geometric transformation preserving lengths (such as a translation or a reflection) leaves a frieze unchanged, it is said to be a *symmetry operation of the frieze* or, simply, a *symmetry*. The complete list of symmetries of a frieze is infinite. Indeed, we would like to distinguish in this list the translation by a distance of one period L from the translations by $2L$, $3L$, etc., and these already account for an infinite number of symmetry operations. Moreover, the list should also contain the inverse of each symmetry operation. The *inverse of a symmetry operation* is the usual inverse of a function: the composition of a function and its inverse is the identity in the plane (or on the subset defined by the frieze as in the present case). The inverse of a translation to the right by a distance L is a translation to the left by the same distance. (Exercise: what is the inverse of a reflection with respect to a given mirror? and that of a rotation by an angle θ?) If translations to the right (respectively to the left) are associated to positive distances (respectively negative distances), then the list of symmetries of a frieze of period L should contain all translations by a distance nL with $n \in \mathbb{Z}$. Instead of listing all symmetries of a frieze, it is common to give only a subset of elements whose compositions and inverses give the whole list. Such a subset is called a *set of generators*. This is what we are going to use from now on. (Mathematicians usually take this subset as small as possible. They call it minimal whenever the subset, after removal of one of its elements, fails to generate the whole set of symmetries.)

The goal for the remainder of this section is to build geometric intuition of key ideas leading to the classification theorem, Theorem 2.12. This theorem gives all possible lists of symmetry generators for friezes of a given period. The reader is urged to make a copy of Figure 2.2 on a transparency and cut it into seven strips, one for each frieze, before reading on. Experimentation is an ideal way to develop intuition!

The three generators t_L, r_h, and r_v. We have already introduced some possible symmetry operations: translations (by any integer multiple of the period), reflections by horizontal and vertical mirrors. We will use the symbol r_h and r_v for the latter. The set of translations of a frieze is generated by the unique translation t_L by a period L. (The inverse of t_L is t_{-L}. Composition of n operations t_L gives $t_L \circ t_L \circ \cdots \circ t_L = t_{nL}$.)

A subtlety should be cleared up right away. For the reflection r_h to leave a frieze unchanged, the horizontal mirror should be located along the middle line of the frieze (the dashed lines in Figure 2.2). Its position is therefore completely determined by the requirement of being a symmetry. This is not the case for reflections through a vertical mirror. Positions of vertical mirrors must be chosen according to the pattern. The frieze **2** (the second from the top in Figure 2.2) has an infinite set of vertical mirrors. All small vertical bars define a position for a vertical mirror. But these are not the only ones. A mirror located halfway between two adjacent vertical bars also defines a symmetry of this frieze. Exercise 7 shows that if a frieze of period L is unchanged under a given vertical mirror, it is also invariant under an infinite number of mirrors, any of those

being at a distance $n\frac{L}{2}$, for $n \in \mathbb{Z}$, from the first. The notation r_v underlies therefore a choice for the position of one mirror and all other vertical mirrors at a distance equal to an integer multiple of $\frac{L}{2}$ from the first one. (Exercise: which other friezes of the figure have a symmetry r_v?)

Notation. Composition of symmetry operations will be used often in the following, and we shall drop the symbol "$\circ$". For example, $r_h \circ r_v$ will be simply noted $r_h r_v$. Soon will also appear the necessity of distinguishing the order of operations. It is important to note that operations are listed from right to left. The composition $r_h r_v$ stands for the operation r_v followed by r_h.

The rotation $r_h r_v$. The frieze **5** introduces a new generator. This frieze has neither r_h nor r_v as a symmetry, but if r_v and then r_h are both performed on it, the frieze remains unchanged. (The vertical mirror is along one of the vertical bars.) (Exercise: check this claim!) It can then happen that neither r_h nor r_v is a symmetry but their composition $r_h r_v$ is. The final result $r_h r_v$ of these two reflections is a rotation by an angle 180°. To see this, note that $r_h r_v$ exchanges the top and bottom, the left and the right, without altering the distances. This is exactly the action of rotation by 180°. (In terms of a coordinate system whose origin is on a vertical bar, a point (x, y) within the frieze is mapped into $(-x, -y)$ under this transformation. This is why this operation is also called *the symmetry through the origin.*) Exercise 8 proposes a geometrical proof of this property.

The following properties of the three generators r_h, r_v, and $r_h r_v$ are easily verified, geometrically or with the use of the copy on transparency that you have made of the figure. They could also be proved using the matrix representation that will be introduced in Section 2.2. (See Exercise 6.)

Proposition 2.1 *1. The operations r_h and r_v commute, that is, the two compositions $r_h r_v$ and $r_v r_h$ are equal.*
2. The inverse of r_h is r_h, that of r_v is r_v, and that of $r_h r_v$ is $r_h r_v$.
3. The composition of r_h and $r_h r_v$ gives r_v. That of r_v and $r_h r_v$ gives r_h. (This allows us to conclude that a frieze that would have any two of the three operations $r_h, r_v, r_h r_v$ as symmetries would automatically have the third also.)

With these properties, it should be easy to determine which of r_h, r_v, and $r_h r_v$ are symmetries of a given frieze of Figure 2.2. (Exercise: do it for all of them!)

The glide reflection symmetry $s_g = t_{L/2} r_h$. After the last proposition, the list of possible generators reads t_L, r_h, r_v, and $r_h r_v$. Any of r_h, r_v, and $r_h r_v$ is a symmetry of at least one frieze in Figure 2.2 and not a symmetry of at least one other frieze. But the frieze **4** shows that this list is not yet complete. None of $r_h, r_v, r_h r_v$ is a symmetry of this frieze. But a reflection r_h followed by a translation by a half-period $\frac{L}{2}$ leaves it unchanged. (See Figure 2.3. Recall that vertical bars *are not* part of the pattern.) We shall refer to this operation as the *glide reflection* and denote it by s_g. Using the composition we can write it as $s_g = t_{L/2} r_h$. (Exercise: only one other frieze among the seven of Figure 2.2 has s_g among its symmetries. Which one?)

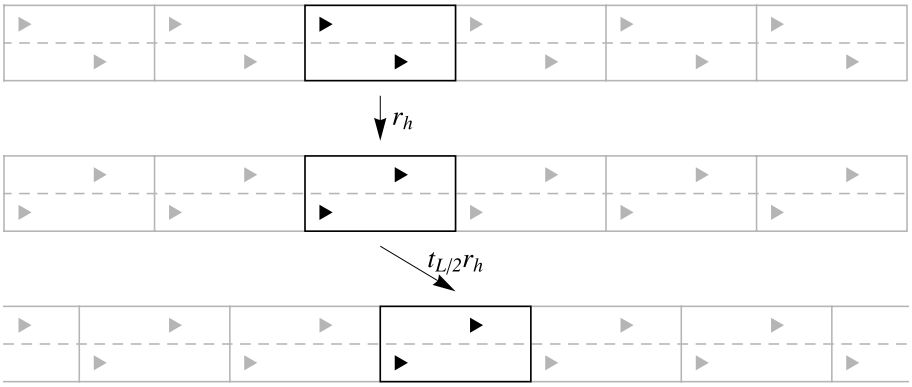


Fig. 2.3. A glide reflection. The frieze **4** as it appears in Figure 2.2 (top line), the same after the operation r_h (middle line), and after a translation by a half-period (bottom line).

Toward the classification theorem. The list of possible generators now contains five operations $(t_L, r_h, r_v, r_h r_v, s_g)$. It was obtained by studying Figure 2.2. To obtain the complete list of symmetry sets of friezes, we need all possible symmetry operations of friezes. What tells us that the list of five operations above is complete? Could there be another frieze that has a symmetry that cannot be obtained from these five? These will be the first questions to answer in order to prove the classification theorem.

Suppose for the time being that this list is complete. We can then enumerate potential sets of symmetries for friezes of period L. As stated above, we shall do this by identifying a set of generators. By definition, all sets will include the translation t_L by a distance L and no shorter ones. Any set may contain either zero or one or two of the three generators $r_h, r_v, r_h r_v$. (If the list contains two, it automatically contains the third one.) These observations lead to the following list.

1. $\langle t_L \rangle$
2. $\langle t_L, r_v \rangle$
3. $\langle t_L, r_h \rangle$
4. $\langle t_L, s_g \rangle$
5. $\langle t_L, r_h r_v \rangle$
6. $\langle t_L, s_g, r_h r_v \rangle$
7. $\langle t_L, r_h, r_v \rangle$
8. $\langle t_L, s_g, r_h \rangle$
9. $\langle t_L, s_g, r_v \rangle$
10. $\langle t_L, s_g, r_h, r_v \rangle$

All of the sets contain t_L. Sets **1** and **4** contain none of $r_h, r_v, r_h r_v$. Set **4** contains s_g, set **1** does not. Sets **2, 3, 5, 6, 8** and **9** contain one and only one of $r_h, r_v, r_h r_v$; **6, 8, 9**

add the glide reflection s_g, but **2, 3, 5** do not. Sets **7** and **10** contain two of $r_h, r_v, r_h r_v$ (and therefore all three). Set **10** has moreover s_g.

The classification theorem will have to resolve two more questions. The first is whether this list contains repetitions. Since we are listing only generators, two in the list above could generate the same list of symmetries. The second question is whether some of the sets do not generate symmetries of friezes of period L. This question might be somewhat surprising. But one can easily see that set **8** needs to be crossed out of the list, since it does not generate symmetries of a frieze of period L.

To see this, it is crucial to remember that the glide reflection s_g is the composition of r_h and $t_{L/2}$. But it can be seen that the set of generators of a frieze of period L cannot contain both s_g and r_h. Why? We have noted that the inverse of r_h is r_h itself. Then the composition of r_h and s_g is $s_g r_h = t_{L/2} r_h r_h = t_{L/2}(\mathrm{Id}) = t_{L/2}$. Because compositions of symmetries are symmetries, the translation $t_{L/2}$ should also be a symmetry of the frieze. But the period of the frieze was assumed to be L, and by definition, this period should be the smallest translation leaving the frieze invariant. The translation $t_{L/2}$ cannot appear, and hence s_g and r_h cannot simultaneously be generators of the same frieze. Set **8** must be rejected. (Note that this set does generate a set of symmetries for a frieze. But that frieze is of period $\frac{L}{2}$ and it is then set **3**, that is, $\langle t_{L/2}, r_h \rangle$.) (Exercise: the classification theorem will end up keeping only seven of the ten lists above. The argument for rejecting **8** was given. Can you guess which other two must be discarded?)

We shall complete the proof of the Classification theorem after having discussed a powerful algebraic tool to study these geometric operations: the matrix representation of affine transformations.[1]

2.2 Symmetry Group and Affine Transformations

We will use affine transformations as the mathematical foundation for describing invariant operations on friezes. (If you have read Chapter 3 or 11, you will have already encountered them.)

Definition 2.2 *An* affine transformation *in the plane is a transformation* $\mathbb{R}^2 \to \mathbb{R}^2$ *of the form* $(x, y) \mapsto (x', y')$*, where*

$$x' = ax + by + p,$$
$$y' = cx + dy + q.$$

An affine transformation is called proper *if it is bijective.*

Such a transformation can be described in matrix form as

[1] It is possible to give a purely geometric proof of this theorem. See, for example, [2] and [5].

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} p \\ q \end{pmatrix}. \tag{2.1}$$

The matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is a *linear transformation*, while p and q represent a *translation* in the plane. For the rest of this chapter we will be considering only proper (or *regular*) affine transformations, that is, affine transformations that are one-to-one. As we shall see soon, this additional condition is equivalent to the invertibility of the linear transformation matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Observe that the following equation describes the same affine transformation:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & p \\ c & d & q \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}. \tag{2.2}$$

In this modified form, a one-to-one correspondence is made between elements (x, y) of the plane $\mathbb{R}^2$ and elements $(x, y, 1)^t$ in the plane at $z = 1$ of $\mathbb{R}^3$. The mapping between affine transformations of the form (2.1) and the 3×3 matrices whose last line is (0 0 1),

$$\begin{pmatrix} a & b & p \\ c & d & q \\ 0 & 0 & 1 \end{pmatrix},$$

is also one-to-one.

If we compose two affine transformations $(x, y) \to (x', y')$ and $(x', y') \to (x'', y'')$ given by

$$\begin{aligned} x' &= a_1 x + b_1 y + p_1, \\ y' &= c_1 x + d_1 y + q_1, \end{aligned}$$

and

$$\begin{aligned} x'' &= a_2 x' + b_2 y' + p_2, \\ y'' &= c_2 x' + d_2 y' + q_2, \end{aligned}$$

the resulting (x'', y'') can be obtained as

$$\begin{aligned} x'' &= a_2 x' + b_2 y' + p_2 \\ &= a_2(a_1 x + b_1 y + p_1) + b_2(c_1 x + d_1 y + q_1) + p_2 \\ &= (a_2 a_1 + b_2 c_1)x + (a_2 b_1 + b_2 d_1)y + (a_2 p_1 + b_2 q_1 + p_2) \end{aligned}$$

and

$$\begin{aligned} y'' &= c_2 x' + d_2 y' + q_2 \\ &= c_2(a_1 x + b_1 y + p_1) + d_2(c_1 x + d_1 y + q_1) + q_2 \\ &= (c_2 a_1 + d_2 c_1)x + (c_2 b_1 + d_2 d_1)y + (c_2 p_1 + d_2 q_1 + q_2). \end{aligned}$$

Note that this compound transformation can itself be described in a 3×3 matrix form:

$$\begin{pmatrix} x'' \\ y'' \\ 1 \end{pmatrix} = \begin{pmatrix} a_2a_1 + b_2c_1 & a_2b_1 + b_2d_1 & a_2p_1 + b_2q_1 + p_2 \\ c_2a_1 + d_2c_1 & c_2b_1 + d_2d_1 & c_2p_1 + d_2q_1 + q_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}.$$

This last example demonstrates the utility of the 3×3 matrix notation, since composed transformations can themselves be expressed as the product of the matrices underlying the individual transformations:

$$\begin{pmatrix} a_2 & b_2 & p_2 \\ c_2 & d_2 & q_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 & b_1 & p_1 \\ c_1 & d_1 & q_1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a_2a_1 + b_2c_1 & a_2b_1 + b_2d_1 & a_2p_1 + b_2q_1 + p_2 \\ c_2a_1 + d_2c_1 & c_2b_1 + d_2d_1 & c_2p_1 + d_2q_1 + q_2 \\ 0 & 0 & 1 \end{pmatrix}.$$

This property allows us to study affine transformations and their *compositions* using this 3×3 representation and simple matrix multiplication. The geometric problem is thus reduced to a linear algebra problem. Because of this correspondence, we shall often use the matrix representation to describe an affine transformation. It should be stressed that an affine transformation can be defined without using a coordinate system, but its matrix representation exists only if one has been chosen.

To show the power of this notation we will now compute the inverse of a proper affine transformation. The inverse is the transform that associates $(x', y') \to (x, y)$, where $x' = ax + by + p$ and $y' = cx + dy + q$. Since the composition of affine transformations is represented by matrix multiplication, it must be that the matrix describing the inverse is the inverse of the matrix describing the original transform. This is easily calculated as

$$\begin{pmatrix} d/D & -b/D & (-dp + bq)/D \\ -c/D & a/D & (cp - aq)/D \\ 0 & 0 & 1 \end{pmatrix},$$

where $D = \det \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = ad - bc$. This is also a matrix describing a proper affine transformation. (Exercise: what must you do to ensure that it actually describes a proper transform? Do it. This exercise confirms the claim that an affine transformation is proper if and only if the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is invertible.) If we write the matrix describing the original transform in the form

$$B = \begin{pmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \mathbf{0} = \begin{pmatrix} 0 & 0 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{t} = \begin{pmatrix} p \\ q \end{pmatrix},$$

then its inverse may be written as

$$B^{-1} = \begin{pmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}\mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}.$$

Note that B^{-1} is of the same form as B: its third row is (0 0 1). Furthermore, note that the linear transformation A^{-1} is also invertible.

The set of all proper affine transformations forms a group.

Definition 2.3 *A set E equipped with a multiplication operation $E \times E \to E$ is a group if it satisfies the following properties:*

1. *associativity: $(ab)c = a(bc), \forall a, b, c \in E$;*
2. *existence of an identity element: there exists an element $e \in E$ such that $ea = ae = a, \forall a \in E$;*
3. *existence of inverses: $\forall a \in E$, $\exists b \in E$ such that $ab = ba = e$.*

The inverse of an element a is usually denoted by a^{-1}.

Groups play an important role in several other chapters. See, for example, Section 1.4 and Section 7.4.

It is easy to verify that the set of matrices representing proper affine transformations forms a group. Thus, the set of affine transformations itself forms a group. This is what we check now.

Proposition 2.4 *The set of matrices representing proper affine transformations forms a group under matrix multiplication. The set of proper affine transformations also forms a group under composition. The latter is called the* affine group*.*

Proof : Consider the matrix

$$B = \begin{pmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}$$

representing a proper affine transformation. Since the affine transformation is proper, A is an invertible 2×2 matrix and therefore the matrix B is itself invertible. Being of the same form as B, the matrix B^{-1} also represents a proper affine transformation, and condition 3 holds. Property 1 holds because matrix multiplication is itself associative, and property 2 holds using the 3×3 identity matrix, which represents the affine transformation

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \longleftrightarrow \qquad \begin{cases} x' = x, \\ y' = y. \end{cases}$$

Therefore the set of matrices representing proper affine transformations forms a group. We have seen that there is a one-to-one correspondence between matrices (with last line (0 0 1)) and affine transformations. Moreover, the composition of affine transformations is represented by matrix multiplication through this correspondence. The verification above automatically holds for the proper affine transformations themselves. □

Earlier, we introduced reflections with respect to horizontal and vertical mirrors. As examples, we now give their matrix representation. To obtain these, we need to fix the origin. We shall place it at equal distance between the top and bottom of the frieze.

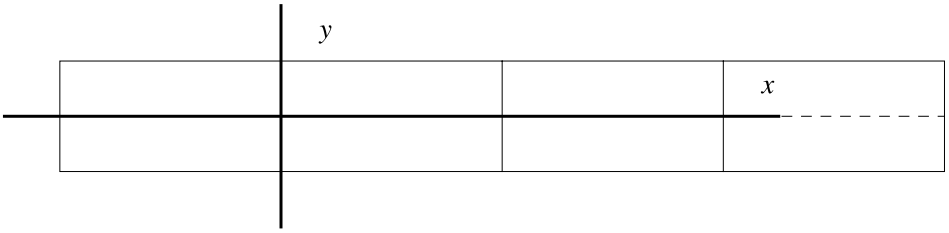


Fig. 2.4. The coordinate system.

(See Figure 2.4.) This still leaves some freedom, since any point on the horizontal axis in the middle of the frieze is a possible choice. (We have already underlined this freedom when discussing the position of vertical mirrors. We shall also use this freedom in the proof of Lemma 2.10.) For a given choice along the horizontal axis, the reflection r_h that exchanges top and bottom (that is, that exchanges the positive vertical axis with the negative one) is represented by the matrix

$$\begin{pmatrix} & & 0 \\ & r_h & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \text{where } r_h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and the reflection r_v that exchanges left and right is

$$\begin{pmatrix} & & 0 \\ & r_v & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \text{where } r_v = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

if the origin is on the mirror. (Exercise: check these claims.) Note that

$$r_h r_v = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We observe again that the rotation by an angle of 180° (or π) can be obtained by a reflection in a vertical mirror followed by a reflection in a horizontal one. (Exercise: determine the 3×3 matrices that represent the translation t_L and the glide reflection s_g.)

The definition of an affine transformation makes it a function from $\mathbb{R}^2$ to $\mathbb{R}^2$. The requirement that these functions leave a frieze invariant restricts the set of affine transformations that we need to consider. But a second restriction is made that limits the affine transformations even more.

Definition 2.5 *An isometry of the plane (or of a region of the plane) is a function $T : \mathbb{R}^2 \to \mathbb{R}^2$ (or $T : F \subset \mathbb{R}^2 \to \mathbb{R}^2$) that preserves lengths. Hence, if (x_1, y_1) and (x_2, y_2) are two points, then the distance between them is equal to the distance between their images $T(x_1, y_1)$ and $T(x_2, y_2)$.*

Definition 2.6 *A symmetry of a frieze is an isometry that maps the frieze onto the frieze.*

Exercise 9 will show that an isometry is an affine transformation. Lemma 2.7 shows that this restriction to isometric affine transformations limits significantly the possible linear transformations A that can play a role.

Lemma 2.7 *Let the isometry represented by the matrix*

$$\begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}$$

be a symmetry of a frieze. Then the 2×2 *block is one of the four matrices*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad r_h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad r_v = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad r_h r_v = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.3}$$

Proof: A linear transformation is completely determined by its action on a basis. We shall use the basis $\{\mathbf{u}, \mathbf{v}\}$, where $\mathbf{u}$ and $\mathbf{v}$ are horizontal and vertical vectors of length equal to half the width of the frieze. With this choice any point of the frieze is of the form $(x, y) = \alpha\mathbf{u} + \beta\mathbf{v}$ with $\alpha \in \mathbb{R}$ and $\beta \in [-1, 1]$. (The constraint $\beta \in [-1, 1]$ ensures that the point (x, y) is within the frieze.) The two basis vectors are perpendicular ($\mathbf{u} \perp \mathbf{v}$) or, equivalently, their inner product vanishes: $(\mathbf{u}, \mathbf{v}) = 0$.

To check whether $\left(\begin{smallmatrix} A & \mathbf{0} \\ \mathbf{0} & 1 \end{smallmatrix}\right)$ represents an isometry, it is sufficient to check that

$$|A\mathbf{u}| = |\mathbf{u}|, \qquad |A\mathbf{v}| = |\mathbf{v}|, \qquad \text{and} \qquad A\mathbf{u} \perp A\mathbf{v}. \tag{2.4}$$

Indeed, if P and Q are two points in the frieze and $Q - P = \alpha\mathbf{u} + \beta\mathbf{v}$ is the vector between them, then the image of $Q - P$ is $A(\alpha\mathbf{u} + \beta\mathbf{v})$ and the square of its length is given by

$$\begin{aligned} |A(\alpha\mathbf{u} + \beta\mathbf{v})|^2 &= (\alpha A\mathbf{u} + \beta A\mathbf{v}, \alpha A\mathbf{u} + \beta A\mathbf{v}) \\ &= \alpha^2 |A\mathbf{u}|^2 + 2\alpha\beta(A\mathbf{u}, A\mathbf{v}) + \beta^2 |A\mathbf{v}|^2 \\ &= \alpha^2 |\mathbf{u}|^2 + \beta^2 |\mathbf{v}|^2 \\ &= (\alpha\mathbf{u} + \beta\mathbf{v}, \alpha\mathbf{u} + \beta\mathbf{v}) \\ &= |\alpha\mathbf{u} + \beta\mathbf{v}|^2, \end{aligned}$$

where we have used, to obtain the third equality, the three relations of (2.4) and, for the fourth, the fact that the basis vectors are perpendicular. Then the distance between any pair of points P and Q is preserved by A if the relations (2.4) are satisfied. (Exercise: show that these relations are also necessary.)

Let $A\mathbf{u} = \gamma\mathbf{u} + \delta\mathbf{v}$ be the image of $\mathbf{u}$ by A. Since the transformation is linear, $A(\beta\mathbf{u}) = \beta(\gamma\mathbf{u} + \delta\mathbf{v})$. If δ is nonzero, then it is possible to choose $\beta \in \mathbb{R}$ sufficiently large that $|\beta\delta| > 1$. This means that the point $A(\beta\mathbf{u})$ is outside the frieze. Since this

must be ruled out, δ has to be set to zero. (In other words, a transformation A such that δ is nonzero is a linear transformation that tilts the frieze out of the horizontal.) Thus $A\mathbf{u} = \gamma\mathbf{u}$, and if $|A\mathbf{u}| = |\mathbf{u}|$, we must have $\gamma = \pm 1$.

Now let $A\mathbf{v} = \rho\mathbf{u} + \sigma\mathbf{v}$ be the image of $\mathbf{v}$ under A. Since $A\mathbf{u}$ must be perpendicular to $A\mathbf{v}$, we must have

$$0 = (A\mathbf{u}, A\mathbf{v}) = (\gamma\mathbf{u}, \rho\mathbf{u} + \sigma\mathbf{v}) = \gamma\rho|\mathbf{u}|^2.$$

Since neither γ nor $|\mathbf{u}|$ is zero, ρ must be set to 0. And again the last condition $|A\mathbf{v}| = |\mathbf{v}|$ fixes σ to be ± 1. The matrix A representing the transformation in the basis $\{\mathbf{u}, \mathbf{v}\}$ is then $\left(\begin{smallmatrix} \gamma & 0 \\ 0 & \sigma \end{smallmatrix}\right)$. There are two choices for each γ and σ and thus four for the matrix A, precisely those appearing in the statement. □

The composition of two isometries and the inverse of an isometry are themselves isometries. Thus the subset of isometric transformations of the affine group itself forms a group, called the *group of isometries*. Finally, the composition of two isometries leaving a frieze unchanged itself leaves the frieze unchanged. The subset of the group of isometries that leaves the frieze invariant is therefore a group. We are led to the following definition.

Definition 2.8 *The group of symmetry of a frieze is the group of all isometries that leave the frieze invariant.*

2.3 The Classification Theorem

Having a formal theory of isometries and affine transformations allows us to create a list of such transformations that could leave a frieze unchanged. This section will first establish a complete list of possible symmetry generators. The second part of this section uses this list of transformations to enumerate and classify all possible types of groups of frieze symmetries.

There are many affine transformations that simply cannot appear in the symmetry group of a frieze. Lemma 2.7 has already rejected the linear transformations that tilt the frieze out of its domain (the constraint $\delta = 0$ excludes these transformations). The following lemmas characterize the transformations that can appear in frieze symmetry groups. The first describes translations along the infinite axis of the frieze.

Lemma 2.9 *The symmetry group of any frieze of period L contains the translations*

$$\begin{pmatrix} 1 & 0 & nL \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad n \in \mathbb{Z}.$$

These are the only translations that appear in the symmetry group.

PROOF: The translation

$$t_L = \begin{pmatrix} 1 & 0 & L \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

leaves any frieze with period L unchanged. Observe that the inverse of this translation is

$$t_{-L} = \begin{pmatrix} 1 & 0 & -L \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and that its composition n times yields

$$t_{nL} = \begin{pmatrix} 1 & 0 & nL \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

(Exercise!) The translation t_{nL} must therefore be in the symmetry group for all $n \in \mathbb{Z}$. No translation of the form

$$\begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix}$$

with $b \neq 0$ can leave a frieze unchanged, since the vertical portion of the translation will map certain points of the frieze outside of its original vertical extent. We are left with possible translations of the form

$$\begin{pmatrix} 1 & 0 & a \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where a is not an integer multiple of L. After performing such a translation by $\left(\begin{smallmatrix} a \\ 0 \end{smallmatrix}\right)$, one can repeatedly perform a translation by $\left(\begin{smallmatrix} L \\ 0 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} -L \\ 0 \end{smallmatrix}\right)$ until the resulting translation is by $\left(\begin{smallmatrix} a' \\ 0 \end{smallmatrix}\right)$, where a' satisfies $0 \leq a' < L$. If $0 < a' < L$, it is a translation by a constant a' smaller than the period L, contradicting the definition of the period. And if $a' = 0$, then the original a was an integer multiple of the period L. The only translations left are therefore $t_{nL}, n \in \mathbb{Z}$. □

Are there any other transformations of the form

$$\begin{pmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}$$

where A is not the identity matrix and $\mathbf{t}$ is nonzero? The next lemma answers this question.

Lemma 2.10 *Consider isometries of the form* $\left(\begin{smallmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{smallmatrix}\right)$*, where* $\mathbf{t}$ *is nonzero. By redefining the origin it is possible to reduce any such transformation to one of the form*

$$(i)\ \begin{pmatrix} & A & & nL \\ & & & 0 \\ 0 & 0 & & 1 \end{pmatrix}, \quad (ii)\ \begin{pmatrix} 1 & 0 & L/2+nL \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad (iii)\ \begin{pmatrix} -1 & 0 & L/2+nL \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $n \in \mathbb{Z}$ *and* A *is one of the four allowed by Lemma 2.7. Form (iii) may occur only if the rotation* $r_h r_v$ *is also a symmetry.*

Proof: By definition of an isometry, lengths must be preserved. Since the distance between two points is the same as the distance between any translation of the same two points, the matrix A must be one of the four given in (2.3). Moreover, if $t_y \neq 0$ in

$$\begin{pmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{pmatrix},$$

then $y' = cx + dy + t_y$ will be outside of the frieze for certain values of x and y. In fact, for the four possible matrices A, the image of the square $[-1,1] \times [-1,1]$ is the square itself. Every translation that has $t_y \neq 0$ moves the square vertically and takes some points of this square out of the frieze. Thus, t_y must be zero.

Since the symmetry group of a frieze contains all horizontal translations by integer multiples of L, the presence of

$$\begin{pmatrix} a & 0 & t_x \\ 0 & d & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

in the group implies the presence of

$$\begin{pmatrix} 1 & 0 & nL \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & 0 & t_x \\ 0 & d & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a & 0 & t_x + nL \\ 0 & d & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for all $n \in \mathbb{Z}$. Out of the set of all such transformations there will be one such that $0 \leq t'_x = t_x + nL < L$.

We now consider the four possibilities for A. If A is the identity matrix, then Lemma 2.9 forces t'_x to be zero, and the resulting matrix is of the form *(i)*.

Let $A = r_h$. Then the square of

$$\begin{pmatrix} & r_h & & t'_x \\ & & & 0 \\ 0 & 0 & & 1 \end{pmatrix}$$

must also be in the symmetry group of the frieze. However,

$$\begin{pmatrix} 1 & 0 & t'_x \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^2 = \begin{pmatrix} 1 & 0 & 2t'_x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is a translation. Thus there exists $m \in \mathbb{Z}$ such that $2t'_x = mL$. Since $0 \leq t'_x < L$, we have that $0 \leq 2t'_x < 2L$. If $t'_x = 0$, the translation is trivial. Otherwise, we must have that $t'_x = L/2$, and the affine transformation becomes

$$\begin{pmatrix} 1 & 0 & L/2 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.5}$$

It remains to consider the two cases $A = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Here we will use our freedom in choosing the origin. (See the remarks after the proof of Proposition 2.9.) Consider translating the origin along the x axis by a distance a. The matrix describing the coordinate change is given by

$$S = \begin{pmatrix} 1 & 0 & -a \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

If T is the matrix representing an affine transformation and S the matrix changing the coordinate system (x, y) to a new one (x', y'), the same affine transformation will be represented by the matrix STS^{-1} in the new system. To see this, we read as usual from right to left. This expression first transforms the coordinates (x', y') of a point into its coordinates (x, y) in the old system using S^{-1}, applies the affine transformation represented in these old coordinates by the matrix T, and transforms the result back with S into the new coordinate system. The affine transformation represented by

$$\begin{pmatrix} -1 & 0 & t'_x \\ 0 & \pm 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.6}$$

will therefore be represented by the matrix

$$\begin{pmatrix} 1 & 0 & -a \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & t'_x \\ 0 & \pm 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} -1 & 0 & t'_x - a \\ 0 & \pm 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 & t'_x - 2a \\ 0 & \pm 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

in the new system. (Exercise: It is crucial to check that this coordinate change does not spoil the form of other symmetry operations. Show that transformations represented by $\left(\begin{smallmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{smallmatrix}\right)$ with A equal to $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ or r_h keep the same matrix representation after a horizontal

translation of the origin.) Thus the affine transformation represented by (2.6) is now represented by

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & \pm 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.7}$$

if we displace the origin by precisely $a = t'_x/2$.

Note that if the symmetry group contains two transformations of the form (2.6) with distinct $t'_{x1}, t'_{x2} \in [0, L)$, then moving the origin assures us that the transformation with t'_{x1} can be written in the form (2.7). The second remains of the form (2.6) with t'_{x2} replaced by $t_{x2} = t'_{x2} - t'_{x1}$. If both transformations have the same A, then their composition will be a translation by t_{x_2}, forcing t_{x2} to be nL for some integer n. In this case both transformations are cast into form *(i)* by the change of origin. If, however, the two transformations have different A's, we may suppose that the first has $A = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ and then it is a rotation $r_h r_v$ by 180°. The composition of the two is then

$$\begin{pmatrix} 1 & 0 & t_{x2} \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and by previous arguments, t_{x2} must be either nL or $nL + \frac{L}{2}$ for some integer n. The second transformation is then of the form *(i)* if t_{x2} is an integer multiple of L or of the form *(iii)* if not. □

The first two forms of isometries allowed by Lemma 2.10 are then *(i)* the composition of one of the linear transformations of Lemma 2.7 and a translation t_{nL} by an integer multiple of the period L and *(ii)* the composition of the glide reflection s_g and a translation t_{nL}. The third form *(iii)* may appear only if $r_h r_v$ is also present, and in this case, one can use $r_h r_v$ and the isometry of the form *(ii)* (with $n = 0$) as generators. Hence the three lemmas together show that the symmetry group of a frieze can be generated by a subset of $\{t_L, r_h, r_v, r_h r_v, s_g\}$. This answers the question of the list of possible generators, a question left open at the end of Section 2.1.

The lemmas will now allow us to finish our classification of the symmetry groups of various friezes, which will provide us with an affirmative answer to our earlier question: *is it possible to classify friezes based on the set of geometric operations under which they are invariant?* When describing the various possible symmetry groups we will simply reference the generators of each group. We recall formally the definition of such a list of generators.

Definition 2.11 *Let $\{a, b, \ldots, c\}$ be a subset of a group G. This set is a set of generators for G, and then we write $G = \langle a, b, \ldots, c \rangle$ if the set of all compositions of a finite number of elements of $\{a, b, \ldots, c\}$ and of their inverses is G.*

Theorem 2.12 (Classification of frieze groups) *The symmetry group of any frieze is one of the following seven groups:*

1. $\langle t_L \rangle$
2. $\langle t_L, r_v \rangle$
3. $\langle t_L, r_h \rangle$
4. $\langle t_L, t_{L/2} r_h \rangle$
5. $\langle t_L, r_h r_v \rangle$
6. $\langle t_L, t_{L/2} r_h, r_h r_v \rangle$
7. $\langle t_L, r_h, r_v \rangle$

Each of these groups is described by a set of generators, and they are presented in the same order as those in Figures 2.1 and 2.2.

PROOF: Let t_L represent translation by a distance L along the horizontal axis. All of the groups contain translations by integer multiples of L, the period of the frieze, and the list of generators must contain t_L. Through an appropriate choice for the origin, the only other generators of the symmetry groups will be the linear transformations denoted by $A = r_h, r_v$ or $r_h r_v$ and the glide reflection s_g allowed by Lemma 2.10. Note that if a symmetry group contains any two of r_h, r_v, and $r_h r_v$ then it must automatically contain all three. The list of all possible combinations of generators therefore consists of the seven given in the statement of the theorem as well as

8. $\langle t_L, t_{L/2} r_h, r_h \rangle$
9. $\langle t_L, t_{L/2} r_h, r_v \rangle$
10. $\langle t_L, t_{L/2} r_h, r_h, r_v \rangle$

(See the discussion at the end of Section 2.1, where this list was first constructed.) We repeat here the argument that forces us to reject the case **8**. The presence of $s_g = t_{L/2} r_h$ and r_h implies that the group must also contain their product $(t_{L/2} r_h) r_h = t_{L/2}(r_h^2) = t_{L/2}$, which is a translation by $L/2$ (since $r_h^2 = \mathrm{Id}$). This contradicts the fact that the frieze is periodic with a minimum period of L, and therefore this set must be rejected.

For case **9**, note that the product of s_g and r_v is of the form $t_{L/2} r_h r_v$ discussed in Lemma 2.10. Through a translation of the origin (by $a = \frac{L}{4}$), this product can be written in the form of (2.7) with $A = r_h r_v$. A simple calculation shows that the generators t_L and s_g are unchanged by this translation but that r_v becomes $s_g = t_{L/2} r_v$. Thus subgroup **9** is equally described by the generators $\langle t_L, t_{L/2} r_h, t_{L/2} r_v, r_h r_v \rangle$. Three of these generators belong to **6**, while the fourth $(t_{L/2} r_v)$ is simply the product of $t_{L/2} r_h$ and $r_h r_v$. Case **9** is in fact identical to case **6** and it may be omitted.

Finally, case **10** contains the generators of case **8** and can be eliminated for the same reason.

Thus the symmetry group of any frieze must be one of the seven listed groups. Is there any redundancy in this list? No, and with the help of Figure 2.2 we can easily convince ourselves of this fact. The full argument is rather tedious, and thus we will restrict ourselves to frieze **4**, whose symmetry group was determined to be $\langle t_L, s_g \rangle$. We first observe that the two generators t_L and s_g are both symmetries of this frieze. The group they generate must therefore be a subgroup of the actual symmetry group of the frieze. Can we add any other generators to these two? A quick inspection shows that

no such addition (from among the remaining possibilities $r_h, r_v, r_h r_v$) is possible. Thus $\langle t_L, s_g \rangle$ is indeed the entire symmetry group of the frieze **4**. Finally, since group **1** is distinct from **4** and the remaining five groups each contain at least one of r_h, r_v, and $r_h r_v$ which group **4** does not have, then group **4** is in fact distinct from the other six. Repeating an argument of this type for each of the remaining friezes and symmetry groups shows that the list is exhaustive and does not contain any redundancy. □

2.4 Mosaics

In architecture, mosaics are as popular, if not more popular, than friezes. For us, a mosaic will be a pattern that can be repeated to fill the plane and that is periodic along two linearly independent directions. Thus, a mosaic has two linearly independent vectors $\mathbf{t}_1$ and $\mathbf{t}_2$ along which it may be translated without change.

As with friezes, mosaics may be studied in terms of the symmetry operations that leave them unchanged. And as with friezes, they may also be classified by their symmetry groups. Due to their importance in the physics and chemistry of crystals, they are referred to as the *crystallographic groups*. There are 17 crystallographic groups. We will not derive this classification. We will limit ourselves to enumerating the rotations that may appear in the symmetry groups of mosaics, and to understanding the description of the classification.

Lemma 2.13 *Any rotation that leaves a mosaic unchanged must have one of the following angles:* $\pi, \frac{2\pi}{3}, \frac{\pi}{2}, \frac{\pi}{3}$.

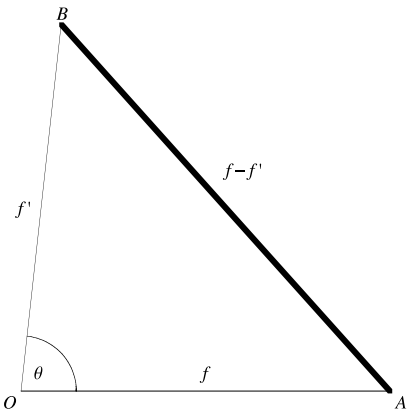


Fig. 2.5. The point $\mathcal{O}$ and two of its images $\mathcal{A}, \mathcal{B}$ under translation.

PROOF: Let $\mathcal{O}$ be the center of a rotation leaving the mosaic unchanged. Let $\theta = \frac{2\pi}{n}$ be the smallest angle describing the rotation about this point. Since the mosaic is periodic in two linearly independent directions, there exists an infinity of such points. Let $\mathbf{f}$ be a vector joining $\mathcal{O}$ to a nearby image $\mathcal{A}$ chosen among the closest images of $\mathcal{O}$ obtained by translations. Then translation along the vector $\mathbf{f}$ belongs to the symmetry group of the mosaic.

By rotating the mosaic about $\mathcal{O}$ by an angle θ, the point $\mathcal{A}$ is mapped to $\mathcal{B}$. The vector $\mathbf{f}'$ joining $\mathcal{O}$ to $\mathcal{B}$ also describes a translation under which the mosaic is invariant (see Figure 2.5). The distance between $\mathcal{A}$ and $\mathcal{B}$ is the length of the vector $\mathbf{f}' - \mathbf{f}$, and since $\mathbf{f}' - \mathbf{f}$ is also a translation leaving the mosaic unchanged, this distance must be greater than or equal to the length of $\mathbf{f}$ by hypothesis. ($\mathcal{A}$ was one of the nearest images of $\mathcal{O}$.) Since $\mathbf{f}$ and $\mathbf{f}'$ are of the same length, it must be that the angle $\theta = \frac{2\pi}{n}$ is greater than or equal to $\frac{2\pi}{6} = \frac{\pi}{3}$ (which is 60°). In fact, $\frac{\pi}{3}$ is the precise angle such that $\mathbf{f}$, $\mathbf{f}'$, and $\mathbf{f}' - \mathbf{f}$ are all the same length. This first argument restricts the possibilities to $\frac{2\pi}{2} = \pi$, $\frac{2\pi}{3}$, $\frac{2\pi}{4} = \frac{\pi}{2}$, $\frac{2\pi}{5}$, and $\frac{2\pi}{6} = \frac{\pi}{3}$.

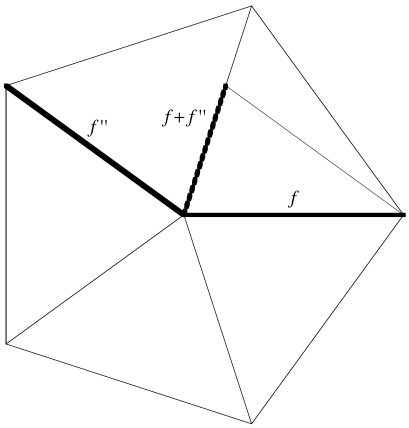


Fig. 2.6. The case of rotation by an angle $\frac{2\pi}{5}$.

However, no mosaic can be left unchanged after rotation by an angle of $\frac{2\pi}{5}$. Figure 2.6 shows $\mathbf{f}$ and its image $\mathbf{f}''$ after a rotation of $\frac{4\pi}{5}$. Translation along $\mathbf{f}+\mathbf{f}''$ must also be an invariant operation, but its length is shorter than that of $\mathbf{f}$, a contradiction. Thus, we can safely reject this angle. □

The elements of the crystallographic groups are similar to those found in the frieze symmetry groups: translations, reflections, reflections followed by translations (that is, glide reflections as for friezes), and rotations. Rather than exhaustively listing the generators for each of the 17 crystallographic groups, we will instead show an example of

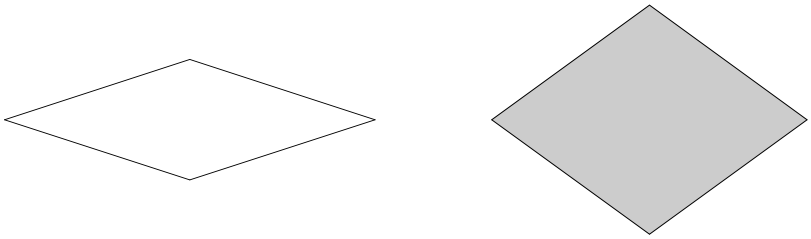


Fig. 2.7. Penrose tiles.

each type and highlight its symmetries (see Figures 2.17 through 2.22, starting on page 77). For each class we illustrate the basic shape of the mosaic at the left, overlaid with a shaded parallelogram whose sides indicate the two linearly independent directions in which the mosaic may be translated. These vectors have been chosen such that the parallelogram encloses the smallest possible area necessary to cover the plane by translations along them. There is usually more than one choice for this parallelogram. On the right, the same mosaic has been drawn again with axes of reflection or glide reflection and points of rotation overlaid. Finally, the legend of each graph identifies the *international symbols* commonly used to designate each crystallographic group [5]. Solid lines indicate that a simple reflection across the axis is a symmetry. Dashed lines indicate glide reflections; the required translations are not explicitly shown but are easily seen nonetheless. Various symbols are used to indicate points about which the mosaic may be rotated. If the center of rotation does not fall on an axis of reflection, the following are used:

$\diamond$ for rotations of angle π,
$\triangle$ for rotations of angle $\frac{2\pi}{3}$,
$\square$ for rotations of angle $\frac{\pi}{2}$,
and hexagons for rotations of angle $\frac{\pi}{3}$.

When the point of rotation lies along an axis of reflection, solid versions of the same symbols (▲, ■, etc.) are employed.

The ancient city of Alhambra, seat of the Moorish government of Granada in the south of modern-day Spain, houses many mosaics that are as stunning in number as they are in complexity. For a long time it was debated whether all 17 crystallographic groups were represented by the Alhambra mosaics. Grünbaum, Grünbaum, and Shephard [4] claim that this is not the case, with only 13 groups being employed. Even with this negative response, it is still natural to ask whether the Moorish artists of the time were aware of such a system of classification.

The precise mathematical formalization of friezes and mosaics allowed mathematicians to study new generalized structures by relaxing certain rules in the definition. Aperiodic tilings are one such structure. All mosaics must fill the plane, meaning that repeating the pattern in all directions covers all points of $\mathbb{R}^2$ without leaving any gaps.

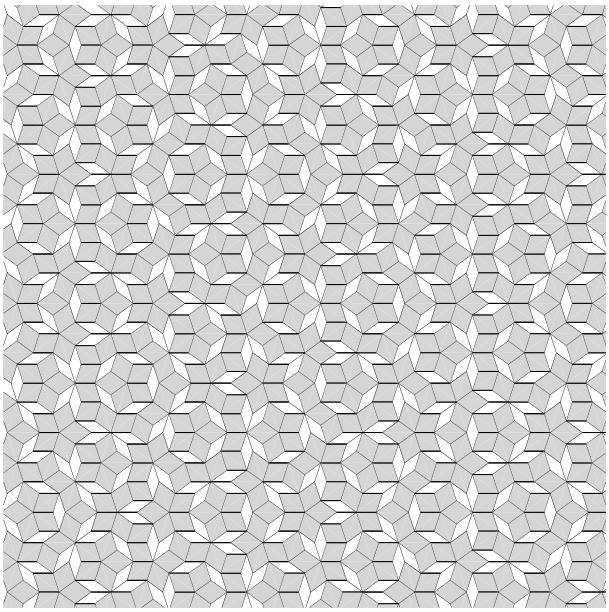


Fig. 2.8. An aperiodic Penrose tiling.

This condition is also satisfied by aperiodic tilings. For example, it is possible to tile the plane $\mathbb{R}^2$ with the two Penrose tiles (referred to as the Penrose rhombs) shown in Figure 2.7 [5]. Even if it is possible to tile the plane in a periodic manner with these tiles, it is also possible to arrange them in such a way that no translational symmetry is present; in other words, they may be used to tile the plane in an *aperiodic* manner. Figure 2.8 shows a fragment of an aperiodic tiling. Maybe these new generalized structures will find their way into architecture... (There are other sets of tiles, constructed by Penrose and others, that may be tiled *only* aperiodically!)

2.5 Exercises

1. We say that two operations $a, b \in E$ commute if $ab = ba$.
(a) Do translation operations commute?
(b) Do r_h, r_v, and $r_h r_v$ all commute with each other?
(c) Do the reflections r_h, r_v, and $r_h r_v$ commute with translations?

2. Find the conditions under which a linear transformation

$$\begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and a translation

$$\begin{pmatrix} 1 & 0 & p \\ 0 & 1 & q \\ 0 & 0 & 1 \end{pmatrix}$$

will commute with each other.

Fig. 2.9. The frieze of Exercise 3.

3. **(a)** Determine the period L of the frieze in Figure 2.9. Indicate it directly on the figure or a copy of it.
(b) Under which of the transformations $t_L, r_h, s_g, r_v, r_h r_v$ is the frieze invariant?
(c) Which of the seven symmetry groups does the frieze belong to?
(d) By drawing a single point per period on the frieze, reduce its symmetry group to $\langle t_L \rangle$ without changing the length of its period.

4. **(a)** Friezes are often used in architecture, with [3] giving several remarkable examples. Select a few such examples, and determine to which of the symmetry groups they belong.
(b) The artist M. C. Escher created several remarkable mosaics, with a large number of them being presented in [6]. Select a few of Escher's mosaics and determine to which of the 17 crystallographic groups they belong.

5. **(a)** Identify the symmetry group of the frieze shown in Figure 2.10.

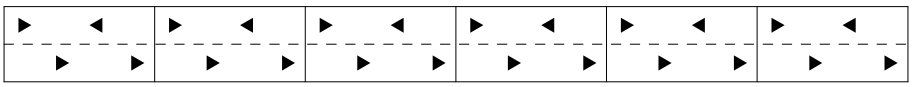


Fig. 2.10. Frieze for Exercise 5.

(b) By removing two triangles from each period of this frieze, construct a frieze belonging to the symmetry group **5**.

6. Prove the three statements of Proposition 2.1. Suggestion: these properties can be proved using only Euclidean geometry or using the matrix representation of affine transformations. Explore both approaches.

7. **(a)** Let m_1 and m_2 be parallel lines at a distance d and let r_{m_1} and r_{m_2} be the reflections through these lines. Show that the composition $r_{m_2}r_{m_1}$ is a translation by a distance $2d$ along a direction perpendicular to the lines (mirrors) m_1 and m_2. Hint: show this using only Euclidean geometry, that is, without use of a coordinate system. You may use the concept of distance or length of a segment.
(b) Let a frieze of period L be invariant under the reflection r_v. Show that it is invariant under reflection through a vertical mirror at distance $\frac{L}{2}$ from the first. Hint: study the composition of r_v and the translation t_L.

8. Let m_1 and m_2 be two lines intersecting at P and let r_{m_1} and r_{m_2} be the reflections through these lines. Show that the composition $r_{m_2}r_{m_1}$ is a rotation of center P by twice the angle between the two lines (mirrors) m_1 and m_2. Hint: show first that the images $r_{m_1}Q$ and $Q' = r_{m_2}r_{m_1}Q$ lie on a circle of center P and of radius $|PQ|$. Then study the angles made by the segments PQ and PQ' with a given line, say m_1.

9. The goal of this exercise is to show that an isometry is the composition of a linear transformation and a translation and therefore is an affine transformation. (Either the linear transformation or the translation could be the identity.) Recall that a linear transformation of the plane is a function $T : \mathbb{R}^2 \to \mathbb{R}^2$ that satisfies the following two conditions: *(i)* $T(\mathbf{u}+\mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ and *(ii)* $T(c\mathbf{u}) = cT(\mathbf{u})$ for all points $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ and constant $c \in \mathbb{R}$.
(a) Show that an isometry $T : \mathbb{R}^2 \to \mathbb{R}^2$ preserves angles. Hint: choose three (non-collinear) points P, Q, R. If P', Q', R' are their images under T, show that the triangles PQR are $P'Q'R'$ are congruent.
(b) Show that a translation is an isometry.
(c) Suppose that an isometry S has no fixed-point and that $S(P) = Q$. Show that the composition TS, where T is the translation that maps Q to P, has at least one fixed-point.
(d) Let S be an isometry that has (at least) one fixed-point O. Let P, Q, R be chosen such that $OPQR$ is a parallelogram. Let P', Q', R' be their image under S. Show that the sum of the vectors OP' and OR' is OQ'. (This amounts to $S(OP + OR) = S(OP) + S(OR)$.)
(e) Let S be an isometry that has (at least) one fixed-point O and let P and Q be two points, distinct and distinct from O, such that O, P, Q are collinear. Show that

$$S(OP) = \frac{|OP|}{|OQ|} S(OQ).$$

(f) Conclude that an isometry is a linear transformation followed by a translation and is therefore an affine transformation. (Either of the two operations could be the identity.)

10. **(a)** The pattern of Figure 2.11 consists of a series of ellipses centered along the x axis at the points $(2^i, 0)$ with principal axes $r_x = 2^{i-2}, r_y = 1$. Thus, this pattern exists over the infinite half-strip $(0, \infty) \times [-\frac{1}{2}, \frac{1}{2}]$. This pattern is not a frieze because it is not periodic. Replace the periodicity condition with another invariance condition such that this pattern is a "frieze."
(b) Describe the transformation that maps one ellipse to the first one on its left. Is it linear? Does the set of such transformations form a group?

Fig. 2.11. A pattern that is not periodic. (For Exercise 10.)

11. Let $r > 1$ be a real number and let

$$A_r = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; \frac{1}{r} \leq \sqrt{x^2 + y^2} \leq r \right\}$$

be the ring with center at the origin of the plane and delimited by the circles with radii r and $\frac{1}{r}$.
(a) Show that the set A_r is invariant under rotations of the form

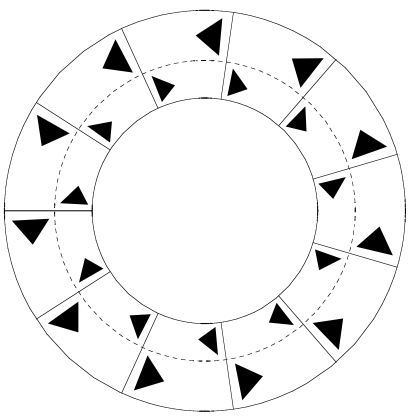


Fig. 2.12. A circular frieze. (See Exercise 11.)

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

for all $\theta \in [0, 2\pi)$. (The invariance of A_r means that the transformation is invertible and that the image of A_r is A_r itself.)

(b) Consider the transformation $\mathbb{R}^2 \setminus \{(0,0)\} \to \mathbb{R}^2 \setminus \{(0,0)\}$ defined by

$$x' = \frac{x}{x^2 + y^2},$$
$$y' = \frac{y}{x^2 + y^2}.$$

This transformation is called an *inversion*. Show that A_r is invariant under this transformation. Show that A_r^2 is the identity transformation. Is this transformation linear?

(c) Figure 2.12 represents a circular frieze drawn on a ring A_r. The dashed line represents the circle of radius 1. Unlike the band friezes discussed earlier, circular friezes are bounded. It is easy to construct a correspondence between the symmetries of a band frieze presented in Section 2.2 and those of a circular frieze. Translations become rotations, and reflection r_h across the horizontal axis becomes inversion as introduced in (b). Define the transformation that corresponds to reflection r_v across a vertical axis. We will call this last transformation *reflection*. Is reflection a linear transformation? (As before, this transformation can be defined only after a suitable origin has been chosen. You will have to carefully choose a particular point of A_r through which the "mirror" will pass.)

(d) Starting from the three operations of rotation, inversion, and reflection, construct a set of generators for the symmetry group of the circular frieze shown in Figure 2.12.

12. **(a)** This exercise continues the previous one. Let n be the largest integer for which a circular frieze is invariant under a rotation of $\frac{2\pi}{n}$. We will suppose that $n \geq 2$. Classify the symmetry groups of a circular frieze for a given n. Does the classification depend on n in any way?

(b) The *order* of a group is the number of elements in the group. The orders of the symmetry groups of regular friezes are infinite, but those of circular friezes are finite. Calculate the orders of the groups you constructed in (a).

13. For each Archimedean tiling shown in Figure 2.13, determine to which of the 17 crystallographic groups it belongs (certain tilings must belong to the same group). An *Archimedean tiling* is a tiling of the plane consisting of regular polygons such that each vertex is of the same *type*. For two vertices to be of the same type, they must be coincident with similar polygons, and the polygons must appear in the same order as we turn about the point in a given direction (clockwise, for example). It is possible that the mirror image of such a tiling is impossible to achieve through rotation and translation alone. If we assume that such tilings are unique up to their mirror image (when such an image is different from the original tiling), there are exactly 11 families of Archimedean

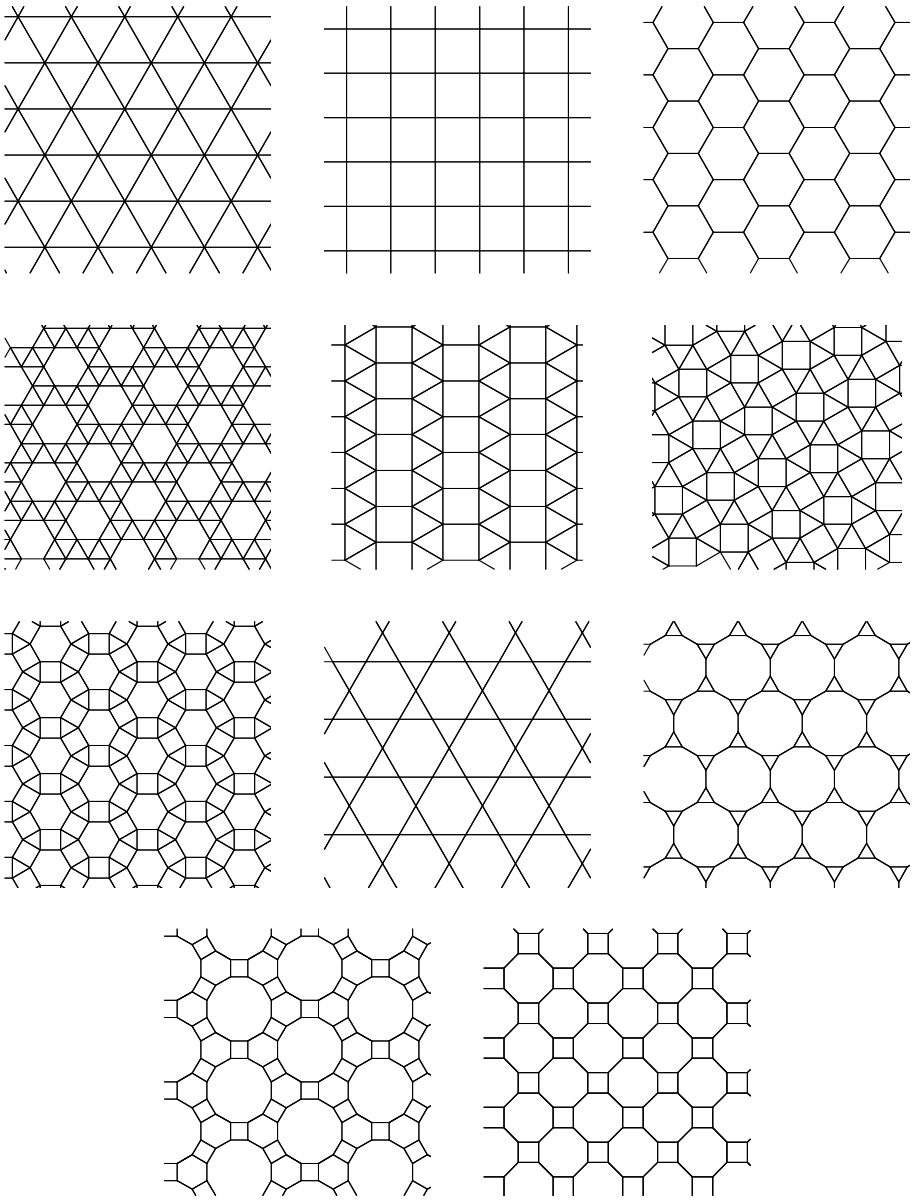


Fig. 2.13. Archimedean tilings. (See Exercise 13.)

tilings. The mirror image is distinct from the original tiling for exactly one of these tilings. Identify it.

14. A small challenge: classify the Archimedean tilings (see Exercise 13).

(a) Denote by n the regular polygon with n sides. Its internal angles are all equal to $\frac{(n-2)\pi}{n}$. (Prove this!) Consider an Archimidean tiling and let $(n_1, n_2, \ldots, n_m)$ be the list of the m polygons that meet at the vertices of this tiling. The sum of the angles at a given vertex must be 2π, and therefore

$$2\pi = \frac{(n_1-2)\pi}{n_1} + \frac{(n_2-2)\pi}{n_2} + \cdots + \frac{(n_m-2)\pi}{n_m}.$$

For example, for the Archimedean tiling of Figure 2.14, the polygons that meet at a vertex are enumerated by the list $(4, 3, 3, 4, 3)$, and as required, they satisfy

$$\frac{(4-2)\pi}{4} + \frac{(3-2)\pi}{3} + \frac{(3-2)\pi}{3} + \frac{(4-2)\pi}{4} + \frac{(3-2)\pi}{3} = 2\pi.$$

Enumerate all possible lists $(n_1, n_2, \ldots, n_m)$ of polygons that may meet at a vertex. Hint: there are 17 such lists if we distinguish between them using only their size, not the order of the n_i's.

(b) Why does the list $(5, 5, 10)$ not correspond to an Archimedean tiling of the plane?

(c) For each of the lists determined in (a), verify whether the set of polygons $(n_1, n_2, \ldots, n_m)$ meeting at a vertex actually describes a tiling of the plane. Caution: the order of the elements in the list $(n_1, n_2, \ldots, n_m)$ is important!

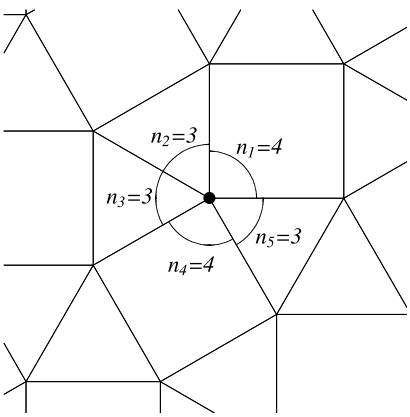


Fig. 2.14. A closer look at an Archimedean tiling (see Exercise 14). The list of polygons meeting at a vertex is denoted by $(4, 3, 3, 4, 3)$.

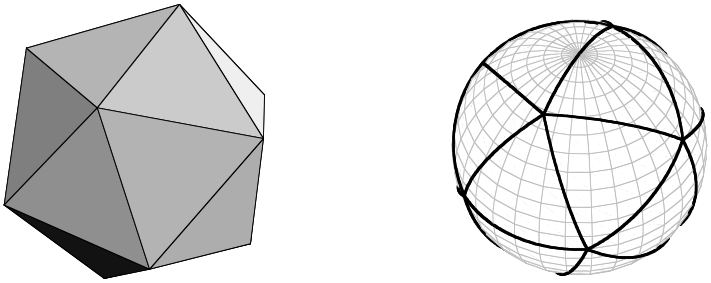


Fig. 2.15. An icosahedron and the corresponding tiling of the sphere (see Exercise 15).

15. A challenge: classify the Archimedean tilings of the sphere. In Section 15.8, we see that each regular polyhedron (the tetrahedron, the cube, the octahedron, the icosahedron, and the dodecahedron) corresponds to a regular tiling of the sphere. This correspondence is constructed as follows:

- the polyhedron is centered at the origin. The distance between the origin and each of the vertices is therefore the same, and we circumscribe a sphere with this radius that passes through all of the vertices;
- for every edge of the polyhedron, we join the vertices by an arc from the great circle between them.

The end result is the desired tiling of the sphere. Figure 2.15 shows such a construction for an icosahedron. The construction can be repeated for any polyhedron whose vertices all lie along the surface of a sphere. This is the case with Archimedean polyhedra: all of their faces are regular polygons with the same side length and all of their corners are incident to the same polygons. Even though regular polyhedra (also called Platonic polyhedra) meet these requirements, we reserve the adjective "Archimedean" for polyhedra whose faces consist of at least two different types of polygons. An example of an Archimedean polyhedron is the familiar shape of a soccer ball, formally called a *truncated icosahedron* (see Figure 2.16). Each vertex is shared by two hexagons and a pentagon. We denote it by the list $(5, 6, 6)$. Archimedean tilings of the sphere are classified as follows: prisms, antiprisms, and the 13 exceptional tilings. (Certain mathematicians prefer to exclude the prisms and antiprisms from the Archimedean tilings, and use the term to refer only to the 13 remaining tilings.)

(a) The list $(n_1, n_2, \ldots, n_m)$ of polygons meeting at a vertex must satisfy two simple conditions. In order for each vertex to be convex (and not planar), the sum of the internal angles meeting at the vertex must be less than 2π:

$$\pi \sum_{i=0}^{m} \frac{n_i - 2}{n_i} < 2\pi.$$

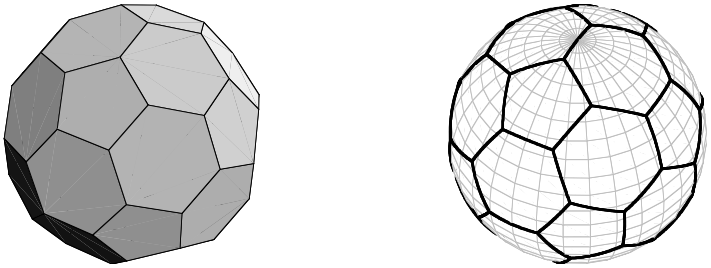


Fig. 2.16. A truncated icosahedron and the corresponding tiling of the sphere (see Exercise 15).

This is the first test. The second condition is based on Descartes's theorem. Each vertex of the polyhedron has associated with it an *angle deficiency* defined as $\Delta = 2\pi - \pi \sum_i (n_i - 2)/n_i$. Descartes's theorem states that the sum of the deficiencies across all vertices of a polyhedron must be equal to 4π. Since all vertices of an Archimedean solid are identical, we must therefore have that $4\pi/\Delta$ is an integer, equal to the number of vertices. This is the second test. Verify that the soccer ball satisfies both of these conditions. (We will see in (d) that these two tests alone are not sufficient to characterize the Archimedean solids.)

(b) A prism is a polyhedron consisting of two identical polygonal faces that are parallel. Each edge of these two faces is then connected by a square. They form an infinite family of solids denoted by $(4, 4, n)$, for $n \geq 3$. Convince yourself that all of the vertices of such a solid are identical and accurately described by the list $(4, 4, n)$. Draw an example of such a prism, for example $(4, 4, 5)$. Verify that the list $(4, 4, n)$ passes both of the tests described in (a) regardless of n. (When n is sufficiently large, these solids begin to resemble stout cylinders.)

(c) An antiprism also consists of two parallel identical polygons with n faces ($n \geq 4$). However, one of the faces is rotated with respect to the other by an angle of $\frac{\pi}{n}$ and the corners joined by equilateral triangles. The antiprisms form an infinite family of solids and are denoted by the list $(3, 3, 3, n)$ for $n \geq 4$. Answer the same questions as for prisms.

(d) Show that the list $(3, 4, 12)$ passes both of the tests described in (a). However, it is impossible to construct a regular polyhedron based on this list. Why? Hint: start by assembling a triangle, a square, and a polygon with twelve sides (a dodecagon) around a single vertex. Consider the other vertices of these three faces. Is it possible for these vertices to have the same configuration described by the list $(3, 4, 12)$? (This is the hardest part of this question!)

(e) Show that there exist 13 Archimedean tilings of the sphere (or, equivalently, 13 Archimedean polyhedra) that are neither prisms nor antiprisms. (The soccer ball is one of these 13 solids.)

16. A difficult challenge: derive the crystallographic groups (shown in Figures 2.17–2.22).

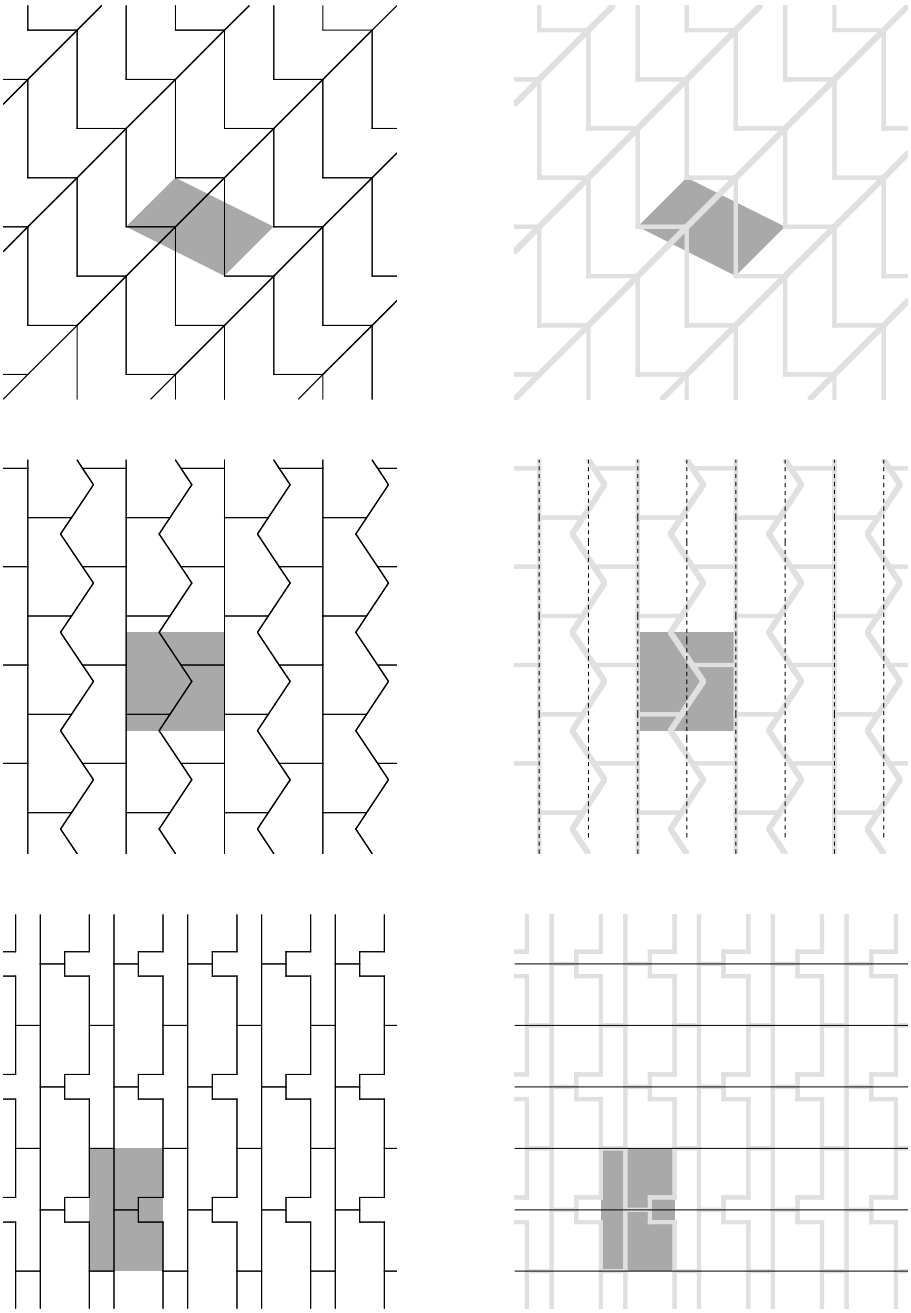


Fig. 2.17. The 17 crystallographic groups. From top to bottom: the groups *p1, pg, pm.*

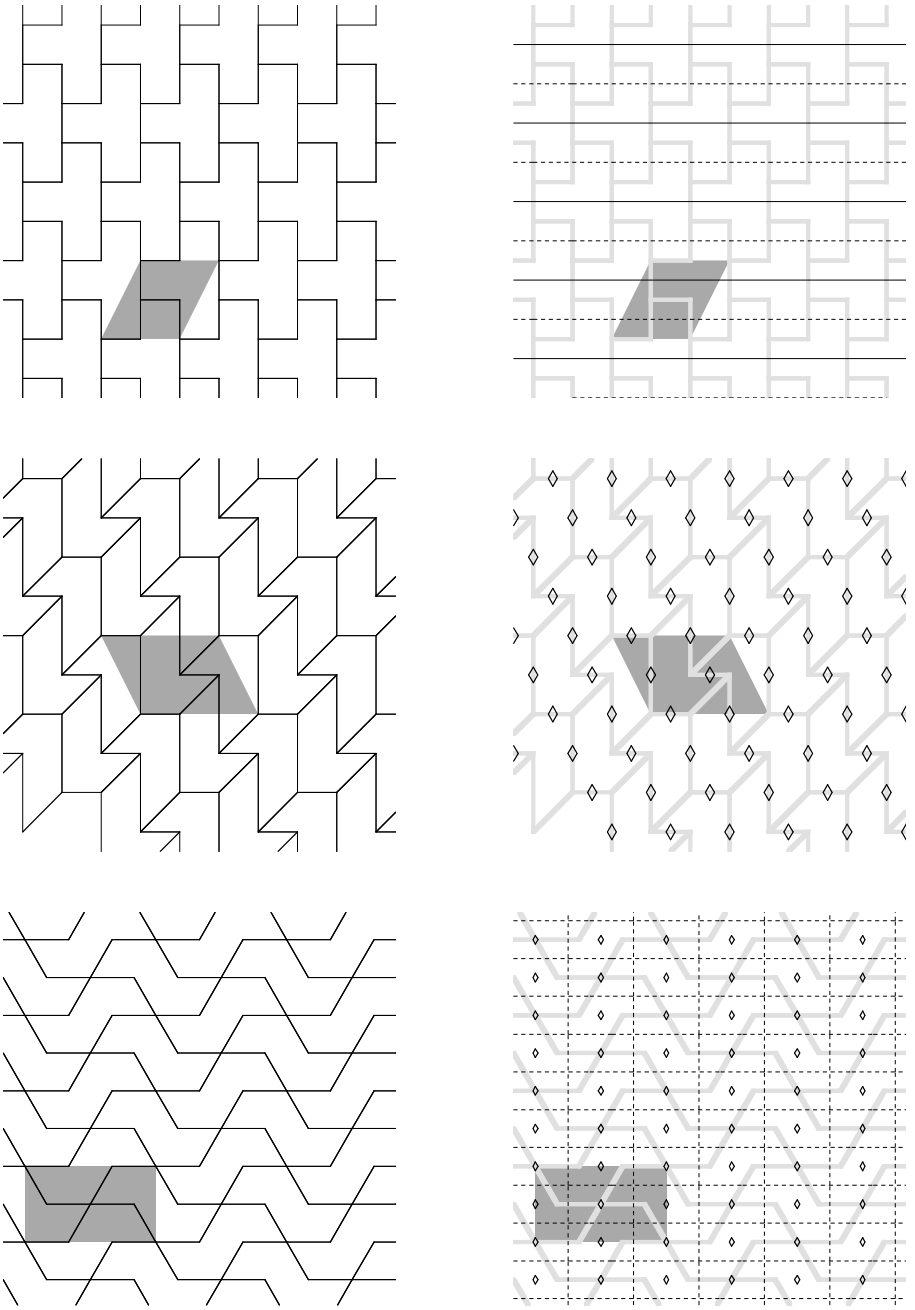


Fig. 2.18. The 17 crystallographic groups (continued). From top to bottom: the groups *cm*, *p2*, *pgg*.

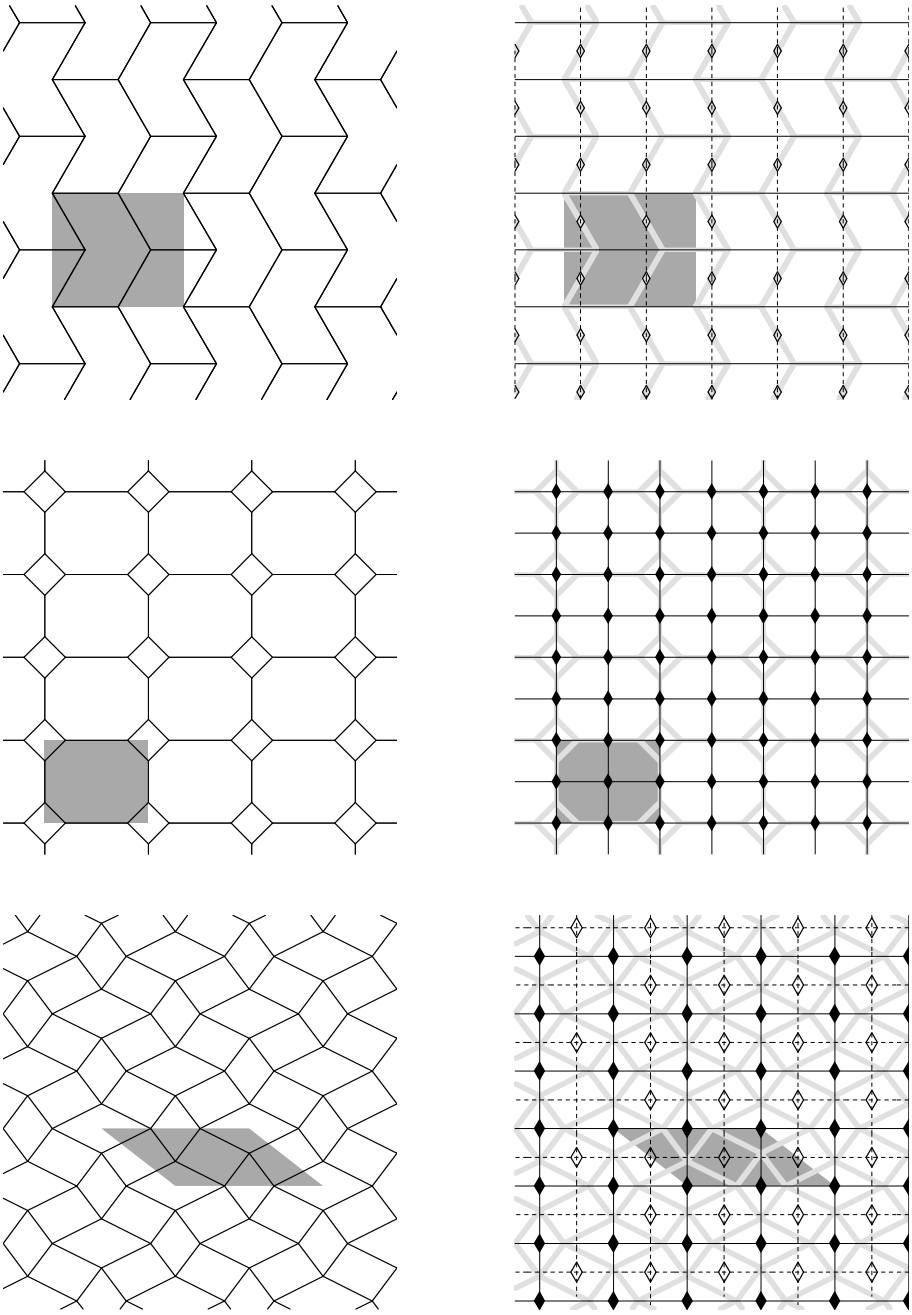


Fig. 2.19. The 17 crystallographic groups (continued). From top to bottom: the groups *pmg, pmm, cmm.*

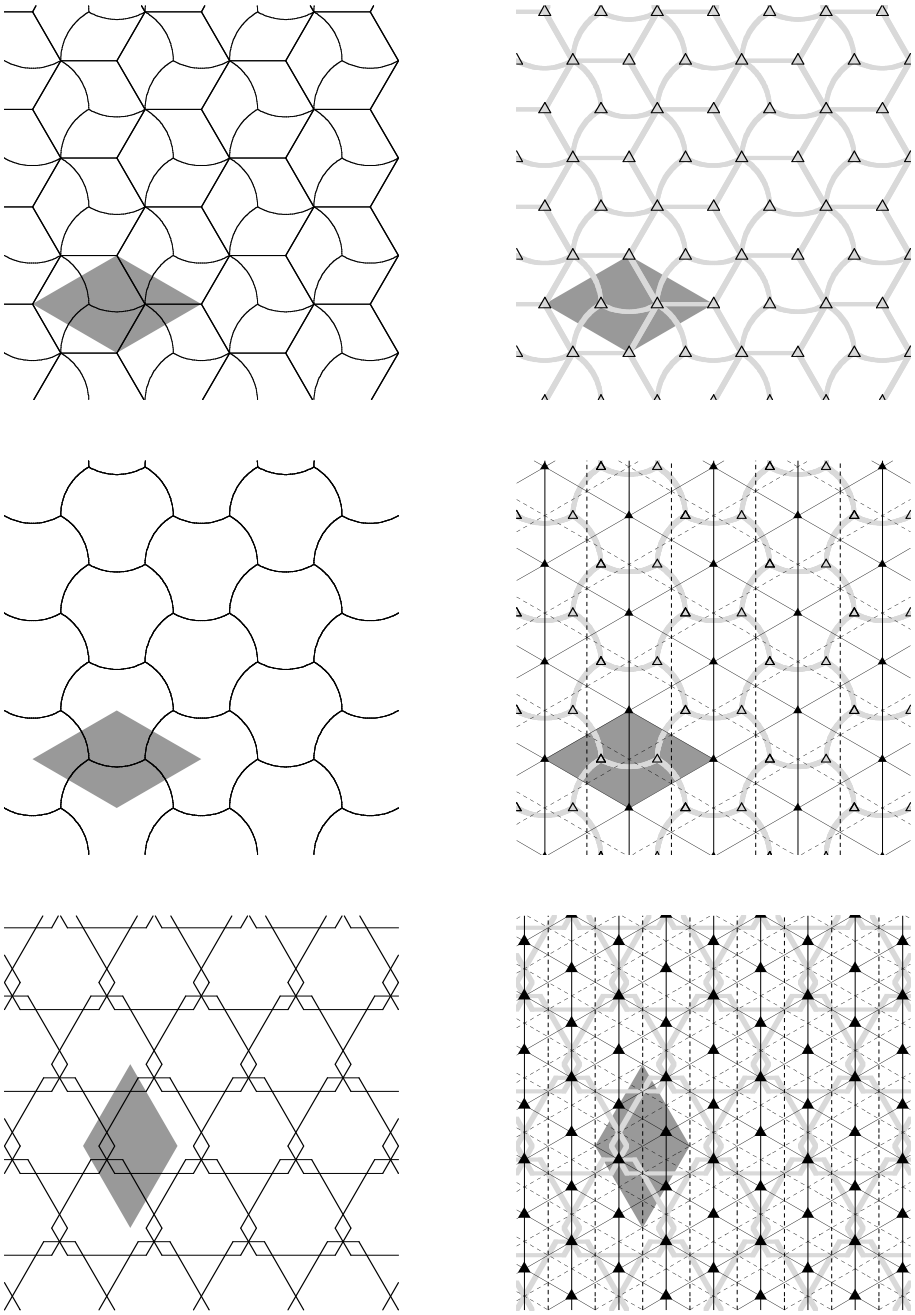


Fig. 2.20. The 17 crystallographic groups (continued). From top to bottom: the groups *p3, p31m, p3m1.*

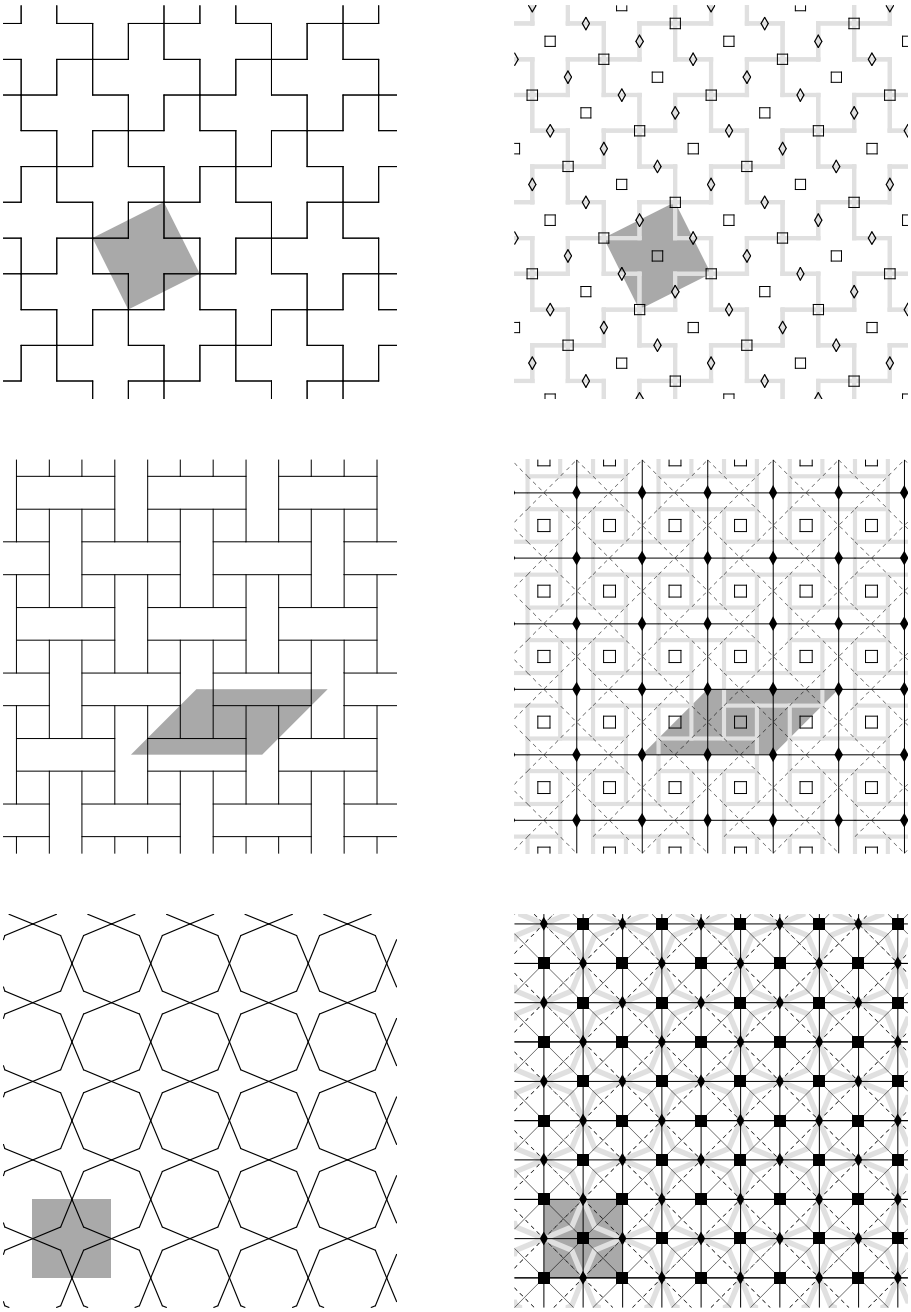


Fig. 2.21. The 17 crystallographic groups (continued). From top to bottom: the groups *p4*, *p4g*, *p4m*.

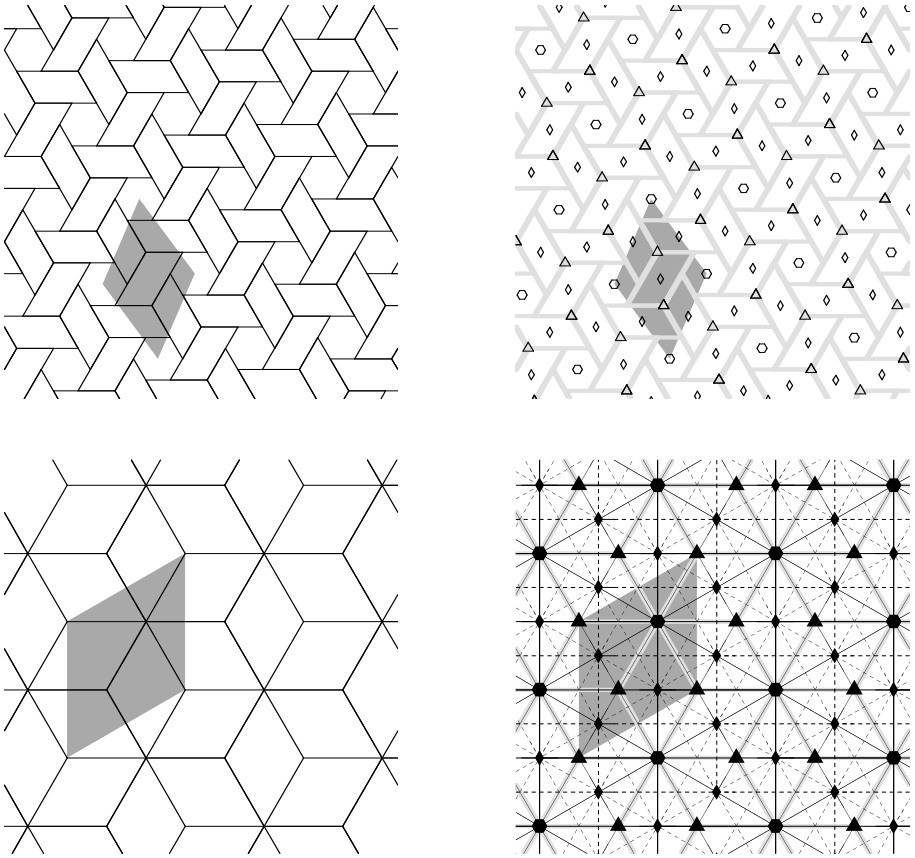


Fig. 2.22. The 17 crystallographic groups (continued). From top to bottom: the groups *p6*, *p6m*.

References

[1] A. Bravais. Mémoire sur les systèmes formés par des points distribués régulièrement sur un plan ou dans l'espace. *Journal de l'École Polytechnique*, 19:1–128, 1850.

[2] H.S.M. Coxeter. *Introduction to Geometry.* Wiley, New York, 1969.

[3] E. Prisse d'Avennes, editor. *Arabic Art in Color.* Dover, 1978. (This book gathers some excerpts of Prisse d'Avennes's monumental work "L'art arabe d'après les monuments du Kaire depuis le VIIe siècle jusqu'à la fin du XVIIe siècle." He assembled this collection between 1869 and 1877. The work was originally published in 1877 in Paris by Morel.)

[4] B. Grünbaum, Z. Grünbaum, and G.C. Shephard. Symmetry in Moorish and other ornaments. *Computers and Mathematics with Applications*, 12:641–653, 1985.

[5] B. Grünbaum and G.C. Shephard. *Tilings and Patterns.* W.H. Freeman, New York, 1987.

[6] D. Schattschneider. *Visions of Symmetry: Notebooks, Periodic Drawings, and Related Work of M.C. Escher.* W.H. Freeman, New York, 1992.

3

Robotic Motion

This chapter can be covered in a week of classes. The first hour is spent describing the robot of Figure 3.1. It is important to make sure that the concept of the "dimension" (number of degrees of freedom) of the problem is well understood by walking through several simple examples. After this, rotations in three-space are presented with their representations as orthogonal matrices by stating and discussing the principal results of Section 3.3. The last hour is devoted to presenting the seven frames of reference associated with the robot of Figure 3.1, and calculating the positions of the various articulations in each frame of reference (see Section 3.5). Since this discussion requires a full hour, it is not possible to cover the entire discussion on orthogonal transformations, nor all of the details of the fundamental theorem (Theorem 3.20), which states that all orthogonal transformations in $\mathbb{R}^3$ *with determinant* 1 *are rotations. So the principal results are only stated and briefly illustrated. The important lesson about orthogonal transformations is that* choosing an appropriate basis facilitates comprehension and visualization of the transformation. *The exact discussion of orthogonal transformations depends on the students' prior experience with linear algebra. It is possible to simply work through a few examples, or instead to choose to work through several proofs.*

3.1 Introduction

Consider the three-dimensional robot in Figure 3.1. It consists of three articulated joints and a claw. On the figure we have indicated six rotations that the robot can perform, numbered 1 through 6. The robot is attached to a wall, with the first segment perpendicular to it. This segment is not fixed, however, and is free to rotate around its central axis as shown by movement 1. At the end of the first segment there is a second segment. The joint between the two segments is similar to an elbow in that its motion is constrained to a plane (as shown by motion 2). However, if we combine this allowed rotation with that of 1, we see that the rotational plane of 2 itself rotates along with

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_3,

the first segment. Thus, the composition of these two rotations allows us to position the second segment in any possible direction. Now consider the third segment. Rotation 3 allows the segment to pivot in a plane (as in rotation 2), while rotation 4 allows the segment to rotate about its axis. This segment can be compared to a shoulder: we can lift our arm (which is equivalent to rotation 3) and we can turn our arm about its axis (which is equivalent to rotation 4). (In reality, a shoulder is not constrained to lifting the arm within a single plane, thus it has yet another degree of freedom as compared to this segment, since we can turn our arm around our body while keeping a fixed angle with the vertical.) Finally, there is a claw attached to the end of the third segment. The claw also has two associated rotations: rotation 5 acts in a plane and varies the angle between the third segment and the claw, while rotation 6 allows the claw to rotate around its axis.

Why was this robot built with six rotational movements? We will see that this was no accident and that if it had even one fewer possible rotation, the robot's movements would be severely limited.

We start with a simple example that considers translations:

Example 3.1 *Let $P = (x_0, y_0, z_0)$ be a point of departure in $\mathbb{R}^3$. We wish to determine which positions Q we can reach if we permit translations along the unit directions $v_1 = (a_1, b_1, c_1)$ and $v_2 = (a_2, b_2, c_2)$. The set of points that may be reached is*

$$\{Q = P + t_1 v_1 + t_2 v_2 \mid t_1, t_2 \in \mathbb{R}\}.$$

This set describes a plane passing through P as long as $v_1 \neq \pm v_2$. (Exercise: prove this!)

If we add a third unit direction v_3 such that $\{v_1, v_2, v_3\}$ are linearly independent, then the set of positions Q that may be reached is the entire space $\mathbb{R}^3$.

Why did we require three translational directions to make the entire space reachable? Because the dimension of the space is three, as evidenced by the fact that we require three coordinates to specify a position in $\mathbb{R}^3$. We say that the problem has three degrees of freedom.

Try adapting this approach to our robot: how many numbers are required to fully describe its exact position? For a worker using the robot to grab an object, precisely positioning the claw is of primary importance. This worker specifies:

- the position of P: it is defined by the three coordinates (x, y, z) of P in space.
- the direction of the axis of the claw. A direction can be specified by a vector, so it looks as if three numbers should be necessary. However, there exist an infinite number of vectors that point in the same direction. Thus, a more efficient manner of providing a direction is to imagine a unit sphere centered at P and indicating a point Q on the surface of the sphere. The ray originating at P and passing through Q specifies a unique direction. If we give ourselves a direction, that is, a ray emanating from P, this will intersect the sphere at exactly one point. Thus, there is a bijection

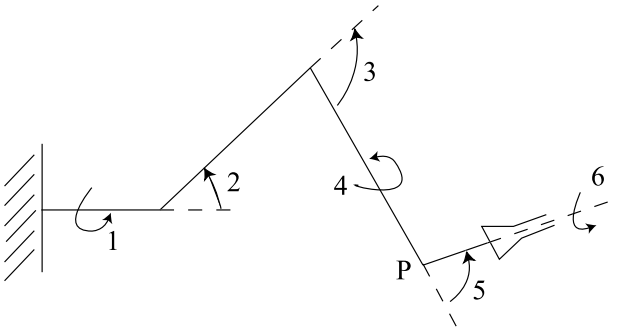


Fig. 3.1. A three-dimensional robot with six degrees of freedom.

between the points on the surface of the sphere and the directions. Specifying a point on the sphere is therefore sufficient to uniquely identify a direction. This can be done most efficiently using spherical coordinates. The points on a sphere of radius 1 are

$$(a, b, c) = (\cos\theta\cos\phi, \sin\theta\cos\phi, \sin\phi),$$

with $\theta \in [0, 2\pi)$ and $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Thus the two numbers θ and ϕ are sufficient to describe the direction of the claw.

- The claw can pivot around its axis by a rotation, the angle of which is specified by a single parameter α.

In total we required six numbers $(x, y, z, \theta, \phi, \alpha)$ in order to specify the position and orientation of the claw. Analogous to Example 3.1, we say that the robot of Figure 3.1 has six *degrees of freedom*. The rotations 1, 2, and 3 are used to place P at the desired position (x, y, z). Rotations 4 and 5 are used to correctly orient the axis of the claw, while rotation 6 rotates the claw to the desired angle about its axis. These six movements correspond to the six degrees of freedom.

Consider the difference between the point Q of Example 3.1 and the claw of our robot. We required only three numbers to specify the position of Q, while we required six to specify the position of the claw. The claw is an example of what is called a "solid" object in $\mathbb{R}^3$, and we will see that we always require six numbers to specify the position of a solid in space. To develop our intuition we will begin by considering a solid in the plane.

3.1.1 Moving a Solid in the Plane

Consider cutting out a triangle from cardboard in such a manner that none of the three angles are the same (and therefore the triangle has no symmetry).

Assume that the triangle is not able to be deformed and that it must rest firmly in the plane; then it is capable only of sliding in the plane. We wish to describe all

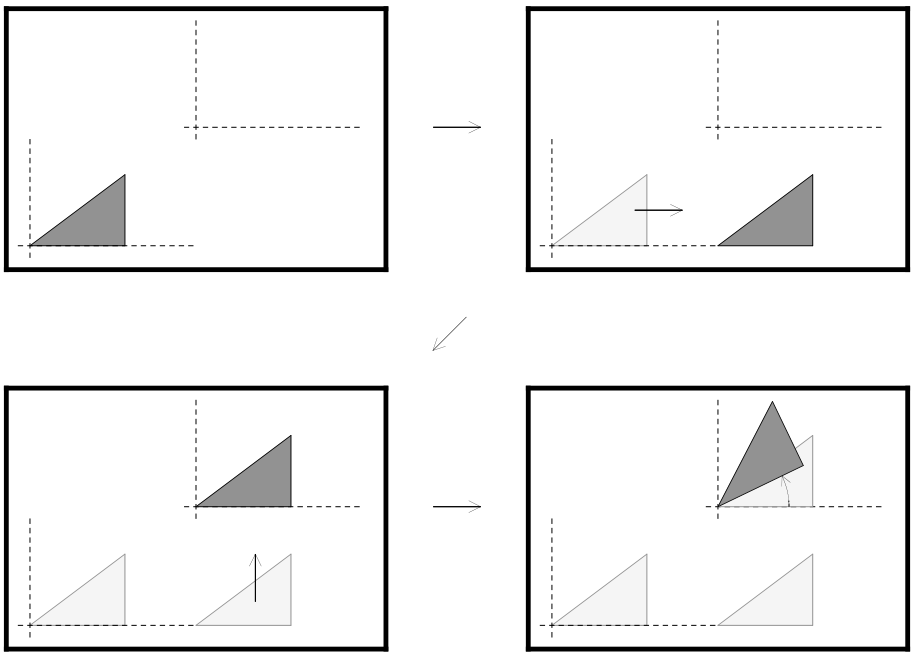


Fig. 3.2. Moving a solid in the plane.

possible positions that the triangle may take (see Figure 3.2). To do this we will choose any one of the corners of the triangle and label it A (but we could have made the same reasoning with any other point).

- We start by specifying the position of A. This requires the two coordinates (x, y) of A in the plane.
- Next we specify the orientation of the triangle with respect to the point A. If A is fixed then the only possible movement of the triangle is rotation about A. If B is a second corner of the triangle then the position of the triangle is determined by the angle α made between the vector $\overrightarrow{AB}$ and some fixed direction, for example the horizontal ray extending to the right from A.

Thus we require three numbers (x, y, α) to fully specify the position of the triangle (and any other asymmetric solid) in the plane.

Consider Figure 3.2 and suppose that we start with A situated at the origin and the vector $\overrightarrow{AB}$ pointing horizontally to the right. To move the triangle to position (x, y, α) we can translate by $(x, 0)$ in the direction $e_1 = (1, 0)$, then translate by $(0, y)$ in the direction $e_2 = (0, 1)$, and finally rotate by an angle of α about (x, y).

We made an equivalence between the numbers (x, y, α) determining the position of the triangle and the movements that bring the triangle to this position from a position of $(0, 0, 0)$. We state the following theorem without proof:

Theorem 3.2 *Movements of a solid in the plane are compositions of translations and rotations. These are movements that preserve lengths, angles, and orientation.*

Example 3.3 *Imagine a robot that is able to realize the motions we just described. Such a robot is shown in Figure 3.3. At the end of the second segment there is a claw that can be rotated perpendicular to the plane of motion of the robot. If a triangle is*

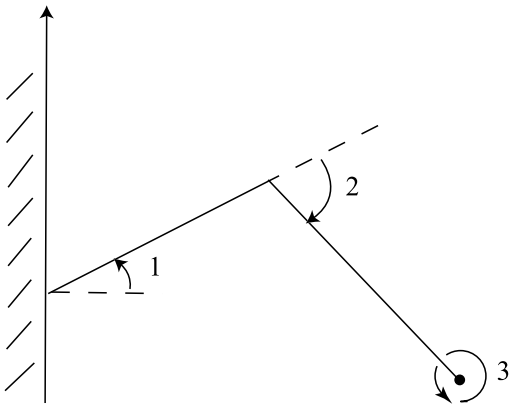


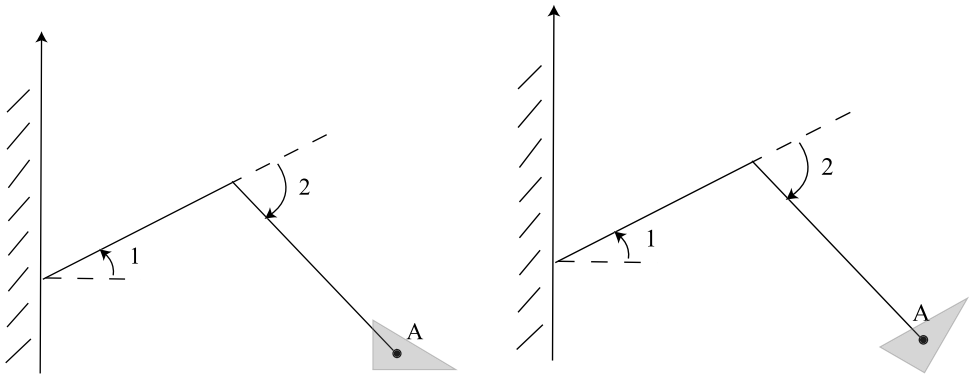
Fig. 3.3. A robot in the plane.

attached to the claw by the corner labeled A, then rotation of the claw will correspond to rotation of the triangle about A (see Figure 3.4). What are the positions that may be reached by the extreme end of the second segment? It is obvious that we cannot reach all of the points in the plane, because we are limited by both the length of the arms and the presence of the wall. But we can reach many positions, described by a 2-dimensional subset of the plane. If the robot had only a single segment we would be limited to a 1-dimensional subset of the plane, specifically an arc of a circle. Finding the exact set of positions reachable by A is the goal of Exercise 13.

This example illustrates that three degrees of freedom are required to move a solid through the plane and demonstrates a robot capable of realizing these motions.

3.1.2 Some Thoughts on the Number of Degrees of Freedom

There are many ways to build a robot in three-space, but *six degrees of freedom* (and thus at least six independent motions) are necessary in order to reach every possible



(a) Initial position (b) Position after rotating the claw by $\frac{\pi}{3}$

Fig. 3.4. Movement along the third degree of freedom of the robot from Figure 3.3.

position with every possible orientation. Thus, six degrees of freedom are also required in the control system that manipulates the robot.

One can imagine adding additional segments to the robotic arm and even installing it on a track. This will possibly enlarge the size and alter the shape of the region that can be reached, but it will not change its "dimension." Such modifications may offer other advantages, which will be discussed a little later.

On the other hand, one can also consider building a robot with only five degrees of freedom. Regardless of how these independent motions are realized and connected, there will always be certain positions or orientations of the claw that are unattainable. In fact, there will be only a small set of reachable positions as compared to an overwhelming majority of unreachable ones.

The robot of Figure 3.1 uses only rotations. These rotations can easily be replaced by other movements such as translation along a track or telescoping arms (segments whose length can alter). Try to think of a few other robotic arms with six degrees of freedom.

The underlying mathematics: If we wish to describe the movements of a robot we must discuss the motion of a solid in $\mathbb{R}^3$. As in the plane, these movements will be compositions of translations and rotations. In general, different rotations will have distinct rotational axes.

- If we choose a coordinate system whose origin is along the axis of rotation, then the rotation is a linear transformation in this frame of reference. Its matrix is simpler if the axis of rotation is one of the coordinate axes.
- Since the rotational axes are distinct, we will need to consider coordinate system changes. If we know the coordinates of a point Q in one coordinate system, such

mappings allow us to calculate the coordinates of the same point in a new coordinate system.

- Considering our example in Figure 3.1, these transformations will allow us to calculate the final position of the claw after applying the rotations $R_i(\theta_i)$ by angles θ_i, for $i \in \{1, 2, 3, 4, 5, 6\}$.

3.2 Movements That Preserve Distances and Angles in the Plane or Space

We begin by considering linear transformations that preserve distances and angles: these are precisely those transformations whose matrices are orthogonal, and they are called *orthogonal transformations*. A rotation about an axis passing through the origin will be of this type.

We will briefly review linear transformations. Although we will initially discuss linear transformations on $\mathbb{R}^n$, we will ultimately focus on the cases $n = 2$ and $n = 3$ that are applicable in practice. Let us start with some notation.

Notation: We will distinguish between the vectors of $\mathbb{R}^n$ that are geometric objects and will be denoted by $v, w, \ldots$ and the column matrices $n \times 1$, which represent their coordinates in the standard basis $\mathcal{C} = \{e_1, \ldots, e_n\}$ of $\mathbb{R}^n$, where

$$\begin{array}{rcl} e_1 & = & (1, 0, \ldots, 0) \\ e_2 & = & (0, 1, 0, \ldots, 0), \\ \vdots & & \\ e_n & = & (0, \ldots, 0, 1). \end{array} \tag{3.1}$$

We will denote the column matrix of coordinates of v by $[v]$ or $[v]_{\mathcal{C}}$. We make this distinction because we will later consider changes of bases.

Theorem 3.4 *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation, in other words one that satisfies the following properties:*

$$\begin{array}{rcll} T(v + w) & = & T(v) + T(w), & \forall v, w \in \mathbb{R}^n, \\ T(\alpha v) & = & \alpha T(v), & \forall v \in \mathbb{R}^n,\ \forall \alpha \in \mathbb{R}. \end{array} \tag{3.2}$$

1. *There exists a unique $n \times n$ matrix A such that the coordinates of $T(v)$ are given by $A[v]$ for all $v \in \mathbb{R}^n$:*
$$[T(v)] = A[v]. \tag{3.3}$$
2. *The transformation matrix A is constructed such that the columns of A are the images of the vectors of the standard basis of $\mathbb{R}^n$.*

PROOF: We begin by proving the second part. Calculate $[T(e_1)]$,

$$[T(e_1)] = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & \alpha_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix},$$

and repeat this for each vector in the standard basis.

For the first part, the matrix A is the matrix whose columns contain the coordinates of $T(e_i)$ expressed in the standard basis. It clearly satisfies (3.3).

The fact that the column vectors of the matrix contain the coordinates of $T(e_i)$ in the standard basis guarantees the uniqueness of the matrix A. □

Definition 3.5 *1. Let $A = (a_{ij})$ be an $n \times n$ matrix. The transpose of A is the matrix $A^t = (b_{ij})$, where*

$$b_{ij} = a_{ji}.$$

2. A matrix A is orthogonal if its inverse is equal to its transpose, in other words, if $A^t = A^{-1}$ or equivalently

$$AA^t = A^tA = I,$$

where I is the $n \times n$ identity matrix.

3. A linear transformation is orthogonal if its matrix in the standard basis is orthogonal.

Definition 3.6 *The scalar product of two vectors $v = (x_1, \dots, x_n)$ and $w = (y_1, \dots, y_n)$ is*

$$\langle v, w \rangle = x_1y_1 + \cdots + x_ny_n.$$

We recall without proof the following classical proposition.

Proposition 3.7 *1. If A is an $m \times n$ matrix and B is an $n \times p$ matrix, then*

$$(AB)^t = B^tA^t.$$

2. The scalar product of two vectors v and w can be calculated as

$$\langle v, w \rangle = [v]^t[w].$$

Theorem 3.8 *1. A matrix is orthogonal if and only if its columns form an orthonormal basis of $\mathbb{R}^n$.*

2. A linear transformation preserves distances and angles if and only if its matrix is orthogonal.

PROOF:

1. Let us remark that the columns of A are given by $X_i = A[e_i]$, $i = 1, \ldots, n$, where the X_i are $n \times 1$ matrices. We write

$$A = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}.$$

Then the transposes $X_1^t, \ldots, X_n^t$ are horizontal $1 \times n$ matrices. If we represent the matrix A^t by its rows, then it has the form

$$A^t = \begin{pmatrix} X_1^t \\ \vdots \\ X_n^t \end{pmatrix}.$$

We calculate the matrix product $A^t A$ using this notation:

$$A^t A = \begin{pmatrix} X_1^t \\ \vdots \\ X_n^t \end{pmatrix} \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix} = \begin{pmatrix} X_1^t X_1 & X_1^t X_2 & \cdots & X_1^t X_n \\ X_2^t X_1 & X_2^t X_2 & \cdots & X_2^t X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_n^t X_1 & X_n^t X_2 & \ldots & X_n^t X_n \end{pmatrix}.$$

Let T be the linear transformation with matrix A. We have

$$X_i^t X_j = (A[e_i])^t A[e_j] = [T(e_i)]^t [T(e_j)] = \langle T(e_i), T(e_j) \rangle.$$

The matrix A is orthogonal if and only if the matrix $A^t A$ is equal to the identity matrix. Saying that the entries on the diagonal are equal to 1 is equivalent to saying that the scalar product of each vector $T(e_i)$ with itself is equal to 1. Since the scalar product is equal to the square of the length of the vector, this is equivalent to saying that they have length 1. So the entries on the diagonal are equal to 1 if and only if all vectors $T(e_i)$ have length 1. All entries not on the diagonal are zero if and only if the scalar product of $T(e_i)$ with $T(e_j)$ is zero when $i \neq j$. Hence, the matrix A is orthogonal if and only if the vectors $T(e_1), \ldots, T(e_n)$ are orthogonal and each has length 1, thus forming an orthonormal basis of $\mathbb{R}^n$.

2. We start by proving the reverse direction, which asserts that if T is a linear transformation with an orthogonal matrix, then T preserves distances and angles. According to the proof of the first part, the images of the vectors of the standard basis (which are the columns of A) form an orthonormal basis. Thus their lengths are preserved as well as the angles between them. We can easily convince ourselves that a linear transformation preserves distances and angles if and only if it preserves scalar products, in other words, if $\langle T(v), T(w) \rangle = \langle v, w \rangle$ for all v, w. Let v, w be two vectors. Observe that their scalar product is preserved if A is orthogonal:

$$\begin{aligned}\langle T(v), T(w)\rangle &= (A[v])^t(A[w]) \\ &= ([v]^t A^t)(A[w]) \\ &= [v]^t(A^t A)[w] \\ &= [v]^t I[w] \\ &= [v]^t[w] \\ &= \langle v, w\rangle.\end{aligned}$$

The other direction makes the hypothesis that T preserves distances and angles. Suppose that $A^tA = (b_{ij})$. Let $v = e_i$ and $w = e_j$. We have $[T(v)] = A[v]$ and $[T(w)] = A[w]$. Then

$$\langle T(v), T(w)\rangle = ([v]^t(A^tA))[w] = (b_{i1}, \ldots, b_{in})[w] = b_{ij}.$$

Moreover, $[v]^t[w] = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j. \end{cases}$$

Thus, $\forall i, j, b_{ij} = \delta_{ij}$, which is equivalent to saying that $A^tA = I$. Hence A is orthogonal. □

Theorem 3.9 *The movements that preserve both distances and angles in $\mathbb{R}^n$ are the compositions of translations and orthogonal transformations. (These movements are called the isometries of $\mathbb{R}^n$.)*

Proof: Consider a movement $F : \mathbb{R}^n \to \mathbb{R}^n$ that preserves distances and angles. Let $F(0) = Q$ and let T be the translation $T(v) = v - Q$. Then $T(Q) = 0$ and therefore $T \circ F(0) = 0$. Let $G = T \circ F$. This is a transformation that preserves distances and angles and that has a fixed point at the origin. If G is to preserve distances and angles it must be linear (for a proof of this fact see Exercise 4), and by the previous theorem it must be an orthogonal transformation. We have also that $F = T^{-1} \circ G$. Since T^{-1} is also a translation, then F has been shown to be the composition of an orthogonal transformation and a translation. □

3.3 Properties of Orthogonal Matrices

Consider the following orthogonal matrix:

$$A = \begin{pmatrix} 1/3 & 2/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \end{pmatrix}. \tag{3.4}$$

Can we describe in geometric terms the orthogonal transformation with matrix A? Looking at this matrix it is rather hard to visualize the action of T on $\mathbb{R}^3$. We know only that it is orthogonal and that it therefore preserves angles and distances. How can we determine the geometry of T? An extremely useful tool for exploring the behavior of T is the technique of *diagonalization*. When we diagonalize a matrix we are in fact changing the coordinate system of the linear transformation. We place ourselves in a coordinate system for which the coefficients of the transformation matrix are extremely simple and the behavior of the transformation is easily understood. Before doing the calculations for this matrix we will recall the relevant definitions.

Definition 3.10 *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation with matrix A. A number $\lambda \in \mathbb{C}$ is an eigenvalue of T (or of A) if there exists a nonzero vector $v \in \mathbb{C}^n$ such that $T(v) = \lambda v$. Any vector v with this property is called an eigenvector of the eigenvalue λ.*

Remarks.

1. In the context of orthogonal transformations it is essential to look at complex eigenvalues. Indeed, when we have a real eigenvector v of a real nonzero eigenvalue λ then the set E of multiples of v forms a subspace of dimension 1 (a line) of $\mathbb{R}^n$ that is invariant by T, thus satisfying $T(E) = E$. Let us consider a rotation in $\mathbb{R}^2$. Obviously there is no invariant line. Hence the eigenvalues and their associated eigenvectors are complex.
2. How do we calculate $T(v)$ if $v \in \mathbb{C}^n$? The standard basis (3.1) is also a basis of $\mathbb{C}^n$. So the following definition makes sense $[T(v)] = A[v]$, yielding that $T(v)$ is the vector of $\mathbb{C}^n$ whose coordinates in the standard basis of $\mathbb{C}^n$ are given by $A[v]$.
3. Consider in $\mathbb{R}^3$ a rotation about an axis: it is an orthogonal transformation whose axis of rotation is an invariant line. So we will find this axis when we will diagonalize the transformation.

We state without proof the following theorem

Theorem 3.11 *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation with matrix A.*

1. *The set of eigenvectors of the eigenvalue λ is a linear subspace of $\mathbb{R}^n$, called the eigenspace of the eigenvalue λ.*
2. *The eigenvalues are the roots of the polynomial*

$$P(\lambda) = \det(\lambda I - A)$$

of degree n. The polynomial $P(\lambda)$ is called the characteristic polynomial *of T (or of A).*
3. *Let $v \in \mathbb{R}^n \setminus \{0\}$. Then v is an eigenvector of λ if and only if $[v]$ is a solution of the homogeneous system of linear equations:*

$$(\lambda I - A)[v] = 0.$$

Example 3.12 *Let T be the orthogonal transformation with matrix A given in* (3.4)*. To diagonalize A we begin by calculating its characteristic polynomial*

$$P(\lambda) = \det(\lambda I - A) = \begin{vmatrix} \lambda - 1/3 & -2/3 & -2/3 \\ -2/3 & \lambda + 2/3 & -1/3 \\ -2/3 & -1/3 & \lambda + 2/3 \end{vmatrix}.$$

We have

$$P(\lambda) = \lambda^3 + \lambda^2 - \lambda - 1 = (\lambda + 1)^2(\lambda - 1).$$

The matrix therefore has the two eigenvalues 1 *and* -1*.*

Eigenvectors of $+1$: *To find them we need to solve the system $(I - A)[v] = 0$. So we transform the matrix $I - A$ into echelon form using Gaussian elimination:*

$$\begin{aligned} I - A &= \begin{pmatrix} 2/3 & -2/3 & -2/3 \\ -2/3 & 5/3 & -1/3 \\ -2/3 & -1/3 & 5/3 \end{pmatrix} \sim \begin{pmatrix} 2/3 & -2/3 & -2/3 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \\ &\sim \begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned}$$

All solutions are multiples of the eigenvector $v_1 = (2, 1, 1)$.

Eigenvectors of -1: *These are the solutions to the system $(-I - A)[v] = 0$, which is equivalent to the system $(I + A)[v] = 0$. To find them we reduce the matrix to echelon form, yielding*

$$I + A = \begin{pmatrix} 4/3 & 2/3 & 2/3 \\ 2/3 & 1/3 & 1/3 \\ 2/3 & 1/3 & 1/3 \end{pmatrix} \sim \begin{pmatrix} 1 & 1/2 & 1/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Here the set of solutions describes a plane. It is generated by the two vectors $v_2 = (1, -2, 0)$ and $v_3 = (1, 0, -2)$.

It is useful to work with an orthonormal basis. Thus, in general we will replace v_3 by a vector $v_3' = (x, y, z)$ that is perpendicular to v_2 but still lies within the plane generated by the two vectors. It must therefore satisfy $2x + y + z = 0$ in order to be an eigenvector of -1, and it must be perpendicular to v_2, meaning it must satisfy $x - 2y = 0$. We can take $v_3' = (-2, -1, 5)$ which is a solution to the system

$$\begin{aligned} 2x + y + z &= 0, \\ x - 2y &= 0. \end{aligned}$$

To make this an orthonormal basis we normalize each vector by dividing it by its length. This yields the orthonormal basis

$$\mathcal{B} = \left\{ \begin{array}{l} w_1 = \left(\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right), w_2 = \left(\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}, 0\right), \\ w_3 = \left(-\frac{2}{\sqrt{30}}, -\frac{1}{\sqrt{30}}, \frac{5}{\sqrt{30}}\right) \end{array} \right\}.$$

In this basis the matrix of the transformation T is given by

$$[T]_{\mathcal{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Geometrically we have that $T(w_1) = w_1$, $T(w_2) = -w_2$, and $T(w_3) = -w_3$. We see that this transformation consists of reflection across the w_1 axis; equivalently, this can be viewed as a rotation of angle π about the w_1 axis. We have now seen how diagonalization allows us to "understand" the transformation.

A few comments on Example 3.12: The two eigenvalues 1 and -1 each have unit absolute values. This is no coincidence, since orthogonal transformations preserve distances, meaning that we could never have $T(v) = \lambda v$ for $|\lambda| \neq 1$. Moreover, all of the eigenvectors associated with eigenvalue -1 are orthogonal to those associated with eigenvalue 1. This is also no coincidence. We will discuss the properties of diagonalizations of orthogonal matrices a little later.

As mentioned, the eigenvalues of an orthogonal transformation are not necessarily real, as shown in the following example.

Example 3.13 *The matrix*

$$B = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

describing a transformation T is orthogonal (exercise!). It represents a rotation of $\frac{\pi}{2}$ about the z axis: this can be verified by looking at the images of the three vectors of the standard basis:

$$T\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad T\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \qquad T\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Under the action of T we see that the third vector e_3 remains fixed, while the two vectors e_1 and e_2 have both rotated by an angle of $\frac{\pi}{2}$ in the plane (x, y). The characteristic polynomial of B is

$$\det(\lambda I - B) = (\lambda^2 + 1)(\lambda - 1),$$

which has as roots 1, i, and $-i$. The two complex eigenvalues i and $-i$ are conjugates of one another and both have a modulus of 1.

We recall without proof the following proposition.

Proposition 3.14 *1. Let A be an $n \times n$ matrix. Then*

$$\det A^t = \det A.$$

2. Let A and B be two $n \times n$ matrices. Then

$$\det AB = \det A \det B.$$

Theorem 3.15 *An orthogonal matrix always has a determinant of $+1$ or -1.*

Proof: Using Proposition 3.14 we have

$$\det AA^t = \det A \det A^t = (\det A)^2.$$

Moreover, $AA^t = I$, which implies $\det AA^t = 1$. Thus $(\det A)^2 = 1$, meaning that $\det A = \pm 1$. □

We see that there are two cases for an orthogonal matrix:

- $\det A = 1$. In this case the orthogonal transformation corresponds to the movement of a solid with one point fixed. We will see that the only movements of this type are rotations about an axis.
- $\det A = -1$. In this case the transformation "reverses the orientation." An example of such a transformation is reflection across a plane. Consider an asymmetric object such as your hand. The mirror image of your right hand is your left hand, and there is no motion that could bring your right-hand to its mirror image. Thus orthogonal transformations with a determinant of -1 cannot be realized by movements of a solid. It can be shown that any orthogonal transformation with a determinant of -1 can be written as the composition of a rotation and a reflection across a plane (see Exercise 10).

A brief review of complex numbers:

- The *conjugate* of a complex number $z = x + iy$ is the complex number $\overline{z} = x - iy$. Moreover, it is easy to verify that if z_1 and z_2 are two complex numbers, then

$$\begin{cases} \overline{z_1 + z_2} = \overline{z}_1 + \overline{z}_2, \\ \overline{z_1 z_2} = \overline{z}_1 \overline{z}_2. \end{cases} \tag{3.5}$$

- z is real if and only if $z = \overline{z}$.
- The *modulus* of a complex number $z = x + iy$ is $|z| = \sqrt{x^2 + y^2} = \sqrt{z\overline{z}}$.

Proposition 3.16 *If A is a real matrix and if $\lambda = a + ib$ with $b \neq 0$ is a complex eigenvalue of A with eigenvector v, then $\overline{\lambda} = a - ib$ is also an eigenvalue of A with eigenvector $\overline{v}$.*

PROOF. Let v be an eigenvector of the complex eigenvalue λ. We have that $A[v] = \lambda[v]$. Taking the conjugate of this expression yields $\overline{A[v]} = \overline{A}[\overline{v}] = \overline{\lambda}[\overline{v}]$. Since A is real we have that $\overline{A} = A$. This implies

$$A[\overline{v}] = \overline{\lambda}[\overline{v}],$$

which shows that $\overline{\lambda}$ is an eigenvalue of A with eigenvector $\overline{v}$. □

The principal result that we are working up to is that any 3×3 orthogonal matrix A with $\det A = 1$ corresponds to a rotation by some angle about some axis. Among the various intermediate results is the corresponding result for 2×2 matrices.

Proposition 3.17 *If A is a 2×2 orthogonal matrix with $\det A = 1$ then A is the matrix of rotation by an angle θ,*

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

for some $\theta \in [0, 2\pi)$. The eigenvalues are $\lambda_1 = a + ib$ and $\lambda_2 = a - ib$, with $a = \cos\theta$ and $b = \sin\theta$. They are both real if and only if $\theta = 0$ or $\theta = \pi$. In the case $\theta = 0$ we obtain $a = 1$, $b = 0$, and A is the identity matrix. In the case $\theta = \pi$ we obtain $a = -1$, $b = 0$, and A is the matrix of rotation by the angle π (also called reflection through the origin).

PROOF. Let

$$A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Since each column vector has length 1 we must have that $a^2 + b^2 = 1$, allowing us to set $a = \cos\theta$ and $b = \sin\theta$. Since the two columns are orthogonal, we must have that

$$c\cos\theta + d\sin\theta = 0.$$

Therefore

$$\begin{cases} c = -C\sin\theta, \\ d = C\cos\theta, \end{cases}$$

for some $C \in \mathbb{R}$. Since the second column is a vector with length 1, then $c^2 + d^2 = 1$, which implies $C^2 = 1$ or equivalently $C = \pm1$. Finally, since $\det A = C$, we must have that $C = 1$.

The characteristic polynomial of this matrix is $\det(\lambda I - A) = \lambda^2 - 2a\lambda + 1$, which has roots $a \pm \sqrt{a^2 - 1}$. The result follows, since

$$\pm\sqrt{a^2 - 1} = \pm\sqrt{\cos^2\theta - 1} = \pm\sqrt{-(1 - \cos^2\theta)} = \pm i\sin\theta = \pm ib.$$

□

Lemma 3.18 *All the real eigenvalues of an orthogonal matrix A are equal to ±1.*

PROOF. Let λ be a real eigenvalue and let v be a corresponding eigenvector. Let T be the orthogonal transformation with matrix A. Since T preserves lengths, we have that $\langle T(v), T(v)\rangle = \langle v, v\rangle$. But $T(v) = \lambda v$. Thus $\langle T(v), T(v)\rangle = \langle \lambda v, \lambda v\rangle = \lambda^2 \langle v, v\rangle$. Finally, $\lambda^2 = 1$. □

Proposition 3.19 *If A is a 3×3 orthogonal matrix with $\det A = 1$, then 1 is always an eigenvalue of A. Moreover, all complex eigenvalues $\lambda = a + ib$ have modulus 1.*

PROOF. The characteristic polynomial of A, $\det(\lambda I - A)$, has degree 3. Therefore it always has one real root λ_1 which can be only 1 or -1 by Proposition 3.18. The other two eigenvalues λ_2 and λ_3 are either both real or both complex and conjugates of each other. The determinant is the product of the eigenvalues. Thus $1 = \lambda_1\lambda_2\lambda_3$. If λ_2 and λ_3 are real, then $\lambda_1, \lambda_2, \lambda_3 \in \{1, -1\}$ by Lemma 3.18. For their product to be 1 it must be that either all three eigenvalues are 1 or two of them are -1 and the remaining eigenvalue is 1. Hence at least one eigenvalue is equal to 1. If λ_2 and λ_3 are complex then $\lambda_2 = a + ib$ and $\lambda_3 = \overline{\lambda}_2 = a - ib$, from which it follows that $\lambda_2\lambda_3 = |\lambda_2|^2 = a^2 + b^2$. Since $1 = \lambda_1\lambda_2\lambda_3 > 0$, then $\lambda_1 = 1$ and $a^2 + b^2 = 1$. □

Theorem 3.20 *If A is a 3×3 orthogonal matrix with $\det A = 1$ then A is the matrix of a rotation T by some angle θ about some axis. If A is not the identity matrix then the axis of rotation corresponds to the eigenvector associated with the eigenvalue $+1$.*

PROOF. Let v_1 be a unit eigenvector of the eigenvalue 1. We consider the subspace orthogonal to v_1:

$$E = \{w \in \mathbb{R}^3 | \langle v_1, w\rangle = 0\},$$

which is a subspace of dimension 2. Let T be the orthogonal transformation with matrix A. Since T preserves scalar products and $T(v_1) = v_1$, if $w \in E$ then $T(w) \in E$, since

$$\langle T(w), T(v_1)\rangle = \langle T(w), v_1\rangle = \langle w, v_1\rangle = 0.$$

Consider the restriction T_E of T on E. Let $\mathcal{B}' = \{v_2, v_3\}$ be an orthonormal basis of E and consider the matrix B of T_E in the basis $\mathcal{B}'$. If

$$B = \begin{pmatrix} b_{22} & b_{23} \\ b_{32} & b_{33} \end{pmatrix},$$

this signifies that

$$\begin{cases} T(v_2) = b_{22}v_2 + b_{32}v_3, \\ T(v_3) = b_{23}v_2 + b_{33}v_3. \end{cases}$$

Since T_E preserves scalar products, then B must be an orthogonal matrix. Now consider the matrix $[T]_{\mathcal{B}}$ of the transformation T expressed in the basis $\mathcal{B} = \{v_1, v_2, v_3\}$ (which is an orthonormal basis of $\mathbb{R}^3$):

$$[T]_{\mathcal{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & B \end{pmatrix}.$$

The determinant of this matrix is equal to $\det B$. (Recall that the determinant of the matrix of a linear transformation does not change when we change the basis.) Thus $\det B = \det A = 1$. By Proposition 3.17 it follows that B is a matrix of rotation, from which it follows that

$$[T]_{\mathcal{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}.$$

Consider this matrix: it tells us that all vectors along the axis described by v_1 are mapped to themselves by T, and that all vectors in the plane E undergo a rotation by the angle θ. If we decompose a vector v as $v = Cv_1 + w$ with $w \in E$, then $T(v) = Cv_1 + T_E(w)$, where T_E corresponds to rotation by the angle θ in the plane E. This corresponds to rotation by an angle θ about the axis described by v_1. □

Corollary 3.21 *Suppose A is a 3×3 orthogonal matrix with $\det A = 1$ and with three real eigenvalues. Then either A is the identity matrix with eigenvalues 1 or A has the three eigenvalues $1, -1, -1$. In the latter case A corresponds to reflection through the axis generated by the eigenvector associated with eigenvalue $+1$. (This transformation can equally be visualized as a rotation by an angle of π about this same axis.)*

Theorem 3.20 states that an orthogonal matrix A with $\det A = 1$ is the matrix of a rotation. How do we calculate the angle of rotation? To do this we introduce the *trace* of a matrix.

Definition 3.22 *Let $A = (a_{ij})$ be an $n \times n$ matrix. The* trace *of A is the sum of the elements along its diagonal:*

$$tr(A) = a_{11} + \cdots + a_{nn}.$$

We state without proof the following property of the trace of a matrix.

Theorem 3.23 *The trace of a matrix is equal to the sum of its eigenvalues.*

Proposition 3.24 *Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be a rotation with matrix A. Then the angle of rotation θ is such that*

$$\cos\theta = \frac{tr(A) - 1}{2}. \tag{3.6}$$

Proof. Consider the proof of Theorem 3.20. In calculating the characteristic polynomial of $[T]_{\mathcal{B}}$ we saw that the eigenvalues of T were 1 and $\cos\theta \pm i\sin\theta$. Thus the sum of the eigenvalues is $1 + 2\cos\theta$. By Theorem 3.23 this is equal to $\mathrm{tr}(A)$. □

Analyzing an orthogonal transformation in $\mathbb{R}^3$. Theorem 3.20 and Proposition 3.24 suggest a strategy:

- We start by calculating $\det A$. If $\det A = 1$ we are sure that 1 is one of the eigenvalues of A and that the transformation is a rotation. If $\det A = -1$ we are sure that -1 is an eigenvalue (see Exercise 10). The rest of this discussion centers on the case $\det A = 1$, with the case $\det A = -1$ being left to Exercise 10.
- To determine the axis of rotation we find the eigenvector v_1 associated with eigenvalue 1.
- We calculate the angle of rotation using equation (3.6). There are two possible solutions, since $\cos\theta = \cos(-\theta)$. We cannot decide between the two without performing a test. To do this we choose a vector w orthogonal to v_1 and we calculate $T(w)$. We then calculate the cross product of w and $T(w)$ (see Definition 3.25 below). It will be a multiple Cv_1 of v_1 with $|C| = |\sin\theta|$. The angle θ is that which satisfies $C = \sin\theta$.

Definition 3.25 *The* cross product *of two vectors $v = (x_1, y_1, z_1)$ and $w = (x_2, y_2, z_2)$ is the vector $v \wedge w$ given by*

$$v \wedge w = \left(\begin{vmatrix} y_1 & z_1 \\ y_2 & z_2 \end{vmatrix}, - \begin{vmatrix} x_1 & z_1 \\ x_2 & z_2 \end{vmatrix}, \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} \right).$$

Remark. The angle of rotation is determined using the *right-hand rule*: with the right hand positioned such that your thumb points along the vector v_1, positive angles are measured in the direction that your fingers curl. Thus, the angle θ depends on the direction that has been chosen for the axis of rotation. Hence, the rotation about an axis determined by v_1 and angle θ is identical to that about the axis determined by $-v_1$ and angle $-\theta$.

We now have all of the elements necessary to define and describe the possible movements of a solid in space.

Definition 3.26 *A transformation F is a movement of a solid in space if F preserves distances and angles, and if for all sets of vectors with the same origin P that form an orthonormal basis $\{v_1, v_2, v_3\}$ of $\mathbb{R}^3$ with $v_3 = v_1 \wedge v_2$, then $\{F(v_1), F(v_2), F(v_3)\}$ is also an orthonormal basis of $\mathbb{R}^3$ with origin at $F(P)$ and such that $F(v_3) = F(v_1) \wedge F(v_2)$.*

The additional condition that F maps $v_1 \wedge v_2$ to $F(v_1) \wedge F(v_2)$ is equivalent to saying that F preserves orientation.

Theorem 3.27 *Any movement of a solid in space is the composition of a translation and a rotation about some axis.*

Proof. Let F be a transformation in $\mathbb{R}^3$ that describes the movement of a solid. It preserves both distances and angles. Consider a point of the solid at an initial position $P_0 = (x_0, y_0, z_0)$ and a final position $P_1 = (x_1, y_1, z_1)$ after the transformation. Let $v = \overrightarrow{P_0P_1}$ and let G be the operation of translation by v. Set $T = F \circ G^{-1}$. Then

$T(P_1) = F(P_1 - v) = F(P_0) = P_1$. Thus P_1 is a fixed point of T. Since T preserves distances and angles and has a fixed point, it is linear (Exercise 4), hence an orthogonal transformation with matrix A. But we have seen that if $\det A = -1$ then A cannot be a transformation of a solid (see Exercise 10). Hence $\det A = 1$, and therefore T is a rotation. □

3.4 Change of Basis

Transformation matrices in a basis $\mathcal{B}$. Consider a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$. We are interested only in the cases $n = 2$ and $n = 3$. Let $\mathcal{B}$ be a basis for $\mathbb{R}^3$. We represent a vector v using its coordinates in the basis $\mathcal{B}$ by a column vector $[v]_{\mathcal{B}} = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ if $n = 2$ and $[v]_{\mathcal{B}} = \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right)$ if $n = 3$. For now, we limit ourselves to the case $n = 3$. If $\mathcal{B} = \{v_1, v_2, v_3\}$, then $[v]_{\mathcal{B}} = \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right)$ signifies that $v = xv_1 + yv_2 + zv_3$. Let A be the matrix describing the transformation T in the basis $\mathcal{B}$, denoted by $A = [T]_{\mathcal{B}}$. The coordinates of $T(v)$ in the basis $\mathcal{B}$ are determined as

$$[T(v)]_{\mathcal{B}} = A[v]_{\mathcal{B}} = [T]_{\mathcal{B}}[v]_{\mathcal{B}}.$$

As is the case with the standard basis, the columns of A are given by the coordinate vectors in the basis $\mathcal{B}$ of the images of the vectors in $\mathcal{B}$ under the transformation T.

Matrices for performing a change of basis

1. If we have two bases $\mathcal{B}_1$ and $\mathcal{B}_2$ of $\mathbb{R}^3$, then

   $$[v]_{\mathcal{B}_2} = P[v]_{\mathcal{B}_1},$$

   where P is the *change of basis matrix* from $\mathcal{B}_1$ to $\mathcal{B}_2$. The matrix P is also sometimes called the *passage matrix* from $\mathcal{B}_1$ to $\mathcal{B}_2$.
2. The columns of P are the coordinates of the vectors of $\mathcal{B}_1$ written in the basis $\mathcal{B}_2$. In the case that the two bases are orthonormal, then P is orthogonal.
3. If Q is the change of basis matrix from $\mathcal{B}_2$ to $\mathcal{B}_1$, then $Q = P^{-1}$. The columns of Q are the coordinates of the vectors of $\mathcal{B}_2$ written in the basis $\mathcal{B}_1$. In the case that the two bases are orthonormal, then $Q = P^t$ and therefore the columns of Q are the rows of P.

Theorem 3.28 *Let T be a linear transformation and let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two bases of $\mathbb{R}^3$. Let P be the change of basis matrix from $\mathcal{B}_1$ to $\mathcal{B}_2$. Then*

$$[T]_{\mathcal{B}_2} = P[T]_{\mathcal{B}_1}P^{-1}.$$

PROOF: Let v be a vector. We have that

$$[T(v)]_{\mathcal{B}_2} = [T]_{\mathcal{B}_2}[v]_{\mathcal{B}_2}.$$

We also have that

$$\begin{aligned} [T(v)]_{\mathcal{B}_2} &= P[T(v)]_{\mathcal{B}_1} \\ &= P([T]_{\mathcal{B}_1}[v]_{\mathcal{B}_1}) \\ &= P[T]_{\mathcal{B}_1}(P^{-1}[v]_{\mathcal{B}_2}) \\ &= (P[T]_{\mathcal{B}_1}P^{-1})[v]_{\mathcal{B}_2}. \end{aligned}$$

The result follows directly from these two equations and from the uniqueness of the matrix $[T(v)]_{\mathcal{B}_2}$ of T in the basis $\mathcal{B}_2$. □

Playing with multiple bases allows us to resolve complicated problems. We have seen how diagonalization allows us to understand the structure of a linear transformation. We can also play the same game in reverse, constructing a transformation matrix from a description of its effect. We illustrate this in the following example.

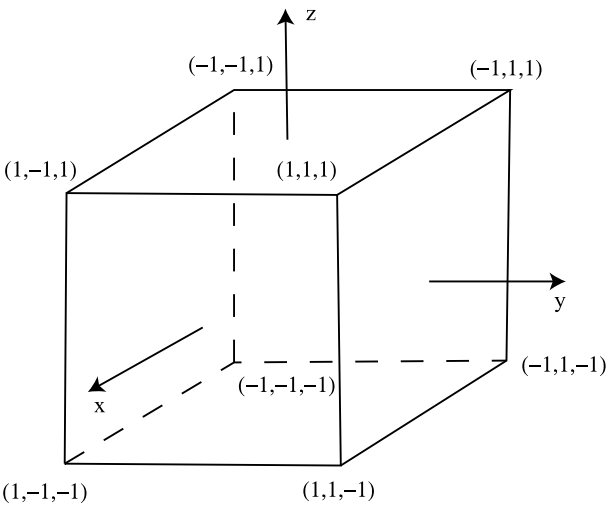


Fig. 3.5. The cube from Example 3.29.

Example 3.29 *Consider a cube whose eight corners are positioned at the points* $(\pm 1, \pm 1, \pm 1)$*, as shown in Figure 3.5. We are looking for the matrices of the two rotations of angles* $\pm\frac{2\pi}{3}$ *about the axis through the corners* $(-1,-1,-1)$ *and* $(1,1,1)$*. Observe that both of these rotations map the cube to itself.*

To do this we start by choosing a basis $\mathcal{B}$ that is suited to the problem. The direction of the first vector will be given by the direction of the axis, $w_1 = (2,2,2)$. The two other vectors w_2 and w_3 of the basis will be taken orthogonal to w_1. Their coordinates (x,y,z) will therefore satisfy $x+y+z=0$. The vector $w_2 = (-1,0,1)$ is easily seen to be of this form. We wish for the third vector to be perpendicular to both w_1 and w_2. Its coordinates must therefore satisfy

$$\begin{cases} x+y+z=0, \\ x-z=0, \end{cases}$$

a possible solution to which is $w_3 = (1,-2,1)$. We would like to work with an orthonormal basis, so we divide each vector by its length: $v_i = \frac{w_i}{||w_i||}$. The final basis is given by

$$\begin{aligned} \mathcal{B} &= \{v_1, v_2, v_2\} \\ &= \left\{ \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right), \left(-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right), \left(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) \right\}. \end{aligned}$$

In this basis the two transformations are simply rotations about the v_1 axis. Note that $\cos(-\frac{2\pi}{3}) = \cos\frac{2\pi}{3} = -\frac{1}{2}$ and $\sin(-\frac{2\pi}{3}) = -\sin\frac{2\pi}{3} = -\frac{\sqrt{3}}{2}$. The two rotations $T_\pm$ are therefore given (in the basis $\mathcal{B}$) by

$$[T_\pm]_{\mathcal{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\frac{2\pi}{3} & \mp\sin\frac{2\pi}{3} \\ 0 & \pm\sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & \mp\frac{\sqrt{3}}{2} \\ 0 & \pm\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

Now we wish to find the matrices $T_\pm$ in the standard basis $\mathcal{C}$. By applying the previous theorem we see that these matrices are given by

$$[T_\pm]_{\mathcal{C}} = P^{-1}[T_\pm]_{\mathcal{B}}P,$$

where P is the passage matrix from $\mathcal{C}$ to $\mathcal{B}$. Thus P^{-1} is the passage matrix from $\mathcal{B}$ to $\mathcal{C}$, whose columns consist of the vectors of $\mathcal{B}$ written in the basis $\mathcal{C}$. These are precisely the vectors v_i, since they are already written in the standard basis. Since $P^{-1} = P^t$ we have that

$$P^{-1} = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix}, \qquad P = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

From this it follows that

$$[T_+]_{\mathcal{C}} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad [T_-]_{\mathcal{C}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

The first transformation T_+ consists of rotation by an angle of $\frac{2\pi}{3}$ about the axis v_1 (see Figure 3.5). It permutes three corners of the cube as $(1,1,-1) \mapsto (-1,1,1) \mapsto (1,-1,1)$. Similarly, it permutes three other corners as $(-1,-1,1) \mapsto (1,-1,-1) \mapsto (-1,1,-1)$, while the two remaining corners $(1,1,1)$ and $(-1,-1,-1)$ remain fixed.

Remark: *$[T_+]_{\mathcal{C}}$ is orthogonal and $T_- = T_+^{-1}$. Thus $[T_-]_{\mathcal{C}} = [T_+]_{\mathcal{C}}^{-1} = [T_+]_{\mathcal{C}}^t$.*

3.5 Different Frames of Reference for a Robot

Definition 3.30 *A* frame of reference *in space consists of a point $P \in \mathbb{R}^3$, called the* origin*, and a basis $\mathcal{B} = \{v_1, v_2, v_3\}$ of $\mathbb{R}^3$.*

Giving ourselves a frame of reference is equivalent to defining a coordinate system centered on the point P whose axes are oriented along the vectors of the basis $\mathcal{B}$. The units of the coordinate system are chosen such that the vectors v_i are the unit vectors $v_1 = (1,0,0)$, $v_2 = (0,1,0)$, and $v_3 = (0,0,1)$ when expressed in this coordinate system.

Consider the robot of Figure 3.1, which we have reproduced in a stretched-out position in Figure 3.6, and after several rotations in Figure 3.8. We have specified seven frames of reference $R_0, \ldots, R_6$, centered at $P_0, \ldots, P_6$ respectively. Each frame of reference has been associated with a set of axes x_i, y_i, and z_i for $i = 0, \ldots, 6$, the directions of which are given by the bases $\mathcal{B}_0, \ldots, \mathcal{B}_6$. The frame of reference $\mathcal{B}_0$ is the base frame of reference. It is fixed and centered at $P_0 = (0,0,0)$. The frame of reference R_i is centered at P_i (Figures 3.6, 3.7, and 3.8). When the robot is stretched out (in its base position), all the frames of reference have parallel axes, as shown in Figure 3.6. The frames of reference themselves will move as the robot moves. In fact, since moving one joint affects all joints attached further along the arm, the frame of reference R_i depends on any motions applied to joints $1, \ldots, i$ and is independent of those applied to joints $i+1, \ldots, 6$.

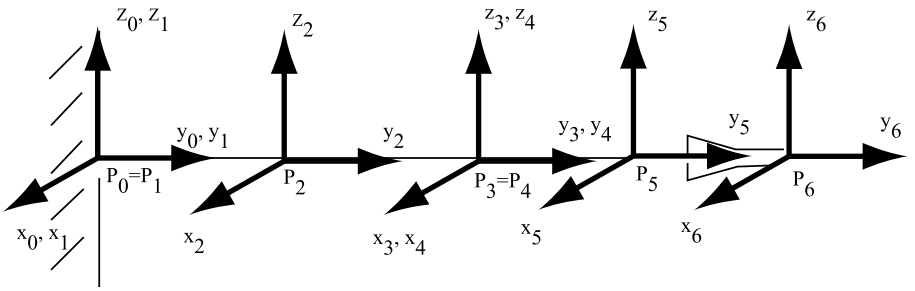


Fig. 3.6. The different frames of reference of the robot.

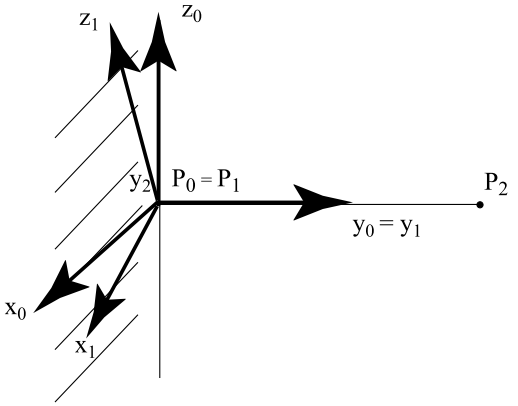


Fig. 3.7. The frame of reference R_1 after a rotation about the axis y_0.

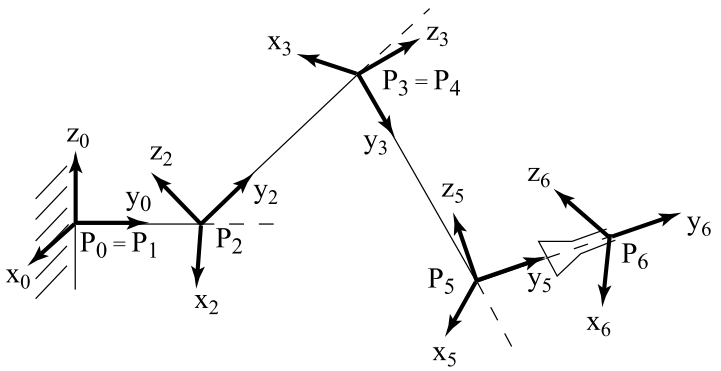


Fig. 3.8. The various frames of reference after several rotations about the joints $2, 3, 5$, and 6. The frame of reference R_1 (respectively R_4) coincides with that of R_0 (respectively R_3) and is not explicitly shown.

We describe the sequence of motions applied to the robot that place it in the position of Figure 3.1 or Figure 3.8.

(i) The first movement consists of a rotation T_1 of angle θ_1 about the axis y_0. In the frame of reference R_0 this is a linear transformation, since the origin is fixed. In the basis $\mathcal{B}_0$ it is described by the matrix

$$A_1 = \begin{pmatrix} \cos\theta_1 & 0 & -\sin\theta_1 \\ 0 & 1 & 0 \\ \sin\theta_1 & 0 & \cos\theta_1 \end{pmatrix}.$$

The second frame of reference R_1 is altered by this motion and obtained by applying T_1 to R_0. In particular, the basis $\mathcal{B}_1$ is given by the image of $\mathcal{B}_0$ under T_1.

(ii) The second movement is a rotation T_2 of angle θ_2 about the axis x_2, described by the matrix

$$A_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2 & -\sin\theta_2 \\ 0 & \sin\theta_2 & \cos\theta_2 \end{pmatrix}.$$

(iii) The third movement is, for instance, a rotation T_3 by angle θ_3 about the axis x_3, described by the matrix

$$A_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & -\sin\theta_3 \\ 0 & \sin\theta_3 & \cos\theta_3 \end{pmatrix}.$$

Looking at Figure 3.1, it is difficult to discern whether this movement is a rotation about x_3 or z_3. What may look like a rotation about x_3 or z_3 actually depends on the earlier applied rotation T_1.

(iv) The fourth movement is a rotation T_4 by angle θ_4 about the axis y_4 as given by the matrix

$$A_4 = \begin{pmatrix} \cos\theta_4 & 0 & -\sin\theta_4 \\ 0 & 1 & 0 \\ \sin\theta_4 & 0 & \cos\theta_4 \end{pmatrix}.$$

(v) The fifth movement consists of a rotation T_5 by angle θ_5 about the axis x_5 and is described by the matrix

$$A_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_5 & -\sin\theta_5 \\ 0 & \sin\theta_5 & \cos\theta_5 \end{pmatrix}.$$

(vi) The sixth movement is a rotation T_6 by angle θ_6 about the axis y_6, given by the matrix

$$A_6 = \begin{pmatrix} \cos\theta_6 & 0 & -\sin\theta_6 \\ 0 & 1 & 0 \\ \sin\theta_6 & 0 & \cos\theta_6 \end{pmatrix}.$$

We wish to calculate the position of a point on the robot with respect to the various frames of reference. To do this we start by calculating how the various axes are modified as we pass from one frame of reference to another. This allows us to find the "orientation" of the basis $\mathcal{B}_{i+k}$ in the basis $\mathcal{B}_i$. The columns of the matrix A_i give the coordinates of the vectors of the basis $\mathcal{B}_{i+1}$ expressed in the basis $\mathcal{B}_i$. This is the change of basis matrix from $\mathcal{B}_{i+1}$ to $\mathcal{B}_i$. We will denote it by M_i^{i+1}.

Change of basis matrix from $\mathcal{B}_{i+k}$ to $\mathcal{B}_i$. We deduce that it is given by

$$M_i^{i+k} = M_i^{i+1} M_{i+1}^{i+2} \cdots M_{i+k-1}^{i+k}.$$

Let Q be a point in space. Finding its position in the frame of reference R_i means to find the vector $\overrightarrow{P_iQ}$ in the basis $\mathcal{B}_i$, in other words, $[\overrightarrow{P_iQ}]_{\mathcal{B}_i}$. Its position in the frame of reference R_{i-1} is given by

$$[\overrightarrow{P_{i-1}Q}]_{\mathcal{B}_{i-1}} = [\overrightarrow{P_{i-1}P_i}]_{\mathcal{B}_{i-1}} + [\overrightarrow{P_iQ}]_{\mathcal{B}_{i-1}} = [\overrightarrow{P_{i-1}P_i}]_{\mathcal{B}_{i-1}} + M_{i-1}^{i}[\overrightarrow{P_iQ}]_{\mathcal{B}_i}.$$

We will use this approach to account for motion at each of the joints $i = 1, \ldots, 6$. We will determine the position and orientation of the extremity of the robot in the basis $\mathcal{B}_0$, accounting for the rotations of the various joints by angles $\theta_1, \ldots, \theta_6$, respectively. Suppose that we know the position of Q in the frame of reference R_6, denoted by $[\overrightarrow{P_6Q}]_{\mathcal{B}_6}$:

- Let l_5 be the length of the claw. Then

$$[\overrightarrow{P_5Q}]_{\mathcal{B}_5} = [\overrightarrow{P_5P_6}]_{\mathcal{B}_5} + [\overrightarrow{P_6Q}]_{\mathcal{B}_5} = \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_5^6[\overrightarrow{P_6Q}]_{\mathcal{B}_6}.$$

- Let l_4 be the length of the third segment of the robot. Then

$$\begin{aligned} [\overrightarrow{P_4Q}]_{\mathcal{B}_4} &= [\overrightarrow{P_4P_5}]_{\mathcal{B}_4} + [\overrightarrow{P_5Q}]_{\mathcal{B}_4} \\ &= \begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_4^5 \left(\begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_5^6[\overrightarrow{P_6Q}]_{\mathcal{B}_6} \right) \\ &= \begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_4^5 \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_4^6[\overrightarrow{P_6Q}]_{\mathcal{B}_6}. \end{aligned}$$

- The frame of reference R_3 has the same origin as R_4: $P_3 = P_4$. Thus, in the frame of reference R_3,

$$\begin{aligned} [\overrightarrow{P_3Q}]_{\mathcal{B}_3} &= [\overrightarrow{P_4Q}]_{\mathcal{B}_3} = M_3^4 \left(\begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_4^5 \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_4^6[\overrightarrow{P_6Q}]_{\mathcal{B}_6} \right) \\ &= M_3^4 \begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_3^5 \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_3^6[\overrightarrow{P_6Q}]_{\mathcal{B}_6}. \end{aligned}$$

- Let l_2 be the length of the second segment of the robot. Then

$$\begin{aligned} [\overrightarrow{P_2Q}]_{\mathcal{B}_2} &= [\overrightarrow{P_2P_3}]_{\mathcal{B}_2} + [\overrightarrow{P_3Q}]_{\mathcal{B}_2} \\ &= \begin{pmatrix} 0 \\ l_2 \\ 0 \end{pmatrix} + M_2^4 \begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_2^5 \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_2^6[\overrightarrow{P_6Q}]_{\mathcal{B}_6}. \end{aligned}$$

- Let l_1 be the length of the first segment of the robot. Then

$$\begin{aligned}[\overrightarrow{P_1Q}]_{\mathcal{B}_1} &= [\overrightarrow{P_1P_2}]_{\mathcal{B}_1} + [\overrightarrow{P_2Q}]_{\mathcal{B}_1} \\ &= \begin{pmatrix} 0 \\ l_1 \\ 0 \end{pmatrix} + M_1^2 \begin{pmatrix} 0 \\ l_2 \\ 0 \end{pmatrix} + M_1^4 \begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_1^5 \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_1^6 [\overrightarrow{P_6Q}]_{\mathcal{B}_6}.\end{aligned}$$

- Finally, in the base frame of reference, since $P_0 = P_1$ we have that

$$[\overrightarrow{P_0Q}]_{\mathcal{B}_0} = M_0^1 \begin{pmatrix} 0 \\ l_1 \\ 0 \end{pmatrix} + M_0^2 \begin{pmatrix} 0 \\ l_2 \\ 0 \end{pmatrix} + M_0^4 \begin{pmatrix} 0 \\ l_4 \\ 0 \end{pmatrix} + M_0^5 \begin{pmatrix} 0 \\ l_5 \\ 0 \end{pmatrix} + M_0^6 [\overrightarrow{P_6Q}]_{\mathcal{B}_6}. \tag{3.7}$$

Setting $l_3 = 0$ allows us to rewrite (3.7) as

$$[\overrightarrow{P_0Q}]_{\mathcal{B}_0} = \sum_{i=1}^{5} M_0^i \begin{pmatrix} 0 \\ l_i \\ 0 \end{pmatrix} + M_0^6 [\overrightarrow{P_6Q}]_{\mathcal{B}_6}.$$

Inversely,

$$[\overrightarrow{P_6Q}]_{\mathcal{B}_6} = M_6^0 [\overrightarrow{P_0Q}]_{\mathcal{B}_0} - \sum_{i=1}^{5} M_6^i \begin{pmatrix} 0 \\ l_i \\ 0 \end{pmatrix},$$

where M_6^i is the change of basis matrix from $\mathcal{B}_i$ to $\mathcal{B}_6$. We have that $M_6^i = (M_i^6)^{-1} = (M_i^6)^t$. If necessary we can also calculate $[\overrightarrow{P_iQ}]_{\mathcal{B}_i}$ as a function of $[\overrightarrow{P_0Q}]_{\mathcal{B}_0}$.

Applications:

1. *The* Canadarm *on the International Space Station.* The *Canadarm* is the robotic arm attached to the International Space Station. Initially it was fixed to the station. It has since been mounted on rails, allowing it to be moved along the length of the station. This facilitates the work of the astronauts as they assemble new space station modules or perform repairs.
   The *Canadarm* (the *Shuttle Remote Manipulator System*, or *SRMS* for short) is a robot with six degrees of freedom. Similar to a human arm, it consists of two segments at the end of which is found a "wrist" of sorts. The first segment is attached to a rail on the station, and can make an arbitrary angle at this attachment, requiring both a *pitch* (up and down) and *yaw* (side to side) motion. The joint between the two segments has only one degree of freedom, allowing only an up and down motion, similar to an elbow. The wristlike joint has three degrees of freedom, allowing pitch, yaw, and *roll* (motion about its axis). (See Exercise 16.) The first segment is 5 m long while the second segment has length 5.8 m.
   Since the original *Canadarm* was built, an improved model has been constructed. The *Canadarm2* is 17 m long and has seven joints, allowing it more flexibility for those hard-to-reach places. It can be controlled from the ground.

2. *Surgical robots.* Such robots allow for noninvasive surgeries, since they can be inserted through small incisions and controlled externally. They have many small segments near the end of the robot, affording it a great degree of flexibility in a small space.

More mathematical problems related to robots. We are far from having considered all mathematical problems related to robots. We present a few other practical problems here:

(i) There exist several sequences of movements that will place a robot in a given final position. Which is better? Certain "small" movements may lead to "large" displacements of the claw, while other "large" movements may cause "small" displacements. The latter are preferable when the robot is being used for work requiring precision, as is the case for surgical robots.
(ii) We can always add more segments and joints to a robot, increasing its flexibility and allowing it to avoid obstacles. What other effects are there in adding more segments and movements?
(iii) What is the effect of changing the lengths of the various segments?
(iv) The inverse problem (difficult!): given a final position for the claw, determine a sequence of movements that will bring the claw to this position. Answering this problem generally involves solving a system of nonlinear equations.
(v) There are many more related problems. It is up to you to think of some.

3.6 Exercises

1. **(a)** Calculate the matrix A of rotation by the angle θ in the plane, using the standard basis $\{e_1 = (1,0), e_2 = (0,1)\}$. Use the fact that the columns of A are the coordinates of the images of the vectors e_1 and e_2.
(b) Let $z = x + iy$. Rotating the vector (x, y) by an angle θ is equivalent to performing the operation $z \mapsto e^{i\theta} z$. Use this formula to determine the matrix A.

2. If two linear transformations T_1 and T_2 described by matrices A_1 and A_2 are composed, then the matrix describing the composed operator $T_1 \circ T_2$ is $A_1 A_2$. In this exercise we will assume that $n = 2$.
(a) Verify that the composition of a rotation by an angle of θ_1 with a rotation by an angle of θ_2 is a rotation by an angle of $\theta_1 + \theta_2$.
(b) Verify that the determinant of a matrix of rotation is equal to 1.
(c) Verify that the inverse of a matrix of rotation A is simply its transpose, A^t.

3. The triangle shown in Figure 3.2 is a right triangle with side lengths 3, 4, and 5. Initially the corner opposing the side of length 3 is at the origin, and at the end of its

movements it is situated at the point $(7, 5)$. Give the coordinates of the corner opposite the side of length 4 if the rotation is by an angle of $\frac{\pi}{7}$.

4. Show that a transformation T of the plane or of the space preserving distances and angles and having a fixed point is a linear transformation. Suggestion:
(a) Start by proving that the transformation preserves the sum of two vectors, using that the sum v_1+v_2 of the two vectors is constructed as the diagonal of the parallelogram with sides v_1 and v_2.
(b) Show now that for any vector v and any $c \in \mathbb{R}$, then $T(cv) = cT(v)$. Make the argument in several steps:

- Prove the assertion for $c \in \mathbb{N}$.
- Prove the assertion for $c \in \mathbb{Q}$.
- Show that T is uniformly continuous. Use this to prove it for $c \in \mathbb{R}$. Indeed, if $c = \lim_{n\to\infty} c_n$ with $c_n \in \mathbb{Q}$, and if T is continuous, then $T(cv) = \lim_{n\to\infty} T(c_n v)$.

5. Show that all orthogonal transformations in $\mathbb{R}^2$ with a determinant of -1 are reflections across an axis passing through the origin.

6. Consider the following orthogonal matrices with determinant 1:

$$A = \begin{pmatrix} 2/3 & -1/3 & -2/3 \\ 2/3 & 2/3 & 1/3 \\ 1/3 & -2/3 & 2/3 \end{pmatrix}, \qquad B = \begin{pmatrix} 1/3 & 2/3 & 2/3 \\ -2/3 & 2/3 & -1/3 \\ -2/3 & -1/3 & 2/3 \end{pmatrix}.$$

For each of these matrices calculate the axis and angle of rotation (up to the sign).

7. Show that the product of two orthogonal matrices A_1 and A_2 with determinant 1 is itself an orthogonal matrix with determinant 1. Deduce that the composition of two rotations in $\mathbb{R}^3$ is also a rotation in $\mathbb{R}^3$ (even if the two axes of rotation are not the same!).

8. Consider a rotation by the angle $+\pi/4$ about the axis v_1 determined by $v_1 = (1/3, 2/3, 2/3)$. Using the basis $\mathcal{B} = \{v_1, v_2, v_3\}$ where $v_1 = (1/3, 2/3, 2/3)$, $v_2 = (2/3, -2/3, 1/3)$, and $v_3 = (2/3, 1/3, -2/3)$, give the matrix describing this rotation expressed in the standard basis.

9. **(a)** Let Π be a plane passing through the origin in $\mathbb{R}^3$ and let v be a unit vector perpendicular to the plane at the origin. Reflection across Π is the operation that maps a vector $x \in \mathbb{R}^3$ to the vector $R_\Pi(x) = x - 2\langle x, v\rangle v$. Show that R_Π is an orthogonal transformation. What is the determinant of the associated matrix?
(b) Show that the composition of two such reflections yields a rotation about some axis passing through the origin. Verify that this axis is the line of intersection between the two planes.

10. **(a)** Show that if an orthogonal 3×3 matrix has determinant -1, then -1 is one of its eigenvalues.

(b) Show that all orthogonal transformations in $\mathbb{R}^3$ with determinant -1 can be described as a composition of a reflection across some plane passing through the origin and a rotation about the axis passing through the origin and perpendicular to the plane. Give a formula for the axis of rotation.

(c) Conclude that an orthogonal transformation in $\mathbb{R}^3$ with determinant -1 cannot describe a movement of a solid in space.

11. Consider the robot of Figure 3.9, which operates in a vertical plane: at the end of the second segment there is a claw that is perpendicular to the plane of operation of the robot and driven by a third rotation (we will ignore this rotation in this question). Assume that the two segments of the robot are of the same length l.

(a) Let Q be the far end of the robot's second segment. Calculate the position of Q if the first segment is rotated through an angle of θ_1 and if the second segment is rotated through an angle of θ_2.

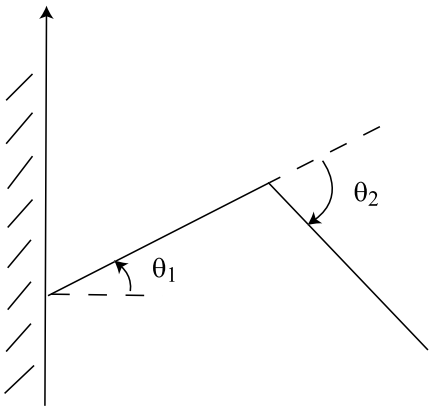


Fig. 3.9. The robot of Exercise 11.

(b) Calculate the (two values of the) angle θ_2 that will position the point Q at a distance of $\frac{l}{2}$ from the point where the robot attaches to the wall.

(c) Calculate the two distinct pairs of angles (θ_1, θ_2) that will position Q at $(\frac{l}{2}, 0)$.

(d) Suppose now that the robot is attached to a vertical rail and can slide up and down the wall. Choose a coordinate system. In this coordinate system calculate the position of Q if we translate the robot by a distance h, rotate the first joint by an angle of θ_1, and rotate the second joint by an angle of θ_2.

12. In $\mathbb{R}^3$ let R_x represent rotation about the x axis by the angle $\pi/2$, let R_y represent rotation about the y axis by the angle $\pi/2$, and let R_z represent rotation about the z axis by the angle $\pi/2$.
(a) The composition $R_y \circ R_z$ is also a rotation. Determine its axis and angle.
(b) Show that $R_x = (R_y)^{-1} \circ R_z \circ R_y$.

13. Consider a robot in the plane attached to a single fixed point. The robot consists of two arms, the first of which has length l_1 and is attached to the fixed point, the second of which has length l_2 and is attached to the end of the first. Both arms are free to rotate completely about their points of attachment. Describe the set of points in the plane that are reachable by the far end of the second segment of the robot as a function of l_1 and l_2.

14. Consider a two-segment robotic arm attached to the wall with segment lengths l_1 and l_2 where $l_2 < l_1$. The first segment is attached to the wall by a universal joint (one that has two degrees of freedom and can make any angle with the wall). Similarly, the second arm is attached to the first by a universal joint. Determine the set of points in space that are reachable by the free end of the second segment as a function of l_1 and l_2.

15. We describe a robot capable of operating in a vertical plane as shown in Figure 3.10:

- The first segment is fixed at $P_0 = P_1$ and has length ℓ_1.
- The second segment is attached to the end of the first segment at P_2. Its length is variable with a minimum of ℓ_2 and a maximum of $L_2 = \ell_2 + d_2$. A claw is attached to its far end.
- The claw has length d_3 such that $d_3 < \ell_1, \ell_2$.

(a) Give the conditions on ℓ_1, ℓ_2, d_2, d_3 such that the extremity P_4 of the claw can grab an object situated at P_0.
(b) Choose a frame of reference centered at P_0. In this coordinate system, give the position of the extremity P_4 of the claw if the rotations θ_1, θ_2, and θ_3 have been applied as in Figure 3.10 and if the second segment has been set to a length of $\ell_2 + r$.

16. The *Canadarm* (the *Shuttle Remote Manipulator System*, or *SRMS* for short) is a robot with six degrees of freedom. Similar to a human arm, it consists of two segments, at the end of which is found a "wrist" of sorts. The first segment is attached to a rail on the station and can make any arbitrary angle at this attachment, requiring both a pitch (up and down) and yaw (side to side) motion. The joint between the two segments has only one degree of freedom, allowing only an up and down motion, similar to an elbow. The wristlike joint has three degrees of freedom allowing pitch, yaw, and roll (motion about its axis).
(a) Ignoring the translational movement on the rails along the station, draw a schematic of the arm and the necessary frames of reference required to calculate the

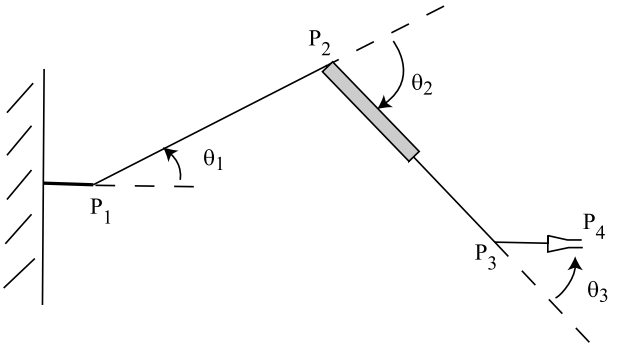


Fig. 3.10. The robot of Exercise 15.

position of the end of the wrist. In the appropriate frame of reference, give the six rotational movements corresponding to the six degrees of freedom of the robot.
(b) Given a set of six rotations with angles $\theta_1, \ldots, \theta_6$ to be applied to each degree of freedom, calculate the position of the end of the wrist in the base frame of reference.

17. Imagine a system of controls for all six degrees of freedom of the robot of Figure 3.1.

18. When an astronomer wishes to make an observation, he or she must first appropriately aim the telescope. Assume that the base of the telescope is fixed.
(a) Show that two independent rotations are sufficient to point the telescope in any direction.
(b) Astronomers face another problem when they want to observe a very distant or very faint object: they must take a photo that has been exposed over many hours. The Earth turns while this photo is being taken; thus the telescope must be continually re-aimed in order to keep it aligned with the targeted celestial body. Here is how such systems function: we install a central axis that is perfectly parallel to the axis of rotation of the Earth. The entire telescope assembly is free to rotate around this axis, and it is called the first axis (see Figure 3.11). For an observatory in the Northern Hemisphere, this axis is essentially lined up with the North Star, Polaris. At the North Pole itself this axis is vertical; otherwise, it is oblique. The telescope itself is mounted on a second axis whose angle between it and the first can be varied. Show that these two degrees of freedom are sufficient to point the telescope in any direction.
(c) Show that a rotation around the first axis is sufficient to keep the telescope aimed at the same celestial object as the Earth rotates.
(d) Show that at the 45th parallel, the angle between the axis of the Earth and the surface is 45 degrees.

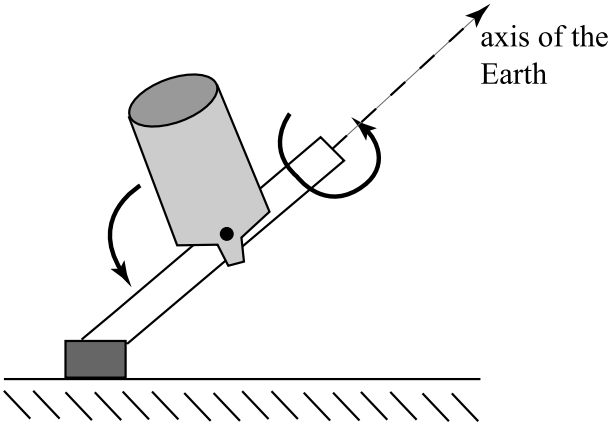


Fig. 3.11. The two degrees of freedom of a telescope (see Exercise 18).

References

[1] R. J. Schilling. *Fundamentals of Robotics.* Prentice Hall, 1990.

4

Skeletons and Gamma-Ray Radiosurgery

The concept of skeletons comes up in the discussion of optimal strategies for performing irradiative surgery, including "gamma knife" techniques ([4] and [5]). They are also an important concept in a variety of scientific problems. If this chapter is to be covered with three hours of theory and two hours of practical work, we recommend formulating the core problem of gamma-ray surgery. Follow this by covering both Sections 4.2 and 4.3, which discuss skeletons in both two and three dimensions with the help of simple examples. Time permitting, Section 4.4 can be discussed briefly in an informative mode. If you have a fourth hour at your disposal there is a choice to be made: there is sufficient time to discuss the numerical algorithms in Section 4.5 or the fundamental property of skeletons in Section 4.7. It may be preferable to concentrate on the algorithmic content for applied math students, for example, or on the fundamental property of skeletons for education majors. The rest of the chapter is enrichment and may be used as a departure point for a semester project.

4.1 Introduction

A "gamma knife" is a surgical device that is used for treating brain tumors. The machine focuses 201 beams of gamma-rays (originating from radioactive cobalt 60 sources distributed evenly around the inner surface of a sphere) into a single small spherical area. The region of intersection is subject to a strong dose of radiation. The beams are focused with the help of a helmet, and may produce focal regions of various sizes (2 mm, 4 mm, 7 mm, or 9 mm radius). Each size of dose requires the use of a different helmet; thus the helmet must be changed when the dose radius needs to be changed. Each helmet weighs roughly 500 pounds. Hence it is important to minimize the number of helmet changes.

The problem presented to mathematicians is to construct an algorithm to create optimal treatment plans, allowing the tumor to be irradiated in a minimum of time. This decreases the cost of the operation, while at the same time improving the quality

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_4,

of treatment for the patient, since long radiotherapy sessions can be quite unpleasant. The problem is quite simple for small tumors, since they can often be treated with a single dose. However, it becomes quite complex for large and irregularly shaped tumors. A good algorithm should be able to limit a treatment to a maximum of 15 individual doses. Similarly, it must be as robust as possible, which is to say that it must return acceptable (if not optimal) treatment plans for nearly all possible shapes and sizes of tumors.

It is easy to see that this problem is somewhat related to the problem of stacking spheres. We wish to fill (as much as possible) a region $R \subset \mathbb{R}^3$ with spheres in such a way that the proportion of volume not covered is less than some threshold of tolerance ϵ. If we use balls (or solid spheres) $B(X_i, r_i) \subset R$, $i = 1, \ldots, N$, with centers X_i and radii r_i, then the irradiated zone is $P_N(R) = \cup_{i=1}^{N} B(X_i, r_i)$. Letting $V(S)$ represent the volume of a region S, we wish to find balls such that

$$\frac{V(R) - V(P_N(R))}{V(R)} \leq \epsilon. \tag{4.1}$$

In order to find an optimal solution, the first task is to wisely choose the centers of the spheres. In fact, we must choose spheres that conform as much as possible to the surface of the region. By definition, these are spheres that have the most points of contact (points of tangency) with the boundary of the region. The centers of the spheres will then be taken along the "skeleton" of the region.

4.2 Definition of Two-Dimensional Region Skeletons

The *skeleton* of a region of $\mathbb{R}^2$ or $\mathbb{R}^3$ is a mathematical concept that is used in shape analysis and automatic shape recognition. We start by giving an intuitive definition.

Suppose that the region is formed of uniformly combustible material (for instance grass) and that we ignite the entire outer surface all at once. As the fire burns inward at a constant rate, it will eventually reach a point where there is no combustible material left. The skeleton of the shape is the set of points at which the fire goes out (see Figure 4.1).

We will return to this intuitive definition of the skeleton a little later, since it will be our guide to developing our intuition. First we will define the formal mathematical notion of skeleton. A *region* is an open subset of the plane $\mathbb{R}^2$ or space $\mathbb{R}^3$. Being open, a region does not include any of the points along its boundary, which we will denote by ∂R. The following definition is equally applicable to two- or three-dimensional regions. However, sometimes the terminology changes depending on the dimension; for example, we typically say "disk" to describe a filled circle in two dimensions, while we say "ball" to describe a solid sphere. In cases where the typical terminology varies, we will place the appropriate word for the three-dimensional definition in parentheses.

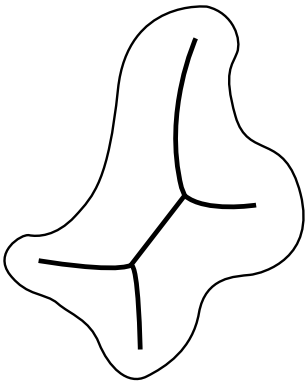


Fig. 4.1. The skeleton of a region.

Definition 4.1 *Let $|X - Y|$ denote the Euclidean distance between two points in the plane or space.*

Thus, if two points X and $Y \in \mathbb{R}^2$ have coordinates (x_1, y_1) and (x_2, y_2) respectively, then the distance between them is

$$|X - Y| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

Definition 4.2 *Let R be a region of $\mathbb{R}^2$ (or $\mathbb{R}^3$) and let ∂R be its boundary. The* skeleton *of R, denoted by $\Sigma(R)$, is the following set of points:*

$$\Sigma(R) = \left\{ X^* \in R \middle| \begin{array}{c} \exists X_1, X_2 \in \partial R \text{ such that } X_1 \neq X_2 \text{ and} \\ |X^* - X_1| = |X^* - X_2| = \min_{Y \in \partial R} |X^* - Y| \end{array} \right\}.$$

This definition is rather opaque; thus we will explain a few elements. The quantity $\min_{Y \in \partial R} |X^* - Y|$ gives the distance between a point X^* and the boundary ∂R of R. Unlike the distance between two points, there is no simple algebraic expression for this distance. Rather, it is expressed as the minimum of the function $f(Y) = |X^* - Y|$, expressed as a function of Y (X^* is constant). Thus, we are looking for the shortest line segment connecting X^* to any point on the boundary. The length of this shortest segment is $\min_{Y \in \partial R} |X^* - Y|$. In the case that R is a region in the plane, Figure 4.2 shows several of the possible line segments, with the shortest being indicated by a bold line.

Suppose that we draw a circle (a sphere) with center X^* and radius

$$d = \min_{Y \in \partial R} |X^* - Y|, \tag{4.2}$$

denoted by

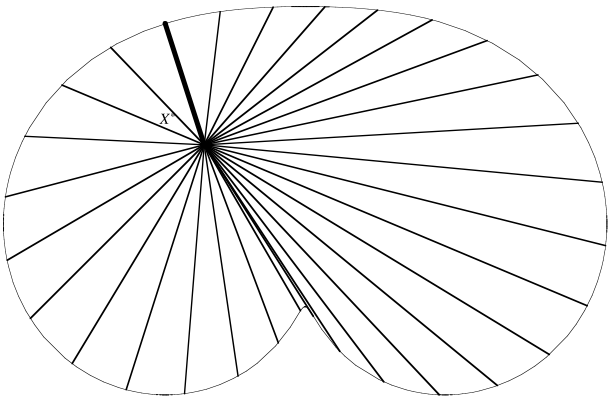


Fig. 4.2. Looking for the shortest distance between a point X^* and the boundary ∂R.

$$S(X, d) = \{Y \in \mathbb{R}^2(\text{or } \mathbb{R}^3) \mid |X - Y| = d\}.$$

In order for X^* to be in the skeleton $\Sigma(R)$, the above definition requires that $S(X^*, d)$ intersect ∂R at (at least) two points X_1 and X_2. Thus $S(X^*, d)$ and the boundary ∂R must have at least two points in common. Since the radius of $S(X^*, d)$ is precisely $\min_{Y \in \partial R} |X^* - Y|$, the interior of $S(X^*, d)$ is contained within R. To see this, choose a point Z in the complement $C(R)$ of the region (in other words, $C(R) = \mathbb{R}^2 \setminus R$ or $C(R) = \mathbb{R}^3 \setminus R$) and draw a line segment between X^* and Z. Since $X^* \in R$ and $Z \in C(R)$, the segment must cross the boundary ∂R at some point, which we will call Y'. By the definition of the distance between X^* and the boundary we have that

$$\min_{Y \in \partial R} |X^* - Y| \leq |X^* - Y'| < |X^* - Z|$$

and the point Z is outside of $S(X^*, d)$. Similarly, no points in the complement of R are in the interior of $S(X^*, d)$, and the interior of $S(X^*, d)$ consists entirely of points of R. If we define the disk (or ball) of center X and radius r by

$$B(X, r) = \{Y \in \mathbb{R}^2(\text{or } \mathbb{R}^3) \mid |X - Y| < r\},$$

then the elements X^* of the skeleton $\Sigma(R)$ satisfy

$$B(X^*, d) \subset R.$$

Even if the radius d is defined as a minimum (see (4.2)), it is also a maximum! It is the maximum radius such that a disk (or ball) centered at X^*, $B(X^*, r)$ lies completely within R. (All disks $B(X^*, r)$ with $r > d$ will contain a point Z in the complement $C(R)$ of R. To see this, draw the line segment between X^* and the nearest point X_1

on the boundary of R.[1] Then $|X_1 - X^*| = d$. If $r > d$ then the segment of length r originating at X^* and passing through X_1 will traverse the boundary ∂R and therefore contain a point outside of R.)

We have thus proved the following proposition, which gives us an alternative but equivalent definition for the skeleton of a region.

Proposition 4.3 *Let $X^* \in R$ and $d = \min_{Y \in \partial R} |X^* - Y|$. Then d is the maximum radius such that $B(X^*, d)$ lies completely within R, i.e., $d = \max\{c > 0 : \ B(X^*, c) \subset R\}$. The point X^* is in the skeleton $\Sigma(R)$ if and only if $S(X^*, d) \cap \partial R$ contains at least two points.*

At this point, it is clear that the distance $d = \min_{Y \in \partial R} |X^* - Y|$ plays a key role in the theory of skeletons. The following definition gives it a name.

Definition 4.4 *Let R be a region of the plane (or space). For each point X in the skeleton $\Sigma(R)$ of R, let $d(X)$ denote the maximum radius of a disk (or ball) centered at X such that it is contained within R. We know that*

$$d(X) = \min_{Y \in \partial R} |X - Y| = \max\{c > 0 : \ B(X, c) \subset R\}.$$

We present another definition, whose utility will soon become obvious.

Definition 4.5 *Let $r \geq 0$. The* r-skeleton *of a region R, denoted by $\Sigma_r(R)$, is the set of points of the skeleton $\Sigma(R)$ that are at least a distance r from the boundary of the region:*

$$\Sigma_r(R) = \{X \in \Sigma(R) | d(X) \geq r\} \subset \Sigma(R).$$

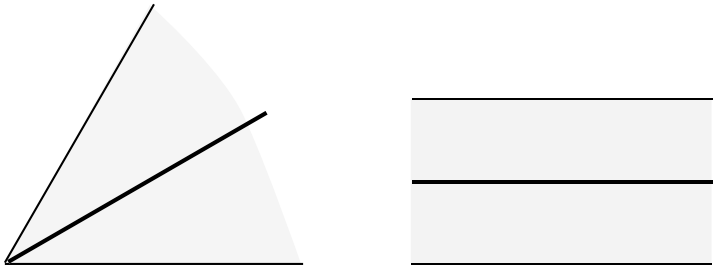
Observe that $\Sigma(R) = \Sigma_0(R)$.

Even with this reformulation, the definition of a skeleton is not easy to use in practice. It presupposes knowledge of the distance between all points in the interior of R to all points in its boundary. However, in its present form it can be used to determine the skeleton of simple geometric shapes. The following lemmas will prove useful.

Lemma 4.6 *1. Consider an angular region R bounded by two half-rays originating at the same point O. Then the skeleton of this region is the bisector of the angle formed by the two half-rays (Figure 4.3(a)).*

2. Consider a strip region R bounded by two parallel rays (D_1) and (D_2) separated by a distance h. Then the skeleton of this region is the parallel line that is equidistant to (D_1) and (D_2) (Figure 4.3(b)).

[1] More advanced readers may have observed that we have implicitly made several assumptions on R. Specifically, we are assuming that the boundary of R is piecewise continuously differentiable. Not to worry, the rest of you can continue to follow your intuition!



(a) Skeleton of an angular region.

(b) Skeleton of a strip region.

Fig. 4.3. The examples of Lemma 4.6.

PROOF. We will give the proof only for the angular region. Let P be a point of the skeleton, and consider Figure 4.4. By hypothesis it must be that $|PA| = |PB|$, since P is equidistant to the two sides of the region. Moreover, $\widehat{PAO} = \widehat{PBO} = \frac{\pi}{2}$. We need to show that $\widehat{POA} = \widehat{POB}$. To do this we will show that the two triangles POA and POB are congruent, by showing that they have three equal sides. Both triangles are

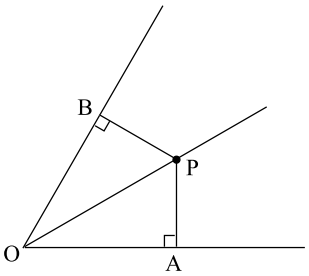


Fig. 4.4. Proof of Lemma 4.6.

right-angled. They both share the same hypotenuse $c = |OP|$. Moreover, $|PA| = |PB|$. Finally, by the Pythagorean theorem, it follows that

$$|OA| = \sqrt{c^2 - |PA|^2} = \sqrt{c^2 - |PB|^2} = |OB|.$$

Since the two triangles are congruent, we can then conclude that the corresponding angles $\widehat{POA}$ and $\widehat{POB}$ are equal. □

Lemma 4.7 *1. A line tangent to a circle O at a point P is perpendicular to the radius OP. As a consequence, if the circle is tangent to the boundary ∂R of a region R in the plane, then the center of the circle is situated along the normal of ∂R at P.*

2. Let P be a point on the circle $S(O,r)$. All lines passing through P other than the tangent line have a segment that lies within $B(O,r)$.

PROOF. To complete the proof we require a precise definition of a tangent line. Consider Figure 4.5. A line that it tangent to a circle at a point P is the limit of the secant lines passing through points A and B as both A and B approach the point P. Since $|OA| = |OB|$, the triangle OAB is isosceles. Thus we conclude that $\widehat{OAB} = \widehat{OBA}$. Since $\widehat{OAB} + \alpha = \pi$ and $\widehat{OBA} + \beta = \pi$, we conclude that $\alpha = \beta$. In the limit that A

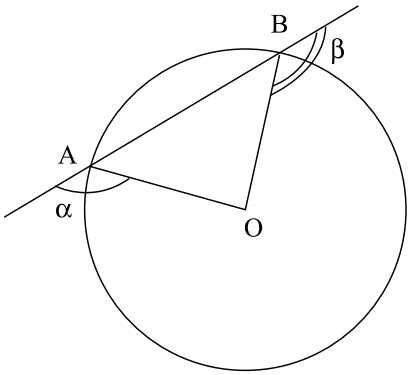


Fig. 4.5. A normal to a circle passes through the center of a circle.

and B approach a single point, the following two conditions hold:

$$\begin{cases} \alpha = \beta, \\ \alpha + \beta = \pi. \end{cases}$$

Thus it follows that $\alpha = \beta = \frac{\pi}{2}$ in the limit. The second part is left as an exercise for the reader. □

Example 4.8 (A rectangle.) *We will determine the skeleton of a rectangle R with base b and height h such that $b > h$. Using Lemma 4.6 we may construct six lines that possibly contain skeleton points of the rectangle by considering two of its sides at a time: the four bisectors, the horizontal parallel equidistant from the two horizontal sides, and the vertical parallel equidistant from the two vertical sides (see Figure 4.6).*

We can rapidly exclude (nearly) all points from the vertical parallel. Consider any point along this parallel that is inside the rectangle. Its distance to the vertical sides will always be greater than its distance to the nearest horizontal side, since $b > h$. Thus, except in the case of the point equidistant to the top and bottom, the circle of largest radius centered at the point will touch only one side of the rectangle. There is, however,

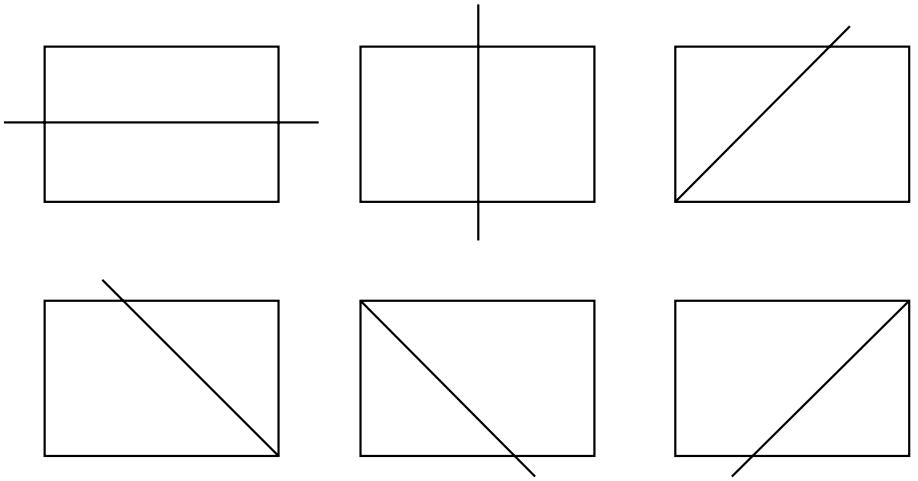
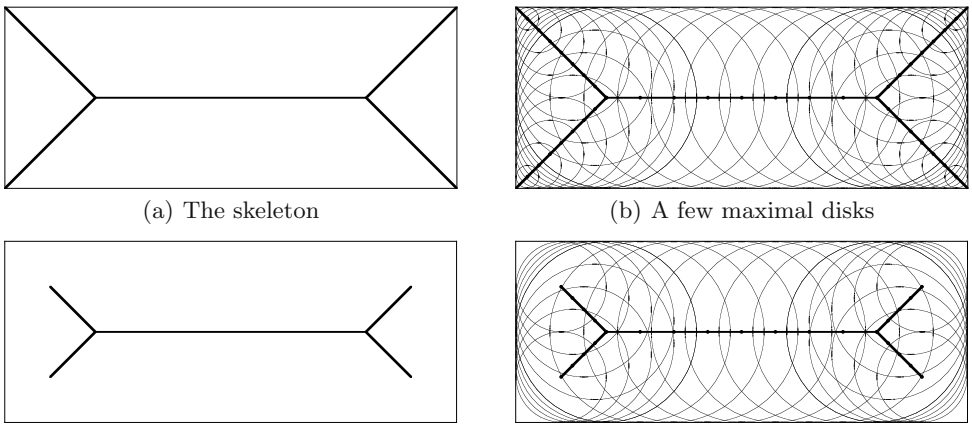


Fig. 4.6. The six lines that can possibly contain the skeleton points of a rectangle.

a segment I of the horizontal parallel that surely belongs to the skeleton. Once again, consider a point on this parallel inside the rectangle. The circle with radius $\frac{h}{2}$ centered at this point will touch both horizontal sides. As long as the point is not so close to one of the vertical sides that the circle of radius $\frac{h}{2}$ centered at that point falls partially outside the rectangle, then it will belong to the skeleton. Thus, it must be at least $\frac{h}{2}$ from the vertical sides. If the origin of the coordinate system corresponds to the bottom left corner of the rectangle, then the two disks of radius $\frac{h}{2}$ with three points of tangency are centered at $(\frac{h}{2}, \frac{h}{2})$ and $(b-\frac{h}{2}, \frac{h}{2})$. We have thus identified a segment that will belong to the skeleton of the rectangle: $I = \{(x, \frac{h}{2}) \in \mathbb{R}^2 \mid \frac{h}{2} \leq x \leq b - \frac{h}{2}\} \subset \Sigma(\text{rectangle})$. Through a similar argument it is relatively simple to convince ourselves that the segments of the bisectors from each corner to I will belong to the skeleton. The skeleton is thus the union of these five segments, as shown in Figure 4.7(a). A few maximal disks are shown in Figure 4.7(b).

Figure 4.7(c) shows an example of an r-skeleton, constructed with $r = \frac{h}{4}$. To obtain the $\frac{h}{4}$-skeleton, we kept only the centers of maximal disks with radius at least $\frac{h}{4}$. Thus, half of the points along each of the bisectors were discarded. The concept of r-skeletons is useful for the following reason: since the doses of radiation in an optimal treatment plan will be centered along the skeleton and the doses have a minimum radius r_0 ($r_0 = 2$ mm with current technology), then these doses will be centered at skeleton points at least a distance r_0 from the boundary. Hence, the doses of an optimal treatment plan will lie along the r_0-skeleton.

Before giving a second example, we return to the earlier intuitive definition of the skeleton, where we described it as the set of points where an inward-burning fire ex-



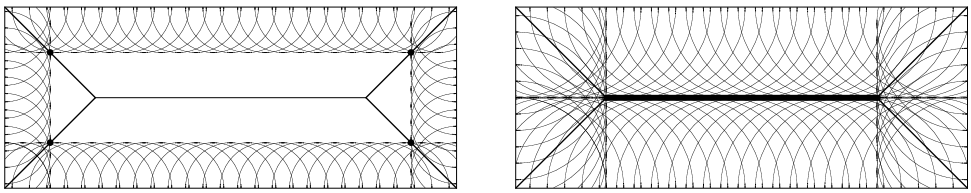
(a) The skeleton (b) A few maximal disks

(c) The $\frac{h}{4}$-skeleton (d) A few maximal disks of the $\frac{h}{4}$-skeleton

Fig. 4.7. The skeleton of a rectangle with base b greater than its height h.

tinguishes itself. Using this analogy, each point along the boundary is the location of a small fire. Each of them burns inward in all directions at a constant speed; thus at each instant in time, the leading edge of each fire is an arc of a circle. We say that a fire extinguishes itself at a point $X \in R$ if this point is reached simultaneously by more than one leading edge. Thus the relationship between this analogy and the formal definition is quite clear. Since X is first reached simultaneously by two leading edges emanating from the points X_1 and X_2 on the boundary, then X is the same distance from both of these points. Hence $|X_1 - X| = |X_2 - X| = \min_{Y \in \partial R} |Y - X|$, which is precisely the condition required to belong to the skeleton. Note that the condition we have chosen to describe, "points where the fire goes out," is only intuitive. For instance, when two fronts meet at a point X in the bisector of an angle of the rectangle, the fire goes out at this point but progresses along the bisector. Figure 4.8 shows the state of the fire at two instants, after having covered a distance $\frac{h}{4}$ in (a) and after having covered a distance $\frac{h}{2}$ in (b). The leading edges of several boundary points of R have been illustrated in both cases. Only the four points indicated in Figure 4.8(a) will burn out at this given moment in time. In contrast, Figure 4.8(b) shows the moment in time where the fire burns out along the entire interval I. The utility of this analogy is quite clear, and it will even allow us to determine the skeleton for any region bounded by a closed continuously differentiable curve.

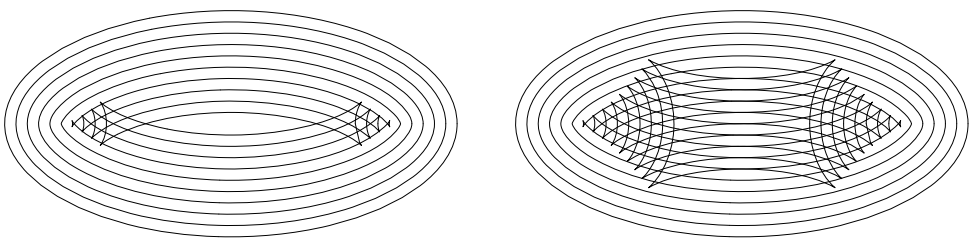
Remark. Even if the fire lit in one point burns in all directions, when we light the fire at all points of the boundary simultaneously we see the fire front advance at constant speed along the normal to the boundary. This comes from the fact that in the other directions, the fire goes out because it meets the fire coming from the other boundary points.



(a) After covering a distance of $\frac{h}{4}$ (b) After covering a distance of $\frac{h}{2}$

Fig. 4.8. Progress of a fire started along the boundary of a rectangle.

Example 4.9 (An ellipse) *We imagine lighting a fire along the entire boundary of an ellipse and observing this fire as it burns inward at a constant velocity. At every moment in time, the fire front advances along the normal line to its leading edge. With the use of mathematical software we have drawn the fire front at several moments in time, as illustrated in Figure 4.9. In the beginning, the fire front is a smooth rounded curve that resembles an ellipse (without being one). After the fire has progressed far enough, we note the appearances of sharp corners to the fire front; the points where the sharp corners first appear are precisely the first points where the fire will burn out.*



(a) (b)

Fig. 4.9. The advancing leading edge of a fire lit on the boundary of an ellipse.

Suppose that the ellipse is described by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

where $a > b$. Then we remark that the points where the fire will burn out are the points where the normal to the ellipse at the point (x_0, y_0) intersects the normal to the ellipse at the point $(x_0, -y_0)$. Due to symmetry, these are precisely the points where the normal lines intersect the x axis (the points are well defined for $y_0 \neq 0$). We wish to determine the set of such points. Let (x_0, y_0) be a point on the ellipse and consider the normal to this point. To do this, we consider the ellipse as the level set $F(x, y) = 1$ of the function

$$F(x,y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}.$$

The gradient vector

$$\nabla F(x_0,y_0) = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}\right)(x_0,y_0) = \left(\frac{2x_0}{a^2}, \frac{2y_0}{b^2}\right)$$

is normal to the ellipse at the point (x_0,y_0)*. (Recall that the gradient of a multivariate function is perpendicular to its level sets!) The normal to the ellipse at* (x_0,y_0) *is therefore the line passing through the point* (x_0,y_0) *in the direction* $\nabla F(x_0,y_0) = \left(\frac{2x_0}{a^2}, \frac{2y_0}{b^2}\right)$*. To find its equation we write that the vector* $(x-x_0, y-y_0)$ *is parallel to the vector* $\left(\frac{2x_0}{a^2}, \frac{2y_0}{b^2}\right)$*, yielding*

$$\frac{2y_0}{b^2}(x-x_0) - \frac{2x_0}{a^2}(y-y_0) = 0.$$

To find the point of intersection with the x *axis we substitute* $y=0$*, giving*

$$x = x_0 - \frac{b^2}{2y_0}\frac{2x_0y_0}{a^2} = x_0\left(1 - \frac{b^2}{a^2}\right) = x_0\frac{a^2-b^2}{a^2}.$$

(Observe that we have implicitly assumed $y_0 \neq 0$*.) If* $x_0 \in (-a,a)$ *then* $x \in \left(-\frac{a^2-b^2}{a}, \frac{a^2-b^2}{a}\right)$*. The skeleton is therefore the segment*

$$y = 0, \quad x \in \left[-\frac{a^2-b^2}{a}, \frac{a^2-b^2}{a}\right].$$

We have added the two extreme points because it is natural that the skeleton is a closed set. However, note that the maximal disk centered at each of these two extreme points touches the ellipse at only one point (one of its extremities along the x *axis). Despite this, these two points are justifiably included in the skeleton* $\Sigma(\text{ellipse})$*, on the basis that they are "multiple tangency points." This will be discussed in Exercise 16.*

It may seem natural to believe that the extreme points of the skeleton should correspond to the focal points of the ellipse, but we will show that this is not the case. To do this we will calculate the positions of the focal points. They are situated along the x *axis at the points* $(\pm c, 0)$*. They have the property that for any* (x_0,y_0) *of the ellipse, the sum of the distances from this point to the two focal points is constant. Consider the points* $(a,0)$ *and* $(0,b)$ *in particular. For* $(a,0)$*, the sum of the distances is*

$$(a+c) + (a-c) = 2a.$$

For the second point we find a sum of distances of

$$2\sqrt{b^2+c^2}.$$

We must have that $2a = 2\sqrt{b^2+c^2}$*, which yields*

$$c = \sqrt{a^2-b^2}.$$

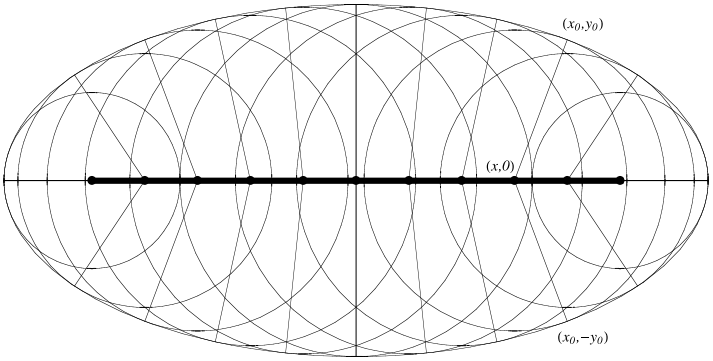


Fig. 4.10. The skeleton of an ellipse. The fire progresses along each line segment from the point of tangency of an inscribed maximal disk to the center of the disk located on $\Sigma(R)$.

4.3 Three-Dimensional Regions

The definition of skeletons given in two dimensions applies directly to three dimensions as well. However, we can distinguish different types of points of a three-dimensional skeleton based on the number of points of tangency between the corresponding maximal ball and the region boundary.

Definition 4.10 *Let R be a region of space and ∂R its boundary. The* linear portion *of the skeleton is defined as*

$$\Sigma_1(R) = \{X^* \in R \mid \exists\ X_1, X_2, X_3 \in \partial R \text{ such that } X_1 \neq X_2 \neq X_3 \neq X_1 \text{ and such that } |X^* - X_1| = |X^* - X_2| = |X^* - X_3| = \min_{X \in \partial R} |X^* - X|\}.$$

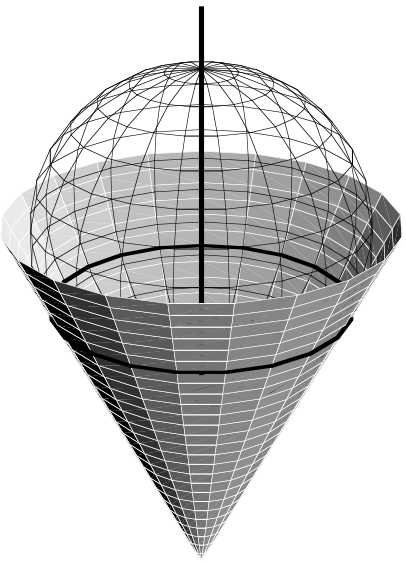
The surface portion *of the skeleton of R is*

$$\Sigma_2(R) = \Sigma(R) \setminus \Sigma_1(R).$$

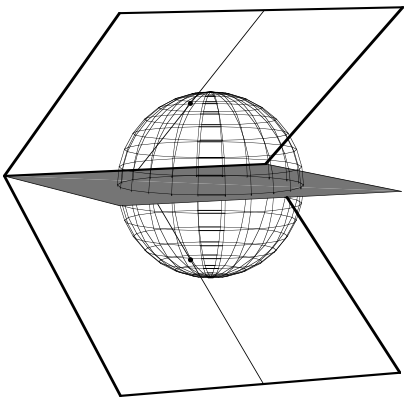
Example 4.11 (A circular cone) *A solid circular cone is described by the following set of points:*

$$\{(x, y, z) \in \mathbb{R}^3 \mid z > x^2 + y^2\}.$$

Any ball inside a cone with two points of tangency to the boundary must have an infinite number of points of tangency, and its center must lie along the central axis of the cone. The skeleton is therefore simply the positive z axis, $\Sigma(\text{cone}) = \{(0, 0, z), z > 0\}$, and contains only a linear part. As we will shortly see, this is a rather unique case. Figure 4.11(a) shows the boundary of a cone, its skeleton, and one maximal ball.



(a) The skeleton of a solid circular cone is given by its central axis



(b) The skeleton of an infinite wedge is given by the bisecting half-plane

Fig. 4.11. The skeletons of two simple regions. (a) While the region is the solid (filled) cone, only the boundary of the cone is shown, as well as one maximal ball and its circle of tangency. (b) An infinite wedge consists of all points between two half-planes emanating from a common axis. A maximal ball is shown with its two points of tangency.

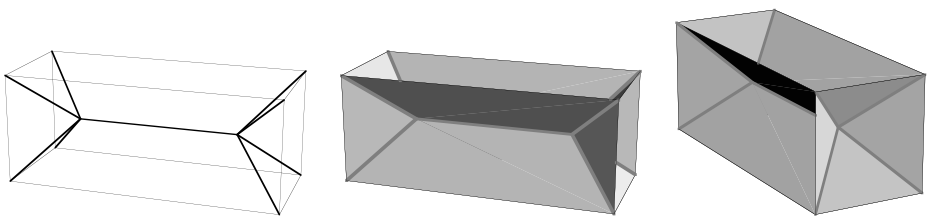
Example 4.12 (An infinite wedge) *Another simple geometric region is the infinite wedge formed by two half-planes emanating from a common axis. The skeleton of this region is the half-plane bisecting the dihedral angle between the bounding half-planes. In this case, the skeleton contains only a surface part. Figure 4.11(b) shows an infinite wedge and its skeleton. A maximal ball and its points of tangency have been indicated.*

The two preceding examples were intuitive and simple. However, neither of them is representative of typical regions. In fact, regions generally have both a linear and surface part. In many of these cases the linear part (or a portion of it) is the boundary of the surface part. We consider an example of this form.

Example 4.13 (A rectangular parallelepiped with two square faces) *We consider the parallelepiped region* $R = [0, b] \times [0, h] \times [0, h] \subset \mathbb{R}^3$ *where* $b > h$. *To simplify the example we have chosen two of the side lengths to be equal. As with our previous examples we must find all balls with at least two points of tangency to the boundary. By necessity, these points of tangency must be on distinct faces. A family of such balls will simultaneously touch the four faces with area* $b \times h$. *These maximal balls have radius*

$\frac{h}{2}$, and their centers lie on the segment $J = \{(x, \frac{h}{2}, \frac{h}{2}) \in \mathbb{R}^3, \frac{h}{2} \leq x \leq b - \frac{h}{2}\}$, which is a subset of the linear portion of the skeleton. Similar to maximal disks in the corner of a rectangle, each corner of R has a family of maximal balls with radius less than or equal to $\frac{h}{2}$ that touch the three adjoining faces. Thus, the linear portion of the skeleton consists of the segment J and the eight segments from the corners to the ends of J. This linear portion of the skeleton is shown in Figure 4.12(a).

We can decrease the radius of a maximal ball touching four faces and ensure that it remains in contact with two faces. Similarly, we can take a ball in contact with three faces in a corner and slide it toward another corner, all the while maintaining contact with two faces. The centers of these families of maximal balls are centered along polygons whose edges are either segments from the linear skeleton or edges of the parallelepiped. Each of these polygons is a portion of the half-plane bisectors between each pair of neighboring faces on R. Figure 4.12 presents the skeleton of R from two points of view. The linear part found earlier is found at the intersections between neighboring polygons.



(a) The linear part of the skeleton (b) The entire skeleton (c) A second view of the skeleton

Fig. 4.12. Skeleton of a rectangular parallelepiped with square faces ($b > h$).

These examples are far from being practical cases. Only computers can hope to tackle the complex regions typically encountered in surgical cases. However, since skeletons are an important concept in science (see Section 4.6), much research effort is focused on finding efficient algorithms for computing them numerically (see Section 4.5).

4.4 The Optimal Surgery Algorithm

In this section we will give an overview of an algorithm for optimal dose planning in gamma-ray surgery. It is based on dynamic programming techniques ([5] and [4]).

To begin with, we recall that we are not required to irradiate the entire region, but only a fraction $1 - \epsilon$ of it (see (4.1)). Why don't we need to irradiate the entire region? The radiation is delivered by focusing an array of 201 beams to a spherical target. However, due to the fact that the overlapping beams come from all directions, it is

clear that the area immediately around the target also receives a relatively large dose of radiation. Experience has shown that we do not need overlapping doses that completely cover the region, provided that neighboring doses are sufficiently close together. Also, it bears repeating that we are only looking for a "reasonably optimal" solution. We are also limited by the four sizes available for the individual doses.

The basic idea of a dynamic programming algorithm is to find the solution step by step, rather than looking for the entire solution at once.

The underlying idea. Suppose that an optimal solution for a region R is given by

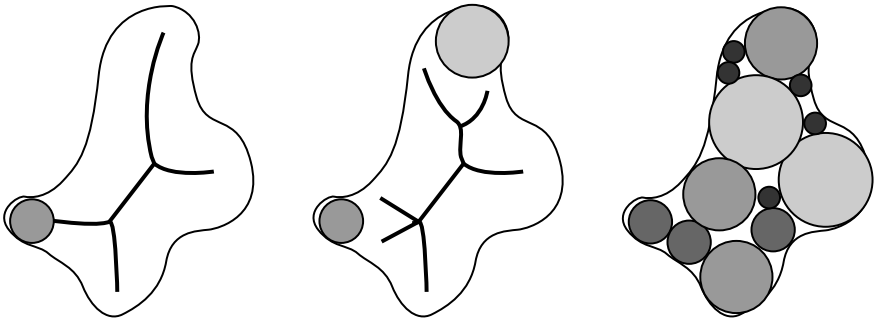
$$\cup_{i=1}^{N} B(X_i^*, r_i).$$

Then if $I \subset \{1, \ldots, N\}$, we must have that $\cup_{i \notin I} B(X_i^*, r_i)$ is an optimal solution for $R \setminus \cup_{i \in I} B(X_i^*, r_i)$ (see Exercise 8).

Although seemingly naive, this concept is very powerful. It allows us to apply an iterative process: rather than determining the entire solution at once, we start with a reasonably optimal initial dose over a subset of the region and optimally plan one dose at a time.

Choosing the first dose. Any dose in an optimal solution must be centered along the skeleton of the region. Recall that the doses may have only one of four sizes $r_1 < r_2 < r_3 < r_4$ and that it is therefore natural to consider r_i-skeletons. Consider a planar region. The initial dose should be placed at an extreme point of one r_i-skeleton or at a point of intersection between various branches of the skeleton (Figure 4.13). (For a three-dimensional region, the equivalent to a point of intersection between various branches is any point along the linear part of the skeleton. It is even possible for there to be points of intersection between branches of the linear part of the skeleton, at which points the maximal ball has at least four points of tangency.) A dose of radius r_i centered at an extreme point of the r_i-skeleton optimally fills a chunk on the boundary of the region. One centered at a point of intersection will irradiate a disk that has at least three points of tangency with the boundary. How do we choose between these two alternatives? In order to cover the region with fewer doses, we favor using larger radius doses. But we have only a small set of sizes to choose from. The second choice is good if we can choose a point of intersection X that can support a reasonable radius: that is, we want the radius $d(X)$ of the maximal ball at point X to be relatively close to one of the r_i. If this is not possible, then we opt for the first choice. In this case, we need to choose an adequate r_i, $i = 1, \ldots, 4$. This is largely dependent on the shape of the boundary at the extreme point. If it is somewhat pointed or narrow, we will need to choose a smaller radius to ensure that the nonirradiated area is not too far from the irradiated one (see Figure 4.14). In contrast, if it is well rounded, then we can choose a larger radius while ensuring adequate coverage.

The rest of the algorithm. Once we have found an initial dose $B(X_1^*, r_1)$ we simply iterate the process. We consider the region $R_1 = R \setminus B(X_1^*, r_1)$, determine its skeleton,

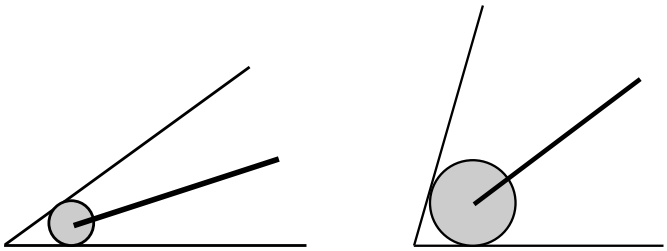


(a) A first dose of radius 4 mm

(b) The skeleton of the remaining region after two doses of radii 4 and 7 mm

(c) The entire region irradiated with doses of radii 2, 4, 7 and 9 mm

Fig. 4.13. Different stages in the irradiation of the region from Figure 4.1.



(a) A small radius

(b) A larger radius

Fig. 4.14. Choosing the radius for a dose centered at an extreme point of a skeleton.

and look for a reasonably optimal dose in the same manner as just described. The tolerance threshold allows us to decide when to stop. If we want to improve the results of the algorithm we can do so by exploring several initial doses, at each step considering a few of the next possible dosage placements.

4.5 A Numerical Algorithm for Finding the Skeleton

It is a nontrivial problem to develop a good algorithm for finding the skeleton of a region. We limit ourselves to discussing the problem in two dimensions. We will take for granted (without proof) that the skeleton of a simply connected region (a single piece without holes) is a particular type of graph: a tree.

The formal definition of a graph varies throughout the literature. In this section we will consider undirected graphs, defined as follows.

Definition 4.14 *1. An (undirected) graph consists of a set of nodes* $\{S_1, \ldots, S_n\}$ *and a set of edges between them. For each distinct pair of nodes* $\{S_i, S_j\}, 1 \leq i < j \leq n$, *we may have at most one edge between them.*

2. We say that two graphs are equivalent *if the following two conditions are satisfied:*

- *we have a bijection* h *between the nodes of the first graph and those of the second;*
- *there is an edge joining nodes* S_i *and* S_j *in the first graph if and only if there is one joining* $h(S_i)$ *and* $h(S_j)$ *in the second.*

Definition 4.15 *1. A graph is* connected *if for all pairs of nodes* S_i *and* S_j, *there exists a sequence of nodes* $S_i = T_1, T_2, \ldots, T_k = S_j$ *such that each pair* $\{T_l, T_{l+1}\}$ *is connected by an edge. In other words, there exists a* path *between every pair of nodes in the graph.*

2. A path $T_1, \ldots, T_k$ *is said to be a* cycle *if* $T_1 = T_k$ *and* $T_i \neq T_j$ *otherwise.*

3. A graph that contains no cycles is called a tree.

We will numerically test to see whether interior points of a region are part of the skeleton. Numerical errors can lead to two problems:

(i) Due to missing certain points that should be included, the skeleton may not be connected.

(ii) Due to falsely including certain points, the skeleton may include extra branches.

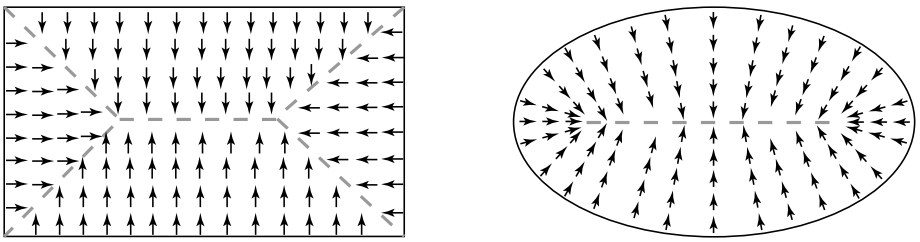
In both of these cases the "topology" of the skeleton has been altered. Thus, it is important to develop a robust algorithm that does not introduce such defects. We describe an algorithm from [2].

The algorithm consists of two parts: the first part makes use of the inward burning fire analogy. The fire propagates along the flow lines in a vector field. This allows the approximate determination of points in the skeleton as points of discontinuity of the vector field along the advancing fire front, but it still suffers from the above errors. The second part of the algorithm seeks to eliminate these errors while preserving the underlying topology of the skeleton.

4.5.1 The First Part of the Algorithm

We consider the analogy of an inward-burning fire lit simultaneously along the entire boundary ∂R. At every point X along the boundary ∂R, the fire will burn inward at a constant velocity (which we will assume equal to 1 unit of distance per unit of time) along the normal vector to the boundary at X. Each point X in the interior of the region will be consumed by the fire originating from a point $X_b \in \partial R$ such that X lies along the normal line through X_b. Thus, when the fire reaches the point X it will continue to travel along the direction of the vector $X - X_b$ at constant speed. Hence

each interior point X may be associated with a vector $V(X)$, the speed vector, creating a vector field over the interior of R (see Figure 4.15). The speed vector $V(X)$ has its origin at X, the direction $X - X_b$, and length one. We must be careful: if a point X is at the intersection of several normal lines to ∂R and at the same distance from the boundary along these normal lines, then $V(X)$ is undefined. Thus $V(X)$ is undefined at points in $\Sigma(R)$ and discontinuous around them. This is the property that we will use to detect points belonging to the skeleton.



(a) A rectangle (b) An ellipse

Fig. 4.15. The vector field $V(X)$ and the skeleton (in dashed lines) for various regions.

Doing this will require the ability to analytically manipulate the vector field $V(X)$. We introduce the function

$$d(X) = \min_{Y \in \partial R} |X - Y|, \tag{4.3}$$

which returns the distance between the point X and the boundary ∂R. Observe that for points along $\Sigma(R)$ the function coincides with the function d introduced in Definition 4.4. This is a two-dimensional function, depending on the coordinates of X. We will show that $V(X) = \nabla d(X)$.

Definition 4.16 *(1) Let U be an open set in $\mathbb{R}^n$ and $r \geq 1$. A function $F = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ is of class C^r (or simply F is C^r) if for all $(i_1, \ldots, i_r) \in \{1, \ldots, n\}^r$ and for all $j \in \{1, \ldots, m\}$ the partial derivative $\frac{\partial^r f_j}{\partial x_{i_1} \cdots \partial x_{i_r}}$ exists and is continuous. In the case $r = 1$ we also say that the function is continuously differentiable.*

(2) We say that a curve $\mathcal{C}$ in $\mathbb{R}^2$ is of class C^r if for every point X_0 on $\mathcal{C}$ there exist an open neighborhood U of X_0 and a function $F : U \to R$ of class C^r such that $\mathcal{C} \cap U = \{X \in U | F(X) = 0\}$ and the gradient of F does not vanish on U.

Proposition 4.17 *Let R be a region such that ∂R is of class C^2. Then the function $d(X)$ is of class C^1 over the points $R \setminus \Sigma(R)$ and the field $\nabla d(X)$ is continuous on the same set. Moreover, if ∂R is of class C^3, then $\nabla d(X)$ is of class C^1 over $R \setminus \Sigma(R)$.*

The proof of this proposition makes use of the implicit function theorem, which is quite advanced. In order to continue with our discussion of the algorithm we defer this proof to Section 4.5.3. You can decide to accept the proposition without proof and to continue with the rest of the algorithm, which is more elementary. In particular, we concentrate on a useful consequence of this result.

Proposition 4.18 *At a point $X \in R \setminus \Sigma(R)$ the vector field $V(X)$ is given by the gradient $\nabla d(X)$ of the function $d(X)$ defined in* (4.3)*. It is a vector of unit length.*

Proof. Consider a point $X_0 \in R \setminus \Sigma(R)$. Then $B(X_0, d(X_0)) \subset R$ and $S(X_0, d(X_0))$ is tangent to ∂R at a single point X_1. The gradient of $d(X)$ at X_0, $\nabla d(X_0)$, is oriented in the direction where the rate of increase of $d(X)$ is the largest. We will convince ourselves that this direction is the inward-pointing normal to ∂R, namely the direction of the line from X_1 to X_0. In fact, the directional derivative of d along the direction of a given unit vector $\mathbf{u}$ is given by $\langle \nabla d(X_0), \mathbf{u} \rangle$, where $\langle ., . \rangle$ is the scalar product. The boundary ∂R in the neighborhood of X_1 can be imagined as an infinitesimally small line segment parallel to the tangent vector $\mathbf{v}(X_1)$ to the boundary at X_1. Indeed, because X_0 is not on the skeleton, for points X in the neighborhood of X_0 then $d(X) = |X - X_2|$ with X_2 in the neighborhood of X_1, so we can forget the other parts of the boundary. Thus, if we move X_0 in a direction parallel to $\mathbf{v}(X_1)$, then the directional derivative of $d(X_0)$ in this direction will be zero, since the function d is constant. Hence $\nabla d(X_0)$ is orthogonal to $\mathbf{v}(X_1)$, and therefore $\nabla d(X_0)$ is a scalar multiple of $X_0 - X_1$. The length of the vector $\nabla d(X_0)$ is given by the directional derivative of $d(X)$ at X_0 in the direction of $X_0 - X_1$. Along this line we have that $d(X) = |X - X_1|$ as long as X is not a point on the skeleton. Since we can assume that X_1 is constant, it is easy to perform the calculation, yielding $\nabla d(X_0) = \frac{X_0 - X_1}{|X_0 - X_1|}$, which has the expected length 1. □

Definition 4.19 *We consider a vector field $V(X)$ defined on a region R, and a circle $S(X_0, r)$ parameterized by $\theta \in [0, 2\pi]$, $X(\theta) = X_0 + r(\cos\theta, \sin\theta)$, such that the disk $B(X_0, r)$ lies within R. Let $N(\theta) = (\cos\theta, \sin\theta)$ be the unit vector normal to $S(X_0, r)$ at $X(\theta)$. The flux of the field $V(X)$ along the circle $S(X_0, r)$ is given by the line integral*

$$I = \int_0^{2\pi} \langle V(X(\theta)), N(\theta) \rangle \, d\theta, \tag{4.4}$$

where $\langle V(X(\theta)), N(\theta) \rangle$ represents the scalar product between $V(X(\theta))$ and $N(\theta)$.

Lemma 4.20 *The flux of a constant vector field $V(X) = (v_1, v_2)$ along a circle $S(X_0, r)$ is zero.*

Proof.

$$\begin{aligned} I &= \int_0^{2\pi} \langle V(X(\theta)), N(\theta) \rangle \, d\theta \\ &= \int_0^{2\pi} (v_1 \cos\theta + v_2 \sin\theta) \, d\theta \\ &= (-v_1 \sin\theta + v_2 \cos\theta)\big|_0^{2\pi} \\ &= 0. \end{aligned}$$

□

Lemma 4.20 gives us the key to finding approximate skeleton points. In fact, when we are at a point X far from the skeleton, the vector field in a small neighborhood of X is approximately constant. Thus, the flux along a small circle around X will be very small. Similarly, we can convince ourselves that the flux will be much larger when the disk contains skeleton points (see Example 4.21 below).

This gives us a test for finding skeleton points: in order to decide whether a point $X \in R$ is on the skeleton we calculate (4.4) along a small circle containing X and lying within R. If the value of this integral is below a certain threshold, then we conclude that X is not on the skeleton. If it exceeds the threshold, we conclude that the disk probably contains some skeleton points, and we refine our search within the disk.

Example 4.21 *At sufficiently small scales, the curves forming the skeleton look like small line segments. Consider the case in which a portion of the skeleton is a line segment along the x axis. Then we can verify that the field $V(X) = \nabla d(X)$ is given by*

$$V(x,y) = \begin{cases} (0,-1), & y > 0, \\ (0,1), & y < 0. \end{cases}$$

If we consider a circle $S(X_0, r)$ centered along the x axis, we find that

$$I = \int_0^{\pi} -\sin\theta \, d\theta + \int_{\pi}^{2\pi} \sin\theta \, d\theta = -4.$$

We can verify that the integral remains nonzero if the circle is not centered on the axis but still contains a portion of the x axis (the calculation is a little more difficult, however). Similarly, we can show that the value of the integral diminishes continually as the center of the circle gets further from the x axis.

Practical implementation of the first part. Suppose that the function d in (4.3) and its gradient have already been calculated. The region R is identified by a set of pixels, and for each one we must decide whether it belongs to the skeleton. Take a pixel P within R, and consider its eight neighboring pixels (those that share a common side or corner), as shown in Figure 4.16(a). Let δ be the side length of a pixel. Consider a circle $S(P, \delta)$ centered at P with radius δ, and take the eight points P_i dividing the circle into eight equal arcs such that point P_i falls within pixel i. We calculate the unit vector N_i normal to $S(P, \delta)$ at P_i. We approximate (up to a constant) the integral of (4.4) with the discrete sum

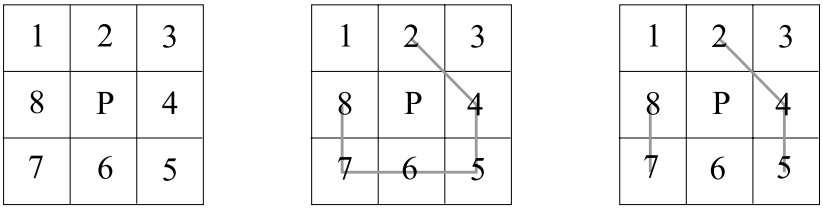
$$\overline{I}(P) = \frac{2\pi}{8} \sum_{i=1}^{8} \langle N_i, \nabla d(P_i) \rangle.$$

The point P is a candidate to be removed if $|\overline{I}(P)| < \epsilon$, where ϵ is an appropriately chosen threshold. If the threshold is sufficiently high, then all of the spurious branches

of the skeleton will be removed. However, if it is too high, we risk removing actual skeleton points and ending up with a skeleton in several disjoint pieces.

4.5.2 Second Part of the Algorithm

How do we prevent the skeleton from fracturing? How can we ensure that the skeleton remains a tree? To do this we construct the skeleton in small steps. For each pixel we decide whether it is in the skeleton. We proceed slowly by removing those points determined not to be in the skeleton. Starting at the boundary, we proceed layer by layer until at the end we are left with only the skeleton (or more precisely, a thickened skeleton visible on screen). Each time we remove a pixel, we ensure that the remaining pixels remain connected and that the implied graph does not contain any cycles.



(a) The eight neighbors of P

(b) We remove P

(c) We do not remove P

Fig. 4.16. The eight neighboring pixels P and the graphs allowing us to decide whether we remove P.

Practical implementation of the second part. We begin by deciding that the pixels along the boundary do not belong to the skeleton. We analyze then the inner pixels one at a time, staring from the boundary. For a given pixel P we begin by calculating $\overline{I}(P)$. If $|\overline{I}(P)| < \epsilon$, then the pixel is a candidate to be removed. In order to decide whether we remove this pixel, we consider its eight neighbors as shown in Figure 4.16(a). If none of the other neighbors of P have been removed, then we do not remove P, since this would create a hole. If some of the neighbors have been removed, then we construct a graph over the remaining neighbors. We connect pixels i and j with an edge if pixels i and j share either an edge or a corner. The possible pairs of connected neighbors are $(1,2)$, $(2,3)$, $(3,4)$, $(4,5)$, $(5,6)$, $(6,7)$, $(7,8)$, $(8,1)$, $(2,4)$, $(4,6)$, $(6,8)$, and $(8,2)$. We want to ensure that we do not have any cycles in this graph. Such cycles will be given by the following triplets of edges:

$$\begin{cases} \{(1,2),(8,1),(8,2)\}, \\ \{(2,3),(3,4),(2,4)\}, \\ \{(4,5),(5,6),(4,6)\}, \\ \{(6,7),(7,8),(6,8)\}. \end{cases}$$

If any of these triplets are present, then we remove the cycle by removing the diagonal edge from the triplet. For example, we would replace the triplet of edges $\{(1,2),(8,1),(8,2)\}$ by the pair $\{(1,2),(8,1)\}$. Once we have constructed the graph over the remaining neighbors of P we will remove P if and only if this graph is a tree (see Figures 4.16(b) and (c)). In this way, neither do we cut the skeleton into disjoint pieces, nor do we create holes in it. Once we have decided for P we study the next pixel in the same manner. As a note, an efficient method for testing whether this graph is a tree is explored in Exercise 15.

Remark. This method can be generalized to deal with three-dimensional regions.

4.5.3 Proof of Proposition 4.17

Recall that Proposition 4.17 stated that if R is a region such that ∂R is of class C^2 (respectively C^3), then the function $d(X)$ is of class C^1 (respectively C^2) over the points $R \setminus \Sigma(R)$, and the field $\nabla d(X)$ is continuous (respectively of class C^1) on $R \setminus \Sigma(R)$. In order to show this, we will have to "calculate" $d(X)$. This can be done using the implicit function theorem, which we state without proof:

Theorem 4.22 *Let $F = (f_1, \ldots, f_n) : U \to R^n$ be a function of class C^r, $r \geq 1$, defined over an open set $U \subset \mathbb{R}^{n+k}$. We represent the points in U as pairs (X, Y), where $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^k$, and we write $X = (x_1, \ldots, x_n)$. Let $(X_0, Y_0) \in U$ be such that $F(X_0, Y_0) = 0$ and such that the partial Jacobian matrix*

$$J(X_0, Y_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} (X_0, Y_0)$$

is invertible. Then there exist a neighborhood V of Y_0, a unique function $g : V \to \mathbb{R}^n$, and a neighborhood W of (X_0, Y_0) such that

(i) g is of class C^r on V and its graph lies within W.
(ii) $g(Y_0) = X_0$.
(iii) For $(X, Y) \in W$ it follows that $F(X, Y) = 0$ if and only if $X = g(Y)$.

Proof of Proposition 4.17. Let $X_0 = (x_0, y_0)$ be a point of R that is not on the skeleton. Its distance from the boundary is given by $d(X_0) = |X_0 - X_1|$, where $X_1 = (x_1, y_1)$ is a point of ∂R such that the vector $X_0 - X_1$ is normal to ∂R. We

wish to show that $d(X)$ is C^1 in a neighborhood of X_0. The biggest difficulty is in calculating $d(X)$. To do this we must identify the boundary point $Y = (\overline{x}, \overline{y})$ that is closest to a point $X = (x, y)$. We will find it using the implicit function theorem. We can suppose that the boundary is the level curve $f_1(Y) = 0$ of a function f_1 of class C^2 with values in $\mathbb{R}$ by Definition 4.16(2). We must have that the vector $X - Y$ is normal to the boundary at Y. Since the normal vector has the same direction as the gradient $\nabla f_1(Y)$ of f_1, then the vector $X - Y$ must be parallel to $\nabla f_1(Y)$, which may be written as

$$f_2(\overline{x}, \overline{y}, x, y) = \begin{vmatrix} \overline{x} - x & \frac{\partial f_1}{\partial \overline{x}}(Y) \\ \overline{y} - y & \frac{\partial f_1}{\partial \overline{y}}(Y) \end{vmatrix} = 0.$$

We are looking for solutions to $F(\overline{x}, \overline{y}, x, y) = 0$ with $F = (f_1, f_2)$. If f_1 is of class C^2, then f_2 and therefore F are of class C^1. By the implicit function theorem (Theorem 4.22), the solutions to $F = 0$ will be given by a unique function $(\overline{x}, \overline{y}) = g(x, y) = g(X)$ of class C^1 if we can show that

$$J(x_1, y_1, x_0, y_0) = \begin{pmatrix} \frac{\partial f_1}{\partial \overline{x}} & \frac{\partial f_1}{\partial \overline{y}} \\ \frac{\partial f_2}{\partial \overline{x}} & \frac{\partial f_2}{\partial \overline{y}} \end{pmatrix} (x_1, y_1, x_0, y_0)$$

is invertible. We have

$$\begin{aligned} &J(X_1, X_0)^t \\ &= \begin{pmatrix} \frac{\partial f_1}{\partial \overline{x}}(X_1) & \frac{\partial f_1}{\partial \overline{y}}(X_1) - (y_1 - y_0)\frac{\partial^2 f_1}{\partial \overline{x}^2}(X_1) + (x_1 - x_0)\frac{\partial^2 f_1}{\partial \overline{x}\partial \overline{y}}(X_1) \\ \frac{\partial f_1}{\partial \overline{y}}(X_1) & -\frac{\partial f_1}{\partial \overline{x}}(X_1) + (x_1 - x_0)\frac{\partial^2 f_1}{\partial \overline{y}^2}(X_1) - (y_1 - y_0)\frac{\partial^2 f_1}{\partial \overline{x}\partial \overline{y}}(X_1) \end{pmatrix}. \end{aligned} \tag{4.5}$$

What does the condition $\det(J(x_1, y_1, x_0, y_0)) = 0$ signify? It is precisely the condition under which the circle $S(X_0, |X_1 - X_0|)$ has a contact of order greater than 1 at X_0, as explored in Exercise 16. Such a point corresponds to an extreme point of the skeleton. We leave the rather delicate proof of this fact to Exercise 17. (A change of variables allows us to consider the easier case of $f_1(\overline{x}, \overline{y}) = \overline{y} - f(\overline{x})$ for a function f of class C^2.) Thus, if X_0 is not on the skeleton, then $J(X_1, X_0)$ is invertible. This ensures the existence of g of class C^1.

We now know that $d(x, y) = |X - g(X)|$ is of class C^1. Thus ∇d is continuous. Similarly, had we had supposed that f_1 is of class C^3, then we would have obtained that ∇d is of class C^1. □

Remark on the proof: Examine the structure of the proof a little further. We started by taking a point $X_0 \in R \setminus \Sigma(R)$. This hypothesis was used only to affirm the existence of a unique point X_1 on the boundary of R closest to X_0. This is also true for extreme points of the skeleton, such as those of the ellipse (see Example 4.9). We wish to show that for each X in a neighborhood of X_0 there exists a unique closest point Y on the boundary of the region. However, this property does not hold for extreme points of the skeleton. In fact, such extreme points have neighbors on the skeleton whose minimum distance to the boundary is realized by more than one point on the boundary. This

obstruction is reflected by the fact that the Jacobian (4.5) vanishes at the extreme points of the skeleton.

Remark on the utility of Proposition 4.17: We have shown many regions whose boundaries are continuous, but only piecewise C^3 (for example, any polygonal region). In these cases the hypothesis of Proposition 4.17 is not satisfied. We could lightly round the corners of such a region so that the boundary of the modified region would be C^3 and the result would apply. We need only convince ourselves that the "rounding" of the boundary will not significantly alter the skeleton of the region. (See Exercise 18.)

4.6 Other Applications of Skeletons

Skeletons in morphology: The notion of the skeleton of a region was first introduced in a biological context by Harry Blum [1] in order to describe the forms of organisms in nature, or *morphology*. Blum called the skeleton the "axis of symmetry" of the form. More specifically, when biologists wish to describe a form they are actually more interested in describing the differences between the forms of two different species. Even within a species there is a large amount of variability in the form of individual organisms. Thus, biologists are interested in finding the *characteristic* properties of the form of *all* individuals of the same species. Recall, for example, that the skeleton of a planar region is a graph (see Definition 4.14). The properties of this graph can be used to describe the form of a species if the graphs of all individuals are equivalent. In that case we say that the graph of the skeleton is an "invariant" of the species.

We may associate a graph to the skeleton of a planar region in the following manner: extremal points and points of intersection of branches of the skeleton become nodes; two nodes are connected by an edge if the points they represent are connected by a portion of a skeleton not containing any other nodes.

In the morphological analysis of planar regions, we are interested in differentiating between forms whose skeleton graphs are not equivalent. Blum's idea was to define a new type of geometry adapted to describing natural shapes and based on the notion of points and "growth." Inward growth from the boundary leads naturally to the definition of the skeleton. Outward growth from the skeleton, coupled with the associated distance function $d(X)$, regenerates the original form.

Blum's ideas are powerful enough that we reserve Section 4.7 for their discussion. We will describe a region not by its boundary, but by its skeleton and the thickness of the region surrounding it. This constitutes the fundamental property of the skeleton.

Some other applications: The concept of the skeleton has long been known by physicists. It appears naturally in the study of wave fronts, particularly in the field of geometric optics. As an example, physicists have long known that the skeleton of an ellipse is a straight line segment.

Skeletons also arise in the study of the shapes of sand dunes. Since sand dunes have roughly constant slopes, the projection of the summit edge onto the base is roughly the skeleton of the base [3].

Skeletons are currently a commonly used concept in the world of three-dimensional modeling. Given a curve in space $X(t) = (x(t), y(t), z(t))$, $t \in [a, b]$, and for each point along the curve a radius $d(t)$, then a volume is described by the union of the balls $B(X(t), d(t))$ along the curve. This volume is in some sense a generalized cylinder, whose axis is a curve rather than a straight line and whose radius is variable. In three-dimensional modeling one tries to approximate a given volume by a finite number of such generalized cylinders. It is relatively easy to see that such a representation provides an economical way of describing complicated volumes.

4.7 The Fundamental Property of the Skeleton of a Region

We will characterize the points in the skeleton of a region R through a fundamental property. All of the proofs in this section will be intuitive, since we will suppose that the boundary ∂R of R possesses a tangent at each point. It is possible to generalize the theorem to less well behaved regions, but at the expense of complicating the proofs.

We define the notion of a maximal disk (ball) in a region R of $\mathbb{R}^2$ ($\mathbb{R}^3$). We will show that skeleton points are precisely the centers of maximal disks (balls).

Definition 4.23 *Let R be a region of the plane $\mathbb{R}^2$ (of the space $\mathbb{R}^3$). Let $B(X, r)$ denote a disk (a ball) of radius r centered at X. Then $B(X, r)$ is* maximal *with respect to the region R if $B(X, r) \subset R$ and $B(X, r)$ is not itself included in any disk (ball) included in R.*

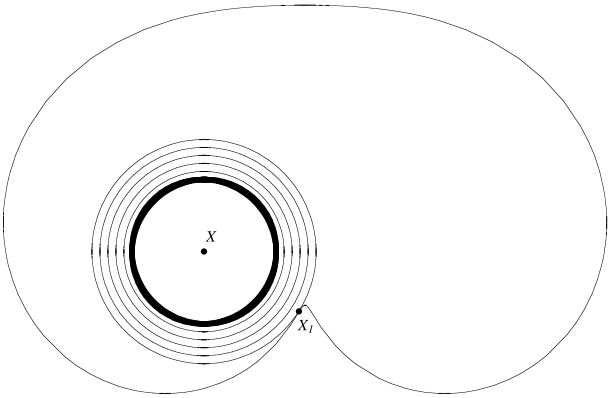
We develop some intuition for this new concept in exploring the following proposition.

Proposition 4.24 *All points X of a region R belong to a maximal disk.*

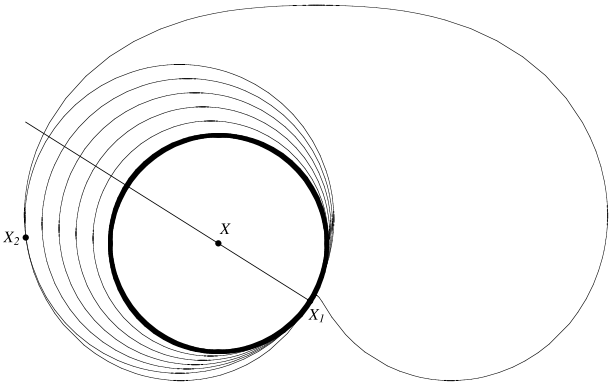
Proof. We give the proof in the case of a region of the plane $\mathbb{R}^2$, and invite the reader to generalize this to higher dimensions.

To do this we will imagine "inflating" a disk around the point X until it is maximal.

Since X is in the interior of R, we can choose a sufficiently small radius ϵ such that the disk $B(X, \epsilon)$ is completely contained in R. We increase the radius of this disk until it touches the boundary of the region. At this point, the radius of the disk is now $\min_{Y \in \partial R} |X - Y|$. A few of the steps in this inflation process are shown in Figure 4.17(a). The initial disk $B(X, \epsilon)$ is shown with a thick line, while several subsequent disks are shown in fine lines. The first point of contact X_1 with the boundary is indicated. The line through X and X_1 contains a diameter of the circle and is normal to the tangent



(a) We increase the radius of the disk until it touches the boundary.



(b) We retreat the center of the disk until it is tangent to ∂R at no fewer than two points.

Fig. 4.17. Constructing a maximal disk in two steps.

of the circle at X_1. Since the circle is itself tangent to the boundary at X_1, the line is also normal to the boundary (see Lemma 4.7).

The disk $B(X, \min_{Y \in \partial R} |X - Y|)$ contains X but is not necessarily maximal. In order to see this, draw the line passing through X and X_1. This line is normal to the boundary, and therefore we know that any circle tangent to the boundary at X_1 must have its center along this line (this follows from the fact that R and the disk have the same tangent at X_1 and from Lemma 4.7). Now consider drawing a few larger disks whose centers remain on the line and that are tangent to the boundary at X_1. This second process of inflation is shown in Figure 4.17(b). The final disk from the previous step is shown with a thick line, while several subsequent disks are shown in fine lines.

We stop this process once a second point of contact X_2 is obtained. (As shown in Exercise 16, this second point of contact may be confounded with X_1.) The final disk $B(X', r)$ must still contain X. The following lemmas will convince us that it is in fact maximal. □

Lemma 4.25 *If $B(X, r) \subset R$ and if its circular boundary $S(X, r)$ contains a point X_1 in ∂R, then X_1 is a point of tangency between $S(X, r)$ and ∂R.*

PROOF: Since $B(X, r) \subset R$ and $S(X, r)$ contains a point X_1 of ∂R, it must be that $r = \min_{Y \in \partial R} |X - Y|$.

Consider the tangent to ∂R at X_1. If it is not the same as the tangent line to the circle $S(X, r)$ at X_1, a portion of it must be included in the disk $B(X, r)$ (see Lemma 4.7). Since the boundary is tangent to this line, a portion of the boundary must also lie within $B(X, r)$. Finally, this implies that $B(X, r)$ must contain some points outside of R (Figure 4.18), which is a contradiction. □

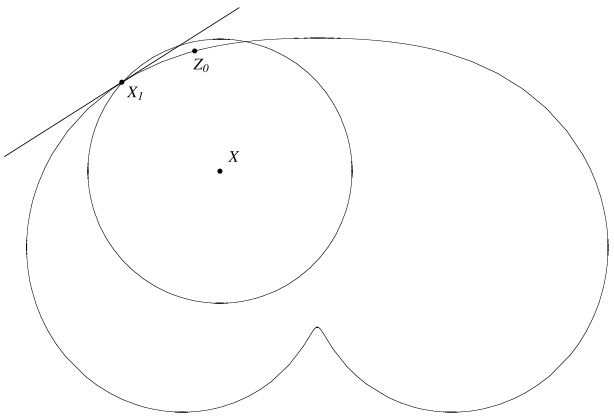


Fig. 4.18. A disk $B(X, r)$ included in R and whose boundary $S(X, r)$ touches ∂R at X_1 must be tangent to ∂R at X_1.

Lemma 4.26 *If $B(X, r) \subset R$ and $S(X, r)$ contains two distinct points X_1 and X_2 of ∂R, then $B(X, r)$ is a maximal disk of R. (We could also generalize this to the case of a single point of contact between $S(X, r)$ and ∂R of order greater than 1. See Exercise 16.)*

PROOF: The question we must answer is the following: does there exist a disk $B(X', r')$ (distinct from $B(X, r)$) such that

$$B(X, r) \subset B(X', r') \subset R? \qquad (\star)$$

If not, then $B(X, r)$ is maximal. We will thus try to construct such a $B(X', r')$.

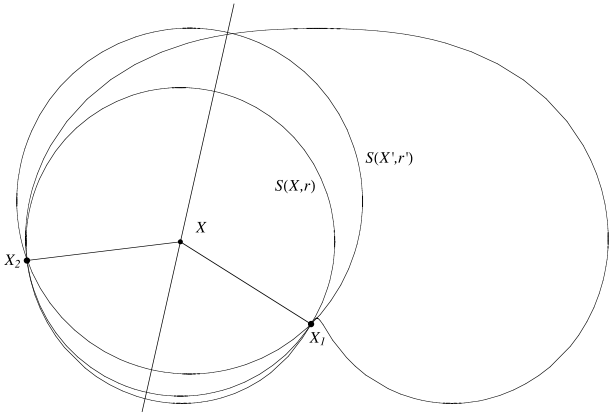


Fig. 4.19. On the hunt for a disk $B(X', r')$ as described in Lemma 4.26.

Since $X_1, X_2 \in S(X, r)$, it must be that the circular boundary $S(X', r')$ of $B(X', r')$ also contains these points. Since they are at the boundary ∂R and since $B(X', r')$ must lie within R, it is impossible to choose $X' = X$ and $r' > r$.

Since X_1 and X_2 must be on the circle $S(X', r')$, they must be the same distance away from the center X'. Thus, the center must lie along the perpendicular bisector of the two points. But by constructing a circle $S(X', r')$ whose center lies along the perpendicular bisector and whose boundary includes both X_1 and X_2, we see that $S(X', r')$ is no longer tangent to ∂R at either X_1 or X_2 (see Figure 4.19) unless $X = X'$ and $r = r'$. So the disk $B(X', r')$ can therefore not lie strictly within R, by the contrapositive of Lemma 4.25. Thus there does not exist a disk $B(X', r')$ that satisfies $(\star)$, and therefore $B(X, r)$ is maximal. □

We are now ready to introduce the fundamental property of the skeleton $\Sigma(R)$ of a region R.

Theorem 4.27 *The skeleton of a region R of the plane $\mathbb{R}^2$ (the space $\mathbb{R}^3$) is the set of centers of all maximal disks (balls) of R.*

Proof: Even though this theorem remains valid for more general regions, we will limit our discussion to two-dimensional regions with continuously differentiable boundaries.

Let E be the set of centers of maximal disks. Proving the equivalence of the two definitions amounts to proving the following two inclusions:

$$\begin{cases} \Sigma(R) & \subset E, \\ \Sigma(R) & \supset E. \end{cases}$$

If $X \in \Sigma(R)$ and $d(X) = \min_{Y \in \partial R} |X - Y|$, then the circle $S(X, d(X))$ contains two points X_1 and X_2 of ∂R, the disk $B(X, d(X))$ is contained within R, and therefore $B(X, d(X))$ is maximal by Lemma 4.26. Thus we have that $\Sigma(R) \subset E$.

To prove the other direction, consider a point $X \in E$ and a radius r such that $B(X, r)$ is maximal. Then $B(X, r) \subset R$. The circle $S(X, r)$ must contain a point $X_1 \in \partial R$ as otherwise we could have applied the first "inflation" step to yield a larger disk containing $B(X, r)$ (see Figure 4.17(a)). Similarly, there must be a second point of tangency, since otherwise we could apply the second "inflation" to again find a larger disk containing $B(X, r)$ (see Figure 4.17(b)). Thus $B(X, r)$ is of maximal radius (that is, $r = \min_{Y \in \partial R} |X - Y|$) and touches ∂R at two points. These are precisely the conditions required for X to be in the skeleton $\Sigma(R)$. □

We leave the proof of the following corollary to the exercises.

Corollary 4.28 *A region R of the plane $\mathbb{R}^2$ (the space $\mathbb{R}^3$) is completely determined by its skeleton $\Sigma(R)$ and the function $d(X)$ defined for $X \in \Sigma(R)$.*

4.8 Exercises

1. **(a)** Find the skeleton of a triangle. Determine its r-skeleton.
(b) Show that the skeleton of the triangle is the union of three line segments. What classical theorem of Euclidean geometry assures us that these three segments meet at a point?

2. This exercise explores the analogy between the r-skeleton and a fire lit simultaneously at all points along the boundary of a region $R \subseteq \mathbb{R}^2$. Let v be the speed of the fire. Describe the points of the r-skeleton in terms of this analogy.

3. Can you construct a region R whose skeleton is
(a) a single point?
(b) a line segment? (Other than an ellipse!)

4. The rectangle example shows that its skeleton consists of five line segments.
(a) What is the skeleton of a square ($b = h$)? Show that this skeleton consists of only two segments.
(b) Are there other regions that have the same skeleton as the square?

5. Determine the skeleton of a parabola (see Figure 4.20). Is the focus of the parabola an extreme point of its skeleton?

6. **(a)** Let R be the region of $\mathbb{R}^2$ represented at the left in Figure 4.21. Both of the curves are semicircles. Draw the skeleton of this region.

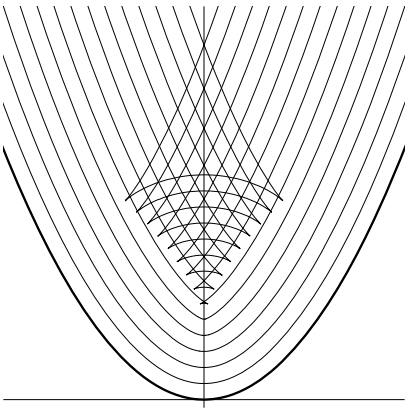


Fig. 4.20. The advancing front of a fire on a parabola (Exercise 5).

(b) Let L be the region of $\mathbb{R}^2$ represented at the right in Figure 4.21. What are the radius and the center of the largest circle that may be inscribed in this region? (Note: the two arms of L have the same width ($h = 1$) and the curves are again semicircles.)
(c) Draw the skeleton of the region L as precisely as possible and explain your answer. (If this skeleton consists of several curves or segments, then their points of intersection should be clearly marked.)

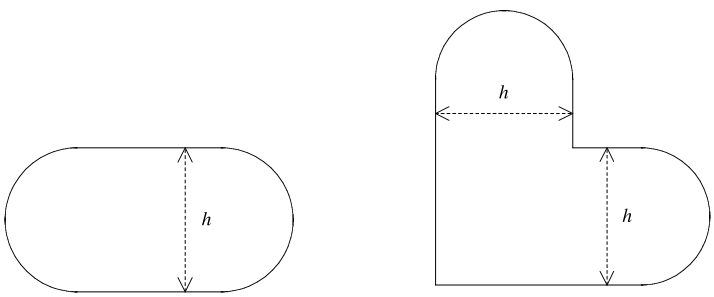


Fig. 4.21. Regions R and L for Exercise 6.

7. Think of an algorithm for drawing the skeleton of a polygon, both convex and not convex. Similarly, think of an algorithm for drawing the r-skeleton of a polygon.

8. In the context of gamma-ray radiosurgery, let us suppose that an optimal solution for a region R is given by $\cup_{i=1}^{N} B(X_i^*, r_i)$. Explain why it is natural that if $I \subset \{1, \ldots, N\}$, then $\cup_{i \notin I} B(X_i^*, r_i)$ is an optimal solution for $R \setminus \cup_{i \in I} B(X_i^*, r_i)$.

9. The proof of Theorem 4.27 does not apply to the skeleton of the triangle, since the tangent vectors at the corners are ill-defined. Show (by some other method) that this theorem still holds for triangles.

10. Find the skeleton of a rectangular parallelepiped whose sides have three distinct lengths. Find its r-skeleton.

11. What is the skeleton of a tetrahedron? What is its r-skeleton?

12. What is the skeleton of a cone with an elliptical cross section?

13. Consider an ellipsoid of revolution, given by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{b^2} = 1,$$

for $b < a$. Describe its skeleton and justify your answer.

14. What is the skeleton of a cylinder with height h and radius r? You will have to consider three cases: (i) $h > 2r$, (ii) $h = 2r$, and (iii) $h < 2r$.

15. **(a)** Show that a connected graph is a tree if and only if its Euler number (defined as the number of nodes minus the number of edges) is 1.
(b) Show that an acyclic graph is connected (in other words, it is a tree) if and only if it has an Euler number of 1.

16. **The extreme points of the ellipse with $b < a$.** This exercise extends Example 4.9. The points of the skeleton were identified as being the points of the interior of the ellipse that are reached simultaneously by two or more fires originating from distinct points on the boundary. The skeleton is a segment of the major axis whose two extremities

$$\left(\frac{a^2 - b^2}{a}, 0\right) \qquad \text{and} \qquad \left(-\frac{a^2 - b^2}{a}, 0\right)$$

are not reached by fires originating from two distinct points. For instance, by studying Figure 4.10 we see that the extreme point $(\frac{a^2-b^2}{a}, 0)$ is first reached by the fire originating at $(a, 0)$. Why do these two extreme points belong to the skeleton? The answer lies in the domain of differential geometry.

Let $\alpha(x) = (x, y_1(x))$ and $\beta(x) = (x, y_2(x))$ be two curves in the plane that touch at $x = 0$:

$$\alpha(0) = \beta(0).$$

We say that α and β have a contact of order at least $p \geq 1$ if

$$\begin{cases} \frac{d}{dx}\alpha(0) = \frac{d}{dx}\beta(0), \\ \frac{d^2}{dx^2}\alpha(0) = \frac{d^2}{dx^2}\beta(0), \\ \vdots \\ \frac{d^p}{dx^p}\alpha(0) = \frac{d^p}{dx^p}\beta(0). \end{cases}$$

The contact is of order exactly p if moreover, $\frac{d^{p+1}}{dx^{p+1}}\alpha(0) \neq \frac{d^{p+1}}{dx^{p+1}}\beta(0)$. Intuitively, a high-order contact between two curves indicates that they stay close to each other "longer" as we distance ourselves from the point of actual contact, or that their "degree of tangency" is higher. A parallel can be drawn to the concept of multiplicity of roots. When we have a root with multiplicity p we treat it as the limiting case of p roots that approach each other. Here we can consider a point of contact of order p as the limiting case of p points of tangency approaching each other.

We will calculate the order of contact between the maximal disk at the end of the minor axis (at $(0, b)$) and then at the end of the major axis (at $(a, 0)$).

(a) Show that the equation of the circle delimiting the boundary of the maximal disk tangent to the ellipse at $(0, b)$ is given by

$$\alpha(x) = \left(x, \sqrt{b^2 - x^2}\right)$$

and that the ellipse is

$$\beta(x) = \left(x, \frac{b}{a}\sqrt{a^2 - x^2}\right).$$

Show that these two curves touch at $x = 0$. Show that the order of the point of contact between these two curves is 1 but not higher.

(b) To study the point of contact at $(a, 0)$ it is useful to change the role of x and y in the above definition. Thus, the equation of the ellipse becomes

$$\beta(y) = \left(\frac{a}{b}\sqrt{b^2 - y^2}, y\right).$$

(Convince yourself of this fact!) Write the equation of the circular boundary of the maximal disk tangent to the ellipse at $(a, 0)$ in the form of $\alpha(y) = (f(y), y)$ for some function $f(y)$. What is the order of contact between the two curves at this point? (The order of contact is determined by taking the derivatives of the curves with respect to y.) Conclude that it is reasonable to include the two extreme points $(\pm(a^2 - b^2)/a, 0)$ in the skeleton Σ(ellipse).

17. In the case that the function $f_1(\overline{x}, \overline{y})$ of the proof of Proposition 4.17 is of the form $f_1(\overline{x}, \overline{y}) = \overline{y} - f(\overline{x})$, show that the condition that J be noninvertible (in other words, $\det(J) = 0$, where J is given by (4.5)) is equivalent to saying that the curve $\overline{y} = f(\overline{x})$ has a contact of order at least 2 at (x_1, y_1) to the circle $(\overline{x} - x_0)^2 + (\overline{y} - y_0)^2 = r^2$, where $r^2 = (x_1 - x_0)^2 + (y_1 - y_0)^2$. (To do this, write the circle in the form $\overline{y} = g(\overline{x})$

and show that $f(x_1) = g(x_1) = y_1$ and $f'(x_1) = g'(x_1)$ implies $\det(J) = 0$ if and only if $f''(x_1) = g''(x_1)$. The concept of "contact of order p" was defined and explored in Exercise 16.

18. We consider a region R_ϵ consisting of a rectangle R whose corners have been replaced by small circles of radius ϵ (see Figure 4.22). Give the skeleton of R_ϵ. Show that it coincides with the r-skeleton of R for a given value r. What is the value?
(Remark: The boundary of R_ϵ is only C^1. In order to obtain a boundary that is piecewise C^3 we would have to replace the quarter-circles by curves with points of contact of order 3 to the sides of the rectangle. However, the exercise still illustrates that in the case of a convex domain, there exists an r_0 such that for $r > r_0$ there is no difference between the r-skeleton of the original region and that of the "smoothed" region. For nonconvex regions the result is not quite so simple, but we can still obtain a reasonable approximation to the skeleton by smoothing the boundary.)

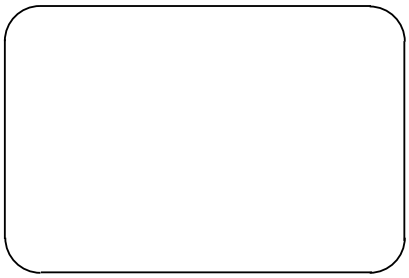


Fig. 4.22. The region R_ϵ of Exercise 18.

19. Relationship to Voronoi diagrams (see Section 15.5). Show that the skeleton of the complement R of a set S of n points is given by the edges of the Voronoi diagram over S. (This means that the boundary of R is given by S.)

References

[1] H. Blum. Biological shape and visual science (part i). *Journal of Theoretical Biology*, 38:205–287, 1973.

[2] P. Dimitrov, C. Phillips, and K. Siddiqi. Robust and efficient skeletal graphs. In *Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition*, 2000.

[3] F. Jamm and D. Parlongue. Les tas de sable. *Gazette des mathématiciens*, 93:65–82, 2002.

[4] Q.J. Wu. Sphere packing using morphological analysis. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 55:45–54, 2000.

[5] Q.J. Wu and J.D. Bourland. Morphology-guided radiosurgery treatment planning and optimization for multiple isocenters. *Medical Physics*, 26:2151–2160, 1999.

5

Savings and Loans

This chapter requires only a familiarity with geometric series, recursive sequences, and limits. It can be covered in two hours of class and contains no advanced part (see the preface).

Nothing seems further from mathematics than buying a home or planning for retirement, especially to a twenty-year-old. However, the saving and borrowing of money is subject to various rules that are amenable to mathematical modeling. In fact, this is one of the oldest uses of mathematics.

Leibniz wrote numerous scientific papers on the subjects of interest, insurance, and financial mathematics [2]. However, our civilization is not the first to have considered these issues. In 1933, an archaeological dig in Iran directed by Contenau and Mecquenem discovered several Babylonian tablets. These tablets were heavily studied over the next few decades, and several of them had mathematical content. In particular, one of the tablets discussed the calculation of compound interest and annuities [1]. These tablets were dated to the end of the first Babylonian dynasty, a little after Hammurabi (1793–1750 BC). As such, the problems discussed in this chapter are surely among the oldest applications of mathematics!

The mathematics used in these financial problems is quite simple. Nonetheless, the average person is not familiar with mortgage terminology and is often suspicious of the seemingly amazing promises of retirement plans. Since these are issues that affect everyone at some point, it is well worth learning the underlying vocabulary and mathematics.

5.1 Banking Vocabulary

As with many subjects in which mathematics is used, the commonly used vocabulary was not created by mathematicians. In these fields, terms are often unclear or even

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_5,

downright confusing. Thankfully, in the financial world they are both simple and precise. Two examples will allow us to introduce the basic vocabulary.

The first example is that of a savings account. Suppose a person deposits \$1000 into a savings account with the intention of withdrawing the money in exactly five years. The bank agrees to pay 5% annually. The *initial deposit*, or *principal*, is the amount that was originally placed into the account. In this example it is \$1000. The 5% paid by the bank is the *interest rate*.[1]

The second example is that of a loan. You have worked several summer jobs but you are \$5000 short of buying your first car. You decide to borrow this money from a bank. The bank requires you to repay the loan by paying \$156.38 monthly for three years because the loan is made with an interest rate of 8%. The *loan amount*, or *initial balance*, is the \$5000 that the bank initially lends you, the *monthly payment* is \$156.38, and the *amortization period* is three years. At every moment during those three years, the precise amount remaining to be paid to the bank (from the original \$5000) is referred to as the *outstanding balance*. At the end of the three years the outstanding balance will be zero and the car will belong completely to you.

5.2 Compound Interest

There are two types of interest: simple and compound. We will start by discussing compound interest, which is by far the most commonly used.

Compound interest does not "add," but rather it "compounds." What precisely does this mean? In the first example of the previous section, the interest rate was 5% (understood to be annual). After the first year the principal of \$1000 will be worth

$$\$1000 + (5\% \text{ of } \$1000) = \left(\$1000 + \frac{5}{100} \times \$1000\right) = (\$1000 + \$50) = \$1050.$$

However, the same interest is not simply added the following year. In fact, the interest in the second year will be calculated based on the "new" balance of \$1050 after the first year. Thus, after two years, the balance is

$$\begin{aligned} \$1050 + (5\% \text{ of } \$1050) &= \left(\$1050 + \frac{5}{100} \times \$1050\right) \\ &= (\$1050 + \$52.50) = \$1102.50. \end{aligned}$$

Is the \$2.50 at all significant? Over time, this small difference will play a large role. Continuing the calculation for each of the remaining anniversaries, we obtain

[1] The expressions 5% and n% signify fractions of 100. Thus, 5% represents $\frac{5}{100}$, and n% represents $\frac{n}{100}$.

3rd anniversary: \$1157.63,
4th anniversary: \$1215.51,
5th anniversary: \$1276.28.

If the interest applied each year remained the same as it was at the end of the first year, the final balance would have been $(\$1000 + 5 \times \$50) = \$1250$. However, since the interest was compounded, the closing balance is instead \$1276.28.

It is time to formalize this concept. Let p_i be the balance after the ith anniversary and let p_0 be the initial balance. Let r be the interest rate, where $r = \frac{5}{100}$ in the above example. The balance p_i at the ith anniversary may be calculated using the balance p_{i-1} from the previous anniversary. In fact, it is given by the simple relation

$$p_i = p_{i-1} + r \cdot p_{i-1} = p_{i-1}(1 + r), \qquad i \geq 1.$$

Expanding this recursive formula, we see that

$$\begin{aligned} p_i &= p_{i-1}(1 + r) \\ &= (p_{i-2}(1 + r))\,(1 + r) = p_{i-2}(1 + r)^2 \\ &= \cdots \\ &= p_0(1 + r)^i, \qquad i \geq 1. \end{aligned} \tag{5.1}$$

This is the formula for compound interest. A mathematician would read this formula by saying that "the balance grows geometrically," meaning that it grows like the power of $1 + r$ (which is greater than 1).

Most banks actually calculate their interest over shorter time periods. Suppose that in the previous example, the interest was applied quarterly, which is to say every three months. Since there are four cycles of three months in a year, the bank would calculate an interest of $\frac{r}{4}\% = \frac{5}{4}\%$ every three months. After one year, there would be four interest deposits, and their compounding would produce an *effective* interest rate greater than the announced 5%. In fact,

$$1 + r_{\text{eff}} = \left(1 + \frac{r}{4}\right)^4$$

and

$$r_{\text{eff}} = 5.095\%,$$

which is to the client's advantage. When a bank calculates interest at intervals smaller than a year, the advertised interest rate is called the *nominal interest rate*. The actual rate of interest observed at the end of a year will be slightly higher than this and is called the *effective interest rate*. In the last example, the nominal interest rate was 5%, while the effective rate was 5.095%.

As we may imagine, the effective interest rate increases as the compounding interval shrinks. For example, if the interest is compounded daily, then the effective rate associated with $r = 5\%$ is

$$r_{\text{eff}} = \left(1 + \frac{r}{365}\right)^{365} - 1 = 5.12675\%.$$

What about interest compounded at every hour? At every second? At every millisecond? Mathematicians are naturally led to pose the following question: does there exist a limit for the effective interest rate as the compounding period tends to zero? If the year is divided into n equal pieces, then the effective interest rate associated with a nominal rate of r is given by

$$1 + r_{\text{eff}}(n) = \left(1 + \frac{r}{n}\right)^{n}.$$

The most generous banker in the world would apply interest continuously, and the effective rate would be

$$1 + r_{\text{eff}}(\infty) = \lim_{n\to\infty} (1 + r_{\text{eff}}(n)) = \lim_{n\to\infty} \left(1 + \frac{r}{n}\right)^{n} = e^{r}.$$

The last step of the above equation uses the formula

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^{n} = e,$$

which is normally shown in a first calculus course. Using the change of variables $m = \frac{n}{r}$, we obtain

$$\lim_{n\to\infty} \left(1 + \frac{r}{n}\right)^{n} = \lim_{m\to\infty} \left(1 + \frac{1}{m}\right)^{mr} = \left(\lim_{m\to\infty} \left(1 + \frac{1}{m}\right)^{m}\right)^{r} = e^{r}.$$

It is somewhat amusing to note the appearance of the base e of the natural logarithms in such a seemingly simple calculation. (Since loans have been around as long as there have been people, bankers could easily have been the first to discover this number.) If $r = 5\%$ as in our earlier examples, then a savings period of 20 years multiplies the initial principal by e. This is seen easily using (5.1), since

$$p_{20} = p_0(1 + r_{\text{eff}}(\infty))^{20} = p_0(e^{r})^{20} = p_0 e^{\frac{5}{100}\times 20} = p_0 e.$$

There is not a large difference between the nominal rate of $r = 5\%$ and the corresponding limiting effective rate $r_{\text{eff}}(\infty)$: $r_{\text{eff}}(\infty) = e^{r} - 1 = 5.127,109\ldots\%$. As such, bankers do not use the limiting effective rate (a somewhat abstract concept) as a marketing tool.

Simple interest is very rare and is almost never used in banking circles. It consists in calculating interest based on the initial deposit regardless of the anniversary. In the case of a initial deposit $p_0 = \$1000$ and an interest rate $r = 5\%$, the (simple) interest applied each year will be $\$1000 \times \frac{5}{100} = \$50$, and the balances at the end of the first five anniversaries will be

$$\begin{aligned} p_1 &= \$1050, \\ p_2 &= \$1100, \\ p_3 &= \$1150, \\ p_4 &= \$1200, \\ p_5 &= \$1250. \end{aligned}$$

This is called an arithmetic progression, and it grows *linearly* with the number of years since the initial deposit:

$$p_i = p_0(1 + ir).$$

If you are looking to put money into a savings account, refuse simple interest. However, if somebody offers you a loan using simple interest, they are being very generous!

5.3 A Savings Plan

Financial institutions recommend starting to save for retirement as early as possible. They propose several savings plans, some of which promise that you can begin your retirement on the day of your 55th birthday, with guaranteed financial comfort. For a young student this may seem quite far off, and it may not seem like such a big deal to delay starting a savings plan by a few years. But the banks are right: the sooner you start, the better!

A savings plan might involve putting aside an amount of Δ dollars annually, for N years. During these N years the bank offers an interest rate r, which we will assume is compounded annually. The variables are as follows:

$$\begin{aligned} \Delta: &\quad \text{annual deposit into the savings account,} \\ r: &\quad \text{constant interest rate during the } N \text{ years,} \\ N: &\quad \text{duration of the savings plan,} \\ p_i: &\quad \text{balance of the account after } i \text{ years}, i = 0, 1, \ldots, N. \end{aligned}$$

We will assume that the client starts the plan by depositing Δ dollars on the first day; thus

$$p_0 = \Delta.$$

After one year, the interest is calculated and deposited into the account, and the client deposits an additional Δ dollars as well. At the end of this first year, the balance is

$$p_1 = p_0 + rp_0 + \Delta = p_0(1 + r) + \Delta.$$

This logic can be repeated for each following year, yielding the recurrence relation

$$p_i = p_{i-1}(1 + r) + \Delta.$$

It is possible to determine p_i as a function of p_0. By experimenting a little, we guess the answer:

$$\begin{aligned} p_2 &= p_1(1+r) + \Delta \\ &= \left(p_0(1+r) + \Delta\right)(1+r) + \Delta \\ &= p_0(1+r)^2 + \Delta(1 + (1+r)) \end{aligned}$$

and

$$\begin{aligned} p_3 &= p_2(1+r) + \Delta \\ &= \left(p_0(1+r)^2 + \Delta(1 + (1+r))\right)(1+r) + \Delta \\ &= p_0(1+r)^3 + \Delta\left(1 + (1+r) + (1+r)^2\right). \end{aligned}$$

It is tempting to propose a general formula of

$$\begin{aligned} p_i &= p_0(1+r)^i + \Delta\left(1 + (1+r) + (1+r)^2 + \cdots + (1+r)^{i-1}\right) \\ &= p_0(1+r)^i + \Delta\sum_{j=0}^{i-1}(1+r)^j. \end{aligned} \tag{5.2}$$

This formula will be proved in Exercise 1.

Recalling that the sum of the first i powers of a number x is given by

$$\sum_{j=0}^{i-1} x^j = \frac{x^i - 1}{x - 1}$$

if $x \neq 1$, then we obtain

$$\begin{aligned} p_i &= \Delta(1+r)^i + \Delta\sum_{j=0}^{i-1}(1+r)^j, \qquad \text{since } p_0 = \Delta \\ &= \Delta\sum_{j=0}^{i}(1+r)^j \\ &= \Delta\frac{(1+r)^{i+1} - 1}{(1+r) - 1} \\ &= \frac{\Delta}{r}\left((1+r)^{i+1} - 1\right) \end{aligned}$$

and therefore

$$p_i = \frac{\Delta}{r}\left((1+r)^{i+1} - 1\right). \tag{5.3}$$

Thus, after N years we have a closing balance of $p_N = \Delta((1+r)^{N+1} - 1)/r$. Observe that if the client begins his retirement after N years, he will not deposit the final amount

of Δ dollars into the account, since this is the day he begins living off his savings. Thus, the actual final balance will be

$$\begin{aligned} q_N &= p_N - \Delta \\ &= \frac{\Delta}{r}\left((1+r)^{N+1} - 1\right) - \Delta \\ &= \frac{\Delta}{r}\left((1+r)^{N+1} - 1 - r\right) \\ &= \frac{\Delta}{r}\left((1+r)^{N+1} - (1+r)\right). \end{aligned} \tag{5.4}$$

Rather than (5.3), we will use (5.4) from now on.

Example 5.1 *(a) We present a numerical example to help give some idea of such a savings plan. Suppose that an annual deposit of* $\Delta = \$1000$ *is deposited over an* $N = 25$ *year period. If the interest rate is* 8%*, then the final balance is*

$$q_N = \frac{\Delta}{r}\left((1+r)^{N+1} - 1\right) - \Delta = \$78{,}954.42,$$

even though the client spent only \$25,000*.*

(b) Suppose that a second client started her savings one year later than the client in the first example, but still retired the same year. What difference will there be in the final balances? For the second client we have that $N = 24$*, while the other variables remain the same. Thus,* $q_{24} = \$72{,}105.94$*, and the difference between the two balances is* \$6848.48*. By having contributed only* \$1000 *less than the first client, the second client finds herself with almost* 10% *less money than the first. As you can see, the banks are right: start your retirement savings early!*

At the beginning of our discussion we made the hypothesis that the interest rate offered over the N years would remain constant. This is not very realistic! Figure 5.3 shows the average interest rate for housing mortgages charged by large Canadian banks over the last fifty years. When banks charge higher interest rates to borrowers, they are able to pay higher rates on savings.

5.4 Borrowing Money

Many people borrow money in order to pay for expensive things like cars, appliances, education, and homes. It is therefore useful to understand how various loans work.

When buying a home, a buyer normally uses a portion of her savings to make a down payment. The rest of the purchase cost is typically borrowed from a bank. The down payment and the borrowed sum are paid directly to the previous owner, and the new owner is left with the responsibility of paying back the bank.

Banks typically let clients choose the *amortization period* of the loan, associated with which will be an interest rate r and a *monthly payment* Δ. Here are the variables involved:

p_i : amount of the borrowed money left to repay after the ith month,
Δ : monthly payment amount,
r_m : effective monthly interest rate,
N : amortization period (in years).

The amount p_0 represents the initial amount of money borrowed from the bank, in other words, the purchase price minus the down payment. It is important to note that the variable i in this section counts months rather than years. At the end of each month, interest is calculated and charged, but the borrower also makes a payment of Δ dollars. Thus, if the borrower owed p_i after i months, after $i+1$ months she owes

$$p_{i+1} = p_i(1 + r_m) - \Delta.$$

The negative sign in front of Δ indicates that the borrower *reduces* her debt with her payment, while the monthly interest $r_m p_i$ increases it. (Thus, it is possible to reduce the debt only if $p_i r_m < \Delta$.) Since she chose to pay back her debt over N years (and therefore $12N$ months), it is required that

$$p_{12N} = 0.$$

Fig. 5.1. The average interest rate for housing mortgages charged by large Canadian banks since 1950. (Source: website of the Bank of Canada.)

Using a similar calculation to that of the previous section (exercise!), it is possible to express p_i as a function of p_0. We find that

$$\begin{aligned} p_i &= p_0(1+r_m)^i - \Delta \sum_{j=0}^{i-1} (1+r_m)^j \\ &= p_0(1+r_m)^i - \Delta \frac{(1+r_m)^i - 1}{r_m}. \end{aligned} \tag{5.5}$$

Since the bank fixes the interest rate (and therefore r_m) and the client chooses the principal p_0 and the amortization period N, the only unknown is Δ. Using the fact that $p_{12N} = 0$, it follows that

$$0 = p_{12N} = p_0(1+r_m)^{12N} - \frac{\Delta}{r_m}\left((1+r_m)^{12N} - 1\right)$$

and therefore

$$\Delta = r_m p_0 \frac{(1+r_m)^{12N}}{((1+r_m)^{12N} - 1)}. \tag{5.6}$$

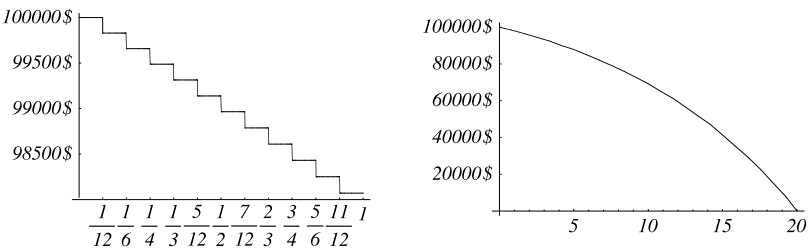


Fig. 5.2. Outstanding balance during the first year (left) and during the 20-year amortization period (right). See Example 5.2.

Example 5.2 *Consider a loan of* \$100,000 *paid over a* 20*-year period with a monthly interest rate of* $\frac{2}{3}\%$ *(and therefore a nominal annual rate of* $12 \times \frac{2}{3}\% = 8\%$*). The borrower must make monthly payments of*

$$\Delta = \frac{2}{300} \times \$100{,}000 \times \frac{(1+\frac{2}{300})^{240}}{\left((1+\frac{2}{300})^{240} - 1\right)} = \$836.44.$$

The 240 monthly payments of \$836.44 *will total* $240 \times \$836.44 = \$200{,}746$*, more than twice the amount borrowed. Using* (5.5)*, we can plot the outstanding balance* p_i *over the course of the 20-year amortization period. Figure 5.2 shows the progress of the debt*

repayment during the first year (at left) and during the entire amortization period (at right). Observe that during the first year of repayment the balance did not even decrease by \$3000*, even though the borrower made* $12 \times \$836.44 = \$10{,}037.28$ *in payments! Mortgages can be pretty frustrating.*

If we wish to pay back the debt in 15 years rather than 20, then the monthly payment will be \$955.65 with a total repayment of \$172,017*. The difference of more than* \$28,000 *between a 20-year and a 15-year mortgage would no doubt make many people think twice when choosing an amortization period. You will surely think about it when making your first home purchase.*

In the first section we saw the difference between nominal interest rates and effective interest rates. A similar distinction appears for mortgage rates. Banks always mention their annual mortgage rate r without explaining how the monthly rate r_m is calculated. Is it

$$r_m = \frac{r}{12}? \tag{r_{m1}}$$

Or is r_m determined by

$$(1+r) = (1+r_m)^{12}? \tag{r_{m2}}$$

In the first case, the effective annual interest rate would be

$$r_{\text{eff1}} = (1+r_{m1})^{12} - 1 = \left(1+\frac{r}{12}\right)^{12} - 1,$$

while in the second case it would be $r_{\text{eff2}} = r$. It is clear that $(1+\frac{r}{12})^{12} - 1 > r$ (why?) and that banks will make more money with a monthly rate of r_{m1} than one of r_{m2}. Thus r_{m1} favors the banks, while r_{m2} favors the borrowers. The question remains, how is the rate calculated?

The answer depends on the country! Even in North America, monthly rates are calculated differently in Canada and the United States. American banks use r_{m1}, while Canadian banks use neither. In fact, in Canada the formula

$$\left(1+\frac{r}{2}\right) = (1+r_m)^6 \tag{$r_{m\text{CAN}}$}$$

is used. In other words, Canadian monthly mortgage rates are calculated such that when compounded over six months they must equal half of the nominal annual rate. Knowing exactly how r_m is calculated is necessary to reproduce the calculations made by bankers.

5.5 Appendix: Mortgage Payment Tables

The following two pages contain monthly payment tables for nominal annual interest rates of 8% and 12%. These are the types of tables that can be found in books called *mortgage payment tables*. The top line gives the amortization period in years, and the

leftmost column the amount borrowed. These tables are provided as an example and are used in several exercises. The effective monthly interest rate has been calculated according to Canadian rules.

	1	2	3	4	5	6	7	8	9	10	15	20	25
1000	86.93	45.17	31.28	24.35	20.21	17.47	15.52	14.07	12.95	12.06	9.48	8.28	7.63
2000	173.86	90.34	62.55	48.70	40.43	34.94	31.04	28.14	25.90	24.13	18.96	16.57	15.26
3000	260.78	135.50	93.83	73.06	60.64	52.41	46.56	42.21	38.85	36.19	28.44	24.85	22.90
4000	347.71	180.67	125.11	97.41	80.86	69.88	62.09	56.28	51.81	48.26	37.93	33.13	30.53
5000	434.64	225.84	156.38	121.76	101.07	87.35	77.61	70.35	64.76	60.32	47.41	41.42	38.16
6000	521.57	271.01	187.66	146.11	121.28	104.82	93.13	84.42	77.71	72.38	56.89	49.70	45.79
7000	608.50	316.18	218.93	170.46	141.50	122.29	108.65	98.49	90.66	84.45	66.37	57.99	53.42
8000	695.43	361.34	250.21	194.81	161.71	139.76	124.17	112.56	103.61	96.51	75.85	66.27	61.06
9000	782.35	406.51	281.49	219.17	181.93	157.23	139.69	126.64	116.56	108.58	85.33	74.55	68.69
10000	869.28	451.68	312.76	243.52	202.14	174.70	155.21	140.71	129.51	120.64	94.82	82.84	76.32
15000	1303.92	677.52	469.15	365.28	303.21	262.05	232.82	211.06	194.27	180.96	142.22	124.25	114.48
20000	1738.57	903.36	625.53	487.04	404.28	349.40	310.43	281.41	259.03	241.28	189.63	165.67	152.64
25000	2173.21	1129.20	781.91	608.80	505.35	436.74	388.04	351.77	323.78	301.60	237.04	207.09	190.80
30000	2607.85	1355.04	938.29	730.56	606.42	524.09	465.64	422.12	388.54	361.92	284.45	248.51	228.96
35000	3042.49	1580.88	1094.67	852.32	707.50	611.44	543.25	492.47	453.30	422.24	331.85	289.93	267.12
40000	3477.13	1806.72	1251.05	974.07	808.57	698.79	620.86	562.82	518.05	482.56	379.26	331.34	305.29
45000	3911.77	2032.56	1407.44	1095.83	909.64	786.14	698.46	633.18	582.81	542.88	426.67	372.76	343.45
50000	4346.41	2258.40	1563.82	1217.59	1010.71	873.49	776.07	703.53	647.57	603.20	474.08	414.18	381.61
60000	5215.70	2710.08	1876.58	1461.11	1212.85	1048.19	931.29	844.24	777.08	723.85	568.89	497.01	457.93
70000	6084.98	3161.76	2189.34	1704.63	1414.99	1222.88	1086.50	984.94	906.59	844.49	663.71	579.85	534.25
80000	6954.26	3613.44	2502.11	1948.15	1617.13	1397.58	1241.72	1125.65	1036.11	965.13	758.52	662.69	610.57
90000	7823.54	4065.12	2814.87	2191.67	1819.27	1572.28	1396.93	1266.36	1165.62	1085.77	853.34	745.52	686.89
100000	8692.83	4516.79	3127.64	2435.19	2021.42	1746.98	1552.14	1407.06	1295.13	1206.41	948.15	828.36	763.21

Table 5.1. Table of mortgage monthly payments for a nominal interest rate of 8%.

	1	2	3	4	5	6	7	8	9	10	15	20	25
1000	88.71	46.94	33.08	26.19	22.10	19.40	17.50	16.09	15.02	14.18	11.82	10.81	10.32
2000	177.43	93.88	66.15	52.38	44.20	38.80	35.00	32.19	30.04	28.36	23.63	21.62	20.64
3000	266.14	140.82	99.23	78.58	66.30	58.20	52.49	48.28	45.06	42.54	35.45	32.43	30.96
4000	354.85	187.75	132.30	104.77	88.39	77.60	69.99	64.38	60.09	56.72	47.26	43.24	41.28
5000	443.57	234.69	165.38	130.96	110.49	97.00	87.49	80.47	75.11	70.90	59.08	54.05	51.59
6000	532.28	281.63	198.46	157.15	132.59	116.40	104.99	96.57	90.13	85.08	70.90	64.86	61.91
7000	620.99	328.57	231.53	183.34	154.69	135.80	122.49	112.66	105.15	99.26	82.71	75.67	72.23
8000	709.71	375.51	264.61	209.54	176.79	155.20	139.99	128.75	120.17	113.44	94.53	86.48	82.55
9000	798.42	422.45	297.69	235.73	198.89	174.60	157.48	144.85	135.19	127.62	106.34	97.29	92.87
10000	887.13	469.38	330.76	261.92	220.98	194.00	174.98	160.94	150.21	141.80	118.16	108.10	103.19
15000	1330.70	704.08	496.14	392.88	331.48	291.00	262.47	241.41	225.32	212.70	177.24	162.15	154.78
20000	1774.27	938.77	661.52	523.84	441.97	388.00	349.97	321.88	300.43	283.61	236.32	216.19	206.38
25000	2217.84	1173.46	826.91	654.80	552.46	485.00	437.46	402.36	375.54	354.51	295.40	270.24	257.97
30000	2661.40	1408.15	992.29	785.76	662.95	582.00	524.95	482.83	450.64	425.41	354.48	324.29	309.57
35000	3104.97	1642.84	1157.67	916.72	773.45	679.00	612.44	563.30	525.75	496.31	413.56	378.34	361.16
40000	3548.54	1877.54	1323.05	1047.68	883.94	776.00	699.93	643.77	600.86	567.21	472.64	432.39	412.76
45000	3992.10	2112.23	1488.43	1178.64	994.43	873.00	787.42	724.24	675.97	638.11	531.72	486.44	464.35
50000	4435.67	2346.92	1653.81	1309.60	1104.92	970.00	874.92	804.71	751.07	709.01	590.80	540.49	515.95
60000	5322.81	2816.30	1984.57	1571.52	1325.91	1164.00	1049.90	965.65	901.29	850.82	708.97	648.58	619.14
70000	6209.94	3285.69	2315.34	1833.44	1546.89	1358.00	1224.88	1126.60	1051.50	992.62	827.13	756.68	722.33
80000	7097.08	3755.07	2646.10	2095.36	1767.88	1552.00	1399.87	1287.54	1201.72	1134.42	945.29	864.78	825.52
90000	7984.21	4224.46	2976.86	2357.27	1988.86	1746.00	1574.85	1448.48	1351.93	1276.22	1063.45	972.88	928.71
100000	8871.34	4693.84	3307.62	2619.19	2209.85	1940.00	1749.83	1609.42	1502.15	1418.03	1181.61	1080.97	1031.90

Table 5.2. Table of mortgage monthly payments for a nominal interest rate of 12%.

5.6 Exercises

Note: Assume that interest is compounded annually unless otherwise stated.

1. Prove formula (5.2). (Hint: by induction, obviously!)

2. **(a)** Is formula (5.4) linear in Δ? In other words, if the annual deposit Δ is multiplied by x, is the balance after i years also multiplied by x?
(b) Is the same formula linear in r?
(c) If the client instead saves $\frac{\Delta}{2}$ every six months, will the sum after N years be different?

3. Most credit card companies advertise annual rates even though they calculate interest monthly. If the effective annual rate of a company is 18%, what monthly rate will they charge? Before finding the precise answer, will it be bigger or smaller than $\frac{18}{12}\% = 1.5\%$?

4. **(a)** A 20-year-old student saves \$1000 into an account with an interest rate of 5%. She intends to leave the money in the account until she retires at age 65. Suppose that the interest rate remains constant throughout her lifetime. What will be the balance in the account at her retirement if the interest is compounded (i) annually and (ii) monthly at a rate of $\frac{5}{12}\%$?
(b) A student of the same age decides not to start saving until he is 45. He wishes to make a deposit that will provide him with the same amount at age 65 as the student in question (a). Considering each of the interest rates in (a), what will this amount be?

5. **(a)** A person invests \$1000 for ten years. What will the values of the investment be after the ten years if the annual rate is 6%, 8%, and 10%?
(b) For each of the interest rates in (a), how long will the investment take to double its initial value?
(c) Same question as (b), but where the interest is simple rather than compound.
(d) What is the answer to (b) if the initial deposit is instead \$2000?

6. A mortgage with a rate of 8% is paid over a 20-year period. How many months will it take to pay back half of the initial principal?

7. **(a)** A 20-year-old student finds a bank that offers a 10% interest rate if she agrees to invest \$1000 per year until she is 65. What will be the value of the investment on her 65th birthday?
(b) What annual deposit is required if she wishes to retire a millionaire?

8. A student wishes to borrow some money. He knows that he will be unable to pay back a single penny for the next five years. He is considering two options. His father has offered to lend him the money with a simple interest rate of 10%. A friend has also

offered to lend him the money with an compound interest rate of 7%. What would you suggest?

9. When negotiating a mortgage the following parameters are established: the mortgage rate, the amount to be borrowed, the amortization period, the payment period (normally monthly, but sometimes weekly or biweekly), and the *mortgage term*. The mortgage term is always less than or equal to the amortization period. At the end of the term, the bank and the borrower renegotiate the terms of the mortgage, with the remaining principal being considered as the borrowed amount.
(a) A couple buys a home and must borrow $100,000 to pay for it. They opt for a 25-year amortization period. Since the interest rates are relatively high at the time of purchase (12%), they decide to choose a relatively short term of three years. What will their monthly payment be during those three years? How much will they owe at the end of the term?
(b) During the first three years, the interest rate has fallen to 8%. They decide that they still wish to pay off their home at the end of 22 more years, and they renew their mortgage for a term of five years. What will their new monthly payment be? What will be the outstanding balance at the end of the second term?

10. Two mortgages are offered for the same amount of money, both with an amortization period of 20 years. If the interest rates are different, which interest rate will have permitted the payment of a larger portion of the outstanding balance after 10 years: the mortgage with the higher interest rate, or that with the lower one?

11. You can buy books of *mortgage payment tables* in nearly any bookstore. In an appendix to this chapter you will find tables corresponding to nominal mortgage rates of 8% and 12% (see Tables 5.1 and 5.2). The monthly rates have been calculated according to Canadian rules.
(a) According to these tables, what will the monthly payment be for a $40,000 mortgage at 8% with an amortization period of 12 years?
(b) What about for a $42,000 loan with same amortization period and rate?
(c) Calculate the answer to question (a) directly, without using the table. You will first need to calculate the effective monthly rate r_m.

12. Several banks offer mortgages with biweekly payments. These banks calculate the payment that must be paid back as if the borrower were making 24 payments per year (two per month), even though the borrower makes 26 payments per year. This allows the mortgage to be paid off more quickly than its full amortization period. Consider a 20-year mortgage at 7%. How many years will it take to fully pay off the mortgage? (You will have to decide on a fair biweekly rate r_{bw}. Try to imitate formula ($r_{m\text{CAN}}$).)

13. Use software of your choice to write a program that reproduces the tables in the appendix.

References

[1] E.M. Bruins and M. Rutten. *Textes mathématiques de Suse*, volume XXXIV of *Mémoires de la Mission archéologique en Iran*. Librairie orientaliste Paul Geuthner, Paris, 1961.

[2] G.W. Leibniz. *Hauptschriften zur Versicherungs- und Finanzmathematik*. Akademie Verlag, Berlin, 2000. (Edited by E. Knobloch and J.-M. Graf von der Schulenburg.)

6

Error-Correcting Codes

The elementary parts of this chapter are found in Sections 6.1 through 6.4. They explain the necessity for error-correcting codes, introduce the finite field $\mathbb{F}_2$, and discuss the Hamming family of error-correcting codes. While the concept of the field $\mathbb{F}_2$ will likely be new to some students, the elementary sections of this chapter use only the concepts of a vector space (over $\mathbb{F}_2$) and basic linear algebra. These sections can be covered in three hours of class. Sections 6.5 and 6.6 constitute the advanced portion of the material. We construct the finite fields $\mathbb{F}_{p^r}$, for p prime, by introducing the notion of multiplication modulo an irreducible polynomial. Several thorough examples help students to digest this initially difficult concept. Reed–Solomon codes are presented in the last section. Covering the advanced material requires at least three additional hours of class time.

6.1 Introduction: Digitizing, Detecting and Correcting

The transmission of information over long distances began very early in human history.[1] The discovery of electromagnetism and its many applications allowed us to send messages through wires and electromagnetic waves in the second half of the nineteenth century. Whether the message is sent in spoken word (in any human language) or an encoded form (using Morse code (1836), for example), the utility of being able to rapidly detect and correct errors is obvious.

An early method for improving the fidelity of a transmission is of historic importance. When telephones were first invented (both wired and wireless), the quality of transmission left much to be desired. Thus, rather than speaking directly, it was quite usual to spell out words phonetically. For example, in order to say the word "error," the caller

[1] According to legend, the soldier charged with reporting the victory of the Athenians over the Persians in 490 BC had to run the distance between Marathon and Athens, dying from exhaustion on his arrival. The distance of the Olympic marathon is now 42.195 km.

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_6,

would instead say "Echo, Romeo, Romeo, Oscar, Romeo." The American and British armies had devised such "alphabets" by the First World War. This method of improving the reliability of transmission works by multiplying the information; the hope is that the receiver can extract the original message, "error," from the code, "Echo, Romeo, Romeo, Oscar, Romeo," even when reception quality is poor. This "multiplication of information" or *redundancy* is the idea underlying all error detectors and correctors.

Our second example is that of an error-detection code: it allows us to detect when an error has occurred in the transmission, but it does not let us correct it. In computer science it is normal to associate each character of our extended alphabet (a, b, c, ..., A, B, C, ..., 0, 1, 2, ..., +, -, :, ;, ...) with a number between 0 and 127.[2] In a binary representation, seven *bits* (a contraction of "binary digits") are required to represent each of the $2^7 = 128$ possible characters. For example, suppose that the letter *a* is associated with the number 97. Because $97 = 64 + 32 + 1 = 1 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^0$, the letter *a* will be encoded as 1100001. The usual encoding is the following "dictionary":

	decimal	binary	parity + binary
A	65	1000001	01000001
B	66	1000010	01000010
C	67	1000011	11000011
⋮	⋮	⋮	⋮
a	97	1100001	11100001
b	98	1100010	11100010
c	99	1100011	01100011
⋮	⋮	⋮	⋮

To detect errors we add an eighth bit to each character, called a *parity bit*. This bit is placed in the leftmost position, and is calculated such that the sum of all eight bits will always be even. For example, since the sum of the seven bits for "A" is $1+0+0+0+0+0+1 = 2$, the parity bit is 0, and "A" will be represented by 01000001. Similarly, the sum of the seven bits of "a" is $1 + 1 + 0 + 0 + 0 + 0 + 1 = 3$ and "a" will be represented by the eight bits 11100001. This parity bit is an error-detection code. It allows us to detect when a single error has occurred in the transmission, but it does not allow us to correct for it, since we have no way of knowing which of the eight bits has been altered. However, once the receiver has determined that an error has occurred, he can simply ask for the affected character to be retransmitted. Note that this error

[2] This is commonly known as a 7-bit ASCII encoding, which is good only for encoding languages using a small number of characters, like English. There is a variety of text encodings for other languages using extensive sets of characters.

detector assumes that at most one bit will be in error. This hypothesis is reasonable if the transmission is nearly perfect and there is a low probability that two in eight consecutive bits will be in error.

Our third example presents a simple idea for constructing an error-correcting code. Such a code allows us both to detect *and* correct errors. It consists in simply sending the entire message several times. For example, we could simply repeat each character in a message twice. Thus, the word "error" could be transmitted as "eerrrroorr." As such, this is only an error-detection code, since we have no way of knowing where the error is if one is detected. Which is the correct message if we receive "aaglee": "age" or "ale"? In order to make this an error-correcting code, we simply need to repeat each letter three times. If it is reasonable to assume that no more than one in three letters will be received in error, then the correct letter can be determined as a simple majority. For example, the message "aaaglleee" would be received as "ale" and not "age." Such a simple error-correcting code is not used in practice, since it is very costly: it triples the cost of sending each message! The codes that we will present in this chapter are much more economical. Note that it is not impossible for two or even three errors to occur in a sequence of three characters; our hypothesis is only that this is *very* unlikely. As Exercise 8 will show, this simple code has a very small advantage as compared to the simplest of Hamming codes, introduced in Section 6.3.

Both error-detecting and error-correcting codes have existed for a long time. In the digital age these codes have become more necessary and easier to implement. Their usefulness can be understood better when one knows the size of usual picture and music files. Figure 6.1 shows a very small digitized photo of the peak of a tower at the Université de Montréal, in Montréal, Canada. At the left, the photo is shown at its intended resolution, while at the right, it has been enlarged eight times, allowing the individual pixels to be seen clearly. The image was divided into 72×72 pixels, each of which is represented by a number between 0 and 255, indicating the intensity of gray from black to white. Each pixel requires 8 bits, meaning that transmitting this tiny black-and-white image requires sending $72 \times 72 \times 8 = 41{,}472$ bits. And this example is far from the current digital cameras, whose sensors capture more than $2{,}000 \times 3{,}000$ pixels in color![3]

Sound, music in particular, is very often stored in digital form. In contrast to images, digitizing sound is harder to visualize. Sound is a type of wave. Waves in the ocean undulate along the surface of the water, light is a wave in the electromagnetic field, and sound is a wave in air density. If we measured the density of the air at a fixed location near a (well-tuned) piano, we would see that the density increases and decreases 440 times a second when the middle A is played. The variation is very small, but out ears are able to detect it and translate it to an electric wave that is then transmitted to and analyzed by our brain. Figure 6.2 shows a representation of this pressure wave.

[3] Those who work regularly with computers are used to seeing file sizes expressed in bytes (1 byte = 8 bits), kilobytes (1 KB = 1,000 bytes), megabytes (1 MB = 10^6 bytes), or even gigabytes (1 GB = 10^9 bytes). Our image therefore consumes 44,472/8 B = 5,184 B = 5.184 KB.)

Fig. 6.1. A digitized photo: the "original" photo is at the left, while the same image is seen eight-times enlarged at the right.

(The horizontal axis indicates time, while the vertical axis indicates the amplitude of the wave.) When the value is positive, this indicates that the density of the air is higher than normal (air at rest), while negative values indicate a decreased density. This wave can be digitized by approximating it with a step function. Each short time period of Δ seconds is approximated by the average value of the wave over the time interval. If Δ is sufficiently small, the step-function approximation to the wave is indistinguishable from the original as heard by the human ear. (Figure 6.3 shows another sound wave and a step function digitization of it.) This digitization having been accomplished, the wave may now be represented by a sequence of integers identifying the heights of the steps along some predefined scale.

On compact discs, the sound wave is cut into 44,100 samples per second (equivalent to a pixel in a photo), and the intensity of each sample is represented by a 16-bit integer

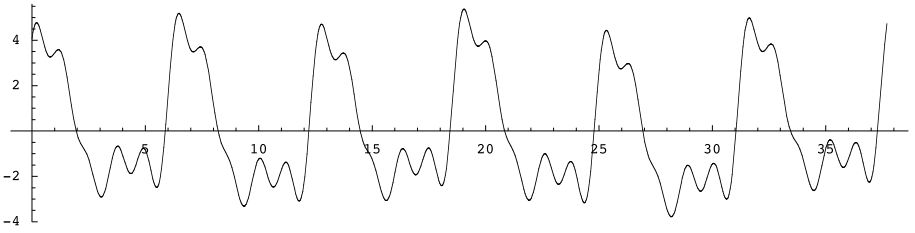


Fig. 6.2. A sound wave measured over a fraction of a second.

($2^{16} = 65{,}536$).[4] Recalling that compact discs store sound in stereo, then we see that each second of music requires $44{,}100 \times 16 \times 2 = 1{,}411{,}200$ bits and 70 minutes of audio requires $1{,}411{,}200 \times 60 \times 70 = 5{,}927{,}040{,}000$ bits = 740,880,000 bytes $\approx$ 740 MB. Given such a large mass of data, it is desirable to be able to automatically detect and correct errors.[5]

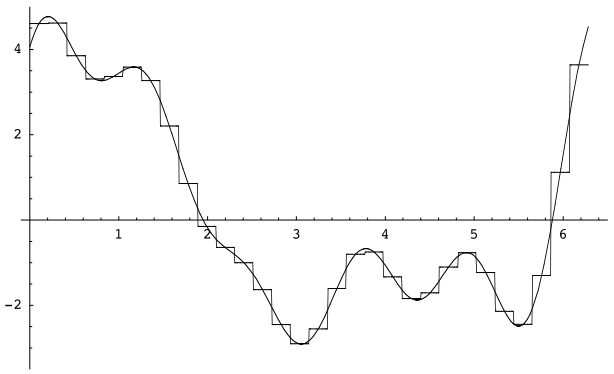


Fig. 6.3. A sound wave and a step-function approximation to it.

This chapter explores two classic families of error-correcting codes: those of Hamming and those of Reed and Solomon. The first of these was used by France-Telecom for the transmission of Minitel, a precursor to the modern Internet. Reed–Solomon codes are used in compact discs. The Consultative Committee for Space Data Systems,

[4] Sony and Philips worked together to establish the Compact Disc standard. After hesitating between a 14-bit and a 16-bit intensity scale, the engineers opted for the finer-grained scale [7]. For much more detail, see Chapter 10.

[5] The scientific development of the field of error-correcting codes and their applications has been followed closely by *Scientific American*. See, for example, [3, 4, 5].

created in 1982 for standardizing the practices of different space agencies, recommended the use of Reed–Solomon codes for information transmitted over satellites.

6.2 The Finite Field $\mathbb{F}_2$

In order to discuss Hamming codes we must first be comfortable working with the finite field of two elements $\mathbb{F}_2$. A field is a collection of elements on which we can define two operations, called "addition" and "multiplication," which must each satisfy properties that are common for rationals and real numbers: associativity, commutativity, distributivity of multiplication with respect to addition, the existence of an identity element for each of addition and multiplication, the existence of an additive inverse, and the existence of a multiplicative inverse for all nonzero elements. The reader will surely recognize the rationals $\mathbb{Q}$, the reals $\mathbb{R}$, and maybe the complex numbers $\mathbb{C}$ as having these properties. These three sets, combined with the normal $+$ and $\times$ operations, are fields. But there exist many others!

Although we will discuss the mathematical structure of fields in more generality in Section 6.5, we begin by providing rules for addition and multiplication over the set of binary digits $\{0, 1\}$. The addition and multiplication tables are given by

$+$	0	1
0	0	1
1	1	0

$\times$	0	1
0	0	0
1	0	1

(6.1)

These operations satisfy the same rules that are satisfied by the fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$: associativity, commutativity, distributivity, and the existence of identity elements and inverses. For example, using both tables above we can verify that for all $x, y, z \in \mathbb{F}_2$, distributivity is satisfied:

$$x \times (y + z) = x \times y + x \times z.$$

Since x, y, and z each take one of two values, this property can be fully proved by considering each of the eight possible combinations of the triplet $(x, y, z) \in \{(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$. Here we show an explicit verification of the distributivity property for the triplet $(x, y, z) = (1, 0, 1)$:

$$x \times (y + z) = 1 \times (0 + 1) = 1 \times 1 = 1$$

and

$$x \times y + x \times z = 1 \times 0 + 1 \times 1 = 0 + 1 = 1.$$

As in $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, 0 is the identity element for addition and 1 is the identity element for multiplication. Inspection shows that all elements have an additive inverse. (Exercise: what is the additive inverse of 1?) Similarly, each element of $\mathbb{F}_2 \setminus \{0\}$ has a multiplicative

inverse. Verifying this last property is very simple, since there is only one element in $\mathbb{F}_2 \setminus \{0\} = \{1\}$, and its multiplicative inverse is itself, since $1 \times 1 = 1$.

Much as we define the vector spaces $\mathbb{R}^3$, $\mathbb{R}^n$, and $\mathbb{C}^2$, it is entirely possible to consider three-dimensional vector spaces in which each of the entries is an element of $\mathbb{F}_2$. It is possible to perform vector addition and scalar multiplication (with coefficients from $\mathbb{F}_2$, obviously!) of these vectors in $\mathbb{F}_2^3$ using the definition of addition and multiplication in $\mathbb{F}_2$. For example,

$$(1,0,1) + (0,1,0) = (1,1,1),$$
$$(1,0,1) + (0,1,1) = (1,1,0),$$

and

$$0 \cdot (1,0,1) + 1 \cdot (0,1,1) + 1 \cdot (1,1,0) = (1,0,1).$$

Since the components must be in $\mathbb{F}_2$ and only linear combinations with coefficients from $\mathbb{F}_2$ are permitted, the number of vectors in $\mathbb{F}_2^3$ (and in any $\mathbb{F}_2^n$ for finite n) is finite! Caution: even though the dimension of $\mathbb{R}^3$ is finite, the number of vectors in $\mathbb{R}^3$ is infinite. On the other hand, there are only $2^3 = 8$ vectors in the vector space $\mathbb{F}_2^3$, given by

$$\{(0,0,0),(0,0,1),(0,1,0),(1,0,0),(0,1,1),(1,0,1),(1,1,0),(1,1,1)\}.$$

(Exercise: recall the formal definition of the dimension of a vector space and calculate the dimension of $\mathbb{F}_2^3$.) Vector spaces over finite fields such as $\mathbb{F}_2$ may seem a little daunting at first because most linear algebra courses do not discuss them, but many of the methods of linear algebra (matrix calculations, among others) apply to them.

6.3 The $C(7,4)$ Hamming Code

Here is a first example of a modern error-correcting code. Rather than using the normal alphabet $(a, b, c, \ldots)$, it uses the elements of $\mathbb{F}_2$.[6] Moreover, we limit ourselves to transmitting "words" containing exactly four "letters" (u_1, u_2, u_3, u_4). (Exercise: does this restriction limit us?) Our vocabulary, or *code* $C = \mathbb{F}_2^4$, therefore contains only 16 "words" or *elements*. Rather than transmitting the four symbols u_i to represent an element, we will instead transmit the seven symbols defined as follows:

[6] This is not really a restriction, since we have already seen ways of encoding the alphabet using only these binary digits.

$$
\begin{aligned}
v_1 &= u_1,\\
v_2 &= u_2,\\
v_3 &= u_3,\\
v_4 &= u_4,\\
v_5 &= u_1 + u_2 + u_4,\\
v_6 &= u_1 + u_3 + u_4,\\
v_7 &= u_2 + u_3 + u_4.
\end{aligned}
$$

Thus, to transmit the element $(1, 0, 1, 1)$ we send the message

$$(v_1, v_2, v_3, v_4, v_5, v_6, v_7) = (1, 0, 1, 1, 0, 1, 0),$$

since

$$
\begin{aligned}
v_5 &= u_1 + u_2 + u_4 = 1 + 0 + 1 = 0,\\
v_6 &= u_1 + u_3 + u_4 = 1 + 1 + 1 = 1,\\
v_7 &= u_2 + u_3 + u_4 = 0 + 1 + 1 = 0.
\end{aligned}
$$

(Note: "+" is the addition operator over $\mathbb{F}_2$.)

Since the first four coefficients of $(v_1, v_2, \ldots, v_7)$ are precisely the four symbols we wish to transmit, what purpose do the other three symbols serve? These symbols are *redundant* and allow us to correct any single erroneous symbol. How can we accomplish this "miracle"?

We consider an example. The receiver receives the seven symbols $(w_1, w_2, \ldots, w_7) = (1, 1, 1, 1, 1, 0, 0)$. We distinguish the received symbols w_i from the sent symbols v_i in case of an error in the transmission. Due to the quality of the transmission link, it is reasonable for us to assume that at most one symbol will be in error. The receiver then calculates

$$
\begin{aligned}
W_5 &= w_1 + w_2 + w_4,\\
W_6 &= w_1 + w_3 + w_4,\\
W_7 &= w_2 + w_3 + w_4,
\end{aligned}
$$

and compares them with w_5, w_6, and w_7 respectively. If there is no error due to the transmission, W_5, W_6, and W_7 should coincide with w_5, w_6, and w_7 that were received. Here is the calculation

$$
\begin{aligned}
W_5 &= w_1 + w_2 + w_4 = 1 + 1 + 1 = 1 = w_5,\\
W_6 &= w_1 + w_3 + w_4 = 1 + 1 + 1 = 1 \neq w_6,\\
W_7 &= w_2 + w_3 + w_4 = 1 + 1 + 1 = 1 \neq w_7.
\end{aligned}
\tag{6.2}
$$

The receiver realizes that an error has occurred, since two of these calculated values (W_6 and W_7) are not in agreement with those received. But where is the error? Is it

in one of the four original symbols or in one of the three redundant ones? It is simple to exclude the possibility that one of w_5, w_6, and w_7 is in error. By changing only *one* of these values, there will remain a second identity that is not satisfied. Thus one of the first four symbols must be in error. Among these letters, which can we change that will simultaneously correct the two incorrect values of (6.2) while preserving the correct value of the first? The answer is simple: we must correct w_3. In fact, the first sum does not contain w_3 and thus is the only one that will not be affected by changing it. The two other relations do contain w_3, and they will both be "corrected" by the change. Thus, even though the first four symbols of the message were received as $(w_1, w_2, w_3, w_4) = (1, 1, 1, 1)$, the receiver determines the correct message as $(v_1, v_2, v_3, v_4) = (1, 1, 0, 1)$.

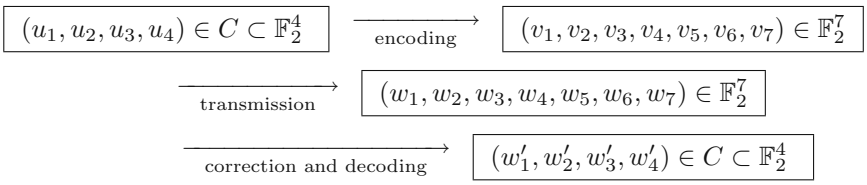
Consider each of the possibilities. Suppose that the receiver received the symbols $(w_1, w_2, \ldots, w_7)$. The only thing the receiver knows for sure (according to our hypothesis) is that these symbols correspond to the seven transmitted symbols $v_i = i, \ldots, 7$, with the exception of at most one error. Thus, there are eight possibilities:

(0) all of the symbols are correct,
(1) w_1 is in error,
(2) w_2 is in error,
(3) w_3 is in error,
(4) w_4 is in error,
(5) w_5 is in error,
(6) w_6 is in error,
(7) w_7 is in error.

Using the redundant symbols, the receiver can determine which of these possibilities is correct. By calculating W_5, W_6, and W_7, he can determine which of the eight possibilities holds with the help of the following table:

(0) if $w_5 = W_5$ and $w_6 = W_6$ and $w_7 = W_7$,
(1) if $w_5 \neq W_5$ and $w_6 \neq W_6$,
(2) if $w_5 \neq W_5$ and $w_7 \neq W_7$,
(3) if $w_6 \neq W_6$ and $w_7 \neq W_7$,
(4) if $w_5 \neq W_5$ and $w_6 \neq W_6$ and $w_7 \neq W_7$,
(5) if $w_5 \neq W_5$,
(6) if $w_6 \neq W_6$,
(7) if $w_7 \neq W_7$.

The hypothesis that at most one symbol is in error is crucial to this analysis. If two letters had been in error, then the receiver would not be able to distinguish, for example, between the cases "w_1 is in error" and "w_5 and w_6 are both in error" and would therefore not be able to perform the appropriate correction. However, in the case of at most one error the receiver can always detect and correct the error. After having discarded the three extra symbols, the receiver is assured of having received the originally intended message. The process can be visualized as



How does the $C(7,4)$ Hamming code compare to other error-correcting codes? This question is a little too vague. In fact, the quality of a code can be judged only as a function of the needs: the error rate of the channel, the average length of messages to be sent, the processing power available for encoding and decoding, etc. Nonetheless, we can compare it to our simple method of repetition. Each of the symbols $u_i, i = 1, 2, 3, 4$, could be repeated until we attained sufficient confidence that the message will be correctly decoded. We again take the hypothesis that at most one bit error can occur every "few" bits (fewer than 15 bits). As we have already seen, if each symbol is sent twice, we are able only to detect an error. Thus, we must transmit each symbol at least 3 times, requiring a total of 12 bits to send this 4-bit message. The Hamming code is able to send the same message with the same confidence in only 7 bits, a significant improvement.

6.4 $C(2^k - 1, 2^k - k - 1)$ Hamming Codes

The $C(7,4)$ Hamming code is the first in a family of $C(2^k - 1, 2^k - k - 1)$ Hamming codes. Each of these codes allows for the correction of at most a single error. The numbers $2^k - 1$ and $2^k - k - 1$ indicate the *length* of a code element and the *dimension* of the subspace formed by the transmitted elements, respectively. Thus, $k = 3$ yields the $C(7,4)$ code, which transmits 7-bit elements in the field $\mathbb{F}_2^7$, and these form a subspace of dimension 4 that is isomorphic to $\mathbb{F}_2^4$.

Two matrices play an important role in the description of Hamming codes (and in the description of all "linear" codes, a family to which Reed–Solomon codes also belong): the *generating matrix* G and the *control matrix* H. The generating matrix G_k is of size $(2^k - k - 1) \times (2^k - 1)$, and its rows form a basis for a subspace that is isomorphic to $\mathbb{F}_2^{(2^k - k - 1)}$. Each element of the code is a linear combination of this basis. For $C(7,4)$ the matrix G_3 can be chosen as

$$G_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

For example, the first line of G_3 corresponds to the element encoding the message $u_1 = 1$ and $u_2 = u_3 = u_4 = 0$. By the rules that we have chosen, it follows that $v_1 = 1, v_2 = v_3 = v_4 = 0$, $v_5 = u_1 + u_2 + u_4 = 1$, $v_6 = u_1 + u_3 + u_4 = 1$, and

$v_7 = u_2 + u_3 + u_4 = 0$. These are the entries of the first row. The 16 elements of the code C are obtained by performing the 16 possible linear combinations of the four rows of G_3. Since G requires only that its rows form a basis, it is not uniquely defined.

The control matrix H is a $k \times (2^k - 1)$ matrix whose k rows form a basis for the orthogonal complement of the subspace spanned by the rows of G. The scalar product is as usual: if $v, w \in \mathbb{F}_2^n$, then $(v, w) = \sum_{i=1}^n v_i w_i \in \mathbb{F}_2$. (The appendix at the end of this chapter formally defines scalar products and explores the important differences between scalar products over the "usual" fields ($\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$) and those over finite fields. A few of these differences are not very intuitive!) For $C(7, 4)$ and our choice of G_3 above, the control matrix H_3 can be chosen as:

$$H_3 = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Since the rows of G and H are pairwise orthogonal, the matrices G and H satisfy

$$GH^t = 0. \tag{6.3}$$

For example, for $k = 3$,

$$G_3 H_3^t = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}}_{4\times 7} \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{7\times 3} = \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{4\times 3}.$$

The general $C(2^k - 1, 2^k - k - 1)$ Hamming code is defined by the control matrix H. The columns of this matrix are precisely all of the nonzero vectors of $\mathbb{F}_2^k$. Since $\mathbb{F}_2^k$ contains 2^k vectors (including the zero vector), H must be a $k \times (2^k - 1)$ matrix. The matrix H_3 given above is an example. As noted earlier, the rows of the generating matrix G form a basis to the orthogonal complement of the span of the rows of H. This concludes the definition of $C(2^k - 1, 2^k - k - 1)$ Hamming codes.

We now discuss the encoding and decoding process.

In our choice of G_3 each of the rows corresponds to the elements $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$ of $\mathbb{F}_2^4$. To obtain a general element (u_1, u_2, u_3, u_4) it suffices to take a linear combination of the four rows of G_3:

$$\begin{pmatrix} u_1 & u_2 & u_3 & u_4 \end{pmatrix} G_3 \in \mathbb{F}_2^7.$$

(Exercise: verify that the matrix product $\begin{pmatrix} u_1 & u_2 & u_3 & u_4 \end{pmatrix} G_3$ yields a 1×7 matrix.) The encoding of $u \in \mathbb{F}_2^{2^k-k-1}$ in the $C(2^k - 1, 2^k - k - 1)$ code is done in exactly the same manner:

$$v = uG \in \mathbb{F}_2^{2^k-1}.$$

Encoding is therefore a simple matrix multiplication over the field $\mathbb{F}_2$.

Decoding is a little more subtle. The following two observations form the heart of this procedure. The first is relatively direct: an element of the code $v \in \mathbb{F}_2^{2^k-1}$ without any errors is annihilated by the control matrix,

$$Hv^t = H(uG)^t = HG^t u^t = (GH^t)^t u^t = 0,$$

due to the pairwise orthogonality between the rows of G and H.

The second observation is a little deeper. Let $v \in \mathbb{F}_2^{2^k-1}$ be an element of the code (without error) and $v^{(i)} \in \mathbb{F}_2^{2^k-1}$ the word obtained from v by adding a 1 to the ith entry of v. Thus $v^{(i)}$ is an encoded element with an error in the ith position. Note that $H(v^{(i)})^t \in \mathbb{F}_2^k$ is independent of v! In fact,

$$v^{(i)} = v + (0, 0, \ldots, 0, \underbrace{1}_{\text{position } i}, 0, \ldots, 0)$$

and

$$H(v^{(i)})^t = Hv^t + H\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = H\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \text{ position } i,$$

since v is an element of the code. Thus $H(v^{(i)})^t$ is the ith column of H. Since all of the columns of H are distinct (by the definition of H), an error in the ith position of the received encoded element w is equivalent to obtaining the ith column of H in the product Hw^t.

Decoding proceeds as follows:

$$\boxed{w \in \mathbb{F}_2^{2^k-1} \text{ received}} \longrightarrow \boxed{\text{calculation of } Hw^t \in \mathbb{F}_2^k}$$

$$\longrightarrow \boxed{\begin{array}{ccc} Hw^t \text{ is zero} & \Rightarrow & w \text{ skips to the next step} \\ Hw^t \text{ is equal to column } i \text{ of } H & \Rightarrow & \text{the entry } w_i \text{ is changed} \end{array}}$$

$$\longrightarrow \boxed{\begin{array}{c} \text{search for the linear combination of the} \\ \text{rows of } G \text{ that yields the corrected } w \end{array}}$$

Although these codes can correct for only a single error, they are very economical for sufficiently large k. For example, for $k = 7$ it suffices to add 7 bits to a message of length 120 in order to be certain that any single error may be corrected. It is precisely this $C(2^k - 1, 2^k - k - 1)$ Hamming code with $k = 7$ that is used in the Minitel system.

6.5 Finite Fields

In order to present the Reed–Solomon code we will need to know several properties of finite fields. This section covers the required background material.

Definition 6.1 *A field $\mathbb{F}$ is a set over which two operations $+$ and $\times$ have been defined, and within which two special elements denoted by 0 and $1 \in \mathbb{F}$ have been identified that satisfy the following five properties:*

(P1) commutativity:
$a + b = b + a \quad$ *and* $\quad a \times b = b \times a, \qquad \forall a, b \in \mathbb{F},$
(P2) associativity:
$(a + b) + c = a + (b + c) \quad$ *and* $\quad (a \times b) \times c = a \times (b \times c), \qquad \forall a, b, c \in \mathbb{F},$
(P3) distributivity:
$(a + b) \times c = (a \times c) + (b \times c), \qquad \forall a, b, c \in \mathbb{F},$
(P4) additive and multiplicative identity:
$a + 0 = a \quad$ *and* $\quad a \times 1 = a, \qquad \forall a \in \mathbb{F},$
(P5) existence of additive and multiplicative inverses:
$\forall\ a \in \mathbb{F},\ \exists\ a' \in \mathbb{F} \quad$ *such that* $\quad a + a' = 0,$
$\forall\ a \in \mathbb{F} \setminus \{0\},\ \exists\ a' \in \mathbb{F} \quad$ *such that* $\quad a \times a' = 1.$

Definition 6.2 *A field $\mathbb{F}$ is called* finite *if the number of elements in $\mathbb{F}$ is finite.*

Example 6.3 *The three most familiar fields are $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, the sets of rational, real, and complex numbers, respectively. They are* not *finite. The above list of properties is probably familiar to most readers. The goal of giving a precise definition of a field is to reduce the properties of these three sets of numbers to a set of axioms. The advantage to this approach is that the entire mechanism of calculation developed over these fields may then be extended to less-intuitive fields that satisfy these same properties.*

Example 6.4 *The set $\mathbb{F}_2$ equipped with $+$ and $\times$ as given in Section 6.2 is a field. The calculations performed in our study of Hamming codes have likely already convinced you of this fact. A systematic verification of this proposition is covered in Exercise 4.*

Example 6.5 *$\mathbb{F}_2$ is only the first among a family of finite fields. Let p be a prime number. We say that two numbers a and b are congruent modulo p if p divides their difference $a - b$. The congruence forms an equivalence relation over the integers. This*

relation induces exactly p distinct classes of equivalence, represented by $\bar{0}, \bar{1}, \ldots, \overline{p-1}$. *For example, for* $p = 3$, *the integers* $\mathbb{Z}$ *are partitioned into three subsets*

$$\begin{aligned}\bar{0} &= \{\ldots, -6, -3, 0, 3, 6, \ldots\},\\ \bar{1} &= \{\ldots, -5, -2, 1, 4, 7, \ldots\},\\ \bar{2} &= \{\ldots, -4, -1, 2, 5, 8, \ldots\}.\end{aligned}$$

The set $\mathbb{Z}_p = \{\bar{0}, \bar{1}, \bar{2}, \ldots, \overline{p-1}\}$ *is the set of these equivalence classes. We define the operations* $+$ *and* $\times$ *over these classes as addition modulo p and multiplication modulo p. In order to perform addition modulo p between two classes* $\bar{a}$ *and* $\bar{b}$, *we choose one element from each of these classes (we will choose a and b). The result of* $\bar{a} + \bar{b}$ *is* $\overline{a+b}$, *the class to which the sum of the chosen elements belongs. (Exercise: why is this result independent of our choice of elements from each of* $\bar{a}$ *and* $\bar{b}$? *Does this definition coincide with that given previously for* $\mathbb{F}_2$ *in Section 6.2?) Multiplication between equivalence classes is defined analogously. It is usual to omit the "*$\bar{}$*" that denotes the equivalence class. Exercise 24 verifies that* $(\mathbb{Z}_p, +, \times)$ *is in fact a field.*

Example 6.6 *The set of integers* $\mathbb{Z}$ *is not a field. For example, the element* 2 *does not have a multiplicative inverse.*

Example 6.7 *Let* $\mathbb{F}$ *be a field. Denote by* $\tilde{\mathbb{F}}$ *the set of all quotients of polynomials in a single variable x with coefficients in* $\mathbb{F}$. *Thus, all elements of* $\tilde{\mathbb{F}}$ *are of the form* $\frac{p(x)}{q(x)}$ *for* $p(x)$ *and* $q(x)$ *polynomials (with finite degree by definition) with coefficients in* $\mathbb{F}$ *such that q is nonzero. If we equip* $\tilde{\mathbb{F}}$ *with the usual operations of addition and multiplication, then* $(\tilde{\mathbb{F}}, +, \times)$ *is a field. The quotient* $0/1 = 0$ *(the quotient with* $p(x) = 0$ *and* $q(x) = 1$*) and the quotient* 1 *(the quotient with* $p(x) = q(x) = 1$*) are the additive and multiplicative identities, respectively. We can easily verify properties (P1) through (P5).*

The set $\mathbb{Z}_p$ mentioned above deserves closer inspection. The addition and multiplication tables for $\mathbb{Z}_3$ are given by

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\times$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

(6.4)

and those of $\mathbb{Z}_5$ are

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\times$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

(6.5)

(Exercise: verify that these tables accurately represent addition and multiplication modulo 3 and 5, respectively.) The example introducing the field $\mathbb{Z}_p$ stipulated that p must be prime. What happens if it is not? Here are the addition and multiplication tables modulo 6 over the set $\mathbb{Z}_5 = \{0, 1, 2, 3, 4, 5\}$:

$+$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\times$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	**0**	2	4
3	0	3	**0**	3	**0**	3
4	0	4	2	**0**	4	2
5	0	5	4	3	2	1

(6.6)

How can we prove that $\mathbb{Z}_6$ equipped with these addition and multiplication tables is not a field? With the help of the bold zeros in the multiplication table! The proof follows.

We know that $0 \times a = 0$ in $\mathbb{Q}$ and in $\mathbb{R}$. Is this true for all nonzero elements a in a given field $\mathbb{F}$? Yes! The proof that follows is elementary. (While reading it, notice that each step follows directly from one of the five defining properties of a field.) Let a be a nonzero element of $\mathbb{F}$. Then

$$\begin{aligned} 0 \times a &= (0+0) \times a && \text{(P4)} \\ &= 0 \times a + 0 \times a && \text{(P3).} \end{aligned}$$

By (P5) all elements of $\mathbb{F}$ possess an additive inverse. Let b be the additive inverse of $(0 \times a)$. Add this element to both sides of the above equation, yielding

$$(0 \times a) + b = (0 \times a + 0 \times a) + b.$$

The left-hand side of the equation is zero (by definition of b), while the right-hand side may be rewritten

$$\begin{aligned} 0 &= 0 \times a + ((0 \times a) + b) && \text{(P2)} \\ &= 0 \times a + 0 \\ &= 0 \times a && \text{(P4)}, \end{aligned}$$

due to our choice of b. Thus $0 \times a$ is zero regardless of our choice of $a \in \mathbb{F}$. We again consider the multiplication table for a field $\mathbb{F}$. Let a and $b \in \mathbb{F}$ be two nonzero elements of $\mathbb{F}$ such that

$$a \times b = 0.$$

By multiplying both sides of this equation by the multiplicative inverse b' of b (which exists by (P5)), we have that

$$a \times (b \times b') = 0 \times b',$$

and by the property we just showed it follows that

$$a \times 1 = 0.$$

By (P4) we have that

$$a = 0,$$

which is a contradiction, since a was chosen to be nonzero. *Thus, in a field $\mathbb{F}$, the product of nonzero elements must be nonzero.* And therefore $(\mathbb{Z}_6, +, \times)$ is not a field, due to the bold zeros in its multiplication table.

If p is not a prime number, there exist q_1 and q_2 different from 0 and 1 such that $p = q_1 q_2$. In $\mathbb{Z}_p$ we would have $q_1 \times q_2 = p = 0 \pmod{p}$. *Thus, if p is not prime then $\mathbb{Z}_p$ equipped with the operation of addition and multiplication modulo p is not a field.* We will use this fact to introduce a result that we will not prove here.

Denote by $\mathbb{F}[x]$ the set of polynomials with coefficients in $\mathbb{F}$ and a single variable x. This set can be equipped with addition and multiplication operations as usual. Note: $\mathbb{F}[x]$ is not a field. For example, the nonzero element $(x + 1)$ does not have a multiplicative inverse.

Example 6.8 *$\mathbb{F}_2[x]$ is the set of all polynomials in x with coefficients in $\mathbb{F}_2$. Here is an example of multiplication in $\mathbb{F}_2[x]$:*

$$(x + 1) \times (x + 1) = x^2 + x + x + 1 = x^2 + (1 + 1)x + 1 = x^2 + 1 \in \mathbb{F}_2[x].$$

In the same way that we can calculate "modulo p" it is possible to calculate "modulo a polynomial $p(x)$." Let $p(x) \in \mathbb{F}[x]$ be a polynomial with degree $n \geq 1$:

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_i \in \mathbb{F}, 0 \leq i \leq n$, and $a_n \neq 0$. Without loss of generality, we will restrict ourselves to polynomials such that $a_n = 1$. The addition and multiplication operations consist in performing normal addition and multiplication of polynomials where individual operations on coefficients are performed in the field $\mathbb{F}$, and then repeatedly removing multiples of the polynomial $p(x)$ until the resulting polynomial has degree less than n. This may sound somewhat complicated, but a few examples will clarify it.

Example 6.9 *Let $p(x) = x^2 + 1 \in \mathbb{Q}[x]$ and let $(x + 1)$ and $(x^2 + 2x)$ be two other polynomials in $\mathbb{Q}[x]$ that we wish to multiply modulo $p(x)$. The equalities that follow are between polynomials that differ only by a multiple of $p(x)$. These are not strict equalities (the polynomials are clearly not equal in the normal sense), as indicated by the "*$\bmod p(x)$*" in the last line:*

$$\begin{aligned}(x+1)\times(x^2+2x) &= x^3+2x^2+x^2+2x\\ &= x^3+3x^2+2x-x(x^2+1)\\ &= x^3-x^3+3x^2+2x-x\\ &= 3x^2+x\\ &= 3x^2+x-3(x^2+1)\\ &= 3x^2-3x^2+x-3\\ &= x-3 \pmod{p(x)}.\end{aligned}$$

You can readily check that $(x-3)$ *is the remainder of the division of* $(x+1)\times(x^2+2x)$ *by* $p(x)$*. This is not a coincidence. This is a general property that actually gives an alternative method to calculate* $q(x) \pmod{p(x)}$*. See Exercise 14.*

Example 6.10 *Let* $p(x) = x^2+x+1 \in \mathbb{F}_2[x]$*. The square of the polynomial* (x^2+1) *modulo* $p(x)$ *is*

$$\begin{aligned}(x^2+1)\times(x^2+1) = x^4+1 &= x^4+1-x^2(x^2+x+1) = x^3+x^2+1\\ &= x^3+x^2+1-x(x^2+x+1) = x+1 \pmod{p(x)}.\end{aligned}$$

Finite fields may be constructed starting from these sets of polynomials $\mathbb{F}[x]$ by copying the construction of $\mathbb{Z}_p$ (for p prime) using equivalence classes. The operations of addition and multiplication will be modulo a polynomial $p(x)$. Will any polynomial do? No! Much as we require p to be prime for $\mathbb{Z}_p$, the polynomial $p(x)$ must satisfy a particular condition: it must be *irreducible*. A nonzero polynomial $p(x) \in \mathbb{F}[x]$ is called irreducible if for all $q_1(x)$ and $q_2(x) \in \mathbb{F}[x]$ such that

$$p(x) = q_1(x)q_2(x),$$

it follows that either $q_1(x)$ or $q_2(x)$ is a constant polynomial. In other words, $p(x)$ does not have any proper polynomial factors with degree less than that of $p(x)$.

Example 6.11 *The polynomial* x^2+x-1 *can be factored over* $\mathbb{R}$*. In fact,*

$$x_1 = \tfrac{1}{2}(\sqrt{5}-1) \quad \textit{and} \quad x_2 = -\tfrac{1}{2}(\sqrt{5}+1)$$

are the roots of this polynomial. These two numbers are in $\mathbb{R}$*, and*

$$x^2+x-1 = (x-x_1)(x-x_2).$$

Thus $x^2+x-1 \in \mathbb{R}[x]$ *is not irreducible over* $\mathbb{R}$*. This same polynomial is irreducible over* $\mathbb{Q}[x]$*, however, since* $x_i \notin \mathbb{Q}$*,* $i = 1, 2$*, and therefore* x^2+x-1 *cannot be factored over* $\mathbb{Q}$*.*

Example 6.12 *The polynomial* x^2+1 *is irreducible over* $\mathbb{R}$*, but over* $\mathbb{F}_2$ *it can be factored as* $x^2+1 = (x+1)\times(x+1)$*. Thus, it is not irreducible over* $\mathbb{F}_2$*.*

We denote by $\mathbb{F}[x]/(p(x))$ the set of polynomials $\mathbb{F}[x]$ equipped with the operation of addition and multiplication modulo $p(x)$. The following is the central result that we need.

Proposition 6.13 *(i) Let $p(x)$ be a polynomial of degree n. The quotient $\mathbb{F}[x]/p(x)$ can be identified with $\{q(x) \in \mathbb{F}[x] \mid \text{degree } q < n\}$ with addition and multiplication modulo $p(x)$.*
(ii) $\mathbb{F}[x]/(p(x))$ is a field if and only if $p(x)$ is irreducible over $\mathbb{F}$.

We do not prove this result, but we will use it to give an explicit construction of a field that is not isomorphic to $\mathbb{Z}_p$ for p prime.

Example 6.14 Construction of $\mathbb{F}_9$, the field with nine elements. *Let $\mathbb{Z}_3$ be the field with three elements whose tables of addition and multiplication were given earlier. Let $\mathbb{Z}_3[x]$ be the set of polynomials with coefficients in $\mathbb{Z}_3$ and define $p(x) = x^2 + x + 2$.*

We first convince ourselves that $p(x)$ is irreducible. If it is not, then there exist two nonconstant polynomials q_1 and q_2 whose product is p. Since the degree of $p(x)$ is 2, these two polynomials must each have degree 1. Thus

$$p(x) = (x + a)(bx + c) \tag{6.7}$$

for some $a, b, c \in \mathbb{Z}_3$. If this is the case, then $p(x)$ will evaluate to zero at the additive inverse of a. However,

$$\begin{aligned} p(0) &= 0^2 + 0 + 2 = 2, \\ p(1) &= 1^2 + 1 + 2 = 1, \\ p(2) &= 2^2 + 2 + 2 = 1 + 2 + 2 = 2, \end{aligned}$$

and thus $p(x)$ is nonzero for each possible value of $x \in \mathbb{Z}_3$. (Note: the calculations are performed in $\mathbb{Z}_3$!) Thus $p(x)$ cannot be written as in (6.7) *and is therefore irreducible.*

Start by finding the number of elements in the field $\mathbb{Z}_3[x]/(p(x))$. Since all the elements of this field are polynomials with degree less than that of $p(x)$, then they are all of the form $a_1x + a_0$. Since $a_0, a_1 \in \mathbb{Z}_3$, they can each take on three distinct values; thus there are $3^2 = 9$ distinct elements in $\mathbb{Z}_3[x]/(p(x))$.

We now construct the multiplication table. Two examples will show how to do this:

$$\begin{aligned} (x + 1)^2 &= x^2 + 2x + 1 = (x^2 + 2x + 1) - (x^2 + x + 2) = x - 1 = x + 2, \\ x(x + 2) &= x^2 + 2x = x^2 + 2x - (x^2 + x + 2) = x - 2 = x + 1. \end{aligned}$$

The complete multiplication table is

$\times$	0	1	2	x	$x+1$	$x+2$	$2x$	$2x+1$	$2x+2$
0	0	0	0	0	0	0	0	0	0
1	0	1	2	x	$x+1$	$x+2$	$2x$	$2x+1$	$2x+2$
2	0	2	1	$2x$	$2x+2$	$2x+1$	x	$x+2$	$x+1$
x	0	x	$2x$	$2x+1$	1	$x+1$	$x+2$	$2x+2$	2
$x+1$	0	$x+1$	$2x+2$	1	$x+2$	$2x$	2	x	$2x+1$
$x+2$	0	$x+2$	$2x+1$	$x+1$	$2x$	2	$2x+2$	1	x
$2x$	0	$2x$	x	$x+2$	2	$2x+2$	$2x+1$	$x+1$	1
$2x+1$	0	$2x+1$	$x+2$	$2x+2$	x	1	$x+1$	2	$2x$
$2x+2$	0	$2x+2$	$x+1$	2	$2x+1$	x	1	$2x$	$x+2$

(6.8)

But this method is tedious. Is there some way to simplify these calculations? Consider enumerating the powers of $q(x) = x$. Taking these powers modulo $p(x)$, we obtain

$$
\begin{aligned}
q &= x,\\
q^2 &= x^2 = x^2 - (x^2 + x + 2) = -x - 2 = 2x + 1,\\
q^3 &= q \times q^2 = 2x^2 + x = 2x^2 + x - 2(x^2 + x + 2) = 2x + 2,\\
q^4 &= q \times q^3 = 2x^2 + 2x = 2x^2 + 2x - 2(x^2 + x + 2) = 2,\\
q^5 &= q \times q^4 = 2x,\\
q^6 &= q \times q^5 = 2x^2 = 2x^2 - 2(x^2 + x + 2) = x + 2,\\
q^7 &= q \times q^6 = x^2 + 2x = x^2 + 2x - (x^2 + x + 2) = x + 1,\\
q^8 &= q \times q^7 = x^2 + x = x^2 + x - (x^2 + x + 2) = 1.
\end{aligned}
$$

By taking the powers of the polynomial $q(x) = x$ we obtain the eight nonzero polynomials of $\mathbb{Z}_3[x]/(p(x))$. Pairwise multiplication between elements in $\{0, q, q^2, q^3, q^4, q^5, q^6, q^7, q^8 = 1\}$ is simplified using $q^i \times q^j = q^k$, where $k = i + j \pmod 8$, since $q^8 = 1$. This gives us a simple manner of calculating the multiplication table. We transform each polynomial into a power of q, and the multiplication of two elements simplifies to an addition of powers modulo 8. We can easily recalculate the above examples as

$$
\begin{aligned}
(x+1)^2 &= q^7 \times q^7 = q^{14} = q^6 = x + 2,\\
x(x+2) &= q \times q^6 = q^7 = x + 1.
\end{aligned}
$$

We can use this second method to verify our earlier calculations. We rewrite the multiplication table replacing each polynomial by its power of q:

$\times$	0	1	q^4	q^1	q^7	q^6	q^5	q^2	q^3
0	0	0	0	0	0	0	0	0	0
1	0	1	q^4	q	q^7	q^6	q^5	q^2	q^3
q^4	0	q^4	1	q^5	q^3	q^2	q	q^6	q^7
q^1	0	q	q^5	q^2	1	q^7	q^6	q^3	q^4
q^7	0	q^7	q^3	1	q^6	q^5	q^4	q	q^2
q^6	0	q^6	q^2	q^7	q^5	q^4	q^3	1	q
q^5	0	q^5	q	q^6	q^4	q^3	q^2	q^7	1
q^2	0	q^2	q^6	q^3	q	1	q^7	q^4	q^5
q^3	0	q^3	q^7	q^4	q^2	q	1	q^5	q^6

(6.9)

With these new names it is more natural to reorder the rows and columns of the table so that the exponents increase. Here is the same table rewritten in this manner:

$\times$	0	q^1	q^2	q^3	q^4	q^5	q^6	q^7	1
0	0	0	0	0	0	0	0	0	0
q^1	0	q^2	q^3	q^4	q^5	q^6	q^7	1	q
q^2	0	q^3	q^4	q^5	q^6	q^7	1	q	q^2
q^3	0	q^4	q^5	q^6	q^7	1	q	q^2	q^3
q^4	0	q^5	q^6	q^7	1	q	q^2	q^3	q^4
q^5	0	q^6	q^7	1	q	q^2	q^3	q^4	q^5
q^6	0	q^7	1	q	q^2	q^3	q^4	q^5	q^6
q^7	0	1	q	q^2	q^3	q^4	q^5	q^6	q^7
1	0	q	q^2	q^3	q^4	q^5	q^6	q^7	1

(6.10)

The addition table may then be obtained in a similar manner. Here are two sample calculations:

$$q^2 + q^4 = (2x + 1) + (2) = 2x + (2 + 1) = 2x = q^5,$$
$$q^3 + q^6 = (2x + 2) + (x + 2) = (2 + 1)x + (2 + 2) = 1 = q^8.$$

The full addition table of $\mathbb{F}_9$ follows. (Exercise: verify a few elements of this table.)

$+$	0	q^1	q^2	q^3	q^4	q^5	q^6	q^7	1
0	0	q^1	q^2	q^3	q^4	q^5	q^6	q^7	1
q^1	q^1	q^5	1	q^4	q^6	0	q^3	q^2	q^7
q^2	q^2	1	q^6	q^1	q^5	q^7	0	q^4	q^3
q^3	q^3	q^4	q^1	q^7	q^2	q^6	1	0	q^5
q^4	q^4	q^6	q^5	q^2	1	q^3	q^7	q^1	0
q^5	q^5	0	q^7	q^6	q^3	q^1	q^4	1	q^2
q^6	q^6	q^3	0	1	q^7	q^4	q^2	q^5	q^1
q^7	q^7	q^2	q^4	0	q^1	1	q^5	q^3	q^6
1	1	q^7	q^3	q^5	0	q^2	q^1	q^6	q^4

(6.11)

Definition 6.15 *A nonzero element whose powers enumerate all other nonzero elements of a field is called primitive or a primitive root.*

Not all elements are primitive. For example, in $\mathbb{F}_9$ the element q^4 is not primitive; the only distinct elements that it enumerates are q^4 and $q^4 \times q^4 = q^8$. In Exercise 13 you will find all of the primitive roots of $\mathbb{F}_9$. In the above example the polynomial $q(x) = x$ is primitive because it allows us to construct the eight nonzero polynomials in the form q^i for $i = 1, \ldots, 8$. But $q(x) = x$ is not a primitive root for all fields modulo a polynomial. We give two examples, the first in Exercise 17 of this chapter and the second in Exercise 6 of Chapter 8.

(If you know the notion of *group*, you may note that a primitive root is a generator of the multiplicative group of nonzero elements of a field, yielding that these elements form a cyclic group. This observation is not used in the present chapter.)

Theorem 6.16 *All finite fields $\mathbb{F}_{p^r}$ possess a primitive root. In other words, there exists a nonzero element α whose powers enumerate the nonzero elements of $\mathbb{F}_{p^r}$:*

$$\mathbb{F}_{p^r} \setminus \{0\} = \{\alpha, \alpha^2, \ldots, \alpha^{p^r-1} = \alpha^0 = 1\}.$$

It is usual to use the symbol α to represent a primitive root. In this section we have often used the letter q, but we will use α in subsequent sections.

Before finishing our introduction to finite fields we will state without proof two important theorems.

Theorem 6.17 *The number of elements in a finite field is a power of a prime number.*

Theorem 6.18 *If two finite fields possess the same number of elements, then they are* isomorphic*. In other words, there exists a reordering of the elements such that the tables of addition and multiplication of the two fields correspond. Such a reordering naturally associates an element from one field with its counterpart in the other, a mapping that is called an* isomorphism*.*

6.6 Reed–Solomon Codes

The codes devised by Reed and Solomon are more complex than those of Hamming. We will start by describing the encoding and decoding process. Afterward, we will prove the three properties that characterize these codes.

Let $\mathbb{F}_{2^m}$ be the field with 2^m elements and let α be a primitive root. The $2^m - 1$ nonzero elements of $\mathbb{F}_{2^m}$ are of the form

$$\{\alpha, \alpha^2, \ldots, \alpha^{2^m-1} = 1\},$$

and therefore for all nonzero elements $x \in \mathbb{F}_{2^m}$ we have that $x^{2^m-1} = 1$.

The words to be encoded will be those of k letters, each letter being an element of $\mathbb{F}_{2^m}$, and where $k < 2^m - 2$. (How to choose this integer k will be explained soon.) Thus, they will be elements $(u_0, u_1, u_2, \ldots, u_{k-1}) \in \mathbb{F}_{2^m}^k$. Each of these words will be associated with the polynomial

$$p(x) = u_0 + u_1 x + u_2 x^2 + \cdots + u_{k-1} x^{k-1} \in \mathbb{F}_{2^m}[x].$$

These words will be encoded in a vector $v = (v_0, v_1, v_2, \ldots, v_{2^m-2}) \in \mathbb{F}_{2^m}^{2^m-1}$ whose entries will be given by

$$v_i = p(\alpha^i), \qquad i = 0, 1, 2, \ldots, 2^m - 2,$$

where α is the primitive root we chose at the outset. Thus, *encoding* consists in calculating

$$\begin{array}{rcrl} v_0 & = & p(1) & = u_0 + u_1 + u_2 + \cdots + u_{k-1}, \\ v_1 & = & p(\alpha) & = u_0 + u_1\alpha + u_2\alpha^2 + \cdots + u_{k-1}\alpha^{k-1}, \\ v_2 & = & p(\alpha^2) & = u_0 + u_1\alpha^2 + u_2\alpha^4 + \cdots + u_{k-1}\alpha^{2(k-1)}, \\ \vdots & = & \vdots & = \quad \vdots \\ v_{2^m-2} & = & p(\alpha^{2^m-2}) & = u_0 + u_1\alpha^{2^m-2} + u_2\alpha^{2(2^m-2)} + \cdots + u_{k-1}\alpha^{(k-1)(2^m-2)}. \end{array} \tag{6.12}$$

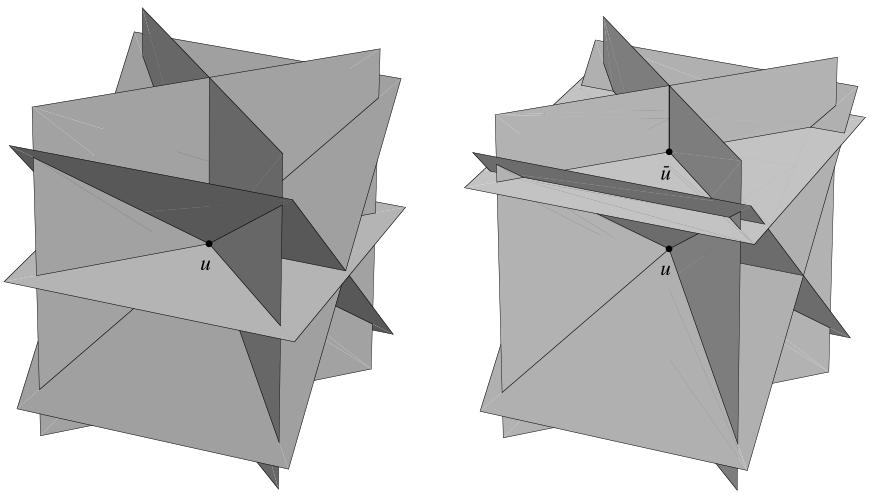
The $C(2^m - 1, k)$ Reed–Solomon code is the set of vectors $v \in \mathbb{F}_{2^m}^{2^m-1}$ obtained in this manner. The basic requirement of any encoding is that different words not get the same encoding. This is the content of the first property.

Property 6.19 *The encoding $u \mapsto v$, where $u \in \mathbb{F}_{2^m}^k$ and $v \in \mathbb{F}_{2^m}^{2^m-1}$, is a linear transformation with a trivial kernel, that is, a kernel equal to $\{0\} \subset \mathbb{F}_{2^m}^k$.*

(The proofs of the Properties 6.19 and 6.20 will be given at the end of this section.)

The transmission might introduce some errors in the encoded message v. The received message $w \in \mathbb{F}_{2^m}^{2^m-1}$ may differ from v at one or more locations. The decoding consists in first replacing, in (6.12), the v_i by the components w_i of w and then extracting from this new linear system the original u, despite the possible errors in w. To understand how this can be achieved, we first describe geometrically the system (6.12). Each of these equations (with v_i replaced by the corresponding w_i) represents a plane in the space $\mathbb{F}_2^k$ with coordinates $(u_0, u_1, \ldots, u_{k-1})$. There are $2^m - 1$ planes, which is more than k, the number of unknowns u_j. Let us use our intuition of $\mathbb{R}^3$ to draw a geometric representation of the situation. Figure 6.4 (a) presents five planes (instead of $2^m - 1$) in $\mathbb{R}^3$ (instead of $\mathbb{F}_2^k$). If there are no mistakes in the transmission (all w_i agree with the original v_i), then all the planes intersect at a single point, the original message u. Moreover, any choice of three planes among the five determines uniquely the solution u. In other words, two of the five planes are redundant, or, in this errorless transmission, there are many distinct ways to reconstruct u. Suppose now that one of the w_i is erroneous. The corresponding equation is then false, and the plane that it represents

will be shifted. This is depicted in Figure 6.4(b), where one plane, the horizontal one, has been moved up. Even though the four correct planes (those with the correct w_i) still intersect at u, a choice of three planes including the wrong one will give a wrong message $\bar{u}$. In $\mathbb{R}^3$, we need three planes to obtain a (correct or false) determination of u. For the system (6.12) we need k planes (= equations) to get one determination of u. We can think of each choice of k planes as "voting" for the value u where they intersect. If some of the w_i are wrong, one may ask whether the correct u will get the largest number of votes. This is the question we now address. (For instance, in our example of Figure 6.4 (b), the correct answer u receives four votes and the wrong $\bar{u}$ gets only one.)



(a) The set of planes without any errors

(b) The set of planes with one error

Fig. 6.4. The planes of system (6.12).

Suppose that once the message has been transmitted, we receive the $2^m - 1$ symbols $w = (w_0, w_1, w_2, \ldots, w_{2^m-2}) \in \mathbb{F}_{2^m}^{2^m-1}$. If all of these symbols are exact, we can recover the original message u by choosing from (6.12) any subset of k rows and resolving the resulting linear system. Suppose that we choose rows $i_0, i_1, \ldots, i_{k-1}$ with $0 \le i_0 < i_1 < \cdots < i_{k-1} \le 2^m - 2$, and that α_j denotes α^{i_j}. Then the resulting linear system is

$$\begin{pmatrix} w_{i_0} \\ w_{i_1} \\ w_{i_2} \\ \vdots \\ w_{i_{k-1}} \end{pmatrix} = \begin{pmatrix} 1 & \alpha_0 & \alpha_0^2 & \alpha_0^3 & \ldots & \alpha_0^{k-1} \\ 1 & \alpha_1 & \alpha_1^2 & \alpha_1^3 & \ldots & \alpha_1^{k-1} \\ 1 & \alpha_2 & \alpha_2^2 & \alpha_2^3 & \ldots & \alpha_2^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_{k-1} & \alpha_{k-1}^2 & \alpha_{k-1}^3 & \ldots & \alpha_{k-1}^{k-1} \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{k-1} \end{pmatrix}, \tag{6.13}$$

and we can obtain the original message by inverting the matrix $\{\alpha_i^j\}_{0\le i,j\le k-1}$, assuming that it is invertible.

Property 6.20 *For all choices of* $0 \le i_0 < i_1 < i_2 < \cdots < i_{k-1} \le 2^m - 2$*, the matrix* $\{\alpha_i^j\}$ *described above is invertible.*

Thus, assuming that the received message does not contain any errors, there are as many ways to recover it as there are ways of choosing k equations from the $2^m - 1$ in (6.12):

$$\binom{2^m - 1}{k} = \frac{(2^m - 1)!}{k!(2^m - 1 - k)!}.$$

Now suppose that s of the $2^m - 1$ coefficients of w are in error. Then only $(2^m - s - 1)$ of the equations of (6.12) will be correct, and only $\binom{2^m - s - 1}{k}$ of the $\binom{2^m - 1}{k}$ possible calculations of u will be correct. The others will be in error, and there will therefore be several candidate vectors u, only one of them correct. Let $\bar{u}$ be one of the incorrect candidates arrived at by choosing false equations from (6.12). How many times can we obtain $\bar{u}$ by changing the equations we use? The solution $\bar{u}$ is obtained as the intersection of the k planes represented by the k chosen equations from (6.12). At most $s + k - 1$ of these planes will intersect at $\bar{u}$, because had there been one more, there would be among them k planes described by valid equations, and $\bar{u} = u$. Thus there are at most $\binom{s+k-1}{k}$ ways to arrive at $\bar{u}$. The correct value u will receive the most "votes" (will be calculated by the most choices of equations) if

$$\binom{2^m - s - 1}{k} > \binom{s + k - 1}{k},$$

or equivalently,

$$2^m - s - 1 > s + k - 1.$$

Thus we deduce that

$$2^m - k > 2s.$$

Because we are interested only in integer values for s, this is equivalent to

$$2^m - k - 1 \ge 2s.$$

In other words, as long as the number of errors is less than or equal to $\frac{1}{2}(2^m - k - 1)$, then the correct value of u will receive the largest number of votes, proving the next property.

Property 6.21 *Reed–Solomon codes can correct* $[\frac{1}{2}(2^m - k - 1)]$ *errors, where* $[x]$ *denotes the integer part of* x*.*

The *decoding* of w consists therefore in choosing from all the determinations of u the one that obtains the most votes.

We finish this section by proving Properties 6.19 and 6.20.

PROOF OF PROPERTY 6.19: Observe that each of the components v_j of $v, j = 0, 1, \ldots, 2^m - 2$, depends linearly on the components u_i. Thus the encoding is a linear transformation from $\mathbb{F}_{2^m}^k$ to $\mathbb{F}_{2^m}^{2^m-1}$.

In order to show that the kernel of this transformation is trivial, it suffices to convince ourselves that only the zero polynomial will be mapped to $0 \in \mathbb{F}_{2^m}^{2^m-1}$. If p is a nonzero polynomial with degree at most $k-1$, then it cannot evaluate to zero at more than $k-1$ values of x. The v_i are evaluations of the polynomial p at the powers $\alpha^i, i = 0, 1, 2, \ldots, 2^m - 2$. Since α is a primitive root, only $k-1$ of the $2^m - 1$ values $v_i = p(\alpha^i)$ can be zero. Thus, every nonzero polynomial p will be mapped to a nonzero vector v. □

Property 6.20 is a consequence of the following lemma which we demonstrate first.

Lemma 6.22 (Vandermonde determinant) *Let $x_1, x_2, \ldots, x_n$ be elements of a field $\mathbb{F}$. Then*

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ 1 & x_3 & x_3^2 & \ldots & x_3^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{vmatrix} = \prod_{1 \le i < j \le n} (x_j - x_i).$$

PROOF: By subtracting row j from row i the value of the determinant is not changed, and row i becomes

$$\begin{pmatrix} 0 & x_i - x_j & x_i^2 - x_j^2 & x_i^3 - x_j^3 & \ldots & x_i^{n-1} - x_j^{n-1} \end{pmatrix}.$$

Since

$$x_i^k - x_j^k = (x_i - x_j) \sum_{l=0}^{k-1} x_i^l x_j^{k-l-1},$$

all the elements of this new row i possess $(x_i - x_j)$ as a factor. The determinant, viewed as a polynomial in the variables $x_1, x_2, \ldots, x_n$, therefore has $(x_i - x_j)$ as a factor for all i and j. The determinant is thus the product of

$$\prod_{1 \le i < j \le n} (x_j - x_i)$$

and of one other polynomial, which remains to be found. Note that in the product $\prod_{1 \le i < j \le n} (x_j - x_i)$, the maximal power of x_n is $n-1$, since there are $(n-1)$ terms with $j = n$. Similarly, in the determinant the maximal power of x_n is also $n-1$, since the

terms with x_n are all in the same row and it is x_n^{n-1} that has the highest power in this row. Hence the polynomial multiplying $\prod_{1\le i<j\le n}(x_j - x_i)$ cannot contain x_n. We can repeat this argument for each of the x_i and conclude that the polynomial multiplying $\prod_{1\le i<j\le n}(x_j - x_i)$ is constant. The term $x_1^0 x_2^1 x_3^2 \cdots x_n^{n-1}$ in the determinant comes from the product of the diagonal terms and therefore has coefficient $+1$. In the product $\prod_{1\le i<j\le n}(x_j - x_i)$, this same term $x_1^0 x_2^1 x_3^2 \cdots x_n^{n-1}$ is obtained by multiplying the *first* term of all of the monomials $(x_j - x_i)$ and also has coefficient $+1$. (Why the *first* terms? There are precisely $n-1$ monomials in the product $\prod_{1\le i<j\le n}(x_j - x_i)$ that contain the term x_n, and in each of these monomials the variable x_n is the first term of $(x_j - x_i)$, since $i < j$. Thus we must choose the first $n-1$ terms of these monomials. Among the remaining monomials there are precisely $n-2$ that contain the term x_{n-1}. Again, in each of these monomials the variable x_{n-1} is the first term. By repeating this argument we arrive at the desired result.) Thus the determinant and the polynomial are equal. □

Proof of Property 6.20: Applying the above lemma to the matrix in (6.13) shows that its determinant is equal to $\prod_{i<j}(\alpha_j - \alpha_i)$. Recall that the α_i are distinct powers of the primitive root α for powers less than $2^m - 1$. Thus each of these α_i is distinct, the determinant is nonzero and the matrix invertible. □

Here is a concrete example of the various parameters k, m, and s of the code. We saw at the beginning of the chapter that it is usual to use 7 or 8 bits to encode common Western typographical symbols (letters, numbers, punctuation, etc). If m is set to 8, then each of the elements ($\in \mathbb{F}_{2^m}$) can directly represent a symbol of the ASCII character set. Thus the correspondence between "ASCII character" and "elements of $\mathbb{F}_{2^m}$" is one-to-one. If $m = 8$ is chosen, then the number k of letters is bounded by $2^m - 2 = 254$. Now suppose that the transmission channel is reliable enough that being able to correct 2 letters is sufficient with high probability. Since the number of correctable errors s is equal to $[\frac{1}{2}(2^m - k - 1)]$, we require that $(2^m - k - 1)$ be greater than or equal to $2s = 4$. Thus we can send text in blocks of $k = 2^8 - 4 - 1 = 251$ letters. The code transforms them into blocks of 255 letters. Note that there can be more than one bit error per letter being corrected. The Reed–Solomon code corrects entire letters, not individual bits.

Compact discs do not store Latin characters but rather digitized sound. However, they use Reed–Solomon codes with the parameters just mentioned: $m = 8$ and a maximum of 2 errors. It should be noted that much more economical decoding methods exist that do not require exploring all of the $\binom{2^m-1}{k}$ possible linear systems of k equations and k unknowns [6, 1]. These algorithms considerably accelerate the decoding process.

6.7 Appendix: The Scalar Product and Finite Fields

It is very likely that you have encountered *scalar products* in your linear algebra courses, where it was defined as function denoted by $(\cdot, \cdot)$ from a vector space V on $\mathbb{R}$ such that

(i) $(x, y) = (y, x)$, for all $x, y \in V$;
(ii) $(x + y, z) = (x, z) + (y, z)$ for all $x, y, z \in V$;
(iii) $(cx, y) = c(x, y)$ for all $x, y \in V$ and $c \in \mathbb{R}$;
(iv) $(x, x) \geq 0$ with $(x, x) = 0$ only for $x = 0$.

If the field $\mathbb{R}$ of real numbers is replaced by a finite field, the same definition applies except for the final property, which becomes

(iv)$_{\text{finite}}$ if $(x, y) = 0$ for all $y \in V$, then $x = 0$.

It is the scalar product with this modification that is used in the present chapter. Note that the original condition *(iv)* does not make sense in a finite field, since we do not have a complete order ("$<$") that is preserved by addition. For example, in $\mathbb{F}_2$ we could propose that $0 < 1$. However, this relation does not satisfy that of the ordering on the reals, which states that if $a < b$ then $a + c < b + c$ for all numbers c. In fact, if the number $1 \in \mathbb{F}_2$ is added to both sides, we then obtain $0 + 1 < 1 + 1$, or $1 < 0$, which clearly contradicts the original statement!

The definition of the orthogonal complement remains the same for both the original and the modified scalar product. We recall it here.

Definition 6.23 *If $W \subset V$ is a subset of V, then the* orthogonal complement $W^\perp$ *is defined by $W^\perp = \{v \in V | (v, w) = 0$ for all $w \in W\}$.*

This is a vector subspace of V. The modification *(iv)* $\rightarrow$ *(iv)*$_{\text{finite}}$ has a nonintuitive consequence. Recall that if $W \subset \mathbb{R}^n$ is a vector subspace, then it and its complement have only the origin in common: $W \cap W^\perp = \{0\}$. In vector spaces over finite fields this is not always the case! For example, consider the subspace W with basis

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \in \mathbb{F}_2^3.$$

The elements $w = (w_1, w_2, w_3)^t \in \mathbb{F}_2^3$ of the orthogonal complement must satisfy

$$\begin{pmatrix} w_1 & w_2 & w_3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = 0$$

and therefore $w_1 + w_2 = 0$. Hence

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

is a basis of $W^\perp$ and

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \in W \cap W^\perp.$$

We must therefore use our intuition about orthogonal complements with caution!

6.8 Exercises

1. **(a)** In the $C(7,4)$ Hamming code, what are the vectors to be sent if we wish to transmit the words $(0,0,0,0)$, $(0,0,1,0)$, and $(0,1,1,1)$?
(b) The receiver receives the words $(1,1,1,1,1,1,1)$, $(1,0,1,1,1,1,1)$, $(0,0,0,0,1,1,1)$, and $(1,1,1,1,0,0,0)$. What were the originally transmitted words?

2. **(a)** We use the $C(15,11)$ Hamming code to correct a message containing at most one bit error. If the control matrix is

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and the received message is

$$w = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix},$$

was there an error in the transmission?
(b) We want to use the $C(2^k - 1, 2^k - k - 1)$ Hamming code for a given k but we do not want to add more than 10% overhead to the original word. What is the minimum length of the original word that must be used and what value of k characterizes the code to be used?

3. The following questions concern the $C(2^k - 1, 2^k - k - 1)$ Hamming code.
(a) Using this code, what is the length of the original words u to be transmitted? How many distinct words may be transmitted?
(b) How many "letters" are there in an encoded word v?
(c) How many distinct received words w (with an error or without) will decode to the same original message u?
(d) Do there exist any received messages that cannot be decoded? (Another way of posing this question is, does there exist a $w \in \mathbb{F}_2^{2^k-1}$ that is not an encoding v of some message $u \in \mathbb{F}_2^{2^k-k-1}$ or within one error of such a v?)

4. Verify that addition $+$ and multiplication $\times$ in $\mathbb{F}_2$ (as defined by the tables in Section 6.2) satisfy the properties of the structure of fields as defined in Section 6.5.

5. Let $(\mathbb{F}, +, \times)$ be a finite field. Show that the multiplication table of the nonzero elements of $\mathbb{F}$ has the following property: all rows and all columns contain all nonzero elements of $\mathbb{F}$ exactly once.

6. **(a)** In the $C(7,4)$ Hamming code, does there exist a received message $(w_1, w_2, w_3, w_4, w_5, w_6, w_7) \in \mathbb{F}_2^7$ that cannot be decoded to one of the 16 elements of $\mathbb{F}_2^4$ under the hypothesis that at most one bit is in error? (See also Exercise 3(d).)
(b) Show that a Hamming type code that lengthens a message from three bits to eight bits cannot correct for two errors.
(c) Construct a Hamming-type code mapping three bits into ten that is able to correct two errors.

7. **(a)** Let H be a $k \times n$ matrix, $n > k$, with entries in $\mathbb{F}_2$. Let G be an $(n-k) \times n$ matrix with entries in $\mathbb{F}_2$, obtained from H by requiring that G be of maximum rank and that its rows be orthogonal to those of H. If H has the form

$$H = \Big(\underbrace{M}_{k\times(n-k)} \mid I_{k\times k} \Big),$$

where M is a $k \times (n-k)$ matrix and $I_{k\times k}$ is the $k \times k$ identity matrix, show that G can be chosen as

$$G = \big(I_{(n-k)\times(n-k)} \mid \underbrace{M^t}_{(n-k)\times k} \big).$$

(b) Write G_4 and H_4 for the $C(15,11)$ Hamming code with $k = 4$. (Start with H_4.)
(c) What is the message u that the sender wished to send if he was using the $C(15,11)$ code and the received message was $(1,1,1,1,1,0,0,0,0,0,1,1,1,1,1)$?

8. Let $p = \frac{1}{1000}$ be the probability that a bit will be transmitted in error.
(a) What is the probability of receiving precisely two bits in error in a transmission of seven bits, when transmitting a word in the $C(7,4)$ Hamming code?
(b) What is the probability of having more than one bit error in a transmission of seven bits?
(c) Rather than using the Hamming code, consider transmitting a bit by repeating it three times. We decode by choosing the bit that is in the majority. Calculate the probability of correctly decoding the sent bit.
(d) We transmit four bits by repeating each one three times. What is the probability that the four bits will be decoded correctly. Comparing the results of this question with part (b), we see that the simple repeating code has a slight advantage over the $C(7,4)$ Hamming code, but at the cost of transmitting 12 bits instead of seven.

9. Most books have an ISBN code (for International Standard Book Number), and this code is unique to each book. This code consists of 10 numbers. For example, ISBN 2-12345-678-0. The first three segments identify the linguistic group, the publishing house, and the volume. The last is an error-detection symbol chosen from $\{0,1,2,3,4,5,6,7,8,9,X\}$, where X represents 10 in Roman numerals. Let a_i, for $i = 1,\ldots,10$, refer to the 10 symbols. The symbol a_{10} is chosen as the remainder of the sum $b = \sum_{i=1}^{9} i a_i$ when divided by 11. In our example we see that

$b = 1\times 2+2\times 1+3\times 2+4\times 3+5\times 4+6\times 5+7\times 6+8\times 7+9\times 8 = 242 = 11\times 22+0$ and $a_{10} = 0$.

(a) Show that this code can detect an error in one digit.

(b) Show that the sum $\sum_{i=1}^{10} ia_i$ is divisible by 11.

(c) Find the last digit of the following ISBN code:

ISBN 0-7267-3514-?.

(d) A common type of error is the inversion of two symbols. The code 0-1311-0362-8 could be erroneously entered as 0-1311-0326-8, for example. Show that the code permits the detection of such an error provided that the consecutive digits are not identical (in which case the inversion does not constitute an error anyway!).

(e) In other references a_{10} is defined as being chosen such that the sum

$$\sum_{i=1}^{10}(11-i)a_i$$

is divisible by 11. Show that this definition is equivalent to that given above.

10. The following method was introduced by IBM for constructing credit card numbers. It is also used in Canada for social insurance numbers. We construct numbers of n digits, $a_1, \ldots, a_n$ with $a_i \in \{0,1,2,3,4,5,6,7,8,9\}$. The number is valid if the number b constructed as follows is a multiple of 10:

- if i is odd we define $c_i = a_i$;
- if i is even and $2a_i < 10$ we define $c_i = 2a_i$;
- if i is even and $2a_i \geq 10$ then $2a_i = 10 + d_i$; we define $c_i = 1 + d_i$, which is the sum of the digits of $2a_i$;
- then

$$b = \sum_{i=1}^{n} c_i.$$

(a) Show that if i is even then c_i is the remainder of the division of $2a_i$ by 9.

(b) The first 15 digits of a credit card are 1234 5678 1234 567. Calculate the 16th digit.

(c) Show that this method can detect an error in one of the digits.

(d) A common error is the inversion of two consecutive digits. The IBM method is not infallible for detecting such errors. Show that the IBM method is capable of detecting such errors if the two consecutive digits are not the same (in which case it is not actually an error) and if they are not both from the set $\{0,9\}$.

11. The following code is constructed using the same principle as the Hamming code. We want to send a message of four bits (x_1, x_2, x_3, x_4), where $x_i \in \{0,1\}$. We lengthen the

message to 11 bits by adding $x_5, \ldots, x_{11}$ defined as follows (where arithmetic is in the field $\mathbb{F}_2$):

$$\begin{aligned} x_5 &= x_1 + x_4, \\ x_6 &= x_1 + x_3, \\ x_7 &= x_1 + x_2, \\ x_8 &= x_1 + x_2 + x_3, \\ x_9 &= x_2 + x_4, \\ x_{10} &= x_2 + x_3 + x_4, \\ x_{11} &= x_3 + x_4. \end{aligned}$$

Show that this code can detect two errors.

12. Construct the finite field $\mathbb{F}_4$ of four elements. (Give the explicit addition and multiplication tables.)

13. Give all the primitive elements of $\mathbb{F}_9$ from Example 6.14, which was constructed with the primitive polynomial $p(x) = x^2 + x + 2$.

14. **(a)** Let $q(x)$ and $p(x)$ be two polynomials in $\mathbb{F}[x]$. Show that there exist polynomials $s(x)$ and $r(x) \in \mathbb{F}[x]$ such that $q(x) = s(x)p(x) + r(x)$ with $0 \leq$ degree $r <$ degree p.
(b) Conclude that $q(x) = r(x) \pmod{p(x)}$.

15. Let $\mathcal{M}_n$ be the set of $n \times n$ matrices and denote by $+$ and $\cdot$ the usual operations of matrix addition and multiplication. Is $(\mathcal{M}_n, +, \cdot)$ a field? Justify your answer.

16. Let $\mathcal{E}$ be a finite set and $U(\mathcal{E})$ the set of its subsets. (For example, if $\mathcal{E} = \{a, b, c\}$, then $U(\mathcal{E}) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$.) We define on $U(\mathcal{E})$ the operations $+$ and $\times$ as the usual operations of set union and intersection, respectively. For $+$ the identity element is $\emptyset$, and for $\times$ it is $\mathcal{E}$. Does the set $U(\mathcal{E})$ equipped with these operations form a field? Prove this statement or show which of the properties is not satisfied.

17. **(a)** Let $\mathbb{F}_3$ be the field of three elements. There are nine degree-2 polynomials of the form $x^2 + ax + b$, where $a, b \in \mathbb{F}_3$. List these nine polynomials and identify the three that are irreducible. (Hint: start by enumerating the polynomials of the form $(x + c)(x + d)$.)
(b) With the goal of constructing the field of nine elements we consider the quotient $\mathbb{F}_3[x]/q(x)$, where $q(x) = x^2 + 2x + 2$. Verify that x is primitive by reducing the powers $x^i, i = 1, 2, \ldots, 8$, to polynomials of degree zero or one.
(c) Using the results from (b), for which i does the equality $x^3 + x^5 = x^i$ hold?
(d) The field $\mathbb{F}_9$ has now been constructed in two different manners, the first in Example 6.14 of Section 6.5 and the second in part (b) of this question. Construct the isomorphism between these two constructions. (See Proposition 6.18 for the definition of an isomorphism.)

(e) In (a) you identified three irreducible polynomials. Let $p(x)$ be the one used in Example 6.14, $q(x)$ the one used in part (b), and $r(x)$ the third. Is the polynomial $i(x) = x$ a primitive root of $\mathbb{F}_3[x]/r(x)$? What could be done to determine the addition and multiplication tables of $\mathbb{F}_3[x]/r(x)$?

18. **(a)** Find the only degree-2 irreducible polynomial over $\mathbb{F}_2$, the two of degree 3, and the three of degree 4.
(b) Construct the addition and multiplication tables of the field $\mathbb{F}_8$ of eight elements.

19. **(a)** For the ambitious: construct $\mathbb{F}_{16}$.
(b) Also for the ambitious: find an irreducible polynomial of degree 8 over $\mathbb{F}_2$. This polynomial allows you to construct a field of how many elements?

20. **(a)** We consider the error-correcting code that consists of repeating each bit three times. In order to send a 7-bit message we must send 21 bits of data. For example, to send 0100111 we transmit

$$000\ 111\ 000\ 000\ 111\ 111\ 111.$$

The code can correct any single bit error. However, it can correct others as well if the errors are sufficiently well placed. What is the maximum number of errors that may be corrected? Under what conditions?
(b) Now consider the $C(7,3)$ Reed–Solomon code. The letters of this code are elements of the field of eight elements, $\mathbb{F}_{2^3}$, identified with $\{0,1\}^3$, on which we have defined addition and multiplication. We write each letter as a sequence of three bits $\underbrace{b_0 b_1 b_2}$. What is the maximum number of bits this code allows us to correct? Under what conditions?

21. Consider the following system of three equations in three unknowns

$$\begin{aligned} 2x - \tfrac{1}{2}y &= 1, \\ -x + 2y - z &= 0, \\ -y + 2z &= 1. \end{aligned} \qquad (\star)$$

(a) Solve this system over the field $\mathbb{F}_3$ of three elements. (The number $\frac{1}{2}$ is the multiplicative inverse of 2.)
(b) Consider the system $(\star)$ over the field $\mathbb{F}_p$ of p elements, where p is a prime number greater than 2. For what values of p does the system possess a unique solution?

22. **(a)** Calculate the following determinant over the reals $\mathbb{R}$:

$$d = \begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{vmatrix}.$$

(b) Explain why the determinant d_2 of the matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \in \mathbb{F}_2^{3\times 3}$$

is equal to $d \pmod 2$.

(c) Calculate in $\mathbb{F}_3$ the determinant d_3 of the matrix

$$\begin{pmatrix} 2 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 2 \end{pmatrix} \in \mathbb{F}_3^{3\times 3}.$$

Could you have arrived at this determinant starting from the answer in (a)?

(d) Consider the system

$$\begin{aligned} 2a \;\; -b \;\;\;\; &= 1, \\ -a + 2b \;\; -c &= 1, \\ -b + 2c &= 1. \end{aligned} \qquad (\star)$$

In which of the fields $\mathbb{R}$, $\mathbb{F}_2$, and $\mathbb{F}_3$ does this system have a unique solution? (The integer coefficients of the system are understood to be taken modulo 2 or 3 if the solution is to be found in $\mathbb{F}_2$ or $\mathbb{F}_3$, respectively.)

(e) Solve $(\star)$ in $\mathbb{F}_3$.

23. This exercise walks through the encoding and decoding of a message using the Reed–Solomon code with $m = 3$ and $k = 3$. You must first have constructed the field $\mathbb{F}_8$ in Exercise 18. The calculations are simple but numerous; thus it is suggested that you work in teams. (All participants must choose the same primitive root α and use the same tables for $\mathbb{F}_8$!)

(a) What is the maximum number of errors that may be corrected by the $C(7,3)$ Reed–Solomon code?

(b) What is the encoding of the word $(0, 1, \alpha) \in \mathbb{F}_2^3$?

(c) Equation (6.12) can be rewritten as

$$p = Cu,$$

where $p \in F_{2^m}^{2^m-1}$, $u \in \mathbb{F}_{2^m}^k$, and $C \in \mathbb{F}_{2^m}^{(2^m-1)\times k}$. Derive the matrix C for the $C(7,3)$ code.

(d) Suppose that the received message is

$$w = (1, \alpha^4, \alpha^2, \alpha^4, \alpha^2, \alpha^4, \alpha^2) \in \mathbb{F}_{2^m}^{2^m-1}.$$

Choose rows $0, 1$, and 4 of the system in equation (6.12) and solve for the vector $u = (u_0, u_1, u_2) \in \mathbb{F}_8^3$.

(e) How many ways are there to choose three distinct equations from those in equation (6.12)? How many more systems would have to resolve to the same answer as (d) before we could be sure that we had recovered the original message?
(f) Is the answer to (d) the original message?

24. Let p be a prime number. This exercise verifies that $\mathbb{Z}_p$ is a field. We say that a and b are congruent modulo p if their difference $a-b$ is an integer multiple of p. (See Example 6.5.)
(a) Show that "being congruent" is an equivalence relation, called *congruence modulo p*.
(b) We identify $\mathbb{Z}_p$ as the set of equivalence classes of integers modulo p. Let $\bar{a}, \bar{b} \in \mathbb{Z}_p$, $i, j \in \bar{a}$, and $m, n \in \bar{b}$. Show that if $i + m \in \bar{c}$ and $j + n \in \bar{d}$, then $\bar{c} = \bar{d}$. Answer the same question for multiplication. This exercise shows that the definitions of $+$ and $\times$ given in Example 6.5 do not depend on the chosen elements of $\bar{a}$ and $\bar{b}$.
(c) Show that the class $\bar{0}$ is the identity element for $+$ and that $\bar{1}$ is the identity element for $\times$.
(d) Let $\bar{a} \in \mathbb{Z}_p$ be an element different from $\bar{0}$. Use Euclid's algorithm (Corollary 7.4) to show that there exists $\bar{b} \in \mathbb{Z}_p$ such that $\bar{a}\bar{b} = \bar{1}$.
(e) Finish verifying that $\mathbb{Z}_p$ is a field.

References

[1] E.R. Berlekamp, editor. *Key Papers in the Development of Coding Theory*. IEEE Press, 1974.

[2] S. Lang. *Undergraduate Algebra*. Springer, New York, 2nd edition, 1990.

[3] R.J. McEliece. The reliability of computer memories. *Scientific American*, pages 88–95, January 1985.

[4] J. Monforte. The digital reproduction of sound. *Scientific American*, pages 78–84, December 1984.

[5] W.W. Peterson. Error-correcting codes. *Scientific American*, pages 96–108, February 1962.

[6] V. Pless. *Introduction to the Theory of Error-Correcting Codes*. Wiley, New York, 3rd edition, 1999.

[7] K.C. Pohlmann. *The Compact Disc Handbook*. A-R Editions, Madison, WI, 2nd edition, 1992.

[8] I.S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 8:300–304, 1960. (This article is also contained in Berlekamp's survey [1].)

7

Public Key Cryptography: RSA (1978)

This chapter contains more material than can be covered in a single week. The review of the number theory surrounding Euclid's algorithm is optional (Section 7.2), depending on the background of the student. As well, a portion of this material can easily be made the subject of a few additional exercises. On the other hand, it is strongly recommended to take the time to discuss arithmetic modulo n. In Section 7.3 we present the RSA algorithm and prove Euler's theorem, allowing us to rigorously justify the workings of RSA. We explain how to sign a message. This first part can be covered in roughly two hours, unless a significant amount of number theory background needs be covered. Finally, the last hour should be dedicated to more advanced material. For example, you could explain the principle of probabilistic primality-testing algorithms (beginning of Section 7.4). While an hour is insufficient to cover all the details of the algorithm, it should allow for a walk-through of several examples.

The rest of this chapter's material is decidedly more advanced. If covered in class, it is preferable that the students have some knowledge of basic group theory. Such notions are required in Section 7.4 on primality testing algorithms and in Section 7.5 on Shor's large integer factorization algorithm. Alternatively, these sections may serve as a starting point for a semester project.

7.1 Introduction

Cryptography is a subject as old as human civilization. Through all of history man has invented secret codes in an attempt to transmit messages that cannot be understood by an interceptor. History has shown that constructing such codes is a very difficult problem and that they all eventually succumb to clever analysis. Take, for example, a code that permutes the letters of the alphabet, replacing each letter with the letter three places further: in other words, a is replaced by d, b by e, c by f, etc. In English, e is the most frequently occurring letter: looking at a scrambled text would quickly lead us to guess that e had been coded by h, and letter by letter, we would finish by breaking the code. The second reason why secret codes are vulnerable is that the sender and

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_7,

the receiver have to share with each other the inner workings of the code they wish to use. As with any exchange of information, it is possible that this communication may be intercepted.

In this chapter we study the RSA algorithm, named after its inventors Rivest, Shamir, and Adleman. It describes a type of *public key cryptosystem*. What is particularly impressive about this algorithm is the fact that it has still not been broken, even after 29 years of scrutiny from the best scientists in the field. This fact is all the more surprising given that the details of exactly how the cryptosystem works are completely public. We will study the operation of this cryptosystem and show that we need only to be able to factor large integers in order to break it. Surprisingly, it is the conceptually simple operation of factorization (splitting a composite integer into a product of its prime factors, as you likely learned in grade school) that has managed to stump the biggest supercomputers and smartest minds (provided the integer is large enough)!

RSA is built on basic number theory, more specifically on $(+, \cdot)$ arithmetic modulo n, and on Fermat's little theorem as generalized by Euler. The whole system works due to three simple properties of integers, well known by theoreticians:

- It is difficult for a computer to factor a large number.
- It is easy for a computer to construct large prime numbers.
- It is easy for a computer to decide whether a given large number is prime.

Advantages of public key cryptosystems: The benefits of public key cryptosystems are quite clear. In order for two people to communicate using a cryptosystem they must both know the details of the system: it is in the sharing of the details of the system where the danger of interception is largest. However, in the case of public key cryptography this danger no longer exists: the entire system is public! Such a system is also effectively the only approach that can work when there are millions of end users, when sending credit card information across the Internet, for example.

We will also learn of another advantage of the RSA cryptosystem: it allows one to "sign" a message such that the receiver can be certain of both its integrity and its sender. Given the relatively common occurrence of identity theft and our increasing dependence on the Internet, such techniques play an important role in protecting one's online identity.

7.2 A Few Tools from Number Theory

Definition 7.1 *(i) Let a and b be two integers. We say that a divides b if there exists an integer q such that $b = aq$. We write $a \mid b$. (This definition is equally as valid for $a, b, q \in \mathbb{N}$ as for $a, b, q \in \mathbb{Z}$.)*

(ii) The greatest common divisor (GCD) of a and b, denoted by (a, b), satisfies the following two properties:

- *$(a, b) \mid a$ and $(a, b) \mid b$,*

- *if* $d \mid a$ *and* $d \mid b$ *then* $d \mid (a, b)$.

(iii) We say that a *is congruent to* b *modulo* n *if* $n \mid (a-b)$*, in other words if there exists* $x \in \mathbb{Z}$ *such that* $(a - b) = nx$*. We write* $a \equiv b \pmod{n}$*, and call this equivalence relation* congruence *modulo* n.

Proposition 7.2 *Consider* $a, b, c, d, x, y \in \mathbb{Z}$*. Then it follows that*

$$\begin{aligned} a \equiv c \pmod{n} \quad \textit{and} \quad b \equiv d \pmod{n} &\implies a + b \equiv c + d \pmod{n}, \\ a \equiv c \pmod{n} \quad \textit{and} \quad b \equiv d \pmod{n} &\implies ab \equiv cd \pmod{n}, \\ a \equiv c \pmod{n} \quad \textit{and} \quad b \equiv d \pmod{n} &\implies ax + by \equiv cx + dy \pmod{n}. \end{aligned}$$

Proof. We show only the second implication, leaving the others as exercises.

Since $a \equiv c \pmod{n}$, it follows that $n \mid a - c$. Thus there exists an integer x such that $a - c = nx$. Similarly, there exists y such that $b - d = ny$. In order to show that $ab \equiv cd \pmod{n}$ we must show that $n \mid ab - cd$:

$$\begin{aligned} ab - cd &= (ab - ad) + (ad - cd) \\ &= a(b - d) + d(a - c) \\ &= nay + nxd \\ &= n(ay + xd). \end{aligned}$$

Thus it follows that $n \mid ab - cd$. □

Euclid's algorithm allows us to find the GCD, (a, b), of two integers a and b. The details of this algorithm are discussed in the following proposition. Of key importance in the algorithm is the notion of integer division with remainder.

Proposition 7.3 *(Euclid's algorithm) Let* a *and* b *be two positive integers with* $a \geq b$*, and let* $\{r_i\}$ *be the sequence of integers constructed in the following manner. Divide* a *by* b*: we call* q_1 *the quotient of this division and* r_1 *the remainder such that*

$$a = bq_1 + r_1, \qquad 0 \leq r_1 < b.$$

In the same manner we now divide b *by* r_1*, yielding*

$$b = r_1 q_2 + r_2, \qquad 0 \leq r_2 < r_1.$$

We iterate such that

$$r_{i-1} = r_i q_{i+1} + r_{i+1}, \qquad 0 \leq r_{i+1} < r_i.$$

The sequence $\{r_i\}$ *is strictly decreasing. Thus there must exist an integer* n *such that* $r_{n+1} = 0$*. It follows that* $r_n = (a, b)$.

PROOF. We start by showing that $r_n \mid a$ and $r_n \mid b$. Since $r_{n+1} = 0$, the last equation may be written $r_{n-1} = q_{n+1}r_n$. Thus $r_n \mid r_{n-1}$. The second-to-last equation is $r_{n-2} = q_n r_{n-1} + r_n$. Since $r_n \mid r_{n-1}$, it follows that $r_n \mid q_n r_{n-1} + r_n$. Thus $r_n \mid r_{n-2}$. We iterate through the equations one by one, obtaining that $r_n \mid r_i$ for all i. Thus $r_n \mid r_1 q_2 + r_2 = b$. Finally, since $r_n \mid b$ and $r_n \mid r_1$, then $r_n \mid bq_1 + r_1 = a$. Thus $r_n \mid a$ and $r_n \mid b$, which immediately implies $r_n \mid (a, b)$.

Let d be a divisor of a and b. We must show that d divides r_n. Now the iteration through our system of equations is downward. Since $d \mid a$ and $d \mid b$, then $d \mid r_1 = a - bq_1$. In the second equation we have that $d \mid b$ and $d \mid r_1$; thus $d \mid r_2 = b - r_1 q_2$. Iterating shows that $d \mid r_i$ for all i. In particular, $d \mid r_n$.

We can thus conclude that $r_n = (a, b)$. □

Corollary 7.4 *Let a and b be integers and let $c = (a, b)$. Then there exist $x, y \in \mathbb{Z}$ such that $c = ax + by$.*

PROOF. Our proof makes use of Proposition 7.3. We know that $c = r_n$. We again iterate upward through the equations. Since $r_{n-2} = q_n r_{n-1} + r_n$, then

$$r_n = r_{n-2} - q_n r_{n-1}. \tag{7.1}$$

Substituting $r_{n-1} = r_{n-3} - q_{n-1} r_{n-2}$, equation (7.1) then becomes

$$r_n = r_{n-2}(1 + q_{n-1}q_n) - q_n r_{n-3}. \tag{7.2}$$

Now substitute $r_{n-2} = r_{n-4} - q_{n-2} r_{n-3}$. Continuing these iterated substitutions yields the equation $r_n = r_1 x_1 + r_2 y_1$, where $x_1, y_1 \in \mathbb{Z}$. We substitute $r_2 = b - r_1 q_2$, yielding

$$r_n = r_1(x_1 - q_2 y_1) + by_1.$$

Finally, we substitute $r_1 = a - bq_1$, yielding our final result

$$r_n = a(x_1 - q_2 y_1) + b(-q_1 x_1 + q_1 q_2 y_1 + y_1) = ax + by,$$

where $x = x_1 - q_2 y_1$ and $y = -q_1 x_1 + q_1 q_2 y_1 + y_1$. □

Remark: The proof of Corollary 7.4 is very important. It gives us an explicit method for finding integers x and y such that $(a, b) = ax + by$. Even if this method may seem tedious when applied by hand, it is easily performed by a computer, even for large a and b. Similarly, it is also easy for a computer to calculate the GCD of two numbers using the algorithm of Proposition 7.3.

Proposition 7.5 *(1) Let $c = (a, b)$. Then c is characterized by the following property:*

$$c = \min\{ax + by \mid x, y \in \mathbb{Z}, ax + by > 0\}.$$

(2) Consider $a, b, m \in \mathbb{N}$. Then

$$(ma, mb) = m(a, b).$$

(3) Consider $a, b, c \in \mathbb{N}$. If $c \mid ab$ and $(c, b) = 1$, then $c \mid a$.
(4) If p is prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.

PROOF.

(1) Define $E = \{ax + by \mid x, y \in \mathbb{Z}, ax + by > 0\}$ and let $c = (a, b)$. Then it follows that $c \in E$ by Corollary 7.4. Now suppose that $d = ax' + by' \in E$ with $d > 0$ and $d < c$. Since $c \mid a$ and $c \mid b$, then $c \mid ax' + by'$. Thus $c \mid d$. However, $0 < d < c$, a contradiction.
(2) By (1) we have that

$$\begin{aligned}(ma, mb) &= \min\{max + mby \mid x, y \in \mathbb{Z}, max + mby > 0\} \\ &= m \min\{ax + by \mid x, y \in \mathbb{Z}, ax + by > 0\} \\ &= m(a, b).\end{aligned}$$

(3) Since $(c, b) = 1$, by Corollary 7.4 there exist $x, y \in \mathbb{Z}$ such that $cx + by = 1$. Multiplying both sides by a yields $acx + aby = a$. We have $c \mid acx$ and $c \mid aby$. Hence $c \mid (acx + aby)$, and finally $c \mid a$.
(4) Apply (3) with $c = p$. If $(p, b) = 1$ we obtain $p \mid a$ by (3). Otherwise, we must have that $(p, b) = d > 1$. Since the only divisors of a prime p are 1 and p, we must have that $d = p = (p, b)$. In other words, $p \mid b$. □

The following corollary is quite useful:

Corollary 7.6 *Let a and n be two integers with $a < n$. If $(a, n) = 1$, then there exists a unique $x \in \{1, \ldots, n-1\}$ such that $ax \equiv 1 \pmod{n}$.*

PROOF. We start with the existence. Since $(a, n) = 1$, Corollary 7.4 ensures the existence of $x, y \in \mathbb{Z}$ such that $ax + ny = (a, n) = 1$. Thus $ax = 1 - ny$ or $ax \equiv 1 \pmod{n}$. If $x \notin \{1, \ldots, n-1\}$, then we can add or remove a multiple of n to bring it into this range, without changing the congruence $ax \equiv 1 \pmod{n}$. So the existence is proved.

Let us now prove the uniqueness. Suppose there exists a second solution $x' \in \{1, \ldots, n-1\}$ with $ax' \equiv 1 \pmod{n}$. Then $a(x - x') \equiv 0 \pmod{n}$, and therefore $n \mid a(x - x')$. Since $(n, a) = 1$, it follows that $n \mid x - x'$. But $x - x' \in \{-(n-1), \ldots, n-1\}$, leaving $x - x' = 0$ as the only possibility. □

7.3 The Idea behind RSA

We present the RSA cryptography system in a manner similar to that taken in the original article [8]. We start by walking through each of the steps in the algorithm. In a second pass, we will revisit each step in greater detail.

A public key cryptography system is initially set up by the person (or organization), which we will call the receiver, that wants to receive messages in a secure manner. It is the receiver that sets up the system and publishes how to send it messages.

Step 1. The receiver chooses two large primes p and q (roughly 100 digits long each), and calculates $n = pq$. The number n, the "public key," will be roughly 200 digits long. Given only n, computers cannot recover p and q in a reasonable amount of time.

Step 2. The receiver calculates $\phi(n)$, where ϕ is the Euler function defined as follows: $\phi(n)$ is the number of integers in $\{1, 2, \ldots, n-1\}$ that are relatively prime to n, for $n > 1$. By convention, we define $\phi(1) = 1$. In Proposition 7.8 we will show that $\phi(n) = (p-1)(q-1)$. Note that this formula requires knowledge of p and q. Thus, calculating $\phi(n)$ without knowing the factorization of n seems to be as hard as factoring n (although there is no rigorous proof that these two problems are in fact equivalently difficult).

Step 3: Choosing the encryption key. The receiver chooses $e \in \{1, \ldots, n-1\}$ relatively prime to $\phi(n)$. The number e is the encryption key. This number is public and is used by the sender to encode the message following the instructions publicly published by the receiver.

Step 4: Constructing the decryption key. There exists $d \in \{1, \ldots, n-1\}$ such that $ed \equiv 1 \pmod{\phi(n)}$. The existence of d follows from Corollary 7.6. The exact method of constructing d is implied by the proofs of Corollary 7.6 and its supporting propositions, including Euclid's algorithm. The integer d, constructed by the receiver, is the decryption key. This key, the "private key," remains secret and allows the receiver to decrypt its received messages.

Step 5: Encrypting a message. The sender wants to send a message that consists of an integer $m \in \{1, \ldots, n-1\}$, where m is relatively prime to n. To encode it, the sender calculates the remainder a from the division of m^e by n. Thus, we have that $m^e \equiv a \pmod{n}$, with $a \in \{1, \ldots, n-1\}$. The calculated integer a is the encrypted message. The sender sends a. As we will see later, it is easy for a computer to calculate a, even when m, e, and n are very large.

Step 6: Decrypting a message. The receiver receives an encrypted message a. To decrypt this message the receiver calculates $a^d \pmod{n}$. In Proposition 7.10, we will show that this will always yield precisely the initial message m.

Before discussing the different steps we consider a simple example with small numbers.

Example 7.7 *We let* $p = 7$ *and* $q = 13$*, and therefore* $n = pq = 91$*. Which integers of* $E = \{1, \ldots, 90\}$ *are not relatively prime to* 91*? These are all the multiples of* p *and* q*, of which there are 18:* 7, 13, 14, 21, 26, 28, 35, 39, 42, 49, 52, 56, 63, 65, 70, 77, 78, 84*. So there are* $90 - 18 = 72$ *integers in* E *that are relatively prime to* 91*, yielding* $\phi(91) = 72$*. We choose* $e = 29$ *from this set. We can easily verify that* $(e, \phi(n)) = 1$*. We use Euclid's algorithm to find* d*:*

$$\begin{aligned} 72 &= 29 \times 2 + 14, \\ 29 &= 14 \times 2 + 1. \end{aligned}$$

We work backward through the set of equations to write 1 *in terms of* 29 *and* 72*:*

$$1 = 29 - 14 \times 2 = 29 - (72 - 29 \times 2) \times 2 = 29 \times 5 - 72 \times 2.$$

Thus we have that $29 \times 5 \equiv 1 \pmod{72}$*, yielding* $d = 5$*. Let* $m = 59$ *be our message. We have* $(59, 91) = 1$*. To encode this message we must calculate* $59^{29} \pmod{91}$*. Since* 59^{29} *is a very large number, we have to be clever in our calculations. We will successively calculate* 59^2*,* 59^4*,* 59^8*, and* 59^{16} *modulo* 91 *and observe that* $59^{29} = 59^{16} \times 59^8 \times 59^4 \times 59$*. Get to it!*

$$\begin{aligned} 59^2 &= 3481 \equiv 23 \pmod{91}, \\ 59^4 &= (59^2)^2 \equiv 23^2 = 529 \equiv 74 \pmod{91}, \\ 59^8 &= (59^4)^2 \equiv 74^2 = 5476 \equiv 16 \pmod{91}, \\ 59^{16} &= (59^8)^2 \equiv 16^2 = 256 \equiv 74 \pmod{91}. \end{aligned}$$

Thus finally

$$\begin{aligned} 59^{29} &= 59^{16} \times 59^8 \times 59^4 \times 59 \pmod{91} \\ &\equiv (74 \times 16) \times 74 \times 59 \pmod{91} \\ &\equiv 1 \times 74 \times 59 = 4366 \pmod{91} \\ &\equiv 89 \pmod{91}. \end{aligned}$$

The method of calculation we have employed is that typically used by most computers. The encoded message is thus $a = 89$*, which we send to the receiver. To decode this message the receiver must calculate the remainder of* 89^5 *divided by* 91*. The same method of computation allows us to easily complete the calculation and recover the initial message. In fact,*

$$\begin{aligned} 89^2 &= 7921 \equiv 4 \pmod{91}, \\ 89^4 &= (89^2)^2 \equiv 4^2 = 16 \pmod{91}, \end{aligned}$$

allowing us to calculate

$$89^5 = 89^4 \times 89 \equiv 16 \times 89 = 1424 \equiv 59 \pmod{91}.$$

We have recovered the message m*!*

Proposition 7.8 *Let* p *and* q *be two distinct primes. Then*

$$\phi(pq) = (p-1)(q-1).$$

PROOF. We need to count the integers in $E = \{1, 2, \ldots, pq-1\}$ that are relatively prime to pq. The only integers which are not relatively prime to pq are the multiples of p, $P = \{p, 2p, \ldots (q-1)p\}$ (there are $q-1$ of them) and the multiples of q, $Q = \{q, 2q, (p-1)q\}$ (there are $p-1$ of them). Note that $P \cap Q = \emptyset$. If not, there would exist n and m such that $np = mq$, where $m < p$; thus $p \mid np = mq$, and by Proposition 7.5(4), either $p \mid m$ or $p \mid q$, both of which lead to a contradiction. Thus the number of integers in E that are relatively prime to pq is

$$pq - 1 - (p-1) - (q-1) = pq - p - q + 1 = (p-1)(q-1).$$

□

Theorem 7.9 *(Euler's theorem and Fermat's little theorem) If $m < n$ is relatively prime to n, then $m^{\phi(n)} \equiv 1 \pmod{n}$. (Fermat proved that $m^{n-1} \equiv 1 \pmod{n}$ when n is prime, this result being called Fermat's little theorem.)*

PROOF: We start by considering the case that n is prime. In this case $\phi(n) = n-1$, since the numbers $1, 2, \ldots, n-1$ are all relatively prime to n. Take $m \in E = \{1, \ldots, n-1\}$, and consider the products

$$1 \cdot m, \quad 2 \cdot m, \quad \ldots, \quad (n-1) \cdot m. \tag{7.3}$$

We will show that when divided by n, the remainders r_k of these products ($k \cdot m \equiv r_k \pmod{n}$) create a permutation of the sequence $1, \ldots, n-1$. To start, the remainder r_k of the division of $k \cdot m$ by n can never be zero if n is prime and $k, m < n$, so it belongs to E. It remains to show that the remainders are distinct. Suppose that $k_1 \cdot m$ and $k_2 \cdot m$ have the same remainder after division by n. Without loss of generality, we assume $k_1 \geq k_2$. Then we see that

$$k_1 \cdot m = q_1 \cdot n + r, \qquad k_2 \cdot m = q_2 \cdot n + r,$$

and therefore

$$(k_1 - k_2) \cdot m = (q_1 - q_2) \cdot n.$$

Thus n must divide $(k_1 - k_2) \cdot m$. Since n is prime, $0 \leq k_1 - k_2 < n$, and $m < n$, the only possibility is that $k_1 = k_2$.

Taking the product of the remainders r_i modulo n of the sequence in (7.3) and working modulo n, we see that

$$\begin{aligned} (n-1)! &= 1 \cdot 2 \cdot 3 \cdots (n-1) = r_1 \cdot r_2 \cdot \cdots \cdot r_{n-1} \\ &\equiv (m \cdot 1) \cdot (m \cdot 2) \cdots (m \cdot (n-1)) \pmod{n} \\ &= m^{n-1} \cdot (n-1)!. \end{aligned}$$

Rewritten, this yields

$$n \mid (m^{n-1} - 1) \cdot (n-1)!.$$

Since n is prime, we know that $(n, (n-1)!) = 1$. Thus $n \mid m^{n-1} - 1$, which is equivalent to the final result $m^{n-1} \equiv 1 \pmod{n}$.

The proof is nearly identical when n is not prime. In this case, instead of taking the numbers $1, \ldots, n-1$, we take only the $\phi(n)$ numbers that are relatively prime to n. As before, we multiply these by m, and consider the remainders after division by n. As before, these do not vanish. Since m and n are relatively prime, the result will again be a permutation of the original sequence. Indeed, if we assume that $k_1 \cdot m$ and $k_2 \cdot m$ are congruent modulo n and $k_1 \geq k_2$, then we can deduce that $(k_1 - k_2) \cdot m = (q_1 - q_2) \cdot n$, and therefore $n \mid (k_1 - k_2) \cdot m$. Since $(n, m) = 1$, then $n \mid k_1 - k_2$. But $0 \leq k_1 - k_2 \leq n-1$. So the only possibility is that $k_1 - k_2 = 0$, thus proving the distinctness of the remainders.

Taking the product of these numbers yields

$$\prod_{\substack{(k,n)=1\\k<n}} k \equiv m^{\phi(n)} \prod_{\substack{(k,n)=1\\k<n}} k \pmod{n}.$$

The result follows by "simplifying" the product $\prod_{(k,n)=1,k<n} k$, which is relatively prime to n by Proposition 7.5(3). Indeed, if $a = \prod_{(k,n)=1,k<n} k$ and $b = m^{\phi(n)} - 1$, we have $n \mid ab$ and $(n, a) = 1$. Hence $n \mid b$, which yields the conclusion. □

Proposition 7.10 *RSA encryption and decryption are inverses one of the other: if we encrypt a message m, where $(m, n) = 1$, as a, where $m^e \equiv a \pmod{n}$, then the decryption always yields the original message m. That is, $a^d \equiv m \pmod{n}$.*

Proof. If $m^e \equiv a \pmod{n}$, then

$$\begin{aligned} a^d &\equiv (m^e)^d = m^{ed} = m^{k\phi(n)+1} = m^{k\phi(n)} \cdot m = (m^{\phi(n)})^k \cdot m \\ &\equiv 1^k \cdot m = 1 \cdot m = m \pmod{n}. \end{aligned}$$

□

Example 7.11 *A company wants to build an online ordering system. To secure the transmission of customer credit card information, they use a public key cryptosystem. The credit card number is a 16-digit number combined with 4 digits describing its expiry date, yielding a total of 20 digits. The company therefore chooses two large primes p and q. In our example we will use primes of 25 digits, yielding $n = pq$ of roughly 50 digits. Let*

$$p = 1234567980199456799008945 9$$

and

$$q = 8369567977777368712343087.$$

This gives

$$n = pq = 10332800633466658218847856400733362485 5622630219933$$

and

$$\begin{aligned} \phi(n) &= (p-1)(q-1) \\ &= 10332800633466658218847854329208584508 3685927787388. \end{aligned}$$

The company chooses

$$e = 115670849$$

such that $(e, \phi(n)) = 1$, and uses Corollary 7.6 to calculate

$$d = 3411393174391092578448356106544218397 7516731202177.$$

The value of d in this example is quite large, which effectively negates any chance of using trial and error to discover it.

The algorithm is constrained to sending only messages m that are relatively prime to n. Fortunately, the only divisors of n have at least 25 digits, thus all 20-digit numbers must be relatively prime to n. Consider a customer with a credit card number of 4540 3204 4567 8231 and an expiration date of 10/02. The customer wishes to securely send the message $m = 45403204456782311002$. Thus, the software calculates

$$\begin{aligned} m^e &\equiv a \\ &\equiv 49329085221791275793017511397395566847998886183308 \pmod{n} \end{aligned}$$

and transmits it to the company. Upon receiving this encrypted transmission the company calculates

$$a^d \equiv 45403204456782311002 = m \pmod{n}.$$

It should be pointed out that the chosen values of p and q, although seemingly large, are not large enough to prevent a computer from easily factoring n.

What would have happened had there been an error in the transmission? With high probability the receiver would be aware of the error, since the decrypted value would have very little chance of being a 20-digit number.

Signing a message: Up until now we have seen how a person, call him Bob, could put in place a public key cryptosystem allowing him to securely receive messages from anybody. Suppose that Bob receives a message from his friend Alice asking him to transfer a large sum of money into her account. Does this prove that the message really came from Alice, and not from someone impersonating Alice? Thus it becomes necessary for Alice to be able to prove that she is in fact the author of the message sent to Bob. This is what we call signing a message.

In this case, both the sender and the receiver construct a public key cryptosystem, consisting of a triplet (n, e, d). Two public keys are necessary.

- The sender (Alice) shares n_A and e_A, while keeping d_A secret.
- The receiver (Bob) publishes n_B and e_B, while keeping d_B secret.

Transmitting a signed message:

- To send a signed message m relatively prime to n_A, the sender starts placing his (her) signature by calculating

  $$m_1 \equiv m^{d_A} \pmod{n_A}.$$

  If m_1 is relatively prime to n_B, she then encodes it with the receiver's public key:

  $$m_2 \equiv m_1^{e_B} \pmod{n_B}.$$

  The sender then sends m_2. If it happens that $(m_1, n_B) \neq 1$, not very likely given that n_B has so few divisors, the sender changes the message m slightly until both $(m, n_A) = 1$ and $(m_1, n_B) = 1$.

- In order to decrypt the signed message the receiver starts by recovering m_1, decrypting it with his secret key d_B:

$$m_1 \equiv m_2^{d_B} \pmod{n_B}.$$

Indeed,

$$m_2^{d_B} \equiv m_1^{e_B d_B} \equiv m_1^{k_1\phi(n_B)+1} = m_1 \cdot (m_1^{\phi(n_B)})^{k_1} \equiv m_1 \pmod{n_B}.$$

Afterward, he recovers the original message using the sender's public key:

$$m \equiv m_1^{e_A} \pmod{n_A}.$$

Indeed,

$$m_1^{e_A} \equiv m^{d_A e_A} \equiv m^{k_2\phi(n_A)+1} = m \cdot (m^{\phi(n_A)})^{k_2} \equiv m \pmod{n_A}.$$

If the message was sent by an impostor, this will become obvious to the receiver after the decryption. In the credit card example, if the message had been sent by an impostor, there would be effectively no chance that the calculated value m would have exactly 20 digits, or correspond to a valid credit card number. In the context of sending a text message, we would initially apply some transform to map a sequence of letters to a sequence of numbers. Had an impostor sent such a message, the decoded text would in all probability be an incoherent jumble.

Applications: The RSA cryptosystem is widely used on the Internet, for example for securing the transmission of sensitive data such as credit card information. The banking system is also protected by RSA encryption. However, the RSA algorithms require long and complex computations. The algorithms therefore lose their allure when we need to send extremely long messages. In this case, other systems are normally used, especially when the message does not need to remain secret for a long period of time. Among the many faster cryptosystems we find DES (the Data Encryption Standard) and the more recent AES (the Advanced Encryption Standard) (see [3]). DES and AES are symmetric key cryptosystems, meaning that the same key is shared by both sender and receiver and used to both encrypt and decrypt the message. The key is typically much shorter than the message itself and may be securely shared using the more costly RSA cryptosystem.

Discussion on the value of the RSA cryptosystem: The RSA cryptosystem was introduced in 1978. It has stimulated much research for improved factorization methods, but without much success: RSA remains unbroken today, provided that n is chosen sufficiently large. It is not even known whether breaking RSA is equivalent to factorization, or whether there exists a cheaper alternative route. However, all efforts to break RSA using techniques other than factoring n have been without success thus far.

In 1978 the original paper [8] estimated that it would take 74 years (using 1978 equipment) to factor a 100-digit number, 3.8×10^9 years to factor a 200-digit number, and 4.2×10^{25} years to factor a 500-digit number. What about using modern equipment? Given the huge advances in computing power, 100-digit keys are to be avoided. As of 2005, 200-digit keys are considered breakable by decryption experts using large supercomputers (see below). The advances in factoring have come on two fronts: better computers and better algorithms. Moore's "law" (named after Gordon Moore, a co-founder of Intel), originally stated in 1965, said that the density of transistors would double every 18 months to two years. Amazingly, this trend has held true until now. How does this relate to the speed of calculations? The following answer comes from Paul Rousseau, an employee of TSMC: the speed of transistors increases by a factor of 1.4 every two to three years. Even if companies announce that the clock speed of a given processor is multiplied by 2, the processor does less work per cycle, and this multiplier is therefore purely artificial. A better measure is therefore the capacity of the processor to do "real work." For an algorithm such as factoring, where the work may be performed in parallel, the real increase in work capacity is roughly 2.8, where a factor of 1.4 comes from the faster transistors and a factor of 2 comes from the increased count of transistors. As of 2005, twenty-seven years have passed since 1978. If we assume that a generation occurs every 2.5 years, then 10.8 generations have passed, yielding a factor of 67,500, which is less than 10^5.

The improvement of algorithms for factoring has been no less spectacular. In the nineteenth century, Gauss had already classified the problem of factoring large numbers as a fundamental problem in number theory. The most important algorithms are:

- the quadratic sieve method of Pomerance,
- the elliptic curve method of Lenstra, and
- the general number field sieve method of Pollard, Adleman, Buhler, Lenstra, and Pomerance.

Carl Pomerance wrote an excellent article on the subject [7].

In 1996 we were factoring numbers of 130 digits, and in 1999 we were factoring numbers of 155 digits. In 2005, F. Bahr, M. Boehm, J. Franke, and T. Kleinjung announced the factoring of a 200-digit number:

$$\begin{aligned} n &= 27997833911221327870829467638722601621070446786955 \\ &\quad 42853756000992932612840010760934567105295536085606 \\ &\quad 18223519109513657886371059544820065767750985805576 \\ &\quad 13579098734950144178863178946295187237869221823983, \end{aligned}$$

factored as $n = pq$, where p and q are primes given by

$$\begin{aligned} p &= 35324619344027701212726049781984643686711974001976 \\ &\quad 25023649303468776121253679423200058547956528088349, \\ q &= 79258699544783330333470858414800596877379758573642 \\ &\quad 19960734330341455767872818152135381409304740185467. \end{aligned}$$

This factorization was obtained using the general number field sieve technique, which as of 2005 remains the best known factoring algorithm.

In his 2000 paper [4], Jean-Paul Delahaye recommends the use of a key of 232 digits for not very important data, a key of 309 digits for commercial use, and a key of 617 digits if the message must remain protected over a long period of time.

Carmichael numbers. The RSA cryptosystem with public key n requires that messages m satisfy $(m, n) = 1$ in order for the encryption and decryption to function. In fact, if we encrypt and decrypt messages m such that $(m, n) \neq 1$, we often find that the method still works. It is therefore natural to ask, is the condition $(m, n) = 1$ really required? The answer to this question is known: the condition is unnecessary if n is a Carmichael number. Unfortunately, Carmichael numbers are hard to find and contain at least three factors. So the condition $(m, n) = 1$ is really required.

7.4 Constructing Large Primes

Earlier, we stated that it is relatively simple to construct large prime numbers. This is a direct consequence of the prime number theorem, which in simple terms tells us the probability that a randomly chosen integer of N digits will be prime. To construct a 100-digit prime number, we simply randomly generate 100-digit numbers and test whether they are prime. The prime number theorem assures us that after an average of 115 tries we should obtain a prime number (assuming that we generate only odd numbers).

Theorem 7.12 *(Prime number theorem) Let $\pi(N) = \#\{p \leq N \mid p \text{ is prime}\}$ (that is, $\pi(N)$ is the number of primes less than or equal to N). Provided N is sufficiently large, then*

$$\pi(N) \sim \frac{N}{\ln N}.$$

Remark: The proof of this theorem is very advanced, and will not be discussed here.

We want to generate large prime numbers. Suppose for the moment that we have an oracle that lets us decide whether a given number is prime. We can then randomly choose a large integer n and test whether it is prime. If it is not, we could test $n + 1$ and so on, until we chance upon a prime number. We will show that this is not really a good approach to take.

Theorem 7.13 *There exist arbitrarily long sequences of consecutive nonprime integers.*

Proof: Consider $n \in \mathbb{N}$. The following sequence of length n consists purely of composite integers:

$$n! + 2, n! + 3, \ldots, n! + n.$$

In fact, for $1 < m \leq n$ we have that $m \mid n!$ and therefore $m \mid n! + m$. □

A better technique is to randomly choose large integers and test them to see whether they are prime. Assuming that our choices are independent, the laws of probability assure us that we will find a prime number after a reasonable number of tries.

Consider the set of integers $F = \{1, \ldots, N\}$ for a given large value of N. If we would like to find primes of 100 digits (or 200), we would take $N = 10^{100}$ ($N = 10^{200}$, respectively). By the prime number theorem there are approximately $\pi(N) = \frac{N}{\ln N}$ prime integers in the set F. Thus, if we are to randomly choose an integer n in the set F, the probability that it will be prime is roughly

$$\text{Prob}(n \text{ prime}) \approx \frac{\frac{N}{\ln N}}{N} = \frac{1}{\ln N}.$$

For $N = 10^{100}$ we obtain $\ln N = 100 \ln 10 \approx 100 \times 2.30259 = 230.259$. Thus, we have roughly a 1 in 230 chance that a randomly chosen integer from F will be prime. We can immediately double our chances if we restrict ourselves to choosing only odd numbers (simply choose the last digit from the set $\{1, 3, 5, 7, 9\}$). Similarly, we can further improve our chances by choosing the last digit from the set $\{1, 3, 7, 9\}$ (eliminating multiples of 5), giving us a final probability of 1 in 92.

We let B be the subset of F of integers that are odd and not divisible by 5. The number of elements of B is approximately $\frac{2}{5}N$. Let $p = \frac{5}{2 \ln N}$. Every time we choose a random-number from B it will be prime with probability p. We consider the "random experiment" of randomly drawing a number from B and testing whether it is prime. We repeat the experiment independently until we chance upon a prime number. Let X be the number of experiments necessary. Then X is a geometric random variable with parameter p. Thus,

$$\text{Prob}(X = k) = (1 - p)^{k-1} p.$$

This formula is a simple expression of the fact that we have probability $(1 - p)$ of drawing a nonprime number on each of the first $k - 1$ experiments, and probability p of drawing a prime on the kth experiment. The expected value of the random variable X is the average number of experiments we would expect to perform before finding a prime number. For our geometric random variable X with parameter p we have that

$$E(X) = \sum_{k=1}^{\infty} k \text{Prob}(X = k) = \sum_{k=1}^{\infty} k(1 - p)^{k-1} p = \frac{1}{p}.$$

(Showing that $\sum_{k=1}^{\infty} k(1 - p)^{k-1} p = \frac{1}{p}$ requires some cleverness. This calculation can be found in any text on probability.)

In our example $p \approx \frac{1}{92}$ if we have chosen the last digit from the set $\{1, 3, 7, 9\}$; thus $E(X) = 92$. Hence, we would expect to perform 92 experiments on average before finding a prime number.

What we have done up until this moment has assumed that testing whether an integer n is prime is significantly easier than just factoring n. Such a test is called a primality test. There exists a wide variety of them in the literature, although most require a certain level of mathematical sophistication. The method we will present here is the one that appeared in the original article on RSA [8]. It is quite technical and uses the nonintuitive Jacobi symbol, introduced below. The underlying principle is that a composite number n leaves its fingerprints everywhere, so much so that roughly half of the numbers in the set $\{1, \ldots, n\}$ "know" that n is composite. If n passes the test with respect to k numbers $m_1, \ldots, m_k \in \{1, \ldots, n\}$, then n is prime with a very high probability, as can be shown using Bayes's formula.

A primality test. Given two relatively prime integers m and n, we may calculate the Jacobi symbol $J(m, n) \in \{-1, 1\}$. The full definition of $J(m, n)$ will be given a little later. Let

$$E = \{1, \ldots, n-1\}.$$

If n is a prime number and if $a \in E$, then

$$\begin{cases} (a, n) = 1, \\ J(a, n) \equiv a^{\frac{n-1}{2}} \pmod{n}. \end{cases} \tag{7.4}$$

If n is not prime, then at least half of the numbers in E will not satisfy (7.4). The instant we find an integer $a \in E$ that fails this test (by not satisfying (7.4)), we know with certainty that n is not prime. If we choose $a \in E$ randomly, then we have that

$$\text{Prob}(a \text{ passes the test} \mid n \text{ is not prime}) \leq \frac{1}{2}.$$

Suppose we have randomly chosen $a_1, \ldots, a_k \in E$ and that n has passed the test with respect to each of them. We wish to calculate the chance that n is in fact prime. We start by labeling each of the events: let A_i be the event "a_i passes the test." Let $P(n)$ be the event "n is prime," and let $Q(n)$ be its complement; that is, $Q(n)$ is the event "n is composite." Let $A = A_1 \cap \cdots \cap A_k$. Therefore, A is the event "all of $a_1, \ldots, a_k$ pass the test." Bayes's formula tells us that

$$\text{Prob}(P(n) \mid A) = \frac{\text{Prob}(A \mid P(n))\text{Prob}(P(n))}{\text{Prob}(A \mid P(n))\text{Prob}(P(n)) + \text{Prob}(A \mid Q(n))\text{Prob}(Q(n))}.$$

Since we have that

$$\begin{aligned} \text{Prob}(A \mid P(n)) &= 1, \\ \text{Prob}(A \mid Q(n)) &\leq \tfrac{1}{2^k}, \end{aligned}$$

and that we may approximately calculate $P(n)$ (and hence $Q(n)$) using the prime number theorem, we may approximately calculate the probability that n is prime given that $a_1, \ldots, a_k$ have each passed the test (more precisely, we may calculate a lower bound to this probability).

In fact, the denominator is given by

$$\begin{aligned}\text{Prob}(A \mid P(n))\text{Prob}(P(n)) + \text{Prob}(A \mid Q(n))\text{Prob}(Q(n)) \\ \leq \text{Prob}(P(n)) + \tfrac{1}{2^k}\text{Prob}(Q(n)),\end{aligned}$$

while the numerator is simply $\text{Prob}(P(n))$. We return to the our earlier example in which n is a 100-digit odd integer not divisible by 5 (that is, an element of B). We have already seen that

$$\text{Prob}(P(n)) \approx \frac{1}{92}$$

and that $\text{Prob}(Q(n)) \approx \frac{91}{92}$. This yields

$$\text{Prob}(P(n) \mid A) \geq \frac{1}{1 + 91\frac{1}{2^k}} = p_k.$$

Consider the following values of p_k, calculated for various k:

$$\begin{aligned} p_{10} &= 0.9184 = 1 - 0.816 \times 10^{-1}, \\ p_{20} &= 0.999913 = 1 - 0.868 \times 10^{-4}, \\ p_{30} &= 0.9999999152 = 1 - 0.848 \times 10^{-7}, \\ p_{40} &= 0.9999999999172 = 1 - 0.828 \times 10^{-10}. \end{aligned}$$

We see that k does not need to be particularly large in order to ensure that n is prime with very high probability.

It remains to fully define the Jacobi symbol and show that fewer than half of the values $a \in E$ pass the test (satisfy (7.4)) when n is composite. We must also show that all integers $a \in E$ will pass the test when n is prime.

The Jacobi symbol. Let $a, b \in \mathbb{N}$ be relatively prime integers. The Jacobi symbol $J(a, b)$ has values in the set $\{-1, 1\}$. If b is prime we define

$$J(a,b) = \begin{cases} 1 & \text{if} \quad \exists x \in \mathbb{N} \quad x^2 \equiv a \pmod{b}, \\ -1 & \text{otherwise.} \end{cases}$$

If b is composite, we may factor b and write it as $b = p_1 \cdots p_r$ (where the p_i are not necessarily distinct). The Jacobi symbol is then defined as

$$J(a,b) = J(a,p_1) \cdots J(a,p_r) = \prod_{i=1}^{r} J(a,p_i).$$

(The Jacobi symbol $J(a, b)$ is denoted by $\left(\frac{a}{b}\right)$ in many number theory texts.) We see quickly that this definition is a little obscure, and worse, quite difficult to manipulate. How are we to determine whether there exists x such that $x^2 \equiv a \pmod{b}$? In other

words, how are we to determine whether a is a square modulo b (we say that a is a *quadratic residue*)? Moreover, this definition requires us to know the factorization of b. This leaves us with the impression of turning in circles, requiring the factorization of an integer in order to test whether it is prime! Thankfully, there exist simpler alternative means of calculating $J(a,b)$. We illustrate with a few examples.

The following theorem, cited without proof, provides an algorithm for easily calculating $J(a,b)$. Observe that in our case, we restrict ourselves to the case $a \leq b$ and b odd.

Theorem 7.14 *If $(a,b) = 1$, $a \leq b$ and b is odd, then*

$$J(a,b) = \begin{cases} 1, & \text{if } a = 1, \\ J(\frac{a}{2},b)(-1)^{\frac{b^2-1}{8}}, & \text{if } a \text{ even}, \\ J(b \pmod a, a)(-1)^{\frac{(a-1)(b-1)}{4}}, & \text{if } a \text{ odd and } a > 1. \end{cases} \tag{7.5}$$

In (7.5), note that the fractions $\frac{b^2-1}{8}$ and $\frac{(a-1)(b-1)}{4}$ are always integers (see Exercise 16).

Example 7.15 *Consider $a = 130$ and $b = 207$. Then*

$$\begin{aligned} J(130,207) &= J(65,207)(-1)^{\frac{42848}{8}} = J(65,207)(-1)^{5356} \\ &= J(65,207) = J(12,65)(-1)^{\frac{64\times 206}{4}} = J(12,65) \\ &= J(6,65)(-1)^{\frac{4224}{8}} = J(6,65)(-1)^{528} = J(6,65) \\ &= J(3,65)(-1)^{528} = J(3,65) = J(2,3)(-1)^{\frac{2\times 64}{4}} \\ &= J(2,3) = J(1,3)(-1)^{\frac{8}{8}} = -J(1,3) = -1. \end{aligned}$$

The calculation may seem long and tedious, but is in fact easily performed by a computer.

To determine whether a passes the test we must calculate $a^{\frac{b-1}{2}} \pmod b$. We have that $\frac{b-1}{2} = 103$. We have already seen how to evaluate $130^{103} \pmod n$. We first decompose $\frac{b-1}{2} = 103$ into powers of 2: $103 = 64+32+4+2+1 = 1+2^1+2^2+2^5+2^6$. *We then calculate*

$$\begin{aligned} 130^2 &= 16900 \equiv 133 \pmod{207}, \\ 130^4 &= (130^2)^2 \equiv 133^2 = 17689 \equiv 94 \pmod{207}, \\ 130^8 &= (130^4)^2 \equiv 94^2 = 8836 \equiv 142 \pmod{207}, \\ 130^{16} &= (130^8)^2 \equiv 142^2 = 20164 \equiv 85 \pmod{207}, \\ 130^{32} &= (130^{16})^2 \equiv 85^2 = 7225 \equiv 187 \pmod{207}, \\ 130^{64} &= (130^{32})^2 \equiv 187^2 = 34969 \equiv 193 \pmod{207}. \end{aligned}$$

Now

$$\begin{aligned} 130^{103} &= 130^{64} \times 130^{32} \times 130^4 \times 130^2 \times 130 \\ &\equiv 193 \times 187 \times 94 \times 133 \times 130 \pmod{207} \\ &\equiv 67 \pmod{207}. \end{aligned}$$

We see that $J(130, 207) \neq 130^{\frac{207-1}{2}}$*, and can thus conclude that* 207 *is not prime. In this case, we could also have seen this directly, since* $207 = 3^2 \cdot 23$.

In the discussion of our primality testing algorithm we asserted that if n is not prime, then fewer than half of the elements of E pass the test. Similarly, we asserted that if n is prime, then all the elements of E will pass the test. We will now provide a sketch of a proof for the first fact and a proof for the second. The following discussions are quite advanced, and make use of group theory.

Definition 7.16 *1. A set* G *combined with an operation* $*$ *is a* group *if*

- *the operation* $*$ *is associative:*

$$\forall a, b, c \in G, \quad (a * b) * c = a * (b * c);$$

- *there exists an identity element* $1 \in G$ *such that*

$$\forall a \in G, \quad 1 * a = a * 1 = a;$$

- *all elements have inverses:*

$$\forall a \in G, \ \exists b \in G, \quad a * b = b * a = 1.$$

2. *A subset* $H \subset G$ *of* G *is a* subgroup *of* G *if* H *is itself a group with respect to the operation* $*$.
3. *A group* G *is* cyclic *if there exists an element* $g \in G$ *such that any element* a *of the group may be expressed in the form* $a = g^m$ *for some integer* $m \in \mathbb{Z}$*, and where we define*

$$g^m = \begin{cases} \underbrace{g * g * \cdots * g}_{m} & m > 0, \\ 1 & m = 0, \\ \underbrace{g^{-1} * g^{-1} * \cdots * g^{-1}}_{|m|} & m < 0. \end{cases}$$

In the case of a finite cyclic group with n *elements we can convince ourselves that the group must be of the form* $G = \{1, g, g^2, \ldots, g^{n-1}\}$ *and that* $g^n = 1$.

Example 7.17 *Let* p *be a prime and* $G = \{1, 2, \ldots, p-1\}$*. We define the operation* $*$ *on* G *as* $a * b = c$*, where* c *is the remainder of* ab *after division by* p*. In other words,* $*$ *is simply the operation of multiplication modulo* p*. Under this operation* G *is a group. We let the reader verify that* $*$ *is associative. It is obvious that* 1 *is the identity element. Finally, we are guaranteed the existence of an inverse for every element by Corollary 7.6. As we will see in Theorem 7.22,* G *is actually a cyclic group.*

Let us verify this for $p = 7$*. We take* $g = 3$*. Then we have* $g^2 = 2$*,* $g^3 = 6$*,* $g^4 = 4$*,* $g^5 = 5$*,* $g^6 = 1$.

Notation. In the above example and those yet to come, the group operation will always be multiplication modulo n. We will thus often omit the $*$ and simply write ab for $a * b$.

Lagrange proved the following theorem:

Theorem 7.18 *(Lagrange's theorem) Let G be a finite group and H a subgroup of G. Then the number of elements in H, written $|H|$, is a divisor of the number of elements in G, written $|G|$.*

PROOF. If $H = G$ then we are finished. Otherwise, there must exist $a_1 \in G \setminus H$.

Let $a_1H = \{a_1 * h \mid h \in H\}$. Then $|a_1H| = |H|$. In fact, given $h, h' \in H$, if $h \neq h'$ then $a_1 * h \neq a_1 * h'$. So the map $f : H \to a_1H$ defined by $h \mapsto a_1h$ is a bijection.

Moreover, $a_1H \cap H = \emptyset$. Indeed, if $h \in a_1H \cap H$, then $h = a_1h'$ for some $h' \in H$. Thus, $a_1 = h * (h')^{-1} \in H$, a contradiction.

There are two cases to consider. Either $a_1H \cup H = G$, in which case $|G| = 2|H|$, or there exists $a_2 \in G \setminus (H \cup a_1H)$. Again, we let $a_2H = \{a_2 * h \mid h \in H\}$ and iterate the preceding argument. Since G is finite we may express it as $G = H \cup a_1H \cup a_2H \cup \cdots \cup a_nH$, where H and the a_iH are pairwise disjoint, and $|H| = |a_1H| = \cdots = |a_nH|$. Thus, $|G| = (n+1)|H|$. □

Theorem 7.19 *If n is not prime, then fewer than half of the integers $a \in E = \{1, \ldots, n-1\}$ pass the test (satisfy (7.4)).*

PROOF SKETCH: The proof uses the following insight. The elements of E that are relatively prime to n form a group G under multiplication modulo n. This can be seen by noticing that the product aa' of two elements $a, a' \in E$ relatively prime to n is itself relatively prime to n; in other words, $(aa', n) = 1$. Let a'' be the remainder of aa' when divided by n. Then a'' must also be relatively prime to n and also a member of E. Our group operation is once again multiplication modulo n (so $a * a' = a''$), and our group G is closed under this operation. It is easy to verify that the operation is associative and that 1 is the identity element. Finally, Corollary 7.6 tells us that each element of G has an inverse in G.

The group G has fewer than $n-1$ elements. The subset of the elements of G that maintain the equality of (7.4) is a subgroup H of G, a fact that we will take for granted here. By Lagrange's theorem the number of elements in H must divide the number of elements in G. Thus, two cases are possible. Suppose $|H| < |G|$. Then $|H|$ is a proper divisor of $|G|$ and in particular, $|H| \leq \frac{|G|}{2}$. Suppose instead that $|H| = |G|$. We can show that this case is impossible by proving the existence of an element $a \in G$ such that $J(a, n)$ is not congruent to $a^{\frac{n-1}{2}}$ modulo n. This proof is also rather advanced, and will not be presented here.

Finally, $|H| \leq \frac{|G|}{2} < \frac{n-1}{2}$. □

Theorem 7.20 *If n is an odd prime, then all $a \in E = \{1, \ldots, n-1\}$ will pass the test (satisfy (7.4)).*

We will present the various pieces of the proof independently, since they will be useful to us in later chapters.

Lemma 7.21 *(1) Let n be prime, $S = \{0, 1, \ldots, n-1\}$, and $P(x)$ a polynomial*

$$P(x) = x^r + a_{r-1}x^{r-1} + \cdots + a_1 x + a_0$$

with $a_i \in S$. Then there exist at most r solutions $x_i \in S$ to the congruence

$$P(x) \equiv 0 \pmod{n}.$$

(2) In the case $P_d(x) = x^d - 1$ where $d \mid n-1$, the congruence $P_d(x) \equiv 0 \pmod{n}$ has exactly d distinct solutions in the set $E = S \setminus \{0\}$.

Proof.

(1) The argument uses induction on r. Clearly, the statement holds for $r = 1$. Suppose now that the statement holds for polynomials of degree r and consider a polynomial $P(x)$ of degree $r+1$. Suppose there exists $a_1 \in E$ such that $P(a_1) \equiv 0 \pmod{n}$. We divide the polynomial $P(x)$ by $x - a_1$, obtaining

$$P(x) = (x - a_1)Q(x) + \beta,$$

where $Q(x)$ is a polynomial of degree r with coefficients in $\mathbb{Z}$. Let

$$Q(x) = x^r + b_{r-1}x^{r-1} + \cdots + b_1 x + b_0,$$

$b_i \equiv c_i \pmod{n}$, and $\beta \equiv \gamma \pmod{n}$, with $c_i, \gamma \in S$. Define

$$Q'(x) = x^r + c_{r-1}x^{r-1} + \cdots + c_1 x + c_0.$$

If $x \in S$ we have that $Q(x) \equiv Q'(x) \pmod{n}$, and therefore

$$P(x) \equiv (x - a_1)Q'(x) + \gamma \pmod{n}.$$

Evaluating this at a_1 we obtain $P(a_1) \equiv \gamma \pmod{n}$. Thus $\gamma = 0$ and

$$P(x) \equiv (x - a_1)Q'(x) \pmod{n}.$$

Thus $P(x) \equiv 0 \pmod{n}$ if and only if $n \mid (x - a_1)Q'(x)$. Since n is prime, this occurs if and only if $n \mid x - a_1$ or $n \mid Q'(x)$, that is, $x \equiv a_1 \pmod{n}$ or $Q'(x) \equiv 0 \pmod{n}$. By the inductive hypothesis, $Q'(x) \equiv 0 \pmod{n}$ has at most r solutions; thus $P(x)$ has at most $r+1$ roots modulo n.

(2) By Fermat's little theorem (Theorem 7.9) all $x \in S \setminus \{0\}$ are solutions to $P_{n-1}(x) \equiv 0 \pmod{n}$. Thus this congruence has exactly $n-1$ distinct solutions. Suppose d is a divisor of $n-1$ such that $n-1 = dk$. Then we may write $P_{n-1}(x) = (x^d - 1)Q(x)$, where $Q(x) = \sum_{i=0}^{k-1} x^{id}$. By (1), $P_d(x)$ has at most d roots modulo n and $Q(x)$ has at most $(k-1)d$ roots. Since P_{n-1} has exactly $n-1$ solutions, each of the $k + (k-1)d = n - 1$ solutions to $P_d(x)$ and $Q(x)$ must exist. Hence, $P_d(x) \equiv 0 \pmod{n}$ has exactly d solutions in $S \setminus \{0\}$. □

Theorem 7.22 *If n is prime then the set $E = \{1, \ldots, n-1\}$ is a cyclic group with respect to multiplication modulo n. If $g \in E$ is such that $E = \{g, g^2, \ldots, g^{n-1} = 1\}$, then g is called a* primitive root *of E.*

PROOF. Begin by observing that $E = \{1, \ldots, n-1\}$ is a group with respect to multiplication modulo n. In fact, since n is prime, all $a \in E$ are relatively prime to n. The conclusion follows from Corollary 7.6.

By Fermat's little theorem (Theorem 7.9), for all $a \in E$ we have $a^{n-1} = 1$. (Note that the equality $a^{n-1} = 1$ is inside the group G. Its meaning is $a^{n-1} \equiv 1 \pmod{n}$.) Let r be the minimum integer such that $a^r = 1$. We are certain that such an r exists, since $a^{n-1} = 1$. This r is called the *order* of the element a. Consider the set $F = \{a, a^2, \ldots, a^r = 1\}$. It is easy to verify that F is in fact a subgroup of E containing r elements. Thus, by Lagrange's theorem if follows that $r \mid n-1$.

We must show that there exists an element $a \in E$ with order $n-1$. Let d be a proper divisor of $n-1$. We will show that there are exactly d elements of G whose orders divide d. In fact, all elements a whose orders divide d are solutions to the congruence $x^d - 1 \equiv 0 \pmod{n}$. The desired result follows from Lemma 7.21(2).

Decompose $n-1$ into prime factors, $n-1 = p_1^{k_1} \cdots p_s^{k_s}$, and consider the polynomials $Q_{p_i^{k_i}}(x) = x^{p_i^{k_i}} - 1$. By Lemma 7.21(2) each congruence $Q_{p_i^{k_i}}(x) \equiv 0 \pmod{n}$ has exactly $p_i^{k_i}$ solutions in E: all of the solutions are the elements of E whose order divides $p_i^{k_i}$. If all solutions to $Q_{p_i^{k_i}}(x) \equiv 0 \pmod{n}$ corresponded to elements of the group with order less than $p_i^{k_i}$, then their orders would divide $p_i^{k_i - 1}$. These elements would thus be solutions to the congruence $Q_{p_i^{k_i-1}}(x) = x^{p_i^{k_i-1}} - 1 \equiv 0 \pmod{n}$. This is a contradiction, since $Q_{p_i^{k_i-1}}(x) \equiv 0 \pmod{n}$ has exactly $p_i^{k_i-1}$ solutions in E. Thus, let $g_i \in E$ be a solution to $Q_{p_i^{k_i}}(x) \equiv 0 \pmod{n}$ corresponding to an element of order $p_i^{k_i}$. We may easily verify that

$$g = g_1 \cdots g_s$$

has order $p_1^{k_1} \cdots p_s^{k_s} = n - 1$. This is a consequence of the following lemma. □

Lemma 7.23 *Let G be a finite group with a commutative group operation. If g_1 has order m_1 and g_2 has order m_2 such that $(m_1, m_2) = 1$, then $g_1 g_2$ has order $m_1 m_2$.*

PROOF. Let m be the order of $g_1 g_2$. We know that $(g_1 g_2)^{m_1 m_2} = (g_1^{m_1})^{m_2} (g_2^{m_2})^{m_1} = 1$, and thus $m \mid m_1 m_2$. Since $m \mid m_1 m_2$, we have that $m = n_1 n_2$, where $n_1 = (m_1, m) \mid m_1$ and $n_2 = (m_2, m) \mid m_2$ (exercise: explain why!). This allows us to write m_i as $m_i = n_i r_i$. We also have

$$g_1^{m r_1} = g_1^{n_1 n_2 r_1} = (g_1^{m_1})^{n_2} = 1.$$

Since $(g_1 g_2)^m = 1$, it follows that $g_1^m = g_2^{-m}$, so we also have $g_2^{-m r_1} = 1$, which yields $g_2^{m r_1} = 1$. But

$$g_2^{m r_1} = g_2^{n_1 n_2 r_1} = g_2^{m_1 n_2}.$$

Therefore we must have that $m_2 \mid m_1 n_2$. Since $(m_1, m_2) = 1$, this implies $m_2 \mid n_2$. Since we also have $n_2 \mid m_2$, we finally conclude that $m_2 = n_2$. Analogously, we may show that $m_1 = n_1$. Thus it follows that $m = m_1 m_2$. □

PROOF OF THEOREM 7.20. It suffices to show that all a satisfy $J(a, n) \equiv a^{\frac{n-1}{2}} \pmod{n}$. For each a we have two possibilities.

If a is a quadratic residue, meaning there exists $x \in E$ such that $x^2 \equiv a \pmod{n}$, then by definition, $J(a, n) = 1$. The other half of the equality, $a^{\frac{n-1}{2}} \equiv x^{n-1} \equiv 1 \pmod{n}$, follows immediately from Fermat's little theorem (Theorem 7.9).

The second case is that a is not a quadratic residue, and it requires a little more work. In this case we have that $J(a, n) = -1$ by definition. We will show that $a^{\frac{n-1}{2}} \equiv -1 \pmod{n}$.

In Theorem 7.22 we showed that there exists $g \in E$ such that $E = \{g, g^2, \ldots, g^{n-1} = 1\}$. Since $g^{n-1} = 1$, each element $a \in E$ satisfies $a^{n-1} = 1$, and is thus a solution to the congruence $x^{n-1} - 1 \equiv 0 \pmod{n}$. Observe that

$$x^{n-1} - 1 = \left(x^{\frac{n-1}{2}} - 1\right)\left(x^{\frac{n-1}{2}} + 1\right).$$

In the proof of Theorem 7.22 we saw that a congruence $P(x) \equiv 0 \pmod{n}$, where $P(x)$ is of degree $\frac{n-1}{2}$, has at most $\frac{n-1}{2}$ solutions in E.

If is obvious that $1, g^2, g^4, \ldots, g^{2k}, \ldots$ are quadratic residues. They are the solutions of $x^{\frac{n-1}{2}} - 1 \equiv 0 \pmod{n}$. Thus the elements $g, g^3, \ldots, g^{2k+1}, \ldots$ are solutions of $x^{\frac{n-1}{2}} \equiv -1 \pmod{n}$. We must verify that these elements may not be quadratic residues. In fact, if $g^{2k+1} \equiv y^2 \pmod{n}$ for $y \in E$, we would have that $(g^{2k+1})^{\frac{n-1}{2}} \equiv (y^2)^{\frac{n-1}{2}} \equiv y^{n-1} \equiv 1 \pmod{n}$. This is a contradiction, since $(g^{2k+1})^{\frac{n-1}{2}} \equiv -1 \pmod{n}$. □

A deterministic algorithm for primality testing. The algorithm that we described for primality testing is a *probabilistic algorithm*. In fact, it lets us prove that some numbers are not prime, but it does not allow us to be certain (in reasonable time) that a number is in fact prime: we would have to complete the test with roughly half of the integers less than n.

In 2003, Agrawal, Kayal, and Saxena announced a new *deterministic algorithm*, called the AKS algorithm, which allows for primality testing in reasonable time. The full article appeared in 2004 [1]. This algorithm remains much slower than the best probabilistic algorithms, but it is of much theoretical interest, since it answers a question originally posed by Gauss more than 200 years ago. It is difficult to summarize succinctly, but a detailed study of the algorithm would make an excellent term project, provided that one had sufficient background in number theory.

7.5 Breaking RSA: Shor's Algorithm for Factoring Large Integers

As we already mentioned, there has been and continues to be a great deal of research toward finding better algorithms for factoring large integers. For computer scientists a good algorithm is one that functions in "polynomial" time (a concept that we will make clear a little later). The 1997 introduction of Shor's polynomial-time factorization algorithm had many repercussions. However, this algorithm requires a quantum computer, and even if they aren't quite the stuff of science fiction anymore, neither are they the stuff of reality.

Before discussing this algorithm we will first discuss algorithmic complexity.

The complexity of an algorithm applied to an m-digit integer n. Suppose that $n \approx 10^m$. The number m is the "size" of the input to the algorithm. The complexity of the algorithm is the number of operations that must be performed by a computer in order to execute the algorithm. This number of operations is dependent on the size of the input.

If the number of operations required is of order Cm^r, where r is a constant, then we say that the algorithm operates in *polynomial-time*.

The classical algorithm for integer factorization requires *exponential time*. In fact, it requires testing each of the numbers $1, 2, \ldots, d \leq \sqrt{n}$ to see whether they are divisors of n. The number of tests required is therefore of order $10^{m/2}$. As m grows, this number of operations quickly becomes too large for a computer.

The probabilistic primality testing algorithm described earlier operates in expected polynomial-time, and the more recent AKS algorithm operates in deterministic polynomial-time. This is why it is significantly easier to choose large primes than it is to factor large integers.

We will start by convincing ourselves that the simple improvements we can bring to the factorization algorithm are not sufficient to allow easy factorization. We consider a 200-digit integer $n \approx 10^{200}$. The classic algorithm requires checking for all potential divisors $d \leq \sqrt{n}$, which means we need to perform roughly 10^{100} trial divisions. Let us try to reduce this complexity by simple tricks:

- If we limit ourselves to dividing only by odd numbers, we have $m_1 \approx \frac{10^{100}}{2}$ tests to perform.
- If we limit ourselves to large potential divisors (numbers with 100 digits), then we have $m_2 = \frac{9}{10} m_1$ tests to perform (exercise!).
- If we use a billion computers working in parallel, each computer must perform $m_3 = 10^{-9} m_2$ tests.
- If each of the billion computers is a supercomputer containing 5000 processors that could perform 5000 operations in parallel (roughly equivalent to the largest individual supercomputer at the end of 2004), we limit the effective number of operations to $m_4 = \frac{m_3}{5000}$.

- Even with these reductions we have $m_5 \geq 10^{86}$ tests to perform. This is still too many!
- Suppose, through other insights and simplifications, that we reduce this to a reasonable amount of computation and are able to factor 200-digit numbers. Then we need only choose public keys with a few dozen more digits in order to make the integer effectively impossible to factor again.

It is easy to see that if we want to be able to factor large numbers, then we require better algorithms. Briefly mentioned earlier, there exist much better algorithms, although they remain at least subexponential in complexity. Shor's 1997 algorithm for integer factorization runs in exponential time on a classic computer. However, it runs in polynomial-time on a quantum computer. It is a probabilistic algorithm: if n is not prime, the algorithm has a very high chance of finding a divisor d of n in polynomial-time. We will content ourselves with only providing a sketch of the algorithm, without all of the details.

The idea behind Shor's algorithm ([6], [9]). The algorithm attempts to find a divisor d of n. Once we have decomposed n such that $n = dm$, we may test whether d and m are prime. If one or both of these factors are not prime, we again apply Shor's algorithm to it until we are finally left with a product of prime factors. As we proceed, the calculations get easier and easier because d and m are much smaller than n.

Step 1: Find an integer r such that $n \mid r^2 - 1$, but such that neither $r - 1$ nor $r + 1$ is divisible by n.

Finding such a value r allows us to find a proper divisor of n. In fact, $r^2 - 1 \equiv 0 \pmod{n}$ implies that $(r-1)(r+1) = mn$ for some integer m. If p is a prime factor of n, then by necessity $p \mid r-1$ or $p \mid r+1$. If $p \mid r-1$ then $(r-1, n) = d > 1$. Since n does not divide $r-1$, then d is a proper divisor of n. Similarly if $p \mid r+1$.
Example: If $n = 65$ and $r = 14$, then $r^2 = 196 = 3 \times 65 + 1 \equiv 1 \pmod{65}$, and $r - 1 = 13$ is a divisor of 65.
On the other hand, if we choose $s = 64 \equiv -1 \pmod{65}$, then $s^2 \equiv -1^2 = 1 \pmod{65}$. We see that $s + 1 = 65$ is divisible by 65, and thus s cannot help us find a proper divisor of n.
Step 2: How to actually find r?

We choose a randomly from the set $E = \{1, \ldots, n-1\}$.

- Calculate (a, n).
- If $(a, n) = d > 1$, then we have found a divisor of n.
- If $(a, n) = 1$, then we calculate the powers of a $(a, a^2, a^3, \ldots)$ reduced modulo n such that $a_k \equiv a^k \pmod{n}$ with $a_k \in E$.
- Since E is finite, there exist l and k such that $a_k = a_l$. Suppose $k > l$. Then $a_{k-l} \equiv a^{k-l} \equiv 1 \pmod{n}$.
- Thus there exists a minimal value $s \leq n$ such that $a^s \equiv 1 \pmod{n}$. This s is the order of a modulo n.

- If s is odd, then our search proves fruitless and we restart with another a chosen randomly from E.
- If s is even, then let $s = 2m$ and $r \equiv a^m \pmod{n} \in E$. Then $r^2 \equiv a^{2m} = a^s \equiv 1 \pmod{n}$.
- If neither $r-1$ nor $r+1$ is divisible by n, then we are finished by Step 1; otherwise, we repeat with another value for a.

It is possible to show that there are many values $a \in E$ with even order that will do the trick; thus the algorithm has good expected performance.

Complexity of the algorithm. the only part of the algorithm that cannot be completed in polynomial-time on a classic computer is the determination of the order of a. It is at this crucial step where a quantum computer can be used.

Calculating the order of a modulo n using a quantum computer. Quantum physics being a rather complicated field, we will provide only a general idea of how the computation works. We begin by writing the number a in base 2. If n may be written with m binary digits (bits), then $n < 2^m$, yielding $a < 2^m$. We write an integer $k \in E = \{1, \ldots, 2^m - 1\}$ in base 2 as

$$k = [j_{m-1} j_{m-2} \cdots j_1 j_0] = j_{m-1} 2^{m-1} + j_{m-2} 2^{m-2} + \cdots + j_1 2^1 + j_0 2^0.$$

So, giving ourselves k is giving ourselves m binary bits $j_{m-1}, \ldots, j_0$. In order to calculate the order of a we wish to calculate a^k simultaneously for each value of $k \in E$. Trying each value of k can be done by trying each of the two values $\{0, 1\}$ for each j_i, amounting to 2^m possibilities altogether. It is at this point where quantum computers come to the rescue. We replace each of the m bits in the calculation with quantum bits (qubits).

Quantum bits. A quantum bit has the ability to be in a *superposition* of states. It is in the state $|0\rangle$ with probability $|\alpha|^2$ and the state $|1\rangle$ with probability $|\beta|^2$, where $|\alpha|^2 + |\beta|^2 = 1$ and both α and β are complex numbers. In quantum mechanics we say that the qubit is in state $\alpha|0\rangle + \beta|1\rangle$. To give an analogy, think of a coin: it has probability 1/2 of coming up heads, and probability 1/2 of coming up tails. Before the coin toss, our coin is thus in a superposed state. However, when we toss it we will observe a single final state: heads or tails. It is the same thing with quantum bits: when we measure them we obtain 0 with probability $|\alpha^2|$ and 1 with probability $|\beta^2|$.

The "super" parallelism of a quantum computer. If we place all m bits $j_{m-1}, \ldots, j_0$ in superposed states simultaneously, then by calculating $a^{|k\rangle} \pmod{n}$ (where $|k\rangle$ is a superposition of all $k \in E$) we effectively compute a^k for all values of $k \in E$ simultaneously! Since quantum calculations are linear and reversible, we can see $a^{|k\rangle} \pmod{n}$ as a superposition of all of the values $a_k \equiv a^k \pmod{n}$. All of the necessary information can be found in this superposed state, but we cannot access it without first measuring it. The difficulty lies in accessing the results. This lies purely in the domain of quantum physics and we avoid discussing details here.

Remark: We have already shown that it is not difficult for a computer to calculate $a^k \pmod n$. In fact, if $k = j_{m-1}2^{m-1} + \cdots + j_0 2^0$ then $a^k = \prod_{\{i|j_i=1\}} a^{2^i}$. Thus it suffices to calculate the $a^{2^i} \pmod n$ for $i = 0, \ldots, m-1$. This calculation is done by jumping from one to the next:

- $a = a_0$,
- $a^2 \equiv a_1 \pmod n$ with $a_1 \in E$;
- $a^4 \equiv (a_1)^2 \equiv a_2 \pmod n$ with $a_2 \in E$;
- $\vdots$
- $a^{2^{m-1}} \equiv (a_{m-2})^2 \equiv a_{m-1} \pmod n$ with $a_{m-1} \in E$.

Finally, $a_k \equiv \prod_{\{i|j_i=1\}} a_i \pmod n$.

How far along are quantum computers? Quantum computers are not yet a serious threat to the RSA cryptosystem. For the moment, real-world functioning quantum computers are able to factor only very small integers: in 2002, the number 15 was factored with the help of a 7-qubit quantum computer by Isaac Chuang and his team of researchers.

7.6 Exercises

1. Let $a, b, c, d, x, y \in \mathbb{Z}$. Show that

$$a \equiv c \pmod n \quad \text{and} \quad b \equiv d \pmod n \implies a + b \equiv c + d \pmod n,$$
$$a \equiv c \pmod n \quad \text{and} \quad b \equiv d \pmod n \implies ax + by \equiv cx + dy \pmod n.$$

2. The Euler function $\phi : \mathbb{N} \to \mathbb{N}$ is defined as follows: if $m \in \mathbb{N}$ then $\phi(m)$ is the number of integers from the set $\{1, 2, \ldots, m-1\}$ that are relatively prime to m.
(a) Show that if $m = p_1 \cdots p_k$ with $p_1, \ldots, p_k$ distinct primes, then $\phi(n) = (p_1 - 1) \cdots (p_k - 1)$.
(b) Let p be a prime. Show that

$$\phi(p^n) = p^n - p^{n-1}.$$

3. Public key cryptography uses an integer $n = pq$, where p and q are two distinct prime integers. Would the same techniques work with a number of the form $n = p_1 p_2 p_3$, where p_1, p_2, and p_3 are three distinct primes?

4. Let p be a prime number.

(a) Calculate $\phi(p^2)$, where ϕ is the Euler function.
(b) Public key cryptography uses an integer $n = pq$, where p and q are two distinct prime integers. Would the same techniques work with a number of the form $n = p^2$? If yes, describe the steps of the algorithm. Why wouldn't we use such a system?

5. An article in the French science magazine *La Recherche* gave the following example of public key cryptography. We choose two prime integers p and q such that $p, q \equiv 2 \pmod 3$, and let $n = pq$. Alice wants to send a message to Bob. Her message is a number x in the set $\{1, \ldots, n-1\}$, where $(x, n) = 1$ (this last important detail didn't actually appear in the article!). To send her message, Alice calculates x^3, and takes the remainder $y \in \{1, \ldots, n-1\}$ of this number modulo n. Bob decodes the message with

$$d = \frac{2(p-1)(q-1)+1}{3}.$$

He calculates y^d and takes the remainder $z \in \{1, \ldots, n-1\}$ modulo n.
(a) Verify that d is in fact an integer.
(b) Explain why y and z cannot be zero, that is, why we will have $y, z \in \{1, \ldots, n\}$.
(c) Show that $z = x$, and therefore that Bob successfully decodes the message.

6. You want to explain to a friend how the RSA code functions. Here is how you do it: you choose a prime integer p, with $p \equiv 2 \pmod 7$, and a prime integer q, with $q \equiv 3 \pmod 7$. Then you calculate $n = pq$. You explain how Alice can send a message to Bob. Her message is an integer number m in $\{1, \ldots, n-1\}$ with $(m, n) = 1$. To send her message, Alice computes m^7 and divides this number by n. Let $a \in \{1, \ldots, n-1\}$ be the remainder of the division of m^7 by n (which means $m^7 \equiv a \pmod n$). You explain that Bob decodes with the decryption key

$$d = \frac{3(p-1)(q-1)+1}{7}.$$

He computes a^d and then the remainder m_1 of the division of a^d by n (which means $a^d \equiv m_1 \pmod n$), where $m_1 \in \{1, \ldots, n-1\}$. You claim that m_1 is the message sent by Alice.
(a) Verify that d is an integer.
(b) Explain why a and m_1 cannot vanish, yielding $a, m_1 \in \{1, \ldots, n-1\}$.
(c) Show that $m_1 = m$, which means that Bob will decode Alice's message.

7. We present a simple cryptography system. The space symbol □ is represented by the number 0. The letters A, ..., Z are represented by the numbers $1, \ldots, 26$, while 27 corresponds to the period and 28 to the comma. The mapping is represented in the following table:

Symbol	□	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Number	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Symbol	O	P	Q	R	S	T	U	V	W	X	Y	Z	.	,
Number	15	16	17	18	19	20	21	22	23	24	25	26	27	28

Here is how we encode a word:

- we replace symbols by their associated numbers;
- we multiply each number by 2;
- we reduce each result modulo 29;
- we map each number back to its corresponding symbol, yielding the encoded word.

For example, to encode the word "THE" we first map it to the sequence $20, 8, 5$. We multiply these numbers by 2 and obtain $40, 16, 10$, leaving $11, 16, 10$ after reduction modulo 29. Replacing the integers by the associated symbols yields the final encoding of "KPJ".

(a) Encode the word "YES".
(b) Explain why the encoding is reversible, and how we go about decoding it.
(c) Decode the word "XMVJ".

8. Here is a refinement of the cryptography system presented in Exercise 7. We use the same 29 symbols, but encrypt a word in the following manner:

- we replace symbols by their associated numbers;
- we multiply each integer by 3 and add 4 to the result;
- we reduce each result modulo 29;
- we map each number back to its corresponding symbol, yielding the encoded word.

(a) Encode the word "MATH".
(b) Explain why the encoding is reversible, and how we go about decoding it.
(c) Decode the word "MTPS".

9. "Casting out nines" is an old trick that can be used to verify the result of the multiplication of two integers. It was widely taught before calculators were common. We multiply two numbers m and n. Let $N = mn$, and we wish to verify the result of our arithmetic. For this, we use the decimal representation of the number. For $M \in \mathbb{N}$, we write $M = a_p \cdots a_0$, where $a_i \in \{0, 1, \ldots, 9\}$, which is equivalent to the summation

$$M = \sum_{i=0}^{p} a_i 10^i.$$

We calculate the value $F(M) \in \{0, 1, \ldots, 8\}$, where $F(M)$ is the remainder of the value

$$\sum_{i=0}^{p} a_i = a_0 + \cdots + a_p$$

modulo 9. Here is an example. Let $M = 2857$. Then $2 + 8 + 5 + 7 = 22 \equiv 4 \pmod 9$, yielding $F(2857) = 4$.

To check the result of our multiplication we calculate $F(N)$, $F(m)$, and $F(n)$, and finally the product $r = F(m)F(n)$. As a last step, we calculate $F(r)$.
(a) Assuming that there were no errors in the calculations, show that we must have

$$F(N) = F(r).$$

If these two "check digits" do not match, we can conclude that an error was made in the original multiplication (assuming that no errors were made in computing the various values of $F(\cdot)$!).
(b) Walk through a simple example.
(c) What can we say when $F(N) = F(r)$? Can we conclude that there were no errors in the calculation of N?

10. Construct a public key cryptosystem with $n = pq$, where n is 60 digits. For this, you should choose distinct primes p and q of 30 digits each.
(a) Using a computer algebra system, generate 30-digit numbers and test whether they are prime.
(b) Verify whether the generated numbers are prime using Jacobi's test with numbers $a_1, \ldots, a_k$ having fewer than 30 digits. Validate the test by running it with a known prime number and a known composite number. The instant the test returns a negative result we can conclude that the number is composite. If the test is positive, continue with the next a_i to obtain a higher degree of certainty that the number is prime.

11. Consider an RSA cryptosystem with $n = 23 \times 37 = 851$ and the encryption key $e = 47$. Find the decryption key d that satisfies

$$e \cdot d \equiv 1 \pmod{\phi(n)}.$$

12. We give ourselves an integer M with N digits. Let $a_{N-1} \cdots a_1 a_0$ be the decimal representation of this number such that

$$M = a_{N-1}10^{N-1} + a_{N-2}10^{N-2} + \cdots + a_1 10 + a_0.$$

(a) Show that M is divisible by 11 if and only if

$$a_0 - a_1 + a_2 - a_3 + \cdots + (-1)^{N-2}a_{N-2} + (-1)^{N-1}a_{N-1} \equiv 0 \pmod{11}.$$

(Hint: consider $10^i \pmod{11}$.) Remark: this simple test can be used to avoid multiples of 11 when searching for prime numbers.
(b) Show that M is divisible by 101 if and only if

$$-(a_0 + 10a_1) + (a_2 + 10a_3) - (a_4 + 10a_5) + (a_6 + 10a_7) + \cdots \equiv 0 \pmod{101}.$$

13. Show that n is prime if and only if

$$(x+1)^n \equiv x^n + 1 \pmod{n}.$$

Remark: this exercise highlights the central idea underlying the AKS algorithm [1].

14. We consider the set $E_n = \{0, 1, \ldots, n-1\}$ for $n \in \mathbb{N}$. Let p and q be such that $(p, q) = 1$. We define the function $F : E_{pq} \to E_p \times E_q$ by $F(n) = (n_1, n_2)$, where $n \equiv n_1 \pmod{p}$ and $n \equiv n_2 \pmod{q}$. Show that F is a bijection. (This result is the modern form of the "Chinese remainder theorem.")

15. Prove Wilson's theorem: n is prime if and only if n divides $(n-1)! + 1$. One of the directions is more difficult than the other. If n is prime we must use the fact that $\{1, \ldots, n-1\}$ is a group under multiplication to show that $n \mid (n-1)! + 1$.
Remark: This theorem provides yet another test for deciding whether n is prime. However, this test is not really of practical interest, since when n is large, the calculation of $(n-1)!$ is out of reach for even the most powerful computers.

16. Show that the fractions $\frac{b^2-1}{8}$ and $\frac{(a-1)(b-1)}{4}$ in equation (7.5) for $J(a, b)$ are always integers when a and b are odd integers.

17. Let $E_n = \{1, \ldots, n-1\}$.
(a) Let $n = 13$. By explicitly calculating $J(a, n)$ and $a^{\frac{n-1}{2}} \pmod{n}$, show that (7.4) holds for all $a \in E_n$.
(b) Now let $n = 15$. How many $a \in E_n$ do not satisfy the test?

18. We wish to use Shor's algorithm to find a divisor of 91. To this end we choose $a = 15$.
(a) Compute the order of a: find the smallest integer exponent s such that $a^s \equiv 1 \pmod{91}$. Verify that s is even.
(b) Compute $r = a^{\frac{s}{2}} \pmod{91}$, and show that neither $r - 1$ nor $r + 1$ is divisible by 91.
(c) Complete Shor's algorithm by using r to find a divisor of 91.

19. We wish to use Shor's algorithm to factor 30. To this end we choose a randomly from $\{1, 2, \ldots, 29\}$ and proceed. List the choices for a that permit the discovery of a proper divisor of 30, and in each case show which method was used.

References

[1] M. Agrawal, N. Kayal, and N. Saxena. PRIMES is in p. *Annals of Mathematics*, 160:781–793, 2004. (See also [2].)

[2] F. Bornemann. PRIMES is in p: A breakthrough for everyman. *Notices of the American Mathematical Society*, 50(5):545–552, 2003.

[3] J. Buchmann. *Introduction to Cryptography*. Springer, New York, 2001.

[4] J.-P. Delahaye. La cryptographie RSA 20 ans après. *Pour la Science*, 2000.

[5] E. Knill, R. Laflamme, H. Barnum, D. Dalvit, J. Dziarmaga, J. Gubernatis, L. Gurvits, G. Ortiz, L. Viola, and W.H. Zurek. From factoring to phase estimation: A discussion of Shor's algorithm. *Los Alamos Science*, 27:38–45, 2002.

[6] E. Knill, R. Laflamme, H. Barnum, D. Dalvit, J. Dziarmaga, J. Gubernatis, L. Gurvits, G. Ortiz, L. Viola, and W.H. Zurek. Quantum information processing: A hands-on primer. *Los Alamos Science*, 27:2–37, 2002.

[7] C. Pomerance. A tale of two sieves. *Notices of the American Mathematical Society*, 43(12):1473–1485, 1996.

[8] R.L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.

[9] P.W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal of Computation*, 26:1484–1509, 1997.

[10] A. Weil. *Number Theory for Beginners*. Springer, New York, 1979.

8

Random-Number Generators

This chapter may be approached in two ways: you can work through one or two hours as though it were a Science Flash[1]*, the only requisite being a basic comfort with arithmetic modulo p; or you can dive into the material in a little more depth. In the latter case it is preferable that you be familiar with finite fields and congruences modulo* 2 *(for instance because you have already worked through Chapter 6 or 7). Sections making reference to these chapters are clearly marked. In this chapter we will thoroughly cover the subject of linear shift registers. Although also discussed in Section 1.4 of Chapter 1, the two discussions are independent. Most exercises are very elementary. Some exercises require a familiarity with probabilities, but these may be safely ignored if you do not have the background knowledge.*

8.1 Introduction

On April 10, 1994, a gambler was arrested by police at the Montreal Casino. His crime? He had just beaten the laws of probability by winning three consecutive jackpots in the game of keno, with winnings totaling more than a half million dollars.[2] He was suspected of having broken gambling laws that prohibit collusion with casino employees, tampering with gambling equipment, etc. An investigation was launched, and after a few weeks, the player was released and his winnings, interest included, returned to him. The Montreal Casino had just learned an expensive lesson about random-number generators.

[1] A *Science Flash* is a small subject to be treated in one or two course hours, as presented in Chapter 15.

[2] In keno, the player has to choose a dozen numbers from the set $\{1, 2, \ldots, 80\}$. The casino then draws at random 20 balls from a set of 80 balls numbered $1, \ldots, 80$. This drawing can be done electronically, as is often the case in most casinos. The winnings of the player depend on the size of his bet and the number of matches between his chosen numbers and the numbers of the balls drawn by the casino.

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_8,

There are very few mechanical devices to be found in a modern casino. In fact, the roulette wheel may very well be the last of them. Most other games have been replaced by computers that simulate randomness. Each of these computers is programmed to generate numbers that *appear* to be random to the user, but which are in fact computed in a completely deterministic manner. These algorithms, random-number generators, play an important role in many of these machines. Video games played on computers or game consoles depend heavily on these algorithms. If these games were to behave the same every time the machine was restarted, players would quickly grow tired of them.

Random-number generators are as important in everyday life as they are in science. Computer simulations of stock exchanges and of virus propagation (both human and computer!) and the selection of those (unlucky) taxpayers whose returns the government will audit all use random-number generators routinely. In science, it is sometimes difficult to model a system whose behavior is known only in the probabilistic sense. An example of such a system is the impartial web surfer described in Section 9.2. The existence of random-number generators is also assumed in the discussion of probabilistic algorithms for cryptography in Chapter 7. Random-number generators are used explicitly (!) in the construction of fractal images by iterated function systems (see Chapter 11) and in the discussion of the GPS satellite signal (see Chapter 1).

These generators find many applications in modern society, and it is therefore not surprising that they are the focus of much research into finding "better" random-number generators. What exactly do we mean by "better"? This depends on the context. For random-number generators being used in casinos, we want to prevent players from being able to take advantage of the games by guessing how the underlying algorithms work. We also require that the generated numbers follow certain laws of probability, chosen a priori, so that the casino cannot be accused of fraud and so that players are offered a fair gaming experience.

To begin with, we introduce a "mechanical" random-number generator. Even though such an approach is impractical on a large scale, it captures the basic challenges that all random-number-generator algorithms face. We can imagine playing a game of heads-or-tails, many times in a row. By noting a 0 every time the coin comes up heads and a 1 when it comes up tails, we generate a random sequence of zeros and ones. That is, we generate a sequence that seems to obey no visible rules. If several people were to repeat this experiment, each would in general generate a sequence that has no resemblance to any of the others.

Now suppose that we wish to generate a random sequence of numbers from the set $S = \{0, \ldots, 31\}$. Given that $32 = 2^5$, each number $n \in S$ may be written in base 2 as

$$n = a_0 + 2a_1 + 2^2 a_2 + 2^3 a_3 + 2^4 a_4 = \sum_{i=0}^{4} a_i 2^i,$$

with $a_i \in \{0, 1\}$. We may also represent the number n by the 5-tuple $(a_0, \ldots, a_4)$. For example, the 5-tuple $(0, 1, 1, 0, 1)$ represents $2 + 4 + 16 = 22$.

If we generate a sequence of zeros and ones by tossing a coin, we can then regroup these numbers as 5-tuples of bits and transform them into numbers in S. For example, suppose we had obtained the binary sequence

$$10000\ 00101\ 11110\ 01001\ 01001\ 11011. \tag{8.1}$$

Converting the 5-tuples yields

$$\underbrace{10000}_{1}\ \underbrace{00101}_{20}\ \underbrace{11110}_{15}\ \underbrace{01001}_{18}\ \underbrace{01001}_{18}\ \underbrace{11011}_{27},$$

or simply

$$1, 20, 15, 18, 18, 27,$$

when represented as numbers from S.

If instead of 31 we had chosen $N = 2^r - 1$ and $S = \{0, \ldots, N\}$, we could still have followed the same approach, transforming a binary sequence into a sequence of numbers from S.

However, when r is large or when we require a particularly long sequence of random-numbers, the method of manually flipping coins quickly becomes cumbersome. The ideal solution is to program a computer to generate a sequence of ones and zeros in such a manner that the results appear as random as those obtained from actually tossing a coin. Such a program is a random-number generator. In reality, since such an algorithm is by its very nature deterministic, it can only generate a sequence of numbers that *appear* to be random. It is for this reason that experts refer to such algorithms as *pseudorandom-number generators*.

Pseudo-random-number generators are used quite often in all sorts of computer simulations. In many cases we simply want to generate random real numbers in the interval $[0, 1]$. In this case, it helps to write real numbers in binary representation. To differentiate between binary and decimal representations we will place a subscript of 2 after all numbers in binary representation. Thus $(0.a_1a_2 \ldots a_n)_2$ represents

$$(0.a_1a_2 \ldots a_n)_2 = a_1 2^{-1} + a_2 2^{-2} + \cdots + a_n 2^{-n} = \sum_{i=1}^{n} \frac{a_i}{2^i}.$$

As a general rule, most real numbers require infinite binary representations. However, given that computers are constrained to finite computations, we limit ourselves to finite representations with a desired degree of precision. Thus, if we look at the sequence of (8.1), we can interpret it as generating a sequence of real numbers in $[0, 1]$ as follows:

$$\begin{cases} 0.10000_2 = 2^{-1} = \frac{1}{2} = 0.5, \\ 0.00101_2 = 2^{-3} + 2^{-5} = 0.15625, \\ 0.11110_2 = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = 0.9375, \\ 0.01001_2 = 2^{-2} + 2^{-5} = 0.28125, \\ 0.01001_2 = 2^{-2} + 2^{-5} = 0.28125, \\ 0.11011_2 = 2^{-1} + 2^{-2} + 2^{-4} + 2^{-5} = 0.84375, \end{cases}$$

the last number on the right being the decimal representation of the number.

What makes a good random number generator? What criteria must it satisfy? When we toss a coin repeatedly the result of each toss is completely independent of those before it, and each of the two outcomes always has probability $\frac{1}{2}$. This implies that if we toss a coin a large number of times, roughly half of the time it should come up heads (noted as 0) and half of the time tails (noted as 1)—a result of the law of large numbers in action. If instead of tossing a coin once we were to toss it twice, each pair of tosses would have one of four possible results:

$$00 \quad 01 \quad 10 \quad 11.$$

If we were to repeat this many times, we would expect each outcome to occur roughly one-quarter of the time. Similarly, if we were to toss a coin three times, we would have $2^3 = 8$ equiprobable outcomes:

$$000 \quad 001 \quad 010 \quad 011 \quad 100 \quad 101 \quad 110 \quad 111.$$

Thus we desire that a random-number generator satisfy these same properties. To ensure that our pseudorandom-number generators do in fact have these properties, we submit them to a battery of statistical tests.

All pseudorandom-number generators are algorithms that generate a deterministic periodic sequence of numbers from a finite set of starting conditions.

Definition 8.1 *A sequence* $\{a_n\}_{n\geq 0}$ *is* periodic *if there exists an integer* $M > 0$ *such that for all* $n \in \mathbb{N}$, $a_n = a_{n+M}$. *The minimum* $N > 0$ *for which this property holds is called the* period *of the sequence. When we want to emphasize this particular aspect of the period, we may refer to* N *as the* minimal period.

Lemma 8.2 *Let* $\{a_n\}_{n\in\mathbb{N}\cup\{0\}}$ *be a periodic sequence with minimal period* N *and let* $M \in \mathbb{N}$ *be such that for all* $n \in \mathbb{N}$, $a_n = a_{n+M}$. *Then it follows that* N *divides* M.

Proof. Divide M by N. Then there must exist integers q and r such that $M = qN + r$ where $0 \leq r < N$. We want to show that for all n, we must have $a_n = a_{n+r}$.

In fact, we easily see that

$$a_n = a_{n+M} = a_{n+qN+r} = a_{n+r}.$$

Since N is the least integer such that $a_n = a_{n+N}$, it must be that $r = 0$. Thus N divides M. □.

Example 8.3 *A linear congruential generator is a very commonly used type of random-number generator. It generates a sequence over the set* $E = \{1, \ldots, p-1\}$ *using the rule*

$$x_n = ax_{n-1} \pmod{p},$$

where p is prime and a is a primitive root of $\mathbb{F}_p$. That is, a is an element of E such that

$$\begin{cases} a^k \not\equiv 1 \pmod{p}, & k < p-1, \\ a^{p-1} \equiv 1 \pmod{p}. \end{cases}$$

Recall that $\mathbb{F}_p$ (also called $\mathbb{Z}_p$ in Chapter 7) is the set of integers $\{0, \ldots, p-1\}$ with addition and multiplication modulo p. Defined in this way, $\mathbb{F}_p$ is a field when p is prime; this implies (see Definition 6.1) that addition and multiplication are both commutative and associative, each operation has an identity element, multiplication is distributive over addition, all elements have additive inverses, and finally all nonzero elements have multiplicative inverses. These properties are explored in Exercise 24 of Chapter 6, but we will use them without proof in the following discussion.

Take as a simple example the case $p = 7$. We see that 2 is not a primitive root since $2^3 = 8 \equiv 1 \pmod{7}$. However, we observe that 3 is a primitive root, since

$$\begin{cases} 3^2 \equiv 2 \pmod{7}, \\ 3^3 \equiv 6 \pmod{7}, \\ 3^4 \equiv 18 \equiv 4 \pmod{7}, \\ 3^5 \equiv 12 \equiv 5 \pmod{7}, \\ 3^6 \equiv 15 \equiv 1 \pmod{7}. \end{cases}$$

The proof that there always exists a primitive root $a \in \mathbb{F}_p$ can be found in Theorem 7.22 of Chapter 7. (Again, you may take this result for granted and continue with the current discussion.)

This generator will create a periodic sequence whose period is exactly $p-1$. Linear congruential generators are commonly used in many pieces of software, where the values $p = 2^{31} - 1$ and $a = 16{,}807$ are often taken. However, experts in the subject do not consider these generators to be very good, since they fail some basic statistical tests (see Exercises 2 and 4.)

Other criteria often come into play, notably those of economy. In many cases we are interested in minimizing the time of computation and memory usage. In these cases we may be content to use a random-number generator that is weaker from a statistical point of view, but sufficient for the task at hand.

8.2 Linear Shift Registers

Linear shift registers (also discussed in Chapter 1) are quite good random-number generators. They can be visualized as an array (or register) of r boxes containing the entries $a_{n-1}, \ldots, a_{n-r}$, where each a_i is in $\{0, 1\}$. Each one of these boxes is multiplied by a value $q_i \in \{0, 1\}$, with the results being summed modulo 2. The q_i are fixed and characterize the particular generator (Figure 8.1). We generate a pseudorandom-number sequence in the following manner:

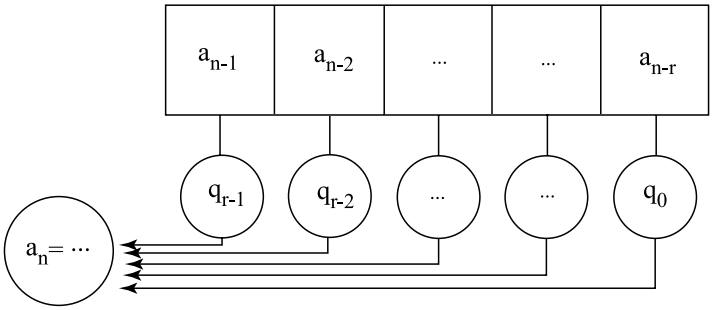


Fig. 8.1. A linear shift register.

- Choose initial values $a_0, \ldots, a_{r-1} \in \{0, 1\}$, not all of which are zero.
- Given the values $a_{n-r}, \ldots, a_{n-1}$, calculate the next value in the sequence, a_n, as

$$a_n \equiv a_{n-r}q_0 + a_{n-r+1}q_1 + \cdots + a_{n-1}q_{r-1} \equiv \sum_{i=0}^{r-1} a_{n-r+i}q_i \pmod 2. \tag{8.2}$$

- Shift each entry to the right, dropping the entry a_{n-r} in doing so. The newly generated a_n now occupies the leftmost box of the register.
- Repeat.

In Section 1.4 of Chapter 1 we showed that if we carefully choose the q_i and the initial conditions $a_0, \ldots, a_{r-1}$, then we will generate a sequence with a period of $2^r - 1$. We will revisit this topic in more detail, discussing exactly how to choose the q_i.

Example 8.4 *We take a linear shift register of length* 4 *and* $(q_0, q_1, q_2, q_3) = (1, 1, 0, 0)$. *Consider also the starting state* $(a_0, a_1, a_2, a_3) = (0, 0, 0, 1)$. *Following the register through its operation, we find that it generates a sequence of period* 15*:*

$$\underbrace{000100110101111}_{15}0001\ldots.$$

In this cycle of 15 *entries, inspection shows that* 0 *was generated seven times and* 1 *was generated eight times. Now consider the* 15 *subsequences of length* 2*:* 00 *occurs three times*

$$\begin{cases} \overbrace{00}0100110101111 \\ 0\overbrace{00}100110101111 \\ 0001\overbrace{00}110101111 \end{cases}$$

while the three other possible sequences of length 2*,* 01*,* 10 *and* 11*, occur exactly four times each. In the case of* 10*, the fourth occurrence straddles two cycles:*

$$\begin{cases} 000\overbrace{10}0110101111 \\ 0001001\overbrace{10}101111 \\ 000100110\overbrace{10}1111 \\ 00010011010111\overbrace{1\ 0}001\ldots \end{cases}$$

We leave it to the reader to verify that each subsequence of length 3 *occurs twice, except the subsequence* 000*, which occurs exactly once. Similarly, all subsequences of length* 4 *will be seen to occur exactly once, except* 0000*, which never occurs. Could we continue with counting subsequences of length* 5 *and longer? The answer is no, since our register is only of length* 4*, which immediately implies that the fifth and subsequent symbols following a given sequence of four are predetermined. We can also easily explain why runs of zeros occur less often: we cannot permit the register to generate a subsequence of the form* 0000*; otherwise, the generating rules will force all following symbols to be zero as well.*

This example shows that linear shift registers generate all subsequences in roughly equal proportion, as long as we do not consider subsequences longer than the register itself (limiting ourselves to 4 *in this example). This is no coincidence, as we will show later in Theorem 8.12.*

If we want our generated sequence to have good statistical properties with respect to longer subsequences, we need only choose a sufficiently large length r for our register.

We will recast the operation of the linear shift register into another form that is more suitable for analysis and can be generalized. At a given moment of time, which we will call the moment j, we consider reading the entries of the register $a_j, \ldots, a_{j+r-1}$. We rewrite these entries as $x_{j,0}, \ldots, x_{j,r-1}$, where $x_{j,i} = a_{j+i}$. The advantage of this notation is that the index j indicates the moment of time, while the index i indicates the entry of the register in box i. Let

$$\mathbf{x}_j = \begin{pmatrix} x_{j,0} \\ \vdots \\ x_{j,r-1} \end{pmatrix}$$

be the column vector of entries at time j. Let A be the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \\ q_0 & q_1 & q_2 & q_3 & \cdots & q_{r-1} \end{pmatrix}. \tag{8.3}$$

With this notation, the vector representing the state of the register at time $j+1$ is given by

$$\mathbf{x}_{j+1} = A\mathbf{x}_j, \tag{8.4}$$

where all operations are modulo 2. (Exercise: verify that this is in fact true!)

Before we further abstract the problem, note the utility of this alternative representation. Suppose we wish to pass directly from $\mathbf{x}_j$ to $\mathbf{x}_{j+k}$ without calculating the intermediate steps. We see that $\mathbf{x}_{j+k} = A^k\mathbf{x}_j$; thus if we calculate A^k, we can pass directly to $\mathbf{x}_{j+k}$ from $\mathbf{x}_j$. The ability to take arbitrarily large steps through the sequence with a reasonable amount of computation is a desirable property for random-number generators.

How do we calculate A^k if k is large? In general, if we take a matrix A over the real numbers, the coefficients of A^k can grow quite large in absolute value. However, in this case we are working over the finite field $\mathbb{F}_2$, where all operations are taken modulo 2. Thus the entries of A^k will also be entries in $\mathbb{F}_2$, and we need not worry about coefficient swell. That still leaves us with the problem of efficiently calculating A^k. Consider writing k in base two as

$$k = b_0 + b_1 2 + b_2 2^2 + \cdots + b_s 2^s.$$

We define $A_0 = A$ and calculate

$$\begin{aligned} A_1 &= A^2, \\ A_2 &= A^4 = A_1^2, \\ &\vdots \\ A_s &= A^{2^s} = A_{s-1}^2. \end{aligned}$$

This lets us calculate the final matrix A^k as

$$A^k = \prod_{\{i|b_i=1\}} A_i.$$

Observe that each A_i is calculated as a product of two matrices thus requiring s matrix products to calculate them. Note also that A^k is calculated as the product of at most $s+1$ of these matrices. Thus the final matrix may be calculated by taking at most $2s = 2\log_2 k$ matrix products.

This notation also makes it clear that we could create other random number generators by generalizing the form of the transition matrix A.

8.3 $\mathbb{F}_p$-Linear Generators

8.3.1 The Case $p = 2$

We start by considering the case $p = 2$, where the finite field $\mathbb{F}_2$ is simply the set $\{0, 1\}$ together with the operations of addition and multiplication modulo 2.

Definition 8.5 *An $\mathbb{F}_2$-linear generator is a generator of the form*

$$\begin{aligned} \mathbf{x}_{n+1} &= A\mathbf{x}_n, \\ \mathbf{y}_n &= B\mathbf{x}_n, \\ u_n &= \textstyle\sum_{i=1}^{k} y_{n,i} 2^{-i}, \end{aligned}$$

where A and B are matrices over $\mathbb{F}_2$, with A of size $r \times r$ and B of size $k \times r$. The matrix A is the transition matrix for passing from $\mathbf{x}_n$ to $\mathbf{x}_{n+1}$, while the matrix B transforms the vector $\mathbf{x}_n$ of length r into a vector $\mathbf{y}_n$ of length k. The final step is to transform the vector $\mathbf{y}_n$ into a number in the range $[0,1]$ by considering the entries of $\mathbf{y}_n$ as the coefficients of u_n in binary representation.

Example 8.6 *We can view the linear shift register as a generator of this form. To do this we need only map sequences of length k, where $k < r$, into elements of the range $[0,1]$. Taking subsequences of length k is equivalent to defining a matrix B as the top k rows of the $r \times r$ identity matrix.*

The matrix B ensures that each subsequence of length k generates exactly one pseudorandom-number, since

$$\mathbf{y}_n = (x_{n,0}, \ldots, x_{n,k-1}) = (a_n, \ldots, a_{n+k-1}).$$

We revisit Example 8.4. This length-4 register has parameters $(q_0, q_1, q_2, q_3) = (1,1,0,0)$, initial conditions $(a_0, a_1, a_2, a_3) = (0,0,0,1)$, and generates the sequence 000100110101111 of period 15. Let $k = 2$. In doing so, the y_n will correspond to each length-2 subsequence $y_n = (a_n, a_{n+1})$. The sequence of y_n will also repeat with a period of 15, yielding

$$00 \quad 00 \quad 01 \quad 10 \quad 00 \quad 01 \quad 11 \quad 10 \quad 01 \quad 10 \quad 01 \quad 11 \quad 11 \quad 11 \quad 10.$$

We now transform each y_n into a number $u_n \in \{0, 1/4, 1/2, 3/4\}$ by letting $u_n = \frac{y_{n,1}}{2} + \frac{y_{n,2}}{4}$. This yields a final sequence of

$$0 \quad 0 \quad \frac{1}{4} \quad \frac{1}{2} \quad 0 \quad \frac{1}{4} \quad \frac{3}{4} \quad \frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{2} \quad \frac{1}{4} \quad \frac{3}{4} \quad \frac{3}{4} \quad \frac{3}{4} \quad \frac{1}{2}.$$

Note that every element of $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ appears four times except 0, which appears three times.

The big advantage of $\mathbb{F}_2$-linear generators is that they are very efficient. On the other hand, if we want to improve their statistical performance we have to lengthen their period, making them more expensive to compute. As it turns out, there are better ways to improve the statistical performance with less loss of efficiency. We will begin by generalizing from $\mathbb{F}_2$-linear generators to $\mathbb{F}_p$-linear generators. Afterward, instead of simply lengthening the period of a given $\mathbb{F}_p$-linear generator, we will build better generators by combining several independent $\mathbb{F}_p$-linear generators of varying periods.

For the moment, let us return to the subject of linear shift registers and discuss how to choose coefficients q_i such that the generated sequences will have length $2^r - 1$. Although not absolutely necessary, it might be useful to have read Section 1.4 of Chapter 1 in order to be able to read the proof of this result (Theorem 8.9 below).

Definition 8.7 *A polynomial*

$$Q(x) = x^r + q_{r-1}x^{r-1} + \cdots + q_1 x + q_0$$

with coefficients $q_i \in \mathbb{F}_p$ is primitive *if and only if it is irreducible and the set of nonzero elements of $\mathbb{F}_{p^r}$, where*

$$\mathbb{F}_{p^r} = \{b_0 + b_1 x + \cdots + b_{r-1}x^{r-1} \mid b_i \in \mathbb{F}_p\},$$

is of the form

$$\mathbb{F}_{p^r} \setminus \{0\} = \{x^i \mid i = 0, \ldots, p^r - 2\},$$

where the powers x^i are taken modulo $Q(x)$.

Example 8.8 *We give an example with $p = 2$. We will show that the polynomial $Q(x) = x^3 + x + 1$ is irreducible. Indeed, suppose $Q(x) = Q_1(x)Q_2(x)$. Because $Q(x)$ is of degree 3, then either $Q_1(x)$ or $Q_2(x)$ is of degree 1, and hence belongs to the set $\{x, x+1\}$. If x divides $Q(x)$, this yields $Q(0) = 0$, which is not true. If $x+1$ divides $Q(x)$, then we should have $Q(1) = 0$, which is also not true. So neither x nor $x+1$ divides $Q(x)$, and $Q(x)$ is irreducible. The nonzero elements of $\mathbb{F}_{2^3}$ are given by $\{1, x, x+1, x^2, x^2+1, x^2+x, x^2+x+1\}$. Let us verify that they are all given by powers of x. Indeed, $Q(x) = 0$ yields $x^3 = x + 1$, so*

$$\begin{aligned}
x^4 &= x(x+1) = x^2 + x,\\
x^5 &= x(x^2+x) = x^3 + x^2 = (x+1) + x^2 = x^2 + x + 1,\\
x^6 &= x(x^2+x+1) = x^3 + x^2 + x = (x+1) + x^2 + x = x^2 + 1,\\
x^7 &= x(x^2+1) = x^3 + x = (x+1) + x = 1.
\end{aligned}$$

Theorem 8.9 *If the coefficients $q_0, \ldots, q_{r-1}$ of a linear shift register are chosen such that the polynomial*

$$Q(x) = x^r + q_{r-1}x^{r-1} + \cdots + q_1 x + q_0 \tag{8.5}$$

is primitive over $\mathbb{F}_2$, then for all initial conditions in which not all a_i are zero, the sequence generated by the linear shift register will have a period of $2^r - 1$.

Proof. We saw in Chapter 6 that

$$\mathbb{F}_{2^r} = \{b_0 + b_1 x + \cdots + b_{r-1}x^{r-1} \mid b_i \in \{0, 1\}\}$$

together with addition and multiplication modulo $Q(x)$ is a field, provided $Q(x)$ is irreducible. In Section 1.4 of Chapter 1, we saw also that it is always possible to

generate the nonzero elements of $\mathbb{F}_{2^r}$ by choosing a primitive polynomial $Q(x)$ and computing

$$\{x^i \mid i = 0, \ldots, 2^r - 2\},$$

and that $x^{2^r-1} = 1$. We introduce a linear mapping $T : \mathbb{F}_{2^r} \to \mathbb{F}_2$ such that

$$T(b_0 + b_1 x + \cdots + b_{r-1}x^{r-1}) = b_{r-1}.$$

We will show in Lemma 8.10 below that for any nonzero sequence $(a_0, \ldots, a_{r-1})$ there exists a unique $b = b_0 + b_1 x + \cdots + b_{r-1}x^{r-1}$ such that $a_i = T(bx^i)$, for $i = 0, \ldots, r-1$. Proposition 1.11 of Chapter 1 tells us that if a_n is a sequence generated by a linear shift register with initial conditions $a_i = T(bx_i)$, then for all n it holds that $a_n = T(bx^n)$. Since $x^{2^r-1} = 1$, a_n is periodic, and for all n it holds that $a_n = a_{n+2^r-1}$.

But is $2^r - 1$ the minimal period? Suppose there exists $m < 2^r - 1$ such that $a_n = a_{n+m}$ for all n. Then it must be that $a_0 = a_m, \ldots, a_{r-1} = a_{r+m-1}$. By Lemma 8.10 below there exists a unique b' such that $a_{i+m} = T(b'x^i)$, for $i = 0, \ldots, r-1$. We have on one side that $b' = b$ and on the other side that $b' = bx^m$, where the equalities are taken modulo $Q(x)$. Thus $b(x^m - 1) = 0$, and since $b \neq 0$, it must be that $x^m = 1$. However, x is a primitive root, and as such, $x^m \neq 1$ for all $m < 2^r - 1$, a contradiction. □

Lemma 8.10 *We consider the field*

$$\mathbb{F}_{2^r} = \{b_0 + b_1 x + \cdots + b_{r-1}x^{r-1} \mid b_i \in \{0, 1\}\}$$

with multiplication and addition taken modulo $Q(x)$, where $Q(x)$ is an irreducible polynomial as in (8.5)*. Then for any nonzero sequence $(a_0, \ldots, a_{r-1})$, there exists a unique $b = b_0 + b_1 x + \cdots + b_{r-1}x^{r-1}$ such that $a_i = T(bx^i)$, for $i = 0, \ldots, r-1$.*

Proof. We consider the linear system of equations $T(bx^i) = a_i$, for $i = 0, \ldots, r-1$, with unknowns $b_0, \ldots, b_{r-1}$. Consider the first equation

$$T(b) = b_{r-1} = a_0,$$

which immediately gives us the value of b_{r-1}. Next

$$\begin{aligned} bx &= (b_0 + b_1 x + \cdots + b_{r-1}x^{r-1})x \\ &= b_0 x + b_1 x^2 + \cdots + b_{r-2}x^{r-1} + b_{r-1}(q_0 + q_1 x + \cdots + q_{r-1}x^{r-1}), \end{aligned}$$

and therefore $T(bx) = b_{r-2} + q_{r-1}b_{r-1} = a_1$. Since b_{r-1} is already known, we may immediately substitute and find b_{r-2}.

We proceed accordingly for each b_i. Suppose $b_{i+1}, \ldots, b_{r-1}$ have already been found and consider bx^{r-1-i}. Then it follows that

$$\begin{aligned} bx^{r-1-i} &= (b_0 + b_1 x + \cdots + b_{r-1}x^{r-1})x^{r-1-i} \\ &= b_0 x^{r-1-i} + b_1 x^{r-i} + \cdots + b_i x^{r-1} + x^r P(x, b_{i+1}, \ldots, b_{r-1}), \end{aligned}$$

where $P(x, b_{i+1}, \ldots, b_{r-1})$ is a polynomial in x with coefficients depending only on the already known $b_{i+1}, \ldots, b_{r-1}$. Thus

$$T(bx^{r-1-i}) = b_i + R(b_{i+1}, \ldots, b_{r-1}).$$

The formula for $R(b_{i+1}, \ldots, b_{r-1})$ is not simple to write, but the important thing is that it depends only on the already known values $b_{i+1}, \ldots, b_{r-1}$. Thus we can easily calculate b_i from equation $T(bx^{r-1-i}) = a_{r-1-i}$, and this process uniquely determines the polynomial b. □

Corollary 8.11 *Consider a linear shift register of length r with coefficients q_i chosen such that the polynomial $Q(x)$ of (8.5) is primitive over $\mathbb{F}_2$. Suppose furthermore that the initial conditions a_i are not all zero. Then in the generated sequence of period $2^r - 1$, each possible subsequence of length r will occur exactly once except for the zero subsequence. (In this context we consider the sequence starting at a_i as being cyclic by identifying the index $n + 2^r - 1$ with the index n, letting us consider subsequences of length r that straddle two periods of the full sequence).*

Proof. Given a cyclic sequence $\{a_n\}$ of length $2^r - 1$, there are $2^r - 1$ subsequences of length r to be considered, one starting at each a_i, $i = 0, \ldots, 2^r - 2$, of the original sequence. (If $i \geq 2^r - r$, then using the periodicity, we may consider the subsequence $a_i, \ldots, a_{2^r-2}, a_0, \ldots, a_{i-2^r+r}$.) Furthermore, there are exactly 2^r possible sequences of length r since there are two choices per entry of such a sequence. Considering only those with at least one nonzero entry, there are exactly $2^r - 1$. Thus, each of these subsequences must appear exactly once if none of them may appear more than once. Suppose that one of the subsequences appears a second time, starting at entries a_i and a_j, where $0 \leq i < j < 2^r - 1$, yielding $0 < j - i < 2^r - 1$. Since the state of the register is the same at a_j as it was at a_i, we would have that $a_n = a_{n-j+i}$ for all $n \geq j$, contradicting the fact that the minimal period of the sequence is $2^r - 1$. Thus each nonzero subsequence of length r appears exactly once in a cyclic sequence of length $2^r - 1$. □

The following theorem shows that a linear shift register has good statistical properties when we consider all subsequences of length k, with $k \leq r$.

Theorem 8.12 *Consider a linear shift register of length r with coefficients q_i chosen such that the polynomial $Q(x)$ of (8.5) is primitive over $\mathbb{F}_2$. Let $k \leq r$. Suppose furthermore that the initial conditions a_i are not all zero. Then in any $2^r - 1$ sequential symbols generated by the register and considered as a cyclic sequence, each possible subsequence of length k will occur exactly 2^{r-k} times, except the null sequence, which will occur exactly $2^{r-k} - 1$ times.*

Proof. In Corollary 8.11 we showed that all subsequences of length r appear exactly once, except for the null subsequence. We consider a subsequence $b_0, \ldots, b_{k-1}$ of length

$k < r$ and count the number of ways this subsequence may be lengthened into a subsequence of length r by adding $b_k, \ldots, b_{r-1}$. Trivially there are 2 choices for each of the remaining $r-k$ symbols, yielding 2^{r-k} distinct ways of lengthening the subsequence. If at least one of the b_i is nonzero, each of the 2^{r-k} lengthened sequences occurs exactly once in the window of length $2^r - 1$, since all nonzero subsequences of length r appear exactly once. Thus, the subsequence $b_0, \ldots, b_{k-1}$ must appear exactly 2^{r-k} times.

In contrast, if all of the b_i are zero, then we cannot count the case in which we lengthen the subsequence with all zeros. However, all of the other lengthened subsequences will still be possible and appear exactly once each in the window of length r. Thus, the null sequence of length $k \leq r$ will appear exactly $2^{r-k} - 1$ times. □

8.3.2 A Lesson on Gambling Machines

Theorems 8.9 and 8.12 are the keys to understanding the story behind the arrested Montreal Casino gambler. Through his work, the gambler in question had an understanding of the basic principles behind random-number generators. He knew that the underlying algorithms were deterministic, and thus for a given algorithm and starting conditions, the sequences of generated numbers are identical. During earlier visits to the casino he had noticed that night after night, the keno machines kept drawing the same numbers in the same order. He thus recorded these numbers, and played them on his next visit with the result as described earlier in this chapter. But knowing that their keno machines had this problem, why did the Montreal Casino reopen them to the public? The official reason given was that the machines had been incorrectly programmed and that the error had been corrected. Another possible reason (more embarrassing for the casino, but equally possible) is that the machines were being turned off each night by an employee, perhaps even by the cleaning staff. The result being that when they were turned back on, the machines kept defaulting to the same initial conditions and thus generated the same sequence of numbers.

This story raises another question. How can the initial conditions be determined such that the sequence of generated numbers is not the same each time the machine is restarted? Do we need to leave the machines on forever? And how about video games? Here are two common solutions to this problem. In the first, we require that the machine be "properly" shut down. When properly shut down (not just by pulling on the power cord!), the machine can save the most recently generated a_i's and use them as initial conditions the next time it is started. A second solution relies on a clock built into the machine. When it is started, the machine retrieves the number of seconds (or even milliseconds) since the first of January 2000, with the last few digits of the time being used to seed the initial-condition a_i's.

8.3.3 The General Case

In this section we assume that the reader is familiar with the field $\mathbb{F}_{p^r}$. For more details regarding this field, refer to Section 6.5 of Chapter 6.

Definition 8.13 *(1) Let p be a prime number. An $\mathbb{F}_p$-linear random-number generator is a generator of the form*

$$a_n = q_0 a_{n-r} + q_1 a_{n-r+1} + \cdots + q_{r-1} a_{n-1} \pmod{p}, \tag{8.6}$$

where the $q_0, \ldots, q_{r-1}$ and the initial conditions $a_0, \ldots, a_{r-1}$ are integers in $\{0, 1, \ldots, p-1\}$ and operations are taken over the field $\mathbb{F}_p$ (in other words, modulo p).

(2) A multiple recursive generator is defined by the linear recurrence

$$\begin{cases} a_n = q_0 a_{n-r} + q_1 a_{n-r+1} + \cdots + q_{r-1} a_{n-1} \pmod{p}, \\ u_n = \frac{a_n}{p}. \end{cases}$$

When $p = 2$ an $\mathbb{F}_2$-linear generator is simply a linear shift register. Additionally, we see that $\mathbb{F}_p$-linear generators generate pseudorandom integers $a_n \in \{0, 1, \ldots, p-1\}$, while multiple recursive generators generate pseudorandom real numbers $u_n \in [0, 1)$.

Theorems 8.9 and 8.12 may be generalized to the case of $\mathbb{F}_p$-linear generators. In the case of $\mathbb{F}_2$, working modulo the polynomial $Q(x)$ given in (8.5) allows us to write

$$x^r = q_0 + q_1 x + \cdots + q_{r-1} x^{r-1} \tag{8.7}$$

because $q_i = -q_i$. Since this is in general no longer true for $\mathbb{F}_p$, we must redefine the polynomial $Q(x)$ such that the relation (8.7) still holds.

Theorem 8.14 *If p is prime and $q_0, \ldots, q_{r-1} \in \{0, 1, \ldots, p-1\}$ are chosen such that the polynomial*

$$Q(x) = x^r - q_{r-1} x^{r-1} - \cdots - q_1 x - q_0$$

is primitive over $\mathbb{F}_p$, then the $\mathbb{F}_p$-linear generator given in (8.6) *generates a sequence with period $p^r - 1$.*

Furthermore, if we take a sequence $(a_0, \ldots, a_{r-1})$ of initial conditions, not all of which are zero valued, then in any window of $p^r - 1$ generated symbols, all subsequences of length k with $k \leq r$ will appear exactly p^{r-k} times, except the null subsequence, which will appear exactly $p^{r-k} - 1$ times. (Again, we treat the window of generated symbols as a cyclic sequence and identify the index $n + p^r - 1$ with the index n.)

Proof. Since the proof is nearly identical to those of Theorems 8.9 and 8.12, we will leave it as an exercise to the reader. □

In practice, we often work with $\mathbb{F}_p$-linear generators in which the polynomial $Q(x)$ has only two nonzero coefficients, q_0 and q_s, for some $0 < s \leq r-1$. This makes modular arithmetic over these polynomials very simple.

Example 8.15 *We consider the case in which $p = 3$ and $Q(x) = x^4 - x - 1$. For the moment, we will assume that this polynomial is primitive and leave the details to*

Exercise 10. If we take initial conditions $(a_0, a_1, a_2, a_3) = (0, 0, 0, 1)$*, then the sequence output by this generator will have a period of* $3^4 - 1 = 80$*, with the first* 80 *values being*

$$\begin{array}{l} 0\,0\,0\,1\,0\,0\,1\,1\,0\,1\,2\,1\,1\,0\,0\,2\,1\,0\,2\,0\,1\,2\,2\,1\,0\,1\,0\,1\,1\,1\,1\,2\,2\,2\,0\,1\,1\,2\,1\,2 \\ 0\,0\,0\,2\,0\,0\,2\,2\,0\,2\,1\,2\,2\,0\,0\,1\,2\,0\,1\,0\,2\,1\,1\,2\,0\,2\,0\,2\,2\,2\,2\,1\,1\,1\,0\,2\,2\,1\,2\,1. \end{array} \tag{8.8}$$

We can verify the statistical properties of the sequence through inspection. The values 1 *and* 2 *each appear* 27 *times, while the value* 0 *appears* 26 *times. Each subsequence of length* 2 *appears nine times, except* 00*, which appears eight times. Each subsequence of length* 3 *appears three times, except* 000 *which appears exactly twice. Finally, each subsequence of length* 4 *appears exactly once, except for the null subsequence* 0000 *which does not appear at all.*

8.4 Combined Multiple Recursive Generators

Restricting ourselves to $\mathbb{F}_p$-linear generators whose polynomials $Q(x)$ have exactly two nonzero coefficients, q_0 and q_s, with $0 < s \leq r - 1$, greatly simplifies calculations. However, the generated sequences do not behave very well from a statistical point of view. In order to mitigate this deficiency we combine several such generators, operating with respect to distinct prime numbers p and distinct polynomials $Q(x)$ of the same degree.

Definition 8.16 *We consider* m *linear recurrences*

$$a_{n,j} = q_{0,j}a_{n-r,j} + q_{1,j}a_{n-r+1,j} + \cdots + q_{r-1,j}a_{n-1,j} \pmod{p_j}, \quad j = 1, \ldots, m,$$

satisfying the hypothesis of Theorem 8.14, where the p_j *are distinct primes. We combine these recurrences with the "output" function*

$$u_n = \left\{ \sum_{j=1}^{m} \frac{\delta_j a_{n,j}}{p_j} \right\},$$

where the δ_j *are arbitrarily chosen integers such that each* δ_j *is relatively prime to* p_j*. Here* $\{x\}$ *represents the fractional part of a real number* x *defined by*

$$\{x\} = x - [x],$$

where $[x]$ *is the integer part of the number* x*. (This means that we consider the* $a_{n,j}$ *both as elements of* $\mathbb{F}_{p_j}$ *and as real numbers!) A random-number generator of this form is called a combined multiple recursive generator.*

Remark. In the literature we also find the notation $x \pmod 1$ instead of $\{x\}$. Even if x and $\{x\}$ are not integers, this definition is similar to the classic definition, where two numbers a and b are congruent modulo an integer n if their difference $a - b$ may be written in the form mn for an integer $m \in \mathbb{Z}$.

Example 8.17 *We consider a combined multiple recursive generator with* $r = 3, m = 2, p_1 = 3, p_2 = 2$*, and* $\delta_1 = \delta_2 = 1$*. We let the reader verify that the polynomial* $Q_1(x) = x^3 - x - 2$ *is primitive over* $\mathbb{F}_3$*. Starting with initial conditions* 001*, the first generator generates the following sequence of period* 26*:*

$$00101211201110020212210222.$$

The second recurrence, when associated with the polynomial $Q_2(x) = x^3 - x - 1$*, which is primitive over* $\mathbb{F}_2$ *as proved in Example 8.8 and initial conditions* 001*, generates a sequence of period* 7*:*

$$0010111.$$

The combined generator will therefore have a period of $7 \times 26 = 182$*. We present the output in three lines, where the first line of each block represents an entire period of the first recurrence. The second line represents outputs from the second recurrence, where vertical lines delimit the boundary of its cycle. The third line represents the combined outputs as generated by the function* $u_n = \frac{a_{n,1}}{3} + \frac{a_{n,2}}{2}$*. Each of the outputs has been written as a fraction over* 6 *to show that the numerators create a sequence of pseudorandom-numbers over the set* $\{0, 1, \ldots, 5\}$*. The first block represents* u_n *for* $n = 0, \ldots, 25$*, while the second represents* $n = 26, \ldots, 51$*, etc.:*

$$\begin{array}{ccccccc|ccccccc|ccccccc|ccccc}
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1\\
0&0&\frac{5}{6}&0&\frac{5}{6}&\frac{1}{6}&\frac{5}{6}&\frac{2}{6}&\frac{4}{6}&\frac{3}{6}&\frac{2}{6}&\frac{5}{6}&\frac{5}{6}&\frac{3}{6}&0&\frac{4}{6}&\frac{3}{6}&\frac{4}{6}&\frac{5}{6}&\frac{1}{6}&\frac{1}{6}&\frac{2}{6}&0&\frac{1}{6}&\frac{4}{6}&\frac{1}{6}\\
\hline
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1\\
\frac{3}{6}&\frac{3}{6}&\frac{2}{6}&0&\frac{5}{6}&\frac{4}{6}&\frac{5}{6}&\frac{5}{6}&\frac{1}{6}&0&\frac{2}{6}&\frac{5}{6}&\frac{2}{6}&\frac{3}{6}&\frac{3}{6}&\frac{1}{6}&0&\frac{4}{6}&\frac{5}{6}&\frac{4}{6}&\frac{1}{6}&\frac{5}{6}&\frac{3}{6}&\frac{4}{6}&\frac{4}{6}&\frac{1}{6}\\
\hline
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0\\
0&\frac{3}{6}&\frac{5}{6}&\frac{3}{6}&\frac{2}{6}&\frac{4}{6}&\frac{5}{6}&\frac{2}{6}&\frac{1}{6}&\frac{3}{6}&\frac{5}{6}&\frac{2}{6}&\frac{2}{6}&\frac{3}{6}&0&\frac{1}{6}&\frac{3}{6}&\frac{1}{6}&\frac{2}{6}&\frac{4}{6}&\frac{1}{6}&\frac{2}{6}&\frac{3}{6}&\frac{1}{6}&\frac{1}{6}&\frac{4}{6}\\
\hline
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1\\
0&\frac{3}{6}&\frac{2}{6}&\frac{3}{6}&\frac{5}{6}&\frac{1}{6}&\frac{2}{6}&\frac{2}{6}&\frac{1}{6}&0&\frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&0&\frac{1}{6}&0&\frac{1}{6}&\frac{5}{6}&\frac{1}{6}&\frac{4}{6}&\frac{2}{6}&\frac{3}{6}&\frac{4}{6}&\frac{1}{6}&\frac{1}{6}\\
\hline
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0\\
\frac{3}{6}&0&\frac{2}{6}&\frac{3}{6}&\frac{2}{6}&\frac{1}{6}&\frac{5}{6}&\frac{5}{6}&\frac{4}{6}&0&\frac{5}{6}&\frac{2}{6}&\frac{5}{6}&\frac{3}{6}&\frac{3}{6}&\frac{4}{6}&0&\frac{1}{6}&\frac{2}{6}&\frac{1}{6}&\frac{1}{6}&\frac{5}{6}&0&\frac{4}{6}&\frac{1}{6}&\frac{4}{6}\\
\hline
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0\\
\frac{3}{6}&\frac{3}{6}&\frac{5}{6}&0&\frac{2}{6}&\frac{1}{6}&\frac{2}{6}&\frac{5}{6}&\frac{1}{6}&\frac{3}{6}&\frac{2}{6}&\frac{2}{6}&\frac{5}{6}&0&\frac{3}{6}&\frac{1}{6}&\frac{3}{6}&\frac{4}{6}&\frac{2}{6}&\frac{1}{6}&\frac{4}{6}&\frac{5}{6}&\frac{3}{6}&\frac{1}{6}&\frac{4}{6}&\frac{4}{6}\\
\hline
0&0&1&0&1&2&1&1&2&0&1&1&1&0&0&2&0&2&1&2&2&1&0&2&2&2\\
1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1&0&0&1&0&1&1&1\\
\frac{3}{6}&0&\frac{5}{6}&\frac{3}{6}&\frac{5}{6}&\frac{4}{6}&\frac{2}{6}&\frac{5}{6}&\frac{4}{6}&\frac{3}{6}&\frac{5}{6}&\frac{5}{6}&\frac{2}{6}&0&\frac{3}{6}&\frac{4}{6}&\frac{3}{6}&\frac{1}{6}&\frac{5}{6}&\frac{4}{6}&\frac{4}{6}&\frac{5}{6}&0&\frac{1}{6}&\frac{1}{6}&\frac{1}{6}
\end{array}$$

We see that when the p_i are small, the outputs and subsequences of outputs seem less regular than those generated by $\mathbb{F}_p$-linear generators.

Such combined generators perform excellently, even when one chooses m as small as 3. We can permit ourselves to choose sparse polynomials $Q_i(x)$ whose nonzero coefficients are simple, allowing efficient computations for each of the individual linear generators and for the combined multiple recursive generator. Examples of good choices for coefficients are given in [2]. Despite the simplicity of the calculations of individual linear generators, the combined generators perform very well from a statistical standpoint. Moreover, since the period of the combined generator is the product of the periods of the underlying linear generators, we can create generators with extremely large periods even though the underlying generators may themselves have short periods. Also, the cost of jumping from u_n to u_{n+N} becomes much cheaper than for a single linear recurrence with complicated coefficients.

8.5 Conclusion

Almost all current programming languages provide a random-number generator. The user thus has no need to delve into the theory of such generators in order to perform probabilistic simulations. However, the field of random-number generators is relatively young, and the number of statistical tests that a "good" random-number generator must pass continues to increase. (See [1] for a listing of basic tests that a decent generator must pass.) It is thus not surprising that the random-number generators made available by several programming languages are rapidly becoming obsolete. The history of the C programming language is interesting in this respect. The language was originally developed in the early 1970s, while the first manual, by Kernighan and Ritchie, the fathers of the language, appeared in 1978. Because of its widespread adoption, the need for a standard was soon felt. The process was arduous, but in 1989 the *American National Standards Institute* (ANSI) established a standard for the language. In the first version, the `rand()` function provided by the language had a period of length $2^{15} - 1 = 32{,}767$. This period is quite short, and certainly too short to be used in a gambling machine. The ANSI standard does not actually define the `rand()` function; it simply limits its output to the range $\{0, 1, \ldots,$ `RAND_MAX`$\}$, where `RAND_MAX` is greater than or equal to 32,767. Thus, the various compilers and C libraries that respect the standard could have different `rand()` functions with different values of `RAND_MAX` and different periods. The same program compiled on different machines could produce different results even given the same initial conditions. The `rand()` functions in several standard C implementations are well known for their poor results, failing some of the fundamental statistical tests. The programmers of these libraries are not necessarily to blame; rather, the whole situation shows that research in this field is still ongoing and active.

8.6 Exercises

1. Consider a string of bits generated by an independent random event (by a coin toss, for example). Consider further grouping the string into blocks of length r, with each r-bit string being interpreted as a number in the range $\{0, 1, \ldots, 2^{r-1}\}$. Show that each of these values will be generated roughly once out of every 2^r such blocks.

2. A linear-congruence generator generates numbers in the set $E = \{1, \ldots, p-1\}$ by the rule

$$x_n = ax_{n-1} \pmod{p},$$

where p is prime and a is chosen such that

$$\begin{cases} a^k \not\equiv 1 \pmod{p}, & k < p-1, \\ a^{p-1} \equiv 1. \end{cases}$$

(The existence of such an a, called a primitive root of $\mathbb{F}_p$ or $\mathbb{Z}_p$, is shown in Theorem 7.22.)

(a) Let $p = 11$. Find the primitive roots of $\mathbb{F}_{11}$ (there are four of them).

(b) Show that this generator will generate a sequence with minimal period $p-1$, regardless of the value of $x_0 \in S$.

3. The linear-congruence generator of Exercise 2 generates a sequence of numbers uniformly distributed over $E = \{1, \ldots, p-1\}$. Describe a method whereby we may transform this sequence into a sequence of 0's and 1's, while maintaining equiprobability for 0 and 1.

4. The following exercise is designed to show that a linear congruence generator (as described in Exercise 2) does not always have good statistical properties. Here we have chosen $p = 151$, the primitive root $a = 30$, and the initial condition $x_0 = 1$. The generated sequence with period 150 will therefore be

30	145	122	36	23	86	13	88	73	76	15	148	61	18	87
43	82	44	112	38	83	74	106	9	119	97	41	22	56	19
117	37	53	80	135	124	96	11	28	85	134	94	102	40	143
62	48	81	14	118	67	47	51	20	147	31	24	116	7	59
109	99	101	10	149	91	12	58	79	105	130	125	126	5	150
121	6	29	115	128	65	138	63	78	75	136	3	90	133	64
108	69	107	39	113	68	77	45	142	32	54	110	129	95	132
34	114	98	71	16	27	55	140	123	66	17	57	49	111	8
89	103	70	137	33	84	104	100	131	4	120	127	35	144	92
42	52	50	141	2	60	139	93	72	46	21	26	25	146	1.

We transform it into a sequence of 1's and 0's through the mapping

$$y_n = \begin{cases} 0, & x_n \leq 75, \\ 1, & 76 \leq x_n, \end{cases}$$

which generates the sequence

0 1 1 0 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 1 0 0 0 0
1 0 0 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 1 0 0 0 0 1 0 0 1 0 0
1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 1 0 1 0 1 0 1 0 1 1 0
1 0 1 0 1 0 1 0 1 0 0 1 1 1 1 0 1 1 0 0 0 0 1 1 0 0 0 0 1 0
1 1 0 1 0 1 1 1 1 0 1 1 0 1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0.

(a) What are the frequencies of 0 and 1, respectively?
(b) What are the frequencies of each of the possible subsequences of length 2: $00, 01, 10, 11$? In a good random-number generator, they should be roughly equal. What can you conclude?
(c) Answer the same question as above, but considering subsequences of length 3.

5. Consider a linear shift register (in other words, an $\mathbb{F}_2$-linear generator) with $(q_0, q_1, q_2, q_3) = (1, 0, 0, 1)$ and initial conditions $(a_0, a_1, a_2, a_3) = (0, 0, 0, 1)$. Verify that the generated sequence has minimal period 15, and that this cycle is not the same as that generated by Example 8.4.

6. Show that the polynomial $x^4 + x^3 + x^2 + x + 1$ is irreducible but *not* primitive over $\mathbb{F}_2$. Furthermore, verify that the linear shift register with $(q_0, q_1, q_2, q_3) = (1, 1, 1, 1)$ does not generate a sequence with minimal period 15.

7. Consider the linear shift register with $(q_0, q_1, q_2, q_3, q_4) = (1, 0, 1, 0, 0)$ and initial conditions $(a_0, a_1, a_2, a_3, a_4) = (0, 0, 0, 0, 1)$.
(a) Verify that the generated sequence has minimal period 31, by explicitly enumerating a_i for $i = 0, \ldots, 35$ (and by ensuring that $a_0 = a_{31}$, $a_1 = a_{32}$, $a_2 = a_{33}$, $a_3 = a_{34}$, and $a_4 = a_{35}$).
(b) Verify that 1 appears 16 times in the period of length 31.
(c) Verify that every subsequence of length 2 appears eight times, except for 00, which appears seven times.
(d) Verify that every subsequence of length 3 appears four times, except for 000, which appears three times.
(e) Verify that every subsequence of length 4 appears twice, except for 0000, which appears once.
(f) Verify that every subsequence of length 5 appears once, except for 00000, which never appears. Deduce that we could have taken any nonzero initial conditions and achieved the same result.
(g) Conclude that if we consider subsequences of length $k \leq r$ and eliminate all those that contain only zeros, then each of the possible outputs $\{1, \ldots, 2^k - 1\}$ is equiprobable.

8. The register of Exercise 7 generates a sequence $\{a_n\}$.
(a) Give the function that calculates a_{n+2} from a_n. (Suggestion: use the matrix form.)
(b) Give the function that calculates a_{n+10} from a_n.

9. Find all irreducible polynomials of degree 2 over $\mathbb{F}_3$. Which of these are primitive?

10. The goal of this exercise is to show that the polynomial $Q(x) = x^4 - x - 1$ is primitive over $\mathbb{F}_3$.
(a) Show that $Q(x)$ is irreducible over $\mathbb{F}_3$. To do this, you will need to have completed Exercise 9.
(b) Show that $Q(x)$ is primitive. That is, show that $x^k \neq 1$ for $k < 80$. To do this, you will have to calculate the powers x^k using the rule $x^4 = x + 1$. For example,

$$\begin{cases} x^5 = x(x+1) = x^2 + x, \\ x^6 = x(x^2 + x) = x^3 + x^2, \\ x^7 = x(x^3 + x^2) = x^4 + x^3 = (x+1) + x^3 = x^3 + x + 1. \end{cases}$$

(This may seem tedious, but it can be greatly simplified using Lemma 8.2, which guarantees that $x^k = 1$ can occur only when k divides 80. This lets us limit ourselves to computing x^k for k a divisor of 80.)

11. Choose a primitive polynomial of degree 2 over $\mathbb{F}_3$ and use its coefficients to construct an $\mathbb{F}_3$-linear generator.
(a) Compute the periodic sequence of pseudorandom numbers generated by this generator.
(b) How many occurrences are there of $0, 1$, and 2 in a period?
(c) Verify that each subsequence of length 2 appears exactly once, except the subsequence 00.

12. Choose a primitive polynomial of degree 3 over $\mathbb{F}_3$ and use its coefficients to construct an $\mathbb{F}_3$-linear generator.
(a) Compute the periodic sequence of pseudorandom numbers generated by this generator.
(b) How many occurrences are there of $0, 1$, and 2 in a period?
(c) Verify that each subsequence of length 2 appears exactly three times, except the subsequence 00, which appears twice.
(d) Verify that each subsequence of length 3 appears exactly once, except the subsequence 000.

13. Choose a degree-2 primitive polynomial Q_1 over $\mathbb{F}_2$ and use its coefficients to construct an $\mathbb{F}_2$-linear generator. Similarly, choose a degree-2 primitive polynomial Q_2 over $\mathbb{F}_3$ to construct an $\mathbb{F}_3$-linear generator.

(a) What is the period of the multiple recursive generator given by taking $\delta_1 = \delta_2 = 1$?
(b) Choose a set of initial conditions and compute the sequence of outputs u_n for a single cycle.

14. Explain how to construct a random-number generator that simulates tossing a die. Remember that such a generator needs to draw numbers uniformly from the set $\{1, \ldots, 6\}$.

15. When performing simulations we often require random-number generators that generate numbers according to a given probability distribution. Up until this point we have considered only generators that are uniform $U[0, 1]$ over $[0, 1]$. Show how to transform such a generator into one that is uniform $U[a, b]$ over a given interval $[a, b]$.

Note: The probability density function of a uniformly random variable over an interval $[a, b]$ is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

16. When we want to generate random numbers that obey more general laws of probability, we need to consider the cumulative distribution function: if X is a random variable, the cumulative distribution function is given by

$$F_X(x) = \text{Prob}(X \leq x).$$

(a) For a given random variable X that is uniform on $[0, 1]$ (we write $X \sim U[0, 1]$), show that the cumulative distribution function is given by

$$F_X(x) = \begin{cases} 0, & x < 0, \\ x, & x \in [0, 1], \\ 1, & x > 1. \end{cases}$$

(The probability density function of X is given in Exercise 15 by taking $a = 0$ and $b = 1$.)
(b) Let $X \sim U[0, 1]$ and let $g(x) : [0, 1] \to \mathbb{R}$ be a strictly increasing function. Consider the random variable $Y = g(X)$. Show that the cumulative distribution function of Y is given by

$$F_Y(y) = F_X(g^{-1}(y)),$$

where g^{-1} denotes the inverse function of g such that $g(g^{-1}(x)) = x$.
(c) Consider a random variable Y that obeys an exponential probability density function with parameter λ:

$$f_Y(y) = \begin{cases} 0, & y < 0, \\ \lambda e^{-\lambda y}, & y \geq 0. \end{cases}$$

Calculate the cumulative distribution function of Y.

(d) Let $X \sim U[0,1]$ and $Y = g(X)$. What function g must we take for Y to obey an exponential distribution with parameter λ? Explain a practical method by which we may generate random numbers observing such an exponential distribution.

17. In the game of bridge, 52 cards are distributed among four players A, B, C, and D.
(a) Explain why there are $\frac{52!}{(13!)^4}$ distinct ways of dealing the cards. (In bridge the players are numbered from 1 to 4, following the order in which they will bid. The order in which cards are played is different and depends on the specific bids. Thus, two games in which the same four hands were dealt to different players should be considered different games.)
(b) How many seconds are there in a year? Compute how many years would be necessary to play every possible game of bridge assuming that we could finish a game every second.
(c) We see that in practice it is impossible to play every possible bridge hand. Does this mean that it is equally impossible to calculate statistics regarding individual bridge games? Statistics permit us to draw conclusions on a population from an analysis of only a sample of the population (in this context, the set of all possible bridge games), provided that the sample is representative. One manner of constructing a sample is to number the cards from 1 to 52. To deal the cards to the first player we first must choose a single card from the deck of 52, corresponding to generating a uniform random number over the set $\{1, \ldots, 52\}$. We then choose a second card from the remaining 51, a third from the remaining 50, and so on, until we choose the first player's last card from among the remaining 40. We continue this process to deal hands to the second and third players, with the fourth player receiving the remaining cards. In order to deal a second game, we repeat the algorithm a second time. What conditions must be satisfied by the different random-number generators involved so that each distinct hand has an equal chance of being generated?
(d) We have seen that it is not sufficient for all possible events (games in this context) to have the same probability of being generated. We require also that each possible subsequence of k events be generated with equal probability. Given that this question is hard to analyze on account of the sheer number of possible subsequences of events, we can instead compute partial statistics. For example, what is the probability that a single bridge hand contains all four aces? We could then randomly generate a thousand games and verify whether the number of times all four aces occur in a single hand is close to the expected number.
(e) Calculate the probability of two other not-too-rare events that could be used as statistical tests.

References

[1] D. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, 1997.

[2] P. L'Ecuyer. Good parameters and implementations for combined multiple recursive random number generators. *Operations Research*, 47:159–164, 1999.

[3] P. L'Ecuyer. Random numbers. In N.J. Smelser and P.B. Baltes, editors, *The International Encyclopedia of the Social and Behavioral Sciences*, pages 12735–12738. Pergamon, Oxford, 2002.

[4] P. L'Ecuyer and F. Panneton. Fast random number generators based on linear recurrences modulo 2: overview and comparison. In *Proceedings of the 2005 Winter Simulation Conference*, pages 110–119, 2005.

9

Google and the *PageRank* Algorithm

The first three sections of this chapter make use of linear algebra (diagonalization, eigenvalues, and eigenvectors) and elementary probability theory (independence of events and conditional probability). These sections provide the basics and can be covered in about three hours. Combined, they give a good idea of how the PageRank algorithm works. Section 9.4 is more advanced, requiring a familiarity with real analysis (accumulation points and convergence of sequences); this section may be covered in one or two hours.

9.1 Search Engines

In the digital world, new problems are generally quickly solved by new algorithms or new hardware. Those who have used the world wide web for more than a few years, say since 1998, will no doubt remember the search engines provided by *AltaVista* and *Yahoo*. More than likely, these same people now use *Google*'s search engine. Surprisingly, among all the general-purpose search engines, *Google* rose to its current supremacy in a matter of months. It did so thanks to its algorithm for ranking search results: the *PageRank* algorithm. The goal of this chapter is to describe this algorithm and the mathematical foundations on which it is built: Markov chains.

Using a search engine is fairly simple. It starts with somebody sitting at a computer connected to the Internet, and a desire to learn about a particular subject. Suppose, for example, that he wants to learn about the annual snowfall in Montreal. He decides to query *Google*[1] with the keywords *precipitation*, *snow*, *Montreal*, and *century*. (Of these, the last word may seem a little strange. However, the user has chosen this word to indicate his desire for long-term statistics.) The search engine responds with a brief list of what it deems to be the best sources of information on the topic (see Figure 9.1). The horizontal bar at the top of the page indicates that the search was performed in

[1] *Google* can be found at `http://www.google.com`

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_9,

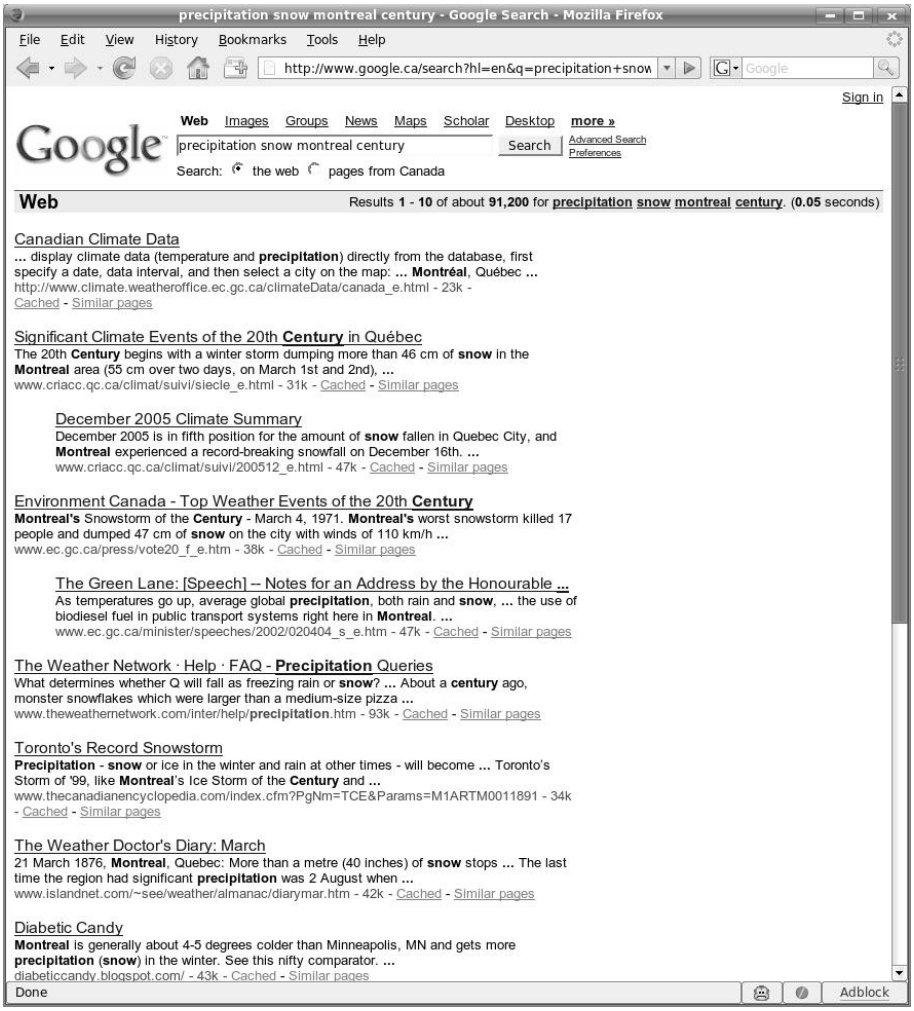


Fig. 9.1. A *Google* search on the keywords *precipitation*, *snow*, *montreal* and *century*.

less than a tenth of a second, and that around 91,200 potentially relevant pages were identified. The first is a link to an online database of Canadian climate data, provided by Environment Canada, which runs the Canadian weather office. (From here we can learn that the most snow seen since accurate record-keeping began was 384.3 cm in 1954! Thankfully, we also learn that the 30-year average is a little more reasonable, at 217.5 cm.) The first search result returned by *Google* often has quite a good chance of answering the user's question. How about the others? As we descend through the list,

the focus of the results tends to wander, with many documents concerning the Montreal Protocol on climate change. These later documents are of very little interest to the user, since they do not speak at all about snow in Montreal. But they are related in some sense, for they effectively all contain at least three of the four search terms.

This anecdote brings up an important point:[2] the pages that *Google* returns first are often exactly those that satisfy the user's needs. The search would definitely be hopeless if the user had to go through the 91,200 pages. The exact keywords entered by the user will obviously have an impact on the pages returned, but how in general can *Google* use a computer to guess the desires of the user?

Automated search tools have been around for a few decades. We can immediately think of several domains with large bodies of knowledge that need to be efficiently navigated: library catalogs, government registries (births, deaths, taxes) and professional databases (legal, dental, medical, parts catalogs). These bodies of information all have a few points in common. First off, they all contain data that lies within a single clearly defined scope. For example, all the books in a library contain a title, one or more authors, a publisher, etc. The *uniformity of the data* to be organized thus makes the database more easily categorized and more easily searched. The *quality* of the information is also very high. For example, books are normally entered into a library's catalog by professionals, and the error rate is thus very low. If and when an error occurs, the simplicity of the database makes it easy for corrections to be made. The *uniformity of the user's needs* is also an advantage in these systems. The goal of a library catalog is above all to maintain a concise listing of exactly what books are on hand. Even though specialized terms may exist (for example in medical or legal databases), the users are typically professionals in the field and will all be familiar with them. Thus, these databases may be searched with relative ease by their users. These databases all evolve relatively slowly. In a library, very few books leave the collection in a year, and a year that sees 10% growth in the catalog would be rare. Add to this the fact that the information already in a library catalog is always accurate, and never changes! The *growth rate* is therefore relatively slow, and such databases are easily maintained by humans. Finally, it is easy to achieve a *consensus* rating on the quality of the items in the database. In most university faculties, committees guide the purchase of new books for the library. Moreover, professors guide students directly toward the best books for their courses.

None of these characteristics exists on the web. The pages on the web have an immense diversity: technical, professional, promotional, commercial, entertainment, etc. The quality to be found is also very inconsistent: we can expect to find many spelling and grammar errors, as well as misinformation (whether these errors are accidental or otherwise). The users of the web are also as numbered and varied as the pages on the web, and their familiarity with search engines is extremely variable. The speed at which

[2] If the user were to repeat this search again today, chances are the results would be vastly different and in all probability there would be many more returned pages. This is due to the constantly changing and expanding nature of the world wide web.

the web evolves is staggering: as of the end of 2005 (when they stopped publishing the size of their database on their front page), *Google* was indexing well over 9 *billion* pages, with others appearing and disappearing daily. Finally, it seems illusory to establish a consensus on the relative quality of web pages given their number, their diversity, and the equally varying interests of the hundreds of millions of users worldwide. It seems that web pages have nothing in common!

In fact, this is a bit of a lie, since most pages on the web *do* have something in common. They are nearly all written in HTML (HyperText Markup Language) or in some related dialect. And the method in which they are related to each other is uniform: links between pages are all encoded in the same manner. These links consist of a few fixed characters preceding the address of the page, otherwise known as its URL (Uniform Resource Locator). These are precisely the links that a human user may follow in surfing the web, and which a computer can differentiate from the text, images, and other elements of a web page. In January 1998, four researchers from Stanford University, L. Page, S. Brin, R. Motwani, and T. Winograd, proposed an algorithm [3] for ranking pages on the web. This algorithm, *PageRank*, does not use the textual or visual content of the page, but rather the structure of the links between them.[3]

9.2 The Web and Markov Chains

The web is composed of billions of individual pages, and even more links between them.[4] As such, the web can be modeled as a directed graph, where pages are nodes, and links are directed edges between them. For example, Figure 9.2 represents a (small) web containing five pages (A, B, C, D, and E). The directed edges between the nodes indicate that

- the only link from page A leads to page B,
- page B links to pages A and C,
- page C links to pages A, B, and E,
- the only link from page D leads to page A, and
- page E links to pages B, C, and D.

In order to determine the ranking to be accorded to each of these five pages, we consider a simple version of the *PageRank* algorithm. Suppose that an impartial web surfer navigates through this web by randomly choosing links to follow. When he has only one choice (for example, if he finds himself on page D), then he will follow that link (leading to page A in this example). If he finds himself on page C, he will follow the link to page A one-third of the time and similarly for the links to pages B and E. In other

[3] The first four letters of *PageRank* refer to the first author's last name, and not to pages of the web.

[4] When Page et al. published their algorithm in 1998, they estimated the size of the web as roughly 150 million pages with 1.7 billion links between them. In early 2006, the web was estimated as containing around 12 billion pages.

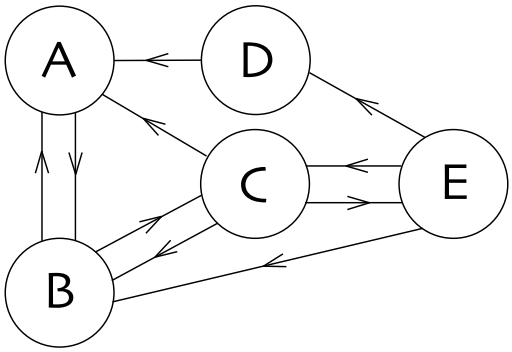


Fig. 9.2. A web of five pages and its links.

words, when he finds himself on a given page, he will randomly choose from among the outbound links, according each an equal probability. If such a web surfer were left to crawl the web in such a manner following one link per minute, where would he find himself in an hour, in two days, or after some large number of jumps? More precisely, given that his path is determined probabilistically, with what probability would he find himself on a given page after a given amount of time?

Figure 9.3 answers this question for the first two steps of an impartial web surfer starting at page C. This page contains three outbound links; thus the web surfer can end up only on one of the pages A, B, E. Thus, after the first step he would find himself on page A with probability $\frac{1}{3}$, on page B with probability $\frac{1}{3}$, and on page E with probability $\frac{1}{3}$. This is indicated in the middle column of Figure 9.3 by the three relations

$$p(A) = \frac{1}{3}, \qquad p(B) = \frac{1}{3}, \qquad p(E) = \frac{1}{3}.$$

Similarly,

$$p(C) = 0 \quad \text{and} \quad p(D) = 0$$

indicate that after one step the web surfer could not possibly be on page C or D, since no links from his previous page can lead him there. Each of the three possible paths is indicated by its probability of being taken. Furthermore, given that he must stay within the web, they satisfy

$$p(A) + p(B) + p(C) + p(D) + p(E) = 1.$$

The results after the first step are rather simple and predictable. However, even after only two steps, things begin to get complicated. The third column of Figure 9.3 gives the possible trajectories after a second step. If the web surfer was on A after the first step, he would be guaranteed to be on B after a second step. Since he had been on A with probability $\frac{1}{3}$, this path contributes $\frac{1}{3}$ to the probability of being on B after

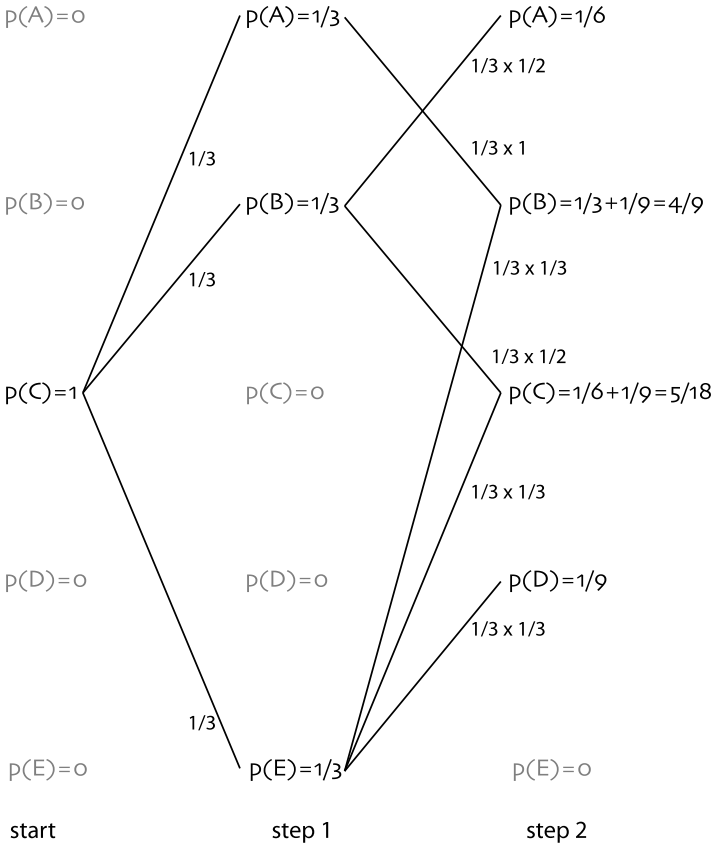


Fig. 9.3. The first two steps of an impartial web surfer starting at page C.

a second step. However, $p(B)$ does not equal $\frac{1}{3}$ after the second step, since there is another independent path that could lead him there: $C \to E \to B$. If the web surfer found himself on page E after the first step, he could choose (with equal probability) from the three links leading to pages B, C, and D. Each of these paths contributes $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$ to the probabilities $p(B)$, $p(C)$, and $p(D)$ after the second step. Although there are more possibilities and the attached probabilities are more complicated, the end result is relatively simple. After two steps, the web surfer finds himself on a given page with the following probabilities:

$$p(A) = \frac{1}{6}, \qquad p(B) = \frac{4}{9}, \qquad p(C) = \frac{5}{18}, \qquad p(D) = \frac{1}{9}, \qquad p(E) = 0.$$

Again, we see that these probabilities satisfy

$$\begin{aligned} p(A)+p(B)+p(C)+p(D)+p(E) &= \frac{1}{6}+\frac{4}{9}+\frac{5}{18}+\frac{1}{9}+0 \\ &= \frac{3+8+5+2+0}{18}=1. \end{aligned}$$

At this point, the method should seem clear, and we could continue to calculate the probabilities after a few more steps. However, it is useful to formalize this impartial walk through the web. The tool best suited to this job is the theory of Markov chains.

A *random process* $\{X_n, n = 0, 1, 2, 3, \dots\}$ is a family of random variables parameterized by the integer n. We assume that each of these random variables X_n takes its values from a finite set T. In the example of the impartial web surfer, T is the set of pages in the web: $T = \{A, B, C, D, E\}$. For each step $n \in \{0, 1, 2, \dots\}$, the position of the web surfer is X_n. Sticking to the language of random processes, we determined earlier the probabilities of the possible outcomes for X_1 and X_2 assuming that the walk started from C. This can be rephrased as a conditional probability $P(I|J)$, which gives the probability that event I occurs given that event J has already occurred. For example, $P(X_1 = A|X_0 = C)$ gives the probability of the web surfer finding himself on page A at step 1 after having been on page C at the beginning (step 0). Thus

$$p(X_1 = A|X_0 = C) = \frac{1}{3}, \quad p(X_1 = B|X_0 = C) = \frac{1}{3}, \quad p(X_1 = C|X_0 = C) = 0,$$
$$p(X_1 = D|X_0 = C) = 0, \quad p(X_1 = E|X_0 = C) = \frac{1}{3},$$

and

$$p(X_2 = A|X_0 = C) = \frac{1}{6}, \quad p(X_2 = B|X_0 = C) = \frac{4}{9}, \quad p(X_2 = C|X_0 = C) = \frac{5}{18},$$
$$p(X_2 = D|X_0 = C) = \frac{1}{9}, \quad p(X_2 = E|X_0 = C) = 0.$$

The random walk followed by the impartial web surfer possesses the defining property of Markov chains. First off, we will define Markov chains.

Definition 9.1 *Let $\{X_n, n = 0, 1, 2, 3, \dots\}$ be a random process taking its values from the set $T = \{A, B, C, \dots\}$. We say that $\{X_n\}$ is a* Markov chain *if the probability $P(X_n = i)$, $i \in T$, depends only on the value of the process at the previous step, X_{n-1}, and not on any of the preceding steps, $X_{n-2}, X_{n-3}, \dots$. We define $N < \infty$ as the number of elements in T.*

In the example of the impartial web surfer, the random variables are the positions X_n after n steps. In thinking back to our earlier calculations we notice that in calculating the probabilities after the first step, $P(X_1)$, we used only the starting point. Similarly, in calculating the probabilities after the second step, $P(X_2)$, we used only the probabilities from the first step. This property of being able to calculate $P(X_n)$ using only the information from $P(X_{n-1})$ is the defining property of Markov chains. Are all random

processes Markov chains? Certainly not. It takes only a slight change to the rules of our impartial web surfer in order to lose the Markov property. Suppose that we want to prevent the web surfer from ever returning immediately to the page where he came from. For example, after the first step, our web surfer found himself on pages A, B, and E with equal probability. He cannot return to page C from page A, but he could possibly do so from pages B and E. Thus, we could prevent the web surfer from following the links to page C from pages B and E. Under these new rules, the web surfer would have only a single choice when arriving at page B from page C (he would have to go to page A), and he would be reduced to two choices at page E (either page B or page D). In prohibiting the web surfer from following links to its previous page we have lost the Markov property: the process has *memory*. In fact, in order to determine the probabilities $P(X_2)$ we need to know not only the probabilities at step 1, but also the page (or pages) where the web surfer was at the start (step zero). The rules that we originally defined are thus rather special in a mathematical sense: Markov chains have no memory of past states, and the future state is completely determined by the current state.

Markov chains are unique in that their behavior may be entirely characterized by their initial state ($p(C) = 1$ in the example of Figure 9.3) and a *transition matrix* given by

$$p(X_n = i \mid X_{n-1} = j) = p_{ij}. \tag{9.1}$$

A matrix P is a Markov chain transition matrix if and only if

$$p_{ij} \in [0,1] \quad \text{for all } i,j \in T \qquad \text{and} \qquad \sum_{i \in T} p_{ij} = 1 \quad \text{for all } j \in T. \tag{9.2}$$

For our impartial web surfer, the elements p_{ij} of the transition matrix P represent the probabilities of finding himself at page $i \in T$ when he is coming from page $j \in T$. However, our rules force the surfer to choose with equal probability from among the available links. Thus, if page j offers m links, then column j of P will contain $\frac{1}{m}$ in the rows corresponding to the m linked pages, and 0 in the remaining rows. The transition matrix for the simple web in Figure 9.2 is thus given by

$$P = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \tag{9.3}$$

The columns of P indicate possible destinations: from page E the web surfer may proceed to pages B, C, and D. Similarly, the nonzero entries in rows indicate possible origins: the single nonzero entry in the fourth row indicates that we may arrive at page D only from page E.

What exactly does the second constraint of (9.2) mean? To clarify, we rewrite it with the help of the transition matrix defined in (9.1):

$$\sum_{i\in T} p_{ij} = \sum_{i\in T} p(X_n = i \mid X_{n-1} = j) = 1,$$

which may be read as follows: if at step $n-1$ the system is in state j (at page $j \in T$), then the probability of being in *any* possible state at step n is 1. Stated even more simply, this means that a web surfer on a given page at step $n-1$ must certainly find himself still in the web at step n. Thus, the constraint is actually rather simple.

This formalization has several advantages. The operation of matrix multiplication suffices to reproduce the multitude of tedious calculations performed as we followed the web surfer through his first two steps. As before, we assume that the web crawler starts at page C. Thus

$$p^0 = \begin{pmatrix} p(X_0 = A) \\ p(X_0 = B) \\ p(X_0 = C) \\ p(X_0 = D) \\ p(X_0 = E) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

The probability vector p^1 after the first step is given by $p^1 = Pp^0$, and therefore

$$p^1 = \begin{pmatrix} p(X_1 = A) \\ p(X_1 = B) \\ p(X_1 = C) \\ p(X_1 = D) \\ p(X_1 = E) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix},$$

the same as we calculated before. In the same manner, applying the transformation matrix again yields $p^2 = Pp^1$; the probability vector after the second step is therefore

$$p^2 = \begin{pmatrix} p(X_2 = A) \\ p(X_2 = B) \\ p(X_2 = C) \\ p(X_2 = D) \\ p(X_2 = E) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ \frac{4}{9} \\ \frac{5}{18} \\ \frac{1}{9} \\ 0 \end{pmatrix}.$$

The same method may be followed to calculate the probability vector after any number of steps: $p^n = Pp^{n-1}$, or alternatively,

$$p^n = Pp^{n-1} = P(Pp^{n-2}) = \cdots = \underbrace{PP\cdots P}_{n \text{ times}} p^0 = P^n p^0.$$

The constraints of (9.2) on the transition matrix P result in several properties of Markov chains that are very important for the *PageRank* algorithm.

This first property we will examine can be seen by taking several powers of the transition matrix P. The powers P^4, P^8, P^{16}, and P^{32}, rounded to three decimal places, are given by

$$P^4 = \begin{pmatrix} 0.333 & 0.296 & 0.204 & 0.167 & 0.420 \\ 0.222 & 0.463 & 0.531 & 0.667 & 0.160 \\ 0.389 & 0.111 & 0.160 & 0.000 & 0.370 \\ 0.056 & 0.000 & 0.031 & 0.000 & 0.019 \\ 0.000 & 0.130 & 0.074 & 0.167 & 0.031 \end{pmatrix}, \quad P^8 = \begin{pmatrix} 0.265 & 0.313 & 0.294 & 0.323 & 0.279 \\ 0.420 & 0.360 & 0.409 & 0.372 & 0.381 \\ 0.217 & 0.233 & 0.191 & 0.201 & 0.252 \\ 0.031 & 0.022 & 0.018 & 0.012 & 0.035 \\ 0.067 & 0.072 & 0.088 & 0.092 & 0.052 \end{pmatrix},$$

$$P^{16} = \begin{pmatrix} 0.294 & 0.291 & 0.293 & 0.291 & 0.294 \\ 0.388 & 0.392 & 0.389 & 0.391 & 0.391 \\ 0.220 & 0.219 & 0.221 & 0.221 & 0.218 \\ 0.024 & 0.025 & 0.025 & 0.025 & 0.024 \\ 0.074 & 0.073 & 0.072 & 0.072 & 0.074 \end{pmatrix}, \quad P^{32} = \begin{pmatrix} 0.293 & 0.293 & 0.293 & 0.293 & 0.293 \\ 0.390 & 0.390 & 0.390 & 0.390 & 0.390 \\ 0.220 & 0.220 & 0.220 & 0.220 & 0.220 \\ 0.024 & 0.024 & 0.024 & 0.024 & 0.024 \\ 0.073 & 0.073 & 0.073 & 0.073 & 0.073 \end{pmatrix}.$$

We observe that P^m seems to converge to a constant matrix as m increases. As it turns out, this is not just by luck, but rather it is a property of most Markov chain transition matrices.

Property 9.2 *The transition matrix P of a Markov chain has at least one eigenvalue equal to* 1.

Proof: Recall that the eigenvalues of a matrix are always equal to the eigenvalues of its transpose. This is a result of the fact that both matrices share the same characteristic polynomial:

$$\Delta_{P^t}(\lambda) = \det(\lambda I - P^t) = \det(\lambda I - P)^t = \det(\lambda I - P) = \Delta_P(\lambda),$$

which itself follows from the fact that the determinant of a matrix is equal to that of its transpose. It is simple to find an eigenvector of P^t. Let $u = (1, 1, \ldots, 1)^t$. Then $P^t u = u$. In fact, expanding the matrix multiplication directly, we see that

$$\begin{aligned} (P^t u)_i &= \sum_{j=1}^{n} [P^t]_{ij} u_j = \sum_{j=1}^{n} p_{ji} \cdot 1, \qquad \text{since all } u_j \text{ are } 1, \\ &= 1, \end{aligned}$$

by (9.2). □

Property 9.3 *If λ is an eigenvalue of an $n \times n$ transition matrix P, then $|\lambda| \leq 1$. Furthermore, there exists an eigenvector associated to the eigenvalue $\lambda = 1$ with all nonnegative entries.*

This property is a direct result of a theorem attributed to Frobenius. Although the proof relies only on elementary linear algebra and analysis, it is far from simple. We will explore this proof in Section 9.4.

Hypotheses Before we continue, we will state three hypotheses that we will assume from now on.

(i) First off, we will suppose that there is exactly one eigenvalue such that $|\lambda| = 1$, and therefore by Property 9.2 this eigenvalue is 1.
(ii) Next, we will suppose that this eigenvalue is not degenerate, which is to say that the associated eigensubspace has dimension 1.
(iii) Finally, we will take for granted that the transition matrix P representing the web is diagonalizable, meaning that its eigenvectors form a basis.

The first two hypotheses are not actually true for all transition matrices, and it is in fact possible to construct valid transition matrices that violate both of them (see the exercises). However, these remain reasonable hypotheses for transition matrices generated by large webs. The third hypothesis is there to simplify the following result.

Property 9.4 *1. If the transition matrix P of a Markov chain satisfies the three hypotheses above, then there exists a unique vector π such that the entries $\pi_i = P(X_n = i), i \in T$, satisfy*

$$\pi_i \geq 0, \qquad \pi_i = \sum_{j \in T} p_{ij} \pi_j, \qquad \text{and} \qquad \sum_{i \in T} \pi_i = 1.$$

We will call the vector π the stationary regime *of the Markov chain.*
2. Regardless of the initial point $p_i^0 = P(X_0 = i)$ (where $\sum_i p_i^0 = 1$), the distribution of probabilities $P(X_n = i)$ will converge to the stationary regime π as $n \to \infty$.

Proof: The first point simply repeats the fact that P has a single eigenvector with eigenvalue 1 whose components sum to 1. In fact, the defining equation for the stationary regime is simply $\pi = P\pi$. In other words, π is the eigenvector of P associated with the nondegenerate eigenvalue 1. Property 2 tells us that π is composed of nonnegative entries. Since an eigenvector is always nonzero, the sum of its entries must be strictly positive. By renormalizing this vector we can therefore always ensure that $\sum_i \pi_i = 1$.

To show the second point we rewrite the initial state vector p^0 in terms of the basis formed by the eigenvectors of P. We index the eigenvalues of P as follows: $1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_N|$. Hypotheses *(i)* and *(ii)* tell us that the first inequality in this ordering is strict (that is, the absolute value of λ_1 is strictly larger than that of λ_2), while hypothesis *(iii)* assures us that the eigenvectors of P form a basis for the space of dimension N where P acts. (For this last step, the eigenvalues must be counted with their multiplicities.) Let v_i be the eigenvector associated with the eigenvalue λ_i. Furthermore, assume that v_1 has been normalized such that $v_1 = \pi$. The set $\{v_i, i \in T\}$ forms a basis, allowing us to write

$$p^0 = \sum_{i=1}^{N} a_i v_i,$$

where the a_i are the coefficients of p^0 in this basis.

We will show that the coefficient a_1 is always 1. For this, we will make use of the vector $u^t = (1, 1, \ldots, 1)$ that was introduced in the discussion of Property 1. If v_i is an eigenvector of P with eigenvalue λ_i (which is to say that $Pv_i = \lambda_i v_i$), then the matrix product $u^t P v_i$ can be simplified in two ways. The first yields

$$u^t P v_i = (u^t P) v_i = u^t v_i,$$

and the second,

$$u^t P v_i = u^t (P v_i) = \lambda_i u^t v_i.$$

These two expressions must be equal by the associativity of matrix multiplication. For $i \geq 2$, the eigenvalue λ_i is not 1, and the equality can only hold if $u^t v_i = 0$, which expands as

$$u^t v_i = \sum_{j=1}^{N} (v_i)_j = 0,$$

where $(v_i)_j$ represents the jth coordinate of the vector v_i. This condition states that the sums of the coordinates of the vectors $v_i, i \geq 2$, must all be zero. If we now sum the components of p^0, we get 1 by hypothesis ($\sum_{i=1}^{N} p_i^0 = 1$). Thus

$$\begin{aligned} 1 &= \sum_{j=1}^{N} p_j^0 = \sum_{j=1}^{N} \sum_{i=1}^{N} a_i (v_i)_j = \sum_{i=1}^{N} a_i \sum_{j=1}^{N} (v_i)_j \\ &= a_1 \sum_{j=1}^{N} (v_1)_j = a_1 \sum_{j=1}^{N} \pi_j = a_1. \end{aligned}$$

(To obtain the second inequality we used the expression p^0 written in the basis of the eigenvectors. For the fourth, we used the fact that the sums of the coefficients of the v_i are all zero-valued except for v_1.)

To obtain the behavior after m steps, repeatedly apply the transition matrix P (m times) starting from the initial state p^0:

$$P^m p^0 = \sum_{j=1}^{N} a_j P^m v_j = \sum_{j=1}^{N} a_j \lambda_j^m v_j = a_1 v_1 + \sum_{j=2}^{N} \lambda_j^m a_j v_j = \pi + \sum_{j=2}^{N} \lambda_j^m a_j v_j.$$

Thus, the distance between the state at the mth step, $P^m p^0$, and the stationary regime π is

$$\|P^m p^0 - \pi\|^2 = \left\| \sum_{j=2}^{N} \lambda_j^m (a_j v_j) \right\|^2.$$

The sum on the right-hand side is a sum over the fixed vectors $a_j v_j$ whose coefficients diminish exponentially like λ_j^m. (Recall that the $\lambda_j, j \geq 2$, all have length less than 1.) This sum is finite, and therefore converges to zero as $m \to \infty$. Thus, $p^m = P^m p^0 \to \pi$ as $m \to \infty$. □

Return to our impartial web surfer. The properties of Markov chains can be interpreted as saying that if the impartial web surfer continues to crawl through the web long enough, he will find himself on each of the pages with a probability that approaches those given by the stationary regime π, where π is the normalized eigenvector associated with eigenvalue 1.

We are now ready to make the connection between the vector π and the *PageRank* ordering of pages.

Definition 9.5 *(1) The score given to page i in the (simplified) PageRank algorithm is the corresponding coefficient π_i from the vector π.*
(2) We sort the pages based on their PageRank scores, with the largest coming first.

The initial example with the web of five pages (Figure 9.2) allows us to obtain an understanding of this score. The norms $|\lambda_i|$ of the eigenvalues of the associated matrix P are 1 with multiplicity 1, and 0.70228 and 0.33563 each with multiplicity 2. Only the eigenvalue 1 is a real number. The eigenvector associated with the eigenvalue 1 is $(12, 16, 9, 1, 3)$, which, when normalized, yields

$$\pi = \frac{1}{41} \begin{pmatrix} 12 \\ 16 \\ 9 \\ 1 \\ 3 \end{pmatrix}.$$

This tells us that given a sufficiently long walk, the impartial web surfer would visit page B the most often, with 16 out of 41 steps leading to it. Similarly, he would nearly completely ignore page D, visiting it once per 41 steps on average.

What is the final order given to the pages? Page B is ranked number 1, which means that it is the most important page. Page A is ranked second, followed by pages C, E, and finally, the least important, page D.

There is an another way in which *PageRank* scores may be interpreted: each page gives its *PageRank* score to all of the pages it links to. Return to the vector $\pi = (\frac{12}{41}, \frac{16}{41}, \frac{9}{41}, \frac{1}{41}, \frac{3}{41})$. Page D is linked to only once, from page E. Since E has a score of $\frac{3}{41}$ and three outbound links that must share this value, D receives a final score of one-third that of E, $\frac{1}{41}$. Three pages point to page B: pages A, C, and E. The three pages have respective scores of $\frac{12}{41}$, $\frac{9}{41}$, and $\frac{3}{41}$. Page A has only one outgoing link, while pages C and E have three each. Thus, the score of page B is

$$\text{score } (B) = 1 \cdot \frac{12}{41} + \frac{1}{3} \cdot \frac{9}{41} + \frac{1}{3} \cdot \frac{3}{41} = \frac{16}{41}.$$

Why does the order implied by the *PageRank* scores give a reasonable ordering of the pages on the web? Mostly because it entrusts the users of the web itself to make the decisions as to which pages are better than others. Similarly, it ignores completely what the creator thinks of the importance of his own page. Moreover, the effect is cumulative. An important page that links to a few other pages can "transmit" its importance to these other pages. Thus, users display their confidence by linking to certain pages, and by doing so they transmit part of their score to these pages in the *PageRank* algorithm. This phenomenon has been named "*collaborative trust*" by the *PageRank* inventors.

9.3 An Improved *PageRank*

The algorithm described in the last section is not quite useable as is. There are two rather evident difficulties that must first be overcome.

The first is the existence of pages that have no outgoing links. The absence of links may come from the fact that *Google*'s web-spider has not yet indexed the destinations of the links, or that the page simply does not have any links. Thus, the impartial web crawler that arrives at this page would be forever caught there. One way of avoiding this problem is simply to ignore such pages, and remove them (and all the links leading to them) from the web. The stationary regime may then be calculated. After this is done, it is possible to assign scores to these pages by "transmitting" importance from all of the pages that link to them, as discussed at the end of the previous section:

$$\sum_{i=1}^{n} \frac{1}{l_i} r_i,$$

where l_i is the number of links issued by the ith page leading to the dead-end page, and r_i is the calculated importance of the ith page. The next problem shows that this somewhat crude approach offers only a partial solution.

The second difficulty resembles the first, but it is not quite so easy to fix. An example is depicted in the web of Figure 9.4. The web consists of the five pages from our original example, plus two others that are connected to the original web by a single link from page D. We saw in the last section that the impartial web surfer did not spend much time on page D. However, all the same, he did occasionally visit it, spending $\frac{1}{41}$ of his time there. What happens in this new modified web? Each time the web surfer visits page D he will choose to go to page A half of the time, while the other half of the time he will choose page F. If he chooses the latter option, then he can never return to the original pages A, B, C, D, or E. It is not surprising then that the stationary regime π of this new web is $\pi = (0, 0, 0, 0, 0, \frac{1}{2}, \frac{1}{2})^t$. In other words, the pages F and G "absorb" all of the importance that should have been divided up among the other pages! (Watch out! In this example, (-1) is also an eigenvalue of P, which means that P^n no longer approaches the matrix with columns π as $n \to \infty$.) Can we solve this problem as before, by simply removing the offending pages from the web? This is

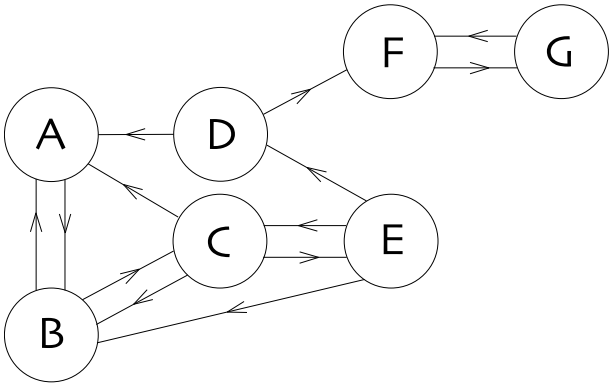


Fig. 9.4. A web of seven pages.

not really the best approach, because in the real world, parts of the graph that act in such a manner may themselves consist of thousands of pages that must also be ranked. Additionally, we can easily imagine that any impartial web surfer caught in such a loop ($F \to G \to F \to G \to \cdots$) would grow bored and decide to visit another part of the web at random. Thus, the inventors of the *PageRank* algorithm suggest adding to P a matrix Q that represents the "taste" of the impartial web surfer. The matrix Q would itself be a transition matrix, and the final transition matrix used in calculations would be

$$P' = \beta P + (1 - \beta)Q, \qquad \beta \in [0, 1].$$

Note that P' is itself a transition matrix: the coefficients of each column in P' still sum to 1. (Exercise!) The balance between the "taste" of the web surfer (represented by the matrix Q) and the structure of the web itself (represented by the matrix P) is controlled by the parameter β. When $\beta = 1$ the tastes of the web surfer are ignored, and the structure of the web may again cause certain pages to absorb all of the importance. Similarly, when $\beta = 0$ the tastes of the web surfer dominate, and the manner in which the web surfer visits pages has absolutely no relation to the structure of the web itself.

But how does *Google* guess the tastes of the web surfer? In other words, how do they choose the matrix Q? In the *PageRank* algorithm the matrix Q is chosen in the most democratic way possible. They give each page in the web an equal probability of transition. If the web consists of N pages, then every element of the matrix Q will be $\frac{1}{N}$: $q_{ij} = \frac{1}{N}$. This means that if the web surfer finds himself stuck in the pair of pages (F, G) from Figure 9.4 he has a probability $\frac{5}{7} \times (1 - \beta)$ of escaping at each step. In their original paper, the inventors of *PageRank* suggested a value of $\beta = 0.85$, forcing the impartial web surfer to ignore the links of the page and choose his next destination using his "taste" roughly 3 times out of 20.

This variation on the algorithm from the previous section, with the matrix Q and the parameter β, is the final algorithm that the inventors called *PageRank*. Several of its properties will be explored in the exercises.

The *PageRank* algorithm first proposed by academics has since been patented. Two of the inventors, Sergey Brin and Larry Page, founded the company *Google* in 1998, while they were both still in their twenties. Since this time, *Google* has gone public and is openly traded on the stock market. It is thus difficult to know what changes and improvements have been made to the algorithm, since it has fallen under commercial secrecy. We can piece together a few bits of information, however. *PageRank* is one of the algorithms for ranking web pages, but it is probably not the only one, or many small changes might have been brought to the original algorithm. *Google* claims to catalog approximately 10 billion web pages, so we can imagine that the number N of rows in the matrix P is of the same order. Thus, in order to determine the *PageRank* of each of these pages, they must calculate an eigenvector of an $N \times N$ matrix, where $N \approx 10{,}000{,}000{,}000$. But solving the equation $\pi = P\pi$ (or more precisely $\pi = P'\pi$), where P is a $10^{10} \times 10^{10}$ matrix is not an easy task. In fact, according to C. Moler, the founder of *Matlab*, it might be one of the largest matrix problems done by computers. (For an up-to-date discussion of search engines and particularly *PageRank* (as of 2006), see [2].) This task is probably done monthly. What is the algorithm used? Is the matrix $(I - P)$ row-reduced first? Or is π obtained by the repeated application $P^m p^0$ of P on some set of initial conditions p^0 (power method)? Or is it by an algorithm targeting first subsets of pages of the web that are connected by many links (method of aggregation)? It seems that the two latter methods are natural for the problem. But the exact details of improvements to *PageRank* and its computation since the founding of *Google* remain secret.[5]

The sequence of events (invention of the *PageRank* algorithm, dissemination of the original article, granting of the patent, creation of *Google*, widespread adoption of the *Google* search engine, ...) was optimal: on one side, the scientific community was made aware of the details of the algorithm, and on the other, the founders of *Google* had several months to get their company started and to reap the rewards of their invention. In knowing the basic details, researchers (with the exception of those that work for *Google* directly and are shrouded in corporate secrecy) can freely discuss improvements to the algorithm and its finer points, for example, how to efficiently take into account personal user preferences, how to benefit from pages that are strongly linked to each other, and how to restrict searches to a particular domain of human activity.

[5]Search requests made to *Google* are filled by a cluster of roughly 22,000 computers (as of December 2003) working with the help of the Linux operating system. Response times are rarely greater than a half-second!

9.4 The Frobenius Theorem

In order to describe and demonstrate the Frobenius theorem, we need to introduce the notion of matrices with nonnegative elements.[6] We will distinguish three cases. If P is an $n \times n$ matrix, then we say that

- $P \geq 0$ if $p_{ij} \geq 0$ for all $1 \leq i, j \leq n$;
- $P > 0$ if $P \geq 0$ and at least one of the p_{ij} is positive;
- $P \gg 0$ if $p_{ij} > 0$ for all $1 \leq i, j \leq n$.

We will use the same notation for vectors $x \in \mathbb{R}^n$. Finally, the notation $x \geq y$ signifies that $x - y \geq 0$. These "inequalities" are likely not very familiar. To help clarify we present a few simple examples of their use. To begin, if $P \geq 0$ and $x \geq y$, then it follows that $Px \geq Py$. This is due to the fact that since $(x - y) \geq 0$ and $P \geq 0$, the matrix product $P(x - y)$ consists only of sums of nonnegative elements. Therefore the entries of the vector $P(x - y) = Px - Py$ are nonnegative, and finally $Px \geq Py$. The second example is proved similarly and left as an exercise: if $P \gg 0$ and $x > y$, then $Px \gg Py$.

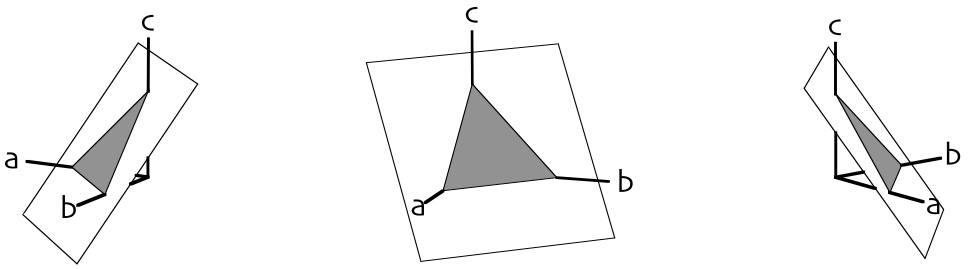


Fig. 9.5. Three points of view of the simplex created by the vectors $x = (a, b, c)$. The plane $a + b + c = 1$ is represented by the white square, while the simplex $(a, b, c \geq 0)$ is represented by the gray triangle.

When $P \geq 0$ we may define a set $\Lambda \subset \mathbb{R}$ of points λ that satisfy the following property: there exists a vector $x = (x_1, x_2, \ldots, x_n)$ such that

$$\sum_{1 \leq j \leq n} x_j = 1, \qquad x > 0, \qquad \text{and} \qquad Px \geq \lambda x. \tag{9.4}$$

For example, if $n = 3$, the condition $x > 0$ places the point $x = (a, b, c)$ in the octant whose points consist of nonnegative coordinates. At the same time, the constraint $a+b+c = 1$ describes a plane surface. Thus the point x is constrained to the intersection of these two sets, as depicted in Figure 9.5. In this figure the octant is depicted by the

[6] Recall that "nonnegative" means "positive or zero."

three axes, and the plane is depicted by a white square. The intersection of the two is depicted by a gray triangle. In the case of finite dimension n, the constructed object is called a simplex. (What does this simplex look like for $n = 2$? And for $n = 4$? Exercise!) The most important property of the simplex is that it is a compact set, in other words, it is both closed and bounded. For each point in the simplex we can calculate Px, which, by our earlier observation, satisfies $Px \geq 0$. Thus it is possible to find $\lambda \geq 0$ such that $Px \geq \lambda x$. (It can also happen that $\lambda = 0$; for example if $P = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and $x = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, then $Px = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right) \geq \lambda \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$ can hold only when $\lambda = 0$.)

Proposition 9.6 *Let $\lambda_0 = \sup_{\lambda \in \Lambda} \lambda$. Then $\lambda_0 < \infty$. Moreover, if $P \gg 0$, then $\lambda_0 > 0$.*

PROOF: Suppose that $M = \max_{i,j} p_{ij}$, the largest element of the matrix P. Then for all x that satisfy $\sum_j x_j = 1$ and $x > 0$, we have that

$$(Px)_i = \sum_{1 \leq j \leq n} p_{ij} x_j \leq \sum_{1 \leq j \leq n} M x_j = M, \qquad \text{for all } i.$$

Since at least one of the entries of x, call it x_i, must satisfy $x_i \geq \frac{1}{n}$, the condition $Px \geq \lambda x$ thus requires that $M \geq (Px)_i \geq \lambda x_i \geq \lambda \frac{1}{n}$. Since this holds for all $\lambda \in \Lambda$, we have that $\lambda_0 = \sup_\Lambda \lambda \leq Mn$. Suppose further that $P \gg 0$, and let $m = \min_{ij} p_{ij}$ be the smallest element of P. Then for $x = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ we have that $(Px)_i = \sum_j p_{ij} \frac{1}{n} \geq (mn)\frac{1}{n} = (mn)x_i$ and therefore $Px \geq (mn)x$ and $\lambda_0 \geq mn > 0$. □

Theorem 9.7 (Frobenius) *Let $P > 0$ and λ_0 be as defined above.*

(a) λ_0 is an eigenvalue of P and it is possible to choose an associated eigenvector x^0 such that $x^0 > 0$;
(b) if λ is another eigenvalue of P, then $|\lambda| \leq \lambda_0$.

PROOF:[7] (a) We will prove this statement in two steps, (a1) and (a2).

(a1) If $P \gg 0$ then there exists $x^0 \gg 0$ such that $Px^0 = \lambda_0 x^0$.
To prove this first statement we consider a sequence $\{\lambda_i < \lambda_0, i \in \mathbb{N}\}$ of elements from Λ that converges to λ_0, and the associated vectors $x^{(i)}, i \in \mathbb{N}$, which satisfy (9.4):

$$\sum_{1 \leq j \leq n} x_j^{(i)} = 1, \qquad x^{(i)} > 0, \qquad \text{and} \qquad Px^{(i)} \geq \lambda_i x^{(i)}.$$

Since the points $x^{(i)}$ all belong to the compact simplex, it must contain an accumulation point, and we may choose a subsequence $\{x^{(n_i)}\}$, with $n_1 < n_2 < \cdots$, that is convergent to this point. Let x^0 be the limit of this subsequence:

$$\lim_{i \to \infty} x^{(n_i)} = x^0.$$

[7] The proof given here is that of Karlin and Taylor, presented in [1].

Note that x^0 is itself in the simplex and therefore satisfies $\sum_j x_j^0 = 1$ and $x^0 > 0$. Finally, since $P(x^{(n_i)} - \lambda_i x^{(n_i)}) \geq 0$, we have that $Px^0 \geq \lambda_0 x^0$. We will now show that $Px^0 = \lambda_0 x^0$. Suppose that $Px^0 > \lambda_0 x^0$. Since $P \gg 0$, by multiplying both sides of $Px^0 > \lambda_0 x^0$ by P and defining $y^0 = Px^0$, we obtain that $Py^0 \gg \lambda_0 y^0$. (Exercise: work through the details of this step.) Since this inequality is strict for all entries, there exists an $\epsilon > 0$ such that $Py^0 \gg (\lambda_0 + \epsilon)y^0$. By normalizing y^0 such that $\sum_j y_j^0 = 1$ we can deduce that $\lambda_0 + \epsilon \in \Lambda$ and that λ_0 *cannot* be the supremum: a contradiction. Thus it must be that $Px^0 = \lambda_0 x^0$. Since $P \gg 0$ and $x^0 > 0$, we have that $Px^0 \gg 0$. In other words, $\lambda_0 x^0 \gg 0$, and finally $x^0 \gg 0$ since $\lambda_0 > 0$.

(a2) If $P > 0$ then there exists $x^0 > 0$ such that $Px^0 = \lambda_0 x^0$.
Consider an $n \times n$ matrix E whose entries are all 1. Observe that if $x > 0$ then $(Ex)_i = \sum_j x_j \geq x_i$ for all i, and therefore $Ex \geq x$. If $P > 0$, then $(P + \delta E) \gg 0$ for all $\delta > 0$, and (a1) can be applied to this matrix. Let $\delta_2 > \delta_1 > 0$, and let $x \in \mathbb{R}^n$ be such that $x > 0$ and $\sum_j x_j = 1$. If $(P + \delta_1 E)x \geq \lambda x$, we have that

$$(P + \delta_2 E)x = (P + \delta_1 E)x + (\delta_2 - \delta_1)Ex \geq \lambda x + (\delta_2 - \delta_1)x,$$

and therefore the function $\lambda_0(\delta)$ whose existence is predicted by applying (a1) to the matrix $(P + \delta E)$ is an increasing function of δ. Moreover, $\lambda_0(0)$ is the λ_0 associated with the matrix P. Construct a decreasing positive sequence $\{\delta_i, i \in \mathbb{N}\}$ converging to 0. By (a1) it is possible to find the $x(\delta_i)$ satisfying $(P + \delta_i E)x(\delta_i) = \lambda_0(\delta_i)x(\delta_i)$, where $x(\delta_i) \gg 0$ and $\sum_j x_j(\delta_i) = 1$. Since all of these vectors lie within the described simplex, there exists a subsequence $\{\delta_{n_i}\}$ such that $x(\delta_{n_i})$ converges toward an accumulation point x^0. This vector must satisfy $x^0 > 0$ and $\sum_j x_j^0 = 1$. Let λ' be the limit of $\lambda_0(\delta_{n_i})$. Since the sequence δ_i is decreasing and $\lambda_0(\delta)$ is an increasing function, $\lambda' \geq \lambda_0(0) = \lambda_0$. Since $P + \delta_{n_i} E \to P$ and $(P + \delta_{n_i} E)x(\delta_{n_i}) = \lambda_0(\delta_{n_i})x(\delta_{n_i})$, taking the limit of both sides yields $Px^0 = \lambda' x^0$, and by the definition of λ_0, it must be that $\lambda' \leq \lambda_0$. Hence $\lambda' = \lambda_0$, completing the proof of (a).

(b) Let $\lambda \neq \lambda_0$ be another eigenvalue of P, and z an associated nonzero eigenvector. Then $Pz = \lambda z$, which is to say

$$(Pz)_i = \sum_{1 \leq j \leq n} p_{ij} z_j = \lambda z_i.$$

In taking the norm of both sides we get

$$|\lambda|\,|z_i| = \left| \sum_{1 \leq j \leq n} p_{ij} z_j \right| \leq \sum_{1 \leq j \leq n} p_{ij} |z_j|$$

and therefore

$$P|z| \geq |\lambda|\,|z|,$$

where $|z| = (|z_1|, |z_2|, \ldots, |z_n|)$. By normalizing $|z|$ appropriately, we can ensure that it lies in the simplex and therefore $|\lambda| \in \Lambda$. Hence, by the definition of λ_0, it follows that $|\lambda| \leq \lambda_0$. □

Corollary 9.8 *If P is a Markov chain transition matrix, then $\lambda_0 = 1$.*

PROOF: Consider $Q = P^t$. Then $\sum_j q_{ij} = 1$ for all i. Since $P > 0$, we have also that $Q > 0$. By part (a) of the Frobenius theorem there exist λ_0 and x_0 (where $x^0 > 0$ and $\sum_j x_j^0 = 1$) such that $Qx^0 = \lambda_0 x^0$. Since $x^0 > 0$, the largest entry of x^0, call it x_k^0, is positive and satisfies

$$\lambda_0 x_k^0 = (Qx^0)_k = \sum_{1 \leq j \leq n} q_{kj} x_j^0 \leq \sum_{1 \leq j \leq n} q_{kj} x_k^0 = x_k^0.$$

From this we may deduce that $\lambda_0 \leq 1$. Property 9.2 showed that 1 is an eigenvalue of P (and of Q as well) and therefore $\lambda_0 \geq 1$, from which the desired result follows immediately. □

Property 9.3 follows directly from the Frobenius theorem and Corollary 9.8.

9.5 Exercises

1. **(a)** For the web given in Figure 9.2, use the transition matrix to calculate the probabilities of the impartial web surfer being on pages A, B, C, D, and E after his third step. Compare these results to the stationary regime π for this transition matrix.
(b) What are the probabilities of being on the pages A, B, C, D, and E after the first step if the impartial web surfer starts at page E? What about after the second step?

2. **(a)** Let

$$P = \begin{pmatrix} 1-a & b \\ a & 1-b \end{pmatrix} \qquad \text{with } a, b \in [0, 1].$$

Show that P is a Markov chain transition matrix.
(b) Calculate the eigenvalues of P as a function of (a, b). (One of the two eigenvalues must be 1 by Property 9.2.)
(c) Which values for a and b lead to a second eigenvalue λ satisfying $|\lambda| = 1$? Draw the corresponding webs.

3. **(a)** Give the transition matrix P associated with the web shown in Figure 9.6.
(b) Show that the three eigenvalues of P have absolute values of 1.
(c) Find (or better yet, intuit) the page ranking that would be assigned by the simplified *PageRank* algorithm.

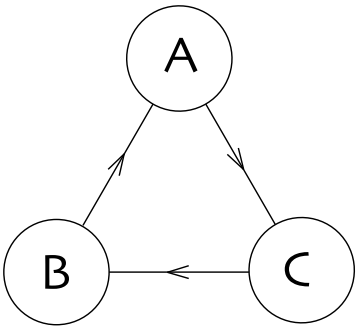


Fig. 9.6. The circular web of Exercises 3 and 4.

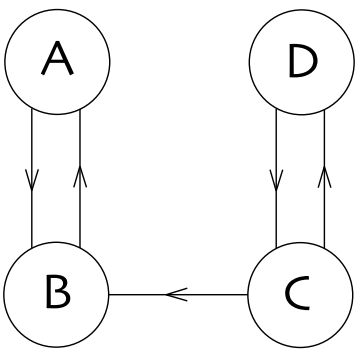


Fig. 9.7. The web of Exercise 5, with two pairs connected by a single link.

Note: We remark that this web does not satisfy hypothesis *(i)*, which was used to obtain Property 9.4.

4. For the web shown in Figure 9.6, an impartial web surfer starts at page A at step $n = 1$. Can you give the probabilities $P(X_n = A)$, $P(X_n = B)$, and $P(X_n = C)$ for all n?

5. **(a)** Consider the web illustrated in Figure 9.7. Intuitively, which of the pairs of pages, (A, B) or (C, D), will be given a greater rank by the simplified *PageRank* algorithm?
(b) Find the page ranking assigned by the simplified *PageRank* algorithm.
(c) Find the stationary regime of the transition matrix used by the full *PageRank* algorithm: $P' = (1 - \beta)E + \beta P$. The matrix E is a 4×4 matrix in which all entries are $\frac{1}{4}$. For which value of β will the impartial web surfer spend one-third of his time visiting the pair (C, D)?

6. **(a)** Find the transition matrix representing the web shown in Figure 9.8.

(b) Assume that at step n, the probabilities of being on each page are equal: $P(X_n = A) = P(X_n = B) = P(X_n = C) = P(X_n = Z) = \frac{1}{4}$. What is the probability of being on page Z at step $n+1$?
(c) Calculate the stationary regime π of this transition matrix. Will an impartial web surfer spend more time on page A or on page Z?

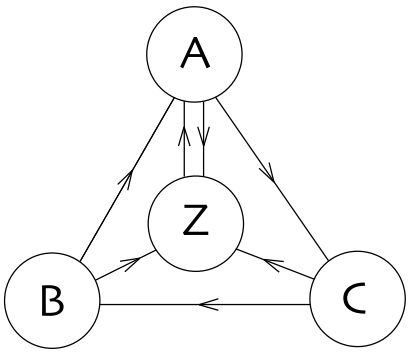


Fig. 9.8. A web of four pages, for Exercise 6.

7. Consider the web of Figure 9.9.
(a) Write out the associated Markov chain transition matrix.
(b) If we start on page B, what is the probability that we will be on page A after 2 steps?
(c) If we start on page B, what is the probability that we will be on page D after 3 steps?
(d) Calculate the stationary regime for this web, and the rank of each page using the simplified *PageRank* algorithm. Which page is the most important?

8. This exercise aims to show that hypothesis *(ii)*, used in obtaining Property 9.4, does not always hold.
(a) Suppose that there are two "parallel" webs in existence. That is, two extremely large webs that never link to each other. Consider the transition matrix for these two webs taken together. This matrix will have a peculiar form. What is it?
(b) Show that the transition matrix P of this pair of parallel webs possesses two distinct eigenvectors with eigenvalue 1.

9. **(a)** Write a program, in *Maple*, *Mathematica*, or *Matlab* for example, that when given n will calculate a random vector $(x_1, x_2, \ldots, x_n)$ satisfying

$$x_i \in [0,1] \quad \text{for all } i \in T \qquad \text{and} \qquad \sum_i x_i = 1.$$

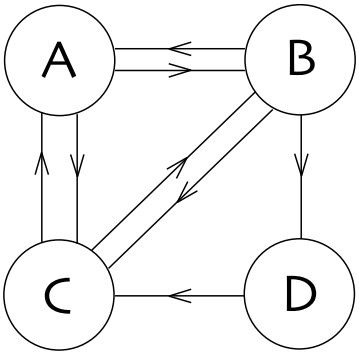


Fig. 9.9. The web for exercise 7

(Most modern programming languages offer functionality for generating pseudorandom numbers.)
(b) Extend your program to compute a random $n \times n$ matrix P such that each column of P sums to 1.
(c) Extend your program to calculate P^m when given an integer m.
(d) Generate several reasonably large matrices P (10×10, 20×20, or even bigger) and check whether the hypotheses of Property 9.4 hold. (Remark: If you are using a language like *C*, *Fortran*, or *Java*, you will have to find a library or write your own code to compute eigenvectors and eigenvalues. Such libraries can be difficult to integrate and use, and writing the code yourself is even harder. As such, you may prefer to use a mathematical computing package like *Maple*, *Mathematica*, or *Matlab*, which natively includes such functionality.)
(e) For a given random matrix P generated as above, at what value of m are all the columns of P^m approximately equal? Start by defining a reasonable criterion for "approximately equal."

10. **(a)** Imagine that you are a slightly villainous businessman who runs an online business. Propose some strategies for ensuring that your site will be assigned a higher importance by the *PageRank* algorithm.
(b) Now imagine that you are a young and ambitious researcher working for *Google*. Your job is to outflank the villainous businessmen of the world by preventing them from obtaining artificially inflated *PageRank* scores. Propose some strategies for countering their ploys.
Note: The original article [3] by Page et al. includes some discussion on the potential impact of commercial interests.

References

[1] S. Karlin and M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2nd edition, 1975.

[2] A.M. Langville and C.D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[4] S.M. Ross. *Stochastic Processes*. Wiley & Sons, 2nd edition, 1996. (A more advanced book than that of Karlin and Taylor [1].)

10

Why 44,100 Samples per Second?

This chapter may be covered in three or four hours, depending on the importance given to the proof in Section 10.4. It has been written for students that have not yet seen any Fourier analysis. As such, the prerequisites are modest: one-variable calculus and a familiarity with the concepts of convergence and, at the end of Section 10.4, of complex numbers. If the students are familiar with Fourier transforms, then the instructor may choose to include a proof of the sampling theorem, which we simply state without proof. (See Sections 8.1 and 8.2 of Kammler [2] or Exercise 60.16 of Körner [3] for a proof.) This subject offers ample opportunity for larger projects: students may continue their exploration through Exercises 13, 14, and 15, supplemented by topics chosen from Benson's book [1]; or if they are good with computers they may explore the many numerical experiments discussed in this chapter.

10.1 Introduction

This chapter explains the choice made by the engineers at Philips and Sony when they were defining the standard for the compact disc. It is possible to digitize sound signals. We have seen an example in Chapter 6: sound is simply a wave of pressure that may be interpreted as a continuous function of pressure versus time. When digitized, this continuous function is replaced by a step function, an example of which is shown in Figure 10.1. More formally, mathematicians call such a function piecewise constant. In digitizing sound, each step has the same width. Thus, the digitized function may be represented simply as the sequence of heights of the steps. The engineers at Philips and Sony decided to make each step have a width of $\frac{1}{44{,}100}$ of a second. This chapter explains why this particular value was chosen.

For somebody with little knowledge of the subject matter, this goal may seem somewhat trivial. However, as is often the case, the choice relied on knowledge from many diverse domains. Of course, the first question is quite basic: what is musical sound? A second equally basic question concerns human physiology: how does the human ear

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_10,

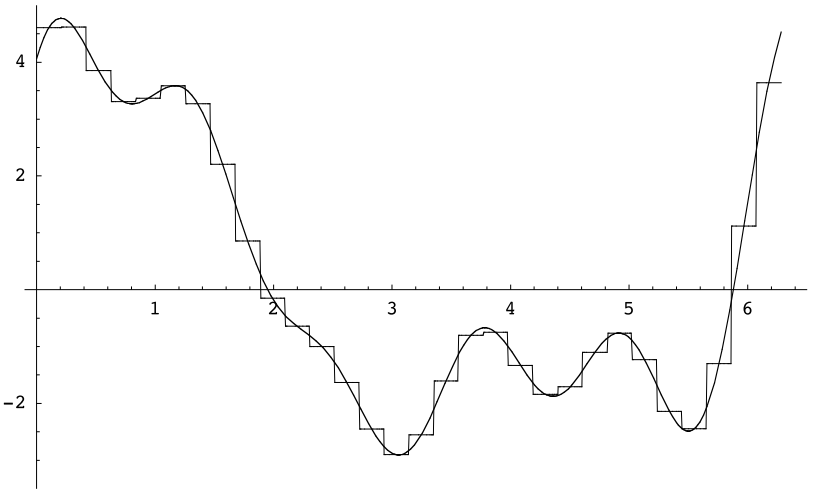


Fig. 10.1. A continuous "wave of pressure" function and a step function approximation of it.

react to sound waves? Finally, mathematics answers the third question: knowing what we do about the nature of sound and how the human ear interprets it, can we show that 44,100 samples per second is sufficient? The answer lies in the domain of mathematics known as *Fourier analysis*.

10.2 The Musical Scale

Sound is simply a wave of pressure. As with all waves, one of the most intuitive ways to represent and describe this wave is through a simple plot (an example of which is given in Figure 10.2). Two mathematical properties of this wave are related to how we perceive it as a sound: the *frequency* of the wave is related to the *pitch* of the sound, while the *amplitude* of the wave is related to the *volume*. Female voices are normally characterized by higher frequencies than those found in male voices. Similarly, the amplitude of the wave representing a song sung by Pavarotti is higher than that of one sung by most other people.

We will discuss the relation between wave amplitude and perceived volume in the next section. For now we will first discuss the relationship between frequency and perceived pitch. Even if many people have never taken a piano lesson, nearly all know that the low notes are on the left end of the keyboard, while the high notes are on the right. Figure 10.3 shows the layout of a modern piano keyboard. The notes C are

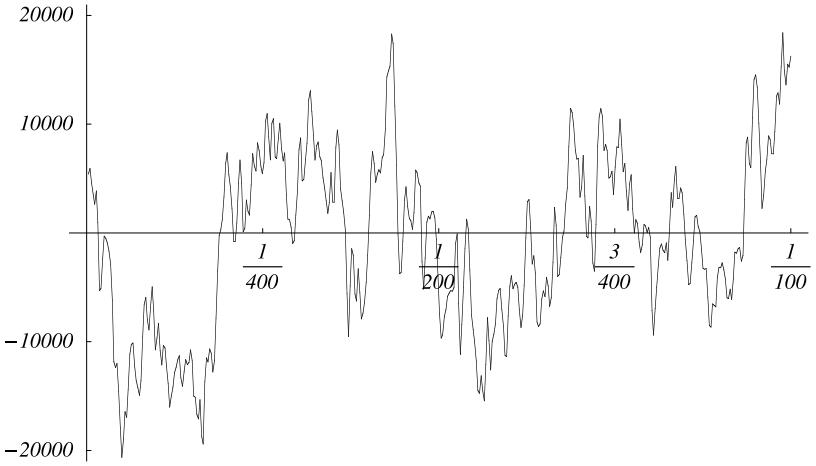


Fig. 10.2. The pressure wave corresponding to $\frac{1}{100}$ of a second of the last note of Beethoven's ninth symphony. On a compact disc each of the 441 steps in this wave is assigned an integer value in the range $[-2^{15}, 2^{15}-1]$, corresponding to the height of the step. The horizontal axis is time while the vertical axis is the height of the step ($2^{15} = 32{,}768$).

indicated. The occidental scale[1] consists of 12 distinct notes. On the white keys we find the notes C, D, E, F, G, A, and B, while on the black keys we find five notes falling between these. Each of these in-between notes can be called either of two names: C♯ or D♭, D♯ or E♭, F♯ or G♭, G♯ or A♭, and A♯ or B♭.[2] Musicians know that the notes D♯ and E♭, like the two notes in each of the other pairs, are not exactly the same sound. The fact that they are considered the same note in the scale of the piano keyboard is a result of a compromise that we will discuss a little later. A modern keyboard has seven sets of these 12 notes. A further C is added at the extreme right, and a few notes are added at the extreme left. In all, there are 88 keys. Note that the ratio of the frequencies of two consecutive D's is 2. (This is actually true for all consecutive notes of the same name.) Later we will be interested in creating a linear representation of all of the frequencies. We will have to deform graphically the keyboard using a logarithmic transformation (see Figure 10.7).

[1] Other cultures have favored other scales. For example, Balinese gamelans are typically based on either a pentatonic or a heptatonic scale, containing five and seven notes respectively as compared to the 12 in the occidental scale.

[2] Why are certain keys white and others black? There is no scientific answer to this question. They are arranged to accommodate the occidental preference for playing in a given key, namely C major. Other cultures, such as the Japanese, prefer other keys, and it is likely that any keyboard-based instruments they would have constructed would have been laid out according to their preference. For our purpose we are not required to understand these cultural differences.

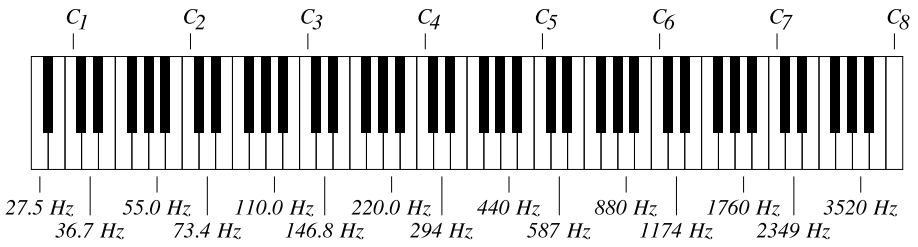


Fig. 10.3. A modern piano keyboard. The eight C keys are indicated as well as the frequencies of each of the D and A notes.

Why aren't there 88 different names for the 88 notes? The answer has mostly to do with physiology, but also a little with physics and mathematics. The physiology of perception shows that two people can sing the same song simultaneously while singing different notes, but still give the impression of singing the same note. We say that they are singing in *unison*. The *interval* between two consecutive notes with the same name is called an *octave*. On a keyboard these two notes are separated by precisely 12 notes, counting the last but not the first one. Notes at intervals of one or several octaves are perceived as almost the same. If these same two people choose to sing two notes with different names, then the result is perceived as slightly strange or discordant. (And if they could hear each other singing, they would quickly perceive this and alter their voice to fall back into unison. It takes a pair of good singers to deliberately maintain a nonoctave interval between their voices throughout an entire song.) The more physical and mathematical reason is that consecutive notes with the same name are arranged such that their frequencies maintain a ratio of two. As said before, the ratio between the frequency of a note and the same note one octave higher is exactly 2. Why do the human ear and brain prefer this factor of 2? Neither physics nor mathematics can answer this question![3]

This preference for powers of two in the ratio of frequencies is quite surprising. Even more surprising is that the ear and brain find a ratio of three equally pleasing. Notes whose frequencies have a ratio of three have an interval of one octave and a fifth. A *fifth* is an interval of seven consecutive notes on the keyboard, not counting the starting note. Notes must be counted consecutively, be they white or black keys. We can thus see that the notes C and G (with another C between them) are separated by an interval of one octave and a fifth.[4] Thus, notes separated by an octave and a fifth have a ratio of

[3]However, physiology does give some insight (see [8] for details).

[4]Why does a fifth correspond to 7 notes while an octave corresponds to 12? After all, the terms fifth and octave seem to suggest 5 and 8 respectively. The reason is again due to the predominant role of the white keys in the key of C major. From C to G there is a fifth: if C

three between their frequencies, while notes separated by only a fifth have a frequency ratio of $\frac{3}{2}$. (Exercise: convince yourself of this fact!)

All deviations from these pleasing ratios between frequencies, even minimal, are perceived easily by experienced musicians. However, tuning a piano while maintaining all of these ideal relationships is a mathematical impossibility. We now describe the root of the problem. The cycle of fifths is an enumeration of all the notes such that a note immediately after another will lie exactly one fifth to the right of the former note on a keyboard. Most good musicians are able to recite the cycle of fifths without even thinking. Starting at C the cycle of fifths is:

$$\begin{array}{cccccccccccc} \mathrm{C}_1, & \mathrm{G}_1, & \mathrm{D}_2, & \mathrm{A}_2, & \mathrm{E}_3, & \mathrm{B}_3, & \mathrm{F}_4\sharp, & \mathrm{C}_5\sharp, & \mathrm{G}_5\sharp, & \mathrm{D}_6\sharp, & \mathrm{A}_6\sharp, & \mathrm{E}_7\sharp, \\ & & & & & & & & (\mathrm{A}_5\flat) & (\mathrm{E}_6\flat) & (\mathrm{B}_6\flat) & (\mathrm{F}_7) \end{array}$$

and after the $\mathrm{E}_7\sharp$ (F_7) the cycle restarts at C_8.[5] (The octave associated with each of the notes, which we have indicated using a subscript, is not normally written. We have done so because it will be useful in the following discussion.)

For each fifth in this cycle, the frequency has been multiplied by $\frac{3}{2}$. From C1 to C8, the factor has been applied 12 times, making for an overall factor of $(\frac{3}{2})^{12}$. Similarly, there are 7 octaves between these two notes and the ratio between their frequencies is 2^7. Thus, we would expect that $(\frac{3}{2})^{12} = 2^7$, or equivalently $3^{12} = 2^{19}$. However, this identity is obviously false. A product of odd numbers remains odd, while a product of even numbers remains even. Thus, 3^{12} is odd and 2^{19} is even, and they cannot possibly be equal. However, the difference is not very large, since

$$3^{12} = 531{,}441 \qquad \text{and} \qquad 2^{19} = 524{,}288.$$

The error, at a little less than 2%, is not enormous considering that it is spread across 8 octaves. Renaissance-era musicians were aware of this difficulty. A well-trained ear is able to hear this error and finds perfect-integer frequency ratios (or ratios in which the denominator is a small integer) to be the most pleasing. This is the source of the error we have described. A solution proposed at the end of the seventeenth century was to tune a keyboard according to the following two rules: *(i)* the frequency ratio between notes separated by one octave is exactly 2, and *(ii)* the frequency ratio between successive notes on the keyboard should be constant. In this *temperament*, commonly called the *equal temperament* in the Western world, all intervals are false except the octave. It is the most democratic choice of distributing the error between possible intervals, and that which has been in common use for nearly three centuries. Thus, a well-tempered

---

is labeled 1, then the nearest G to its right would be labeled 5, counting only the white keys. Similarly, the next C to the right would be labeled with an 8.

[5] On a keyboard, the notes in parentheses coincide with the notes above them. Violinists, who can choose the exact frequencies of their notes with their left-hands, actually distinguish between these notes. Instead of restarting the cycle at C they continue it with $\mathrm{B}\sharp$, which pianists identify with a C.

tuning of a piano is perfectly and precisely false.[6] (For a discussion of the history of temperaments by a mathematician, see Benson [1].)

Can we determine the frequencies of the notes on a modern piano? Not yet, because we are still missing one important piece of information. In fact, the entire discussion up until this point has been concerned only with ratios of frequencies. We must still specify the frequency of a single note so that the rest of them may be determined. It has been traditional for nearly a century to tune the first A right of the center of the piano to 440 Hz,[7] meaning that the fundamental vibration of the note oscillates 440 times per second. An octave sees a doubling of the frequency. Since there are 12 intervals in an octave (for example between two C's or two A's) and the ratios of the frequencies for all must be equal, each of the 12 intervals must represent a frequency increase by a factor of $\sqrt[12]{2}$. Between the A vibrating at 440 Hz and the E just above it the frequency ratio is therefore $\sqrt[12]{2}^7 \approx 1.49831$, which is very close to the ideal of $\frac{3}{2} = 1.5$. The frequency of this E is therefore $\sqrt[12]{2}^7 \times 440 \text{ Hz} \approx 659.26$ Hz, which is very close to the "true" value of 660 Hz.

10.3 The Last Note of Beethoven's Last Symphony: A Quick Introduction to Fourier Analysis

Can we know what notes are on a compact disc without listening to it? Is it possible to read the 44,100 integers in a second of music and determine what notes are being played? This is what we aim to do in this section.

We will focus our attention on a quarter of a second of music taken from the last note of the last movement of Ludwig van Beethoven's ninth symphony. (In most performances of this piece this note is just slightly longer than a quarter of a second.) This choice is particularly appropriate. As the story is told, in establishing the standard for the compact disc, engineers made every possible effort to ensure that this symphony would fit on a single disc [6]. Although the length of this piece varies from performance to performance, some of them last as long as 75 minutes, such as that conducted by Karajan. This is why a compact disc can hold just a little more than 79 minutes. Another reason for our choice is that the last note of this symphony is particularly easy to study mathematically, since the entire orchestra plays the same note, D, at the same time. (Even though the musicians are all playing the same note (these notes are all D), they are actually being played in a variety of octaves.) For those who can read music, the

[6] The title of Johann Sebastian Bach's two books of preludes and fugues (The well-tempered clavier) highlights the fact that temperament was a hot topic at the beginning of the eighteenth century.

[7] The measure of frequency is the hertz, abbreviated Hz. One hertz corresponds to one vibration per second. The choice of 440 Hz for A is arbitrary. Certain musicians and orchestras are distancing themselves from this standard, with most of them increasing the reference frequency.

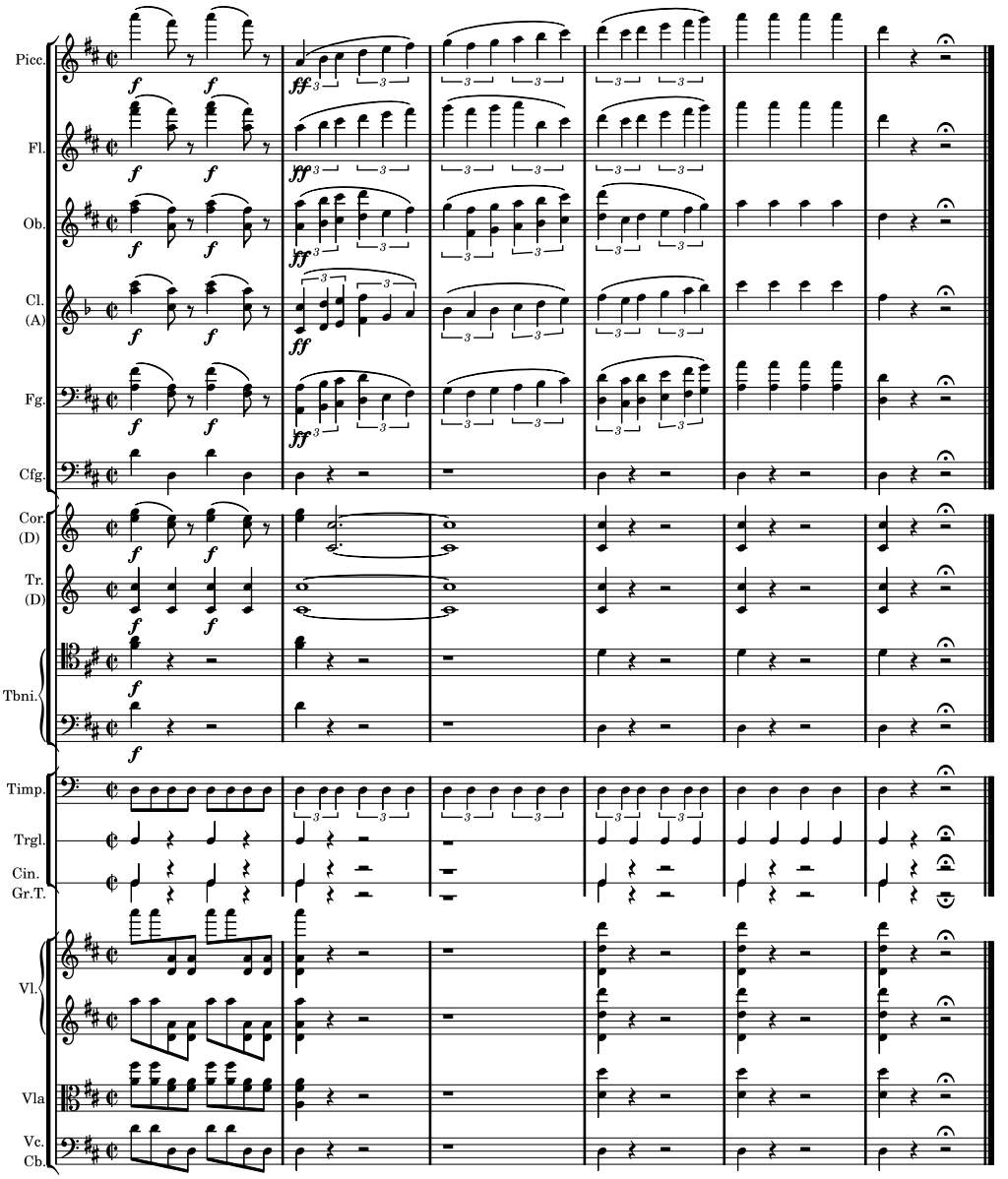


Fig. 10.4. The last page of Beethoven's ninth symphony.

last page can be found in Figure 10.4. Each line represents a group of instruments, with piccolo and flutes at the top, and cellos and contrabass at the bottom. The triangles and cymbals are capable of producing only one note (or sound); thus they are accorded only a single line on the score. All of the other instruments, including the timpani (marked "Timp." on the score), can produce a variety of notes, thus they use a five-line staff. Time flows from left to right, and all notes appearing along a given vertical line are played simultaneously. The last note is found in the rightmost column. In this column we will find only D notes, covering every D on a piano except for the lowest two. (Certain families of instruments seem to play notes other than D. For example, the note written for the clarinets ("Cl." on the score) is an F. But this will sound as a D! The reason for the discrepancy between the note written and that produced lies in the history of the development of the instrument. After much experiment, it was agreed that a given length for the tube of the clarinet gave the best sound quality over all its register (all its spectrum). Unfortunately, it also gave queer fingerings for the most common notes. The solution was to relabel the notes: when a clarinet plays the note written as a C, the frequency of the sound emitted is that of the B♭. We must therefore ask the clarinets to play an F in order that we hear a D. Composers routinely do this transposition for these instruments.)

Recall that stereo recordings contain two tracks, allowing the listener to perceive the spatial spread of the sound. We will limit ourselves to a single one of these two tracks. The quarter of a second that we will study contains $\frac{44{,}100}{4} = 11{,}025$ samples. The first 10 of these 11,025 integers are 5409, 5926, 4634, 3567, 2622, 3855, 948, −5318, −5092, and −2376, and the first 441 samples (giving one-hundredth of a second of music) are shown in Figure 10.2. How can we possibly mathematically "listen" to the note being played?

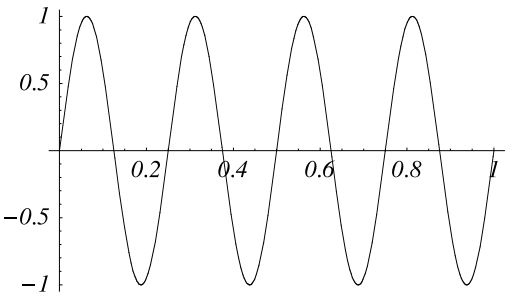


Fig. 10.5. A simple sound wave (a pure sound without harmonics).

Example 10.1 *We begin with a very simple example. Suppose that rather than analyzing the wave shown in Figure 10.2 we consider a sound $f(t)$ containing only a single frequency, as shown in Figure 10.5. We remark that there are exactly four complete*

cycles of this sinusoidal wave occurring in our one-second sample; thus the wave corresponds to a frequency of 4 *Hz. Thus,* $f(t) = \sin(4 \cdot 2\pi t)$. *It is easy to see this by looking at the figure, but how to do so mathematically? The answer to this question is given by Fourier analysis. The basic idea is to compare the sound wave* $f(t)$ *to all of the cosine and sine waves with integer frequencies (those whose frequencies are integer multiples of* 1 *Hz).*

Fourier analysis. Fourier analysis allows us to calculate the component of the sound wave that has a frequency of k Hz and to reconstruct the original wave from this set of components. The component with frequency k is given by the pair of coefficients c_k and s_k. The formula for these Fourier coefficients is given by:

$$c_k = 2\int_0^1 f(t)\cos(2\pi kt)\,dt, \qquad k = 0, 1, 2, \ldots, \tag{10.1}$$

$$s_k = 2\int_0^1 f(t)\sin(2\pi kt)\,dt, \qquad k = 1, 2, 3, \ldots. \tag{10.2}$$

(Exercise 3 explains why we require two coefficients to describe the component for a single frequency.)

EXAMPLE 1 (CONTINUED) *We start to calculate the coefficients* c_k *and* s_k *for the function* $f(t)$ *of Example 10.1. The coefficient* c_0 *is calculated by multiplying* $\cos 2\pi kt$ *(with* $k = 0$*) and* $f(t)$*, and then integrating the resulting function over the one-second interval. Since* $\cos 2\pi kt = 1$ *for* $k = 0$*, the coefficient* c_0 *is given by*

$$c_0 = 2\int_0^1 f(t)\,dt.$$

However, $f(t)$ *is a sinusoidal curve and the area under this curve between* $t = 0$ *and* $t = 1$ *is clearly zero. (Recall that the area between the* t*-axis and the curve is negative when* $f(t)$ *is negative.) Thus*

$$c_0 = 0.$$

Now consider s_1*:*

$$s_1 = 2\int_0^1 f(t)\sin 2\pi t\,dt.$$

The product of $\sin 2\pi t$ *and* $f(t)$ *is shown in Figure 10.6. Observe that* $f(t) = f(t + \frac{1}{2})$ *and that* $\sin 2\pi t = -\sin 2\pi(t + \frac{1}{2})$*, implying* $f(t)\sin 2\pi t = -(f(t + \frac{1}{2})\sin 2\pi(t + \frac{1}{2}))$ *for* $t \in [0, \frac{1}{2}]$*. Thus, the integral of* $f(t)\sin 2\pi t$ *will be zero:*

$$s_1 = 0.$$

Can we repeat this same procedure for all of the $c_k, k = 0, 1, 2, \ldots$*, and all of the* $s_k, k = 1, 2, 3, \ldots$*? It seems that we need a more efficient method for calculating these coefficients, since the graphical method will be difficult to use for all* k*.*

The following proposition gives us the necessary tools for this calculation.

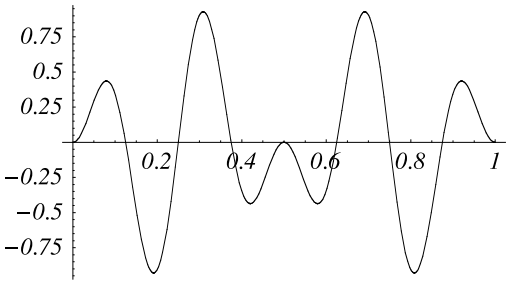


Fig. 10.6. The product of $f(t)$ and $\sin 2\pi t$.

Proposition 10.2 *Let $m, n \in \mathbb{Z}$. The Kronecker delta function $\delta_{m,n}$ is defined as follows: it takes the values* 1 *if $m = n$ and* 0 *otherwise. Thus*

$$2\int_0^1 \cos(2\pi mt)\ \cos(2\pi nt)\,dt = \delta_{m,n} + \delta_{m,-n}; \tag{10.3}$$

$$\int_0^1 \cos(2\pi mt)\ \sin(2\pi nt)\,dt = 0; \tag{10.4}$$

$$2\int_0^1 \sin(2\pi mt)\ \sin(2\pi nt)\,dt = \delta_{m,n} - \delta_{m,-n}. \tag{10.5}$$

PROOF: Let

$$I_1 = \int_0^1 \cos(2\pi mt)\cos(2\pi nt)\,dt,$$
$$I_2 = \int_0^1 \cos(2\pi mt)\sin(2\pi nt)\,dt,$$
$$\text{and} \qquad I_3 = \int_0^1 \sin(2\pi mt)\sin(2\pi nt)\,dt.$$

To calculate these integrals recall the identities

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta,$$
$$\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta,$$
$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta,$$
$$\sin(\alpha-\beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta.$$

By adding the first two of these we find that

$$2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta).$$

Thus

$$\begin{aligned} 2I_1 &= 2\int_0^1 \cos(2\pi mt)\cos(2\pi nt)\,dt, \\ &= \int_0^1 (\cos(2\pi(m+n)t) + \cos(2\pi(m-n)t))\,dt, \end{aligned}$$

which is simple to integrate. If $m+n \neq 0$ and $m-n \neq 0$, then

$$2I_1 = \left(\frac{\sin(2\pi(m+n)t)}{2\pi(m+n)} + \frac{\sin(2\pi(m-n)t)}{2\pi(m-n)}\right)\Bigg|_0^1 = 0,$$

since m and n are integers and $\sin \pi p = 0$ if p is an integer. On the other hand, if $m+n=0$ or $m-n=0$, the above evaluation is false, since one of the denominators is zero. (If m and n are nonnegative integers, then $m+n=0$ can happen only when $m=n=0$.) But if $m-n=0$ then the second term $\cos(2\pi(m-n)t)$ of the integral is equal to 1, and therefore

$$\int_0^1 \cos 2\pi(m-n)t\,dt = 1.$$

Hence

$$2I_1 = 2\int_0^1 \cos(2\pi mt)\ \cos(2\pi nt)\,dt = \delta_{m,n} + \delta_{m,-n},$$

where $\delta_{m,n}$ is the Kronecker delta. Similarly, we find that (see Exercise 2)

$$I_2 = \int_0^1 \cos(2\pi mt)\ \sin(2\pi nt)\,dt = 0$$

and

$$2I_3 = 2\int_0^1 \sin(2\pi mt)\ \sin(2\pi nt)\,dt = \delta_{m,n} - \delta_{m,-n},$$

which completes the proof. □

EXAMPLE 1 (CONTINUED) *We are now able to easily calculate the coefficients c_k and s_k for the function from Example 10.1. For the sound wave $f(t) = \sin(4 \cdot 2\pi t)$, all of the coefficients c_k and s_k are zero* except *s_4, which is*

$$s_4 = 1.$$

The fact that s_4 is nonzero tells us that $f(t)$ contains a component vibrating at 4 Hz and that its amplitude is 1. The fact that all of the other coefficients are zero indicates that $f(t)$ contains no other frequencies.

This calculation reveals a bit about the meaning of Fourier coefficients:

Fourier coefficients describe the wave function $f(t)$ in terms of its underlying frequencies and their respective amplitudes.

It may now seem obvious how to calculate the Fourier coefficients of the last quarter second of the last note of the ninth symphony. However, we do not actually know $f(t)$; we know only its value at $N = 11{,}025$ equidistant points in time. We will therefore suppose that these samples accurately describe the function $f(t)$, and we will replace the integrals by discrete sums. If $f_i, i = 1, 2, \ldots, N$, are the numbers stored on the compact disc, then we will calculate the coefficients

$$C_k = \frac{1}{N}\sum_{i=1}^{N} f_i \cos\left(2\pi k \frac{i}{N}\right) \quad \text{and} \quad S_k = \frac{1}{N}\sum_{i=1}^{N} f_i \sin\left(2\pi k \frac{i}{N}\right). \tag{10.6}$$

The continuous time t has been replaced by a discrete time $t_i = \frac{i}{N}, i = 1, 2, \ldots, N$. Be careful: the k are no longer exactly the frequencies, since k describes the number of cycles of the cosine and sine functions during $\frac{1}{4}$ second. To obtain the actual frequency we must multiply it by 4, obtaining $(4k)$ Hz. Note finally that in discretizing the integral $\int f(t)\,dt$ as a sum $\sum f(t_i)\Delta t$, we have introduced a numeric factor Δt, which in our case is $\Delta t = \frac{1}{N} = \frac{1}{11{,}025}$. This factor appears in front of the above two sums.

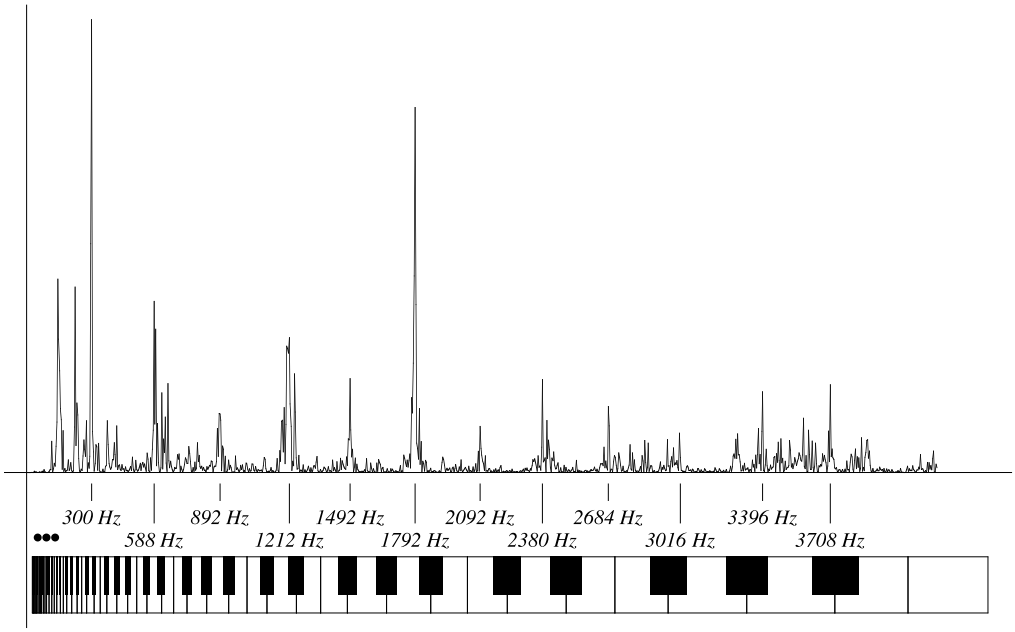


Fig. 10.7. The function $e_k = k(C_k^2 + S_k^2)$ as a function of the frequency $(4k)$ Hz.

The work involved in calculating the Fourier coefficients may seem tedious, but a computer is particularly well suited to this task. The results of these N-term sums are

shown in Figures 10.7 (for the higher frequencies) and 10.8 (for the lower frequencies). These figures contain the numbers $e_k = k(C_k^2 + S_k^2)$ for each of the frequencies $(4k)$ Hz for $k = 1$ to 1000, and thus for the frequencies 4 to 4000 Hz. The points $(4k, e_k)$ have been joined by line segments, and the graphs therefore appear to show a continuous function. Since the coefficients C_k and S_k represent waves with the same frequencies, it is natural to join them together into a single number. The sum of squares $(C_k^2 + S_k^2)$ is related to the amount of energy present in a sound wave of frequency $(4k)$ Hz. Many authors prefer to plot this single value (or its square root), and it is this sum of squares that we will use in the exercises. In this example the function $(C_k^2 + S_k^2)$ decreases so fast as k increases that we have chosen (somewhat arbitrarily) to apply a multiplier of k to the usual sum of squares. The image of a keyboard has been added to make it easier to identify the notes associated to a given frequency. Since we have shown the frequencies on a linear scale, the keyboard appears deformed.

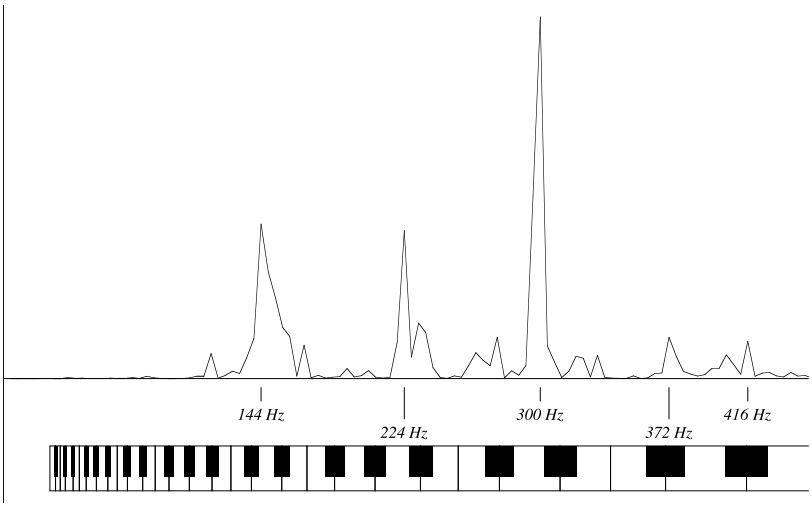


Fig. 10.8. The function $e_k = k(C_k^2 + S_k^2)$ as a function of frequency $(4k)$ Hz for frequencies below 450 Hz.

Below these graphs we have indicated the frequencies of the local maxima of e_k. Observe that the peaks of e_k are sometimes quite wide (for example around 1212 Hz) and that characterizing them by the local maximum is somewhat arbitrary.

What are the most audible frequencies? We find local maxima occurring at 144, 300, 588, 1212, and 2380 Hz, which are quite close to the frequencies associated to the various D notes (see Figure 10.3), and further local maxima at 224, 892, and 1792 Hz which correspond to A notes. There are also a few other frequencies strongly present, such as 1492, 2092, 2684, 3016, 3396, and 3708 Hz, which almost seem to have been added

simply to make the space between peaks a little more regular. Before we can understand where the A notes and other assorted frequencies come from (after all, Beethoven asked only for D's to be played) we must delve into the domain of physics.

Fundamental frequencies and harmonics. The wave equation describing the motion of a vibrating string, such as those on a violin, can be resolved by finding all possible movements of the string such that each segment of the string moves with the same frequency. These solutions are all of the form

$$f_k(x,t) = A \sin \frac{\pi k x}{L} \cdot \sin(\omega_k t + \alpha),$$

where A is the amplitude of the wave, L is the length of the string, t is the time, and $x \in [0, L]$ is the position on the string. The function f_k gives the transverse displacement of the string relative to its position when at rest. (Here the word "transverse" means perpendicular to the axis of the string.) There is an infinite number of such solutions f_k, enumerated by $k = 1, 2, \ldots$. The phase[8] α is arbitrary but the frequency ω_k is completely determined by k and by two properties of the string: its density and its tension. (Since it is rather difficult to change the density of a string, musicians tune strings by adjusting their tension.) The relation describing ω_k is simply

$$\omega_k = k\omega_1,$$

where ω_1 is the fundamental frequency of the string, depending only on its physical properties (density and tension). This frequency is called *fundamental*. All of the other solutions (the other "pure" frequencies of the string) vibrate at frequencies that are integer multiples of the fundamental frequency. These other frequencies are called *harmonics*. In general, the fundamental frequency is the dominant one (although this is not always the case) and it is therefore easy to hear "the" note being played by the instrument. This does not stop the harmonics from being present, however. Each type of instrument emits certain harmonic frequencies more than others; it is the relative importance of particular harmonics that plays a large part in determining the timbre of an instrument. The presence of these harmonics is thus one of the features used by the human ear and brain to differentiate individual instruments.[9] These are not the only characteristics used in perceiving sound; for example, another crucial element is the *attack* (the first few fractions of a second when a sound is being produced).

The expected presence of harmonics as explained by the physics of sound helps us to better understand the graph in Figure 10.7. In fact, starting at 300 Hz (which is

[8] The human ear does not perceive phase. More precisely, two sources of sound emitting the same pure frequency out of phase with each other will be perceived identically.

[9] A student learning to play an instrument is normally advised on how to produce the best quality of sound. If the teacher and student are well versed in mathematics, the teacher could ask, "Can you adjust the Fourier coefficients of this note?" The spectrum of an instrument, in other words, the frequencies and associated amplitudes emitted by the instrument, is one of the tools used by synthesizers.

close to 293.7 Hz, one of the D's on a piano) we find peaks close to every multiple of 293.7 up until $9 \times 293.7 = 2643$ Hz, which is very close to 2684. The larger peaks of the graph are distributed among the integer multiples of the fundamental frequency. We observe the same phenomenon in Figure 10.8, which shows the bass frequencies. The first peak occurs at 144 Hz, very close to the D at 146.8 Hz (the lowest one indicated on the score), and several of the first few integer multiples of this frequency are equally visible. Figure 10.8 indicates a peak close to the note A at 220.2 Hz. This frequency is three times the frequency of the D at 73.4 Hz. However, this D is not actually played by the orchestra; thus the presence of this A is not so easily explained.

Fourier analysis goes much further than just extracting the intensity of the frequencies in a given function f. In fact, the following theorem by Dirichlet tells us that the numbers c_k and s_k *completely* describe the function f, provided it is sufficiently well behaved.

Theorem 10.3 (Dirichlet) *Let $f : \mathbb{R} \to \mathbb{R}$ be a once continuously differentiable periodic function with period 1 (that is, such that $f(x+1) = f(x), \forall x \in \mathbb{R}$). Let c_k and s_k be the Fourier coefficients as given by equations* (10.1)–(10.2)*. Then*

$$f(x) = \frac{c_0}{2} + \sum_{k=1}^{\infty} \left(c_k \cos 2\pi kx + s_k \sin 2\pi kx \right), \qquad \forall x \in \mathbb{R}. \tag{10.7}$$

More precisely, the series on the right-hand side converges uniformly to f.

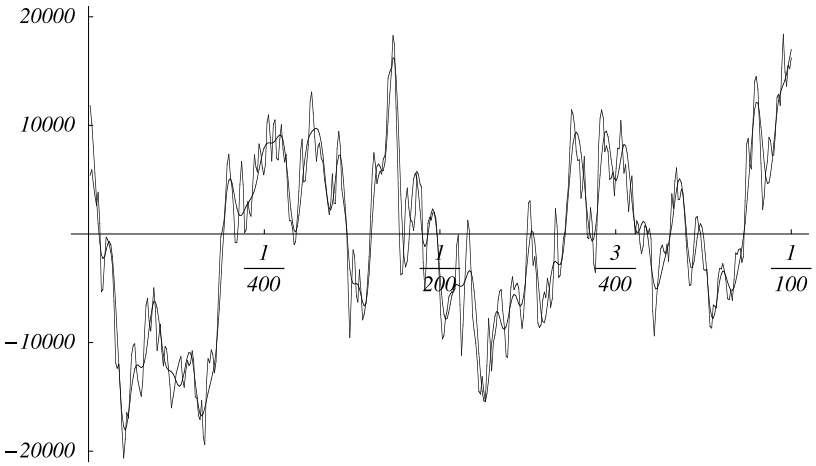


Fig. 10.9. The first hundredth of a second from Figure 10.2 and its reconstruction using the Fourier coefficients $C_k, k = 0, 1, \ldots, 800$, and $S_k, k = 1, 2, \ldots, 800$.

Does this mean that the numbers C_k and S_k that we have calculated can be used to reconstruct the sound wave? Yes, and to convince ourselves Figure 10.9 shows the superposition of the first hundredth of a second from Figure 10.2 and its partial reconstruction

$$\frac{C_0}{2} + \sum_{k=1}^{800} \left(C_k \cos 2\pi kt + S_k \sin 2\pi kt \right).$$

Note that we have limited the sum to the values of k from 1 to 800 rather than using all of them, as required by Dirichlet's theorem. Even though the number of terms is finite, the agreement between the two functions is quite good, *but* the rapid oscillations have been somewhat flattened. This is not surprising; we would have to continually add more terms to the above sum to capture higher and higher frequencies. Furthermore, recall that the coefficients C_k and S_k used in the sum are only approximate values obtained by discretizing the integral defining c_k and s_k. Does there exist a discrete form of Dirichlet's theorem? And if so, how many terms are required to exactly reproduce the discretized step function given by Figure 10.2? The following section answers these questions.

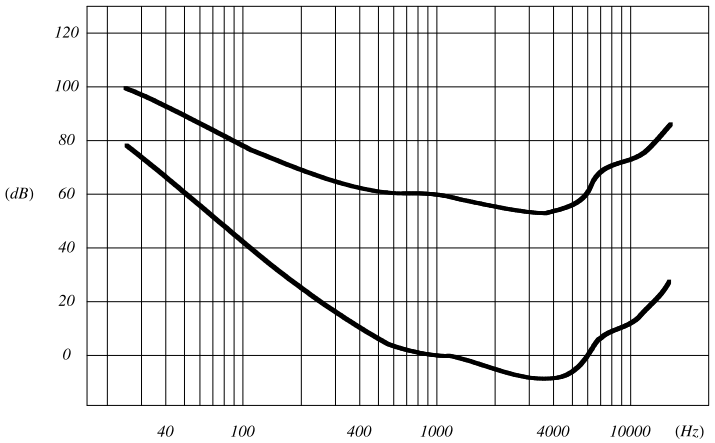


Fig. 10.10. The hearing threshold curve (bottom) and the 60-dB equal-loudness curve (top) as a function of frequency.

We finish this study of the last note of the ninth symphony by discussing an important bit of physiology. The sounds with frequencies of 144, 224, and 300 Hz from Figure 10.7 dominate all of the others by a large margin. (Recall that we plotted the quantity $e_k = k(C_k^2 + S_k^2)$ in Figures 10.7 and 10.8, while it is usual to plot $(C_k^2 + S_k^2)$. Without this factor k, the peak at 1792 Hz would be roughly six times smaller than that near 300 Hz.) How is it that these three sounds do not completely drown out all of the others? Human physiology explains this phenomenon. In 1933, two researchers

named H. Fletcher and W. Munson proposed a method to relate the physical measure of sound-wave pressure to average perceived volume by humans. The bottom curve of Figure 10.10 represents the *hearing threshold* as a function of frequency. (Each person has his or her own proper hearing threshold curve, with this one representing an average.) Note first of all that the frequency scale is logarithmic. The vertical scale, measured in dB (decibels), is also a logarithmic scale. In fact, decibels are scaled such that an increase of 10 dB corresponds to a 10-fold increase in intensity, while an increase of 20 dB corresponds to a 100-fold increase in intensity. Table 10.1 presents a list of common sounds and noises and their typical intensities on the decibel scale. The hearing threshold is the minimum intensity required in order for the human ear to perceive a sound, with its precise values depending on the frequency. As indicated by Figure 10.10, human hearing is the most sensitive (has the lowest threshold) between 2 and 5 kHz. It is harder for us to perceive lower frequencies between 20 and 200 Hz and higher frequencies above 8 kHz. Although these figures are approximate and depend on the individual (including age!), the vast majority of humans are unable to perceive sounds below 20 Hz and sounds above 20 kHz. These physiological measures help to explain why the sounds occurring between 100 and 300 Hz of Figure 10.7 do not deafen us and drown out the others. Moreover, they give us a crucial piece of information for the next section. Figure 10.10 contains a second curve passing through 60 dB at 1000 Hz. This curve is the equal-loudness curve at 60 dB. It indicates the intensities at which given frequencies must be played in order for them to be perceived as having a constant 60 dB volume. Thus somebody listening to a sound at 200 Hz and 70 dB would say it has the same intensity as another at 1000 Hz and 60 dB. Such a curve is clearly subjective and makes sense only when taken as an average over many individuals. Since the earliest work of Fletcher and Munson these definitions have been refined and the experiments repeated. However, the general shape and nature of the curves has not changed: it is between 2000 Hz and 5000 Hz that the human ear is most sensitive.

10.4 The Nyquist Frequency and the Reason for 44,100

The previous section took an intuitive approach to describing how mathematicians and engineers understand sound: *sound waves are a sum of many "pure sounds" of given frequencies and intensities. These pure sounds are trigonometric curves (*sin *and* cos*) oscillating at a single frequency, and their superposition (sum) weighted by their intensity (the Fourier coefficients) yields the sound wave.*

This section asks the following question: at what interval do we need to sample a sound wave in order to accurately reproduce all audible frequencies? We answer this question in two steps.

For the first step we will make the hypothesis that the music we wish to digitize contains only pure sounds with integer frequencies $(1, 2, 3, \ldots$ Hz). The human ear can perceive frequencies between 20 Hz and 20 kHz. How often must we sample the sound wave such that the human ear is unable to perceive the digitization of the sound?

Sound	Intensity in watt/m^2	Intensity in dB
hearing threshold	10^{-12}	0 dB
rustling of leaves in a tree	10^{-11}	10 dB
whispering	10^{-10}	20 dB
normal conversation	10^{-6}	60 dB
busy street	10^{-5}	70 dB
vacuum cleaner	10^{-4}	80 dB
large orchestra	6.3×10^{-3}	98 dB
walkman at full volume	10^{-2}	100 dB
rock concert (close to the stage)	10^{-1}	110 dB
threshold of pain	10^{+1}	130 dB
military jet taking off	10^{+2}	140 dB
perforation of eardrum	10^{+4}	160 dB

Table 10.1. Various sources of sound and their intensities.

With the above hypothesis the sound wave may be described by pure sound waves with frequencies between 20 Hz and 20 kHz:

$$f(t) = \sum_{k=20}^{20{,}000} (c_k \cos 2\pi kt + s_k \sin 2\pi kt)\,. \tag{10.8}$$

The coefficients c_k, s_k for $k = 20, 21, \ldots, 20{,}000$ thus completely determine the function. (For notational simplicity, we will start our sum at $k = 0$ instead of $k = 20$.) Is it possible to replace the Fourier coefficients c_k and s_k by a number of samples $f_i = f(i\Delta), i = 1, 2, \ldots$, of f at regular intervals without losing information? If so, what interval Δ should be used?

Rather than attacking the general case immediately, we will begin with a simple example illustrating the mechanics of the calculation.

Example 10.4 *Rather than considering frequencies from* 20 *Hz to* 20 *kHz we will restrict ourselves to three discrete frequencies and consider the sum*

$$f(t) = \tfrac{1}{2}c_0 + c_1 \cos 2\pi t + c_2 \cos 4\pi t + c_3 \cos 6\pi t + s_1 \sin 2\pi t + s_2 \sin 4\pi t \tag{10.9}$$

for $t \in [0, 1]$. *The term* c_0 *has been added to simplify the discussion; it does not play much of a role when we start considering sums with* 20,000 *terms. Finally, we remark that the term* $\sin 6\pi t$ *has been omitted; we will explain why a little later.*

This sound wave is completely determined by the six real coefficients $c_0, c_1, c_2, c_3, s_1,$ *and* s_2. *We will see shortly that the relationship between these coefficients and the sampled values* $f_i = f(i\Delta)$ *of the function* f *is linear. Thus, we will require at least six*

sampled f_i in order to uniquely determine the coefficients $c_0, c_1, c_2, c_3, s_1, s_2$ from the samples f_i. This motivates our choice of $\Delta = \frac{1}{6}$, leading to

$$f_i = f(\tfrac{i}{6}), \qquad i = 0, 1, 2, 3, 4, 5.$$

These values may be explicitly calculated using (10.9)*. For example, f_1 is given by*

$$\begin{aligned} f_1 &= \tfrac{1}{2}c_0 + c_1 \cos 2\pi(\tfrac{1}{6}) + c_2 \cos 4\pi(\tfrac{1}{6}) \\ &\quad + c_3 \cos 6\pi(\tfrac{1}{6}) + s_1 \sin 2\pi(\tfrac{1}{6}) + s_2 \sin 4\pi(\tfrac{1}{6}) \\ &= \tfrac{1}{2}c_0 + \tfrac{1}{2}c_1 - \tfrac{1}{2}c_2 - c_3 + \tfrac{\sqrt{3}}{2}s_1 + \tfrac{\sqrt{3}}{2}s_2. \end{aligned}$$

Repeating this calculation for the five other values, we obtain

$$\begin{aligned} f_0 &= \tfrac{1}{2}c_0 + c_1 + c_2 + c_3, \\ f_1 &= \tfrac{1}{2}c_0 + \tfrac{1}{2}c_1 - \tfrac{1}{2}c_2 - c_3 + \tfrac{\sqrt{3}}{2}s_1 + \tfrac{\sqrt{3}}{2}s_2, \\ f_2 &= \tfrac{1}{2}c_0 - \tfrac{1}{2}c_1 - \tfrac{1}{2}c_2 + c_3 + \tfrac{\sqrt{3}}{2}s_1 - \tfrac{\sqrt{3}}{2}s_2, \\ f_3 &= \tfrac{1}{2}c_0 - c_1 + c_2 - c_3, \\ f_4 &= \tfrac{1}{2}c_0 - \tfrac{1}{2}c_1 - \tfrac{1}{2}c_2 + c_3 - \tfrac{\sqrt{3}}{2}s_1 + \tfrac{\sqrt{3}}{2}s_2, \\ f_5 &= \tfrac{1}{2}c_0 + \tfrac{1}{2}c_1 - \tfrac{1}{2}c_2 - c_3 - \tfrac{\sqrt{3}}{2}s_1 - \tfrac{\sqrt{3}}{2}s_2. \end{aligned}$$

We can rewrite this system in matrix form as

$$\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 1 & 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -1 & \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & -1 & 1 & -1 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 & -\frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -1 & -\frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ s_1 \\ s_2 \end{pmatrix}.$$

As we stated earlier, the relationship between the Fourier coefficients and sampled values in linear. Whether we can recover the Fourier coefficients from the sample values f_i is therefore equivalent to asking whether the matrix is invertible. The matrix will be invertible if and only if its determinant is nonzero. Several of the rows of this matrix are very similar, and the determinant may be easily calculated through a few simple row and column operations. It is easier to perform the reductions yourself, but we present here a possible sequence of intermediate results (if you do this yourself, your intermediate steps will likely be different!). Using row operations the determinant may be simplified to

$$2 \begin{vmatrix} \frac{1}{2} & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{3} & \sqrt{3} \\ 0 & 0 & 0 & 0 & \sqrt{3} & -\sqrt{3} \\ 0 & -1 & 0 & -1 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -1 & 0 & 0 \end{vmatrix},$$

which may be further simplified using column operations:

$$\begin{vmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\sqrt{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\sqrt{3} \\ 0 & 0 & 0 & -3 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \end{vmatrix}.$$

The remainder of the calculation is now straightforward, yielding a determinant of 27. *Thus, the matrix is invertible and a sound wave of the form* (10.9) *can be completely recovered starting from its six sampled values* $f_i = f(i/6)$.

We can now understand why we did not use the wave $\sin 6\pi t$ *in this example. If we had done so, we would have been presented with two options. The first would have been to omit* c_0 *in order to keep the number of constants to six. We would still have sampled* f *using* $\Delta = \frac{1}{6}$, *but* $\sin 6\pi(\frac{i}{6}) = \sin i\pi$ *is zero for* $i = 0, \ldots, 5$. *The matrix would then have contained a null column and would not have been invertible. The second possibility would have been to leave* c_0 *and to take seven samples using the interval* $\Delta = \frac{1}{7}$. *Although it would have worked, the example would have been significantly more complicated, since trigonometric functions do not take simple values at multiples of* $\frac{2\pi}{7}$.

The general case is equally simple at the conceptual level. However, the most direct proof uses the complex exponential representation of trigonometric functions. The advantage to this representation is that the inverse of the matrix may be explicitly calculated.

Recall that

$$\left.\begin{array}{rcl} e^{i\alpha} & = & \cos\alpha + i\sin\alpha \\ e^{-i\alpha} & = & \cos\alpha - i\sin\alpha \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{rcl} \cos\alpha & = & \frac{1}{2}(e^{i\alpha} + e^{-i\alpha}) \\ \sin\alpha & = & \frac{1}{2i}(e^{i\alpha} - e^{-i\alpha}) \end{array}\right.$$

where $i = \sqrt{-1}$. Then the sum of trigonometric functions with the same frequency

$$c_k \cos 2\pi kt + s_k \sin 2\pi kt$$

may be replaced by

$$\begin{aligned} c_k \cos 2\pi kt + s_k \sin 2\pi kt &= \frac{1}{2}c_k(e^{2\pi ikt} + e^{-2\pi ikt}) + \frac{1}{2i}s_k(e^{2\pi ikt} - e^{-2\pi ikt}) \\ &= \frac{1}{2}(c_k - is_k)e^{2\pi ikt} + \frac{1}{2}(c_k + is_k)e^{-2\pi ikt}. \end{aligned}$$

By introducing new complex Fourier coefficients

$$d_k = \frac{1}{2}(c_k - is_k), \qquad d_{-k} = \frac{1}{2}(c_k + is_k), \qquad k \neq 0,$$

this becomes

$$c_k \cos 2\pi kt + s_k \sin 2\pi kt = d_k e^{2\pi ikt} + d_{-k} e^{-2\pi ikt}.$$

Finally, we define $d_0 = \frac{1}{2}c_0$. A sound wave containing all of the pure sounds with frequencies from 0 to N Hz has the form

$$\frac{c_0}{2} + \sum_{k=1}^{N} \left(c_k \cos 2\pi kt + s_k \sin 2\pi kt\right).$$

When using the new coefficients this becomes

$$\sum_{k=-N}^{N} d_k e^{2\pi ikt}.$$

To keep things simple we will ignore the pure sound corresponding to $e^{2\pi iNt}$ in order to maintain exactly $2N$ coefficients d_k in the above expression. In fact, the index k takes on the $(2N+1)$ values $-N, -N+1, \ldots, -1, 0, 1, \ldots, N-1, N$. The omission of one frequency from the sum does not affect the generality of the result: if the wave contains a component with frequency N Hz, it suffices to use a sum with $(N+1)$ frequencies. We will therefore suppose that

$$f(t) = \sum_{k=-N}^{N-1} d_k e^{2\pi ikt}. \tag{10.10}$$

Since there are $2N$ coefficients d_k in equation (10.10), it is reasonable, as demonstrated in the simplified example above, to use a sampling with interval $\Delta = \frac{1}{2N}$. The sampled values f_l will then be

$$f_l = f(l\Delta) = \sum_{k=-N}^{N-1} d_k e^{2\pi ikl/2N}, \qquad l = 0, 1, \ldots, 2N-1. \tag{10.11}$$

Can the set of coefficients d_k be recovered from the set of samples $f_l, l = 0, 1, \ldots, 2N-1$? In other words, is the matrix

$$\left\{e^{2\pi ikl/2N}\right\}_{-N \le k \le N-1, 0 \le l \le 2N-1} \tag{10.12}$$

invertible?

The answer to this question depends on the following simple observation. Let p be a rational number and n an integer such that $e^{2\pi ipn} = 1$. Then

$$\sum_{l=0}^{n-1} e^{2\pi ipl} = \begin{cases} 0, & \text{if } e^{2\pi ip} \neq 1, \\ n, & \text{if } e^{2\pi ip} = 1. \end{cases} \tag{10.13}$$

To prove this we will use the formula for partial geometric sums

$$\sum_{l=0}^{n-1} e^{2\pi ipl} = \frac{1-e^{2\pi ipn}}{1-e^{2\pi ip}} \qquad \text{if } e^{2\pi ip} \neq 1$$
$$= \frac{1-1}{1-e^{2\pi ip}} = 0.$$

If $e^{2\pi ip} = 1$, then

$$\sum_{l=0}^{n-1} e^{2\pi ipl} = \sum_{l=0}^{n-1} (1)^l = n.$$

Equation (10.13) suggests taking linear combinations of the equations in (10.11) as follows. Multiply both sides of the equation for f_l by $e^{-2\pi iml/2N}$ and sum for $l = 0, 1, \ldots, 2N-1$. Here m will be an integer. The left-hand side of the equation becomes

$$A_m = \sum_{l=0}^{2N-1} e^{-2\pi iml/2N} f_l,$$

while the right-hand side may be simplified to

$$A_m = \sum_{l=0}^{2N-1} \sum_{k=-N}^{N-1} d_k e^{-2\pi iml/2N} e^{2\pi ikl/2N}$$
$$= \sum_{k=-N}^{N-1} d_k \sum_{l=0}^{2N-1} e^{2\pi il(k-m)/2N}.$$

The index k of the coefficients d_k is an integer in the range $[-N, N-1]$. Restricting the integer m to this same interval, the difference $k-m$ will be an integer in the interval $[-(2N-1), 2N-1]$, and the number $e^{2\pi ip}$ with $p = (k-m)/2N$ will never be 1 unless $k = m$. Hence, using (10.13),

$$A_m = 2N \sum_{k=-N}^{N-1} d_k \delta_{k,m}.$$

Whatever the value of $m \in [-N, N-1]$, one (and only one) of the terms in this last sum will satisfy $k = m$ and hence

$$A_m = 2N d_m.$$

The set of coefficients $d_k, k = -N, -N+1, \ldots, N-1$, can be obtained from the samples $f_l, l = 0, 1, \ldots, 2N-1$, through the relation

$$d_k = \frac{1}{2N} A_k = \frac{1}{2N} \sum_{l=0}^{2N-1} f_l e^{-2\pi ikl/2N}. \tag{10.14}$$

Thus, in order to reproduce all of the (integer) frequencies up to the maximal frequency N, we must sample the function at least $2N$ times per second. Conversely, if a wave is sampled at an interval of Δ seconds, then we may extract the amplitudes of each component frequency for frequencies up to

$$f_{\text{Nyquist}} = \frac{1}{2\Delta}. \tag{10.15}$$

The maximal frequency, called the *Nyquist frequency* or *Nyquist limit*, is named after an engineer who studied problems relating to transmission quality and the reproduction of analog signals [5]. Although an immediate result in Fourier analysis, it is of key importance in transforming an analog (continuous) signal into a digital (discrete) one.

Recall that this calculation was made under the assumption that the component frequencies are integer-valued. The invertibility of the linear transform $\{f_l, 0 \le l \le 2N-1\} \mapsto \{d_k, -N \le k \le N-1\}$ assures us that the coefficients can reconstruct the signal and vice versa. However, there is one detail left to discuss. Dirichlet's theorem stated that the reconstruction of a (sufficiently nice) function f is perfect if the coefficients c_k and s_k defined in equations (10.1) and (10.2) are used. In the exercises we will show that the complex coefficients d_k are given by

$$d_k = \int_0^1 f(t)e^{-2\pi ikt}dt.$$

However, in our discretization this integral is replaced by a finite sum, as shown in equation (10.14). Thus it seems there are two ways to calculate the coefficients d_k, provided the component frequencies of f are bounded. Exercise 11 will show that these two methods are equivalent. In practice, compact disc players do not use any of the d_k, c_k, and s_k coefficients to reconstruct the analog sound wave. Rather, they use the samples f_l to generate a smooth and continuous version of the implied step function.

Knowing that the overwhelming majority of people are unable to discern frequencies higher than 20 kHz, the engineers at Philips and Sony chose a sampling rate of 44,100 samples per second, just a little greater than the Nyquist limit $2 \times 20{,}000 = 40{,}000$ for reproducing 20 kHz signals. Thus here is the answer to the question asked at the beginning of this chapter. The exact value (44,100 rather than 40,000) was chosen by taking into consideration other technologies existing at the time [6]. Early video recorders used cassette tapes as storage. The European PAL image standard uses 294 lines of video per frame, each one containing 3 separate color components and being refreshed 50 times per second. This standard thus required $294 \times 3 \times 50 = 44{,}100$ "lines" per second. Thus, the reason for choosing precisely 44,100 was more about making the new standard easier to integrate into existing ones; the only constraint the engineers had to satisfy was that $\frac{1}{\Delta} \le 2f_{\text{Nyquist}} = 2 \times 20{,}000$ Hz.

The case of noninteger frequencies. The second part of this section briefly considers the case in which the component frequencies are no longer integer-valued. The sound wave can therefore now contain any frequency ω between 0 and some maximum

frequency σ, for example 20,000 Hz. (If we continue to use complex component waves $e^{2\pi i\omega t}$, then the frequency ω can be in the interval $[-\sigma, \sigma]$.) This situation is markedly more difficult: the representation of the sound wave through the use of a finite sum such as equation (10.8) will no longer work and must be replaced by an integral over all possible frequencies, such as

$$f(t) = \int_0^\sigma \mathcal{C}(\omega)\cos(2\pi\omega t)d\omega + \int_0^\sigma \mathcal{S}(\omega)\sin(2\pi\omega t)d\omega,$$

or

$$f(t) = \int_{-\sigma}^{\sigma} \mathcal{F}(\omega)e^{2\pi i\omega t}d\omega \tag{10.16}$$

if complex component waves are used. The three functions $\mathcal{C}(\omega)$, $\mathcal{S}(\omega)$, and $\mathcal{F}(\omega)$ play the role of the coefficients c_k and s_k in Dirichlet's theorem (equation (10.7)) and d_k in equation (10.10). They describe the frequency and amplitude content of the sound wave $f(t)$. Despite this additional complexity, the following theorem shows that Nyquist's limit plays a key role in selecting an appropriate sampling rate.

We begin by introducing two definitions. Let $\text{sinc} : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$\text{sinc}(x) = \begin{cases} 1, & \text{if } x = 0, \\ \frac{\sin \pi x}{\pi x}, & \text{if } x \neq 0. \end{cases} \tag{10.17}$$

The amplitude of each frequency ω in the sound wave is given by the *Fourier transform* $\mathcal{F}$ of the function f, defined by

$$\mathcal{F}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i\omega x}dx.$$

(In order for the Fourier transform to exist, the function f must satisfy certain conditions. For example, its absolute value must decrease sufficiently fast as $t \to \pm\infty$. We will assume that these conditions are satisfied.) As stated earlier, it is the function $\mathcal{F}$ that will play the role of the Fourier coefficients c_k and s_k in Dirichlet's representation (10.7). Note that the domain of $\mathcal{F}$ is $\mathbb{R}$, in contrast to the coefficients c_k and s_k, which are enumerated by an integer k. It is thus possible to differentiate $\mathcal{F}$ with respect to ω. Here is the sampling theorem.

Theorem 10.5 (Sampling theorem) *Let f be a function such that the Fourier transform $\mathcal{F}$ is zero-valued outside of the interval $[-\sigma, \sigma]$ for some given fixed σ. Let Δ be chosen such that $\Delta \leq \frac{1}{2\sigma}$. If $\mathcal{F}$ is continuously differentiable, then the series*

$$g(t) = \sum_{n=-\infty}^{\infty} f(n\Delta)\ \text{sinc}\left(\frac{t - n\Delta}{\Delta}\right) \tag{10.18}$$

converges uniformly toward f on $\mathbb{R}$, where the function sinc *is given by* (10.17).

We are not going to prove this theorem. But we can at least provide an intuitive explanation for the curious function sinc. Since the theorem assumes that the Fourier transform $\mathcal{F}$ is nonvanishing only on the interval $[-\sigma, \sigma]$, the reconstruction of f with (10.16) follows the elementary steps

$$\begin{aligned}
f(t) &= \int_{-\sigma}^{\sigma} \mathcal{F}(\omega) e^{2\pi i \omega t} d\omega \\
&= \int_{-\sigma}^{\sigma} \left(\int_{-\infty}^{\infty} f(x) e^{-2\pi i \omega x} dx \right) e^{2\pi i \omega t} d\omega \\
&= \int_{-\sigma}^{\sigma} \left(\int_{-\infty}^{\infty} f(x) e^{2\pi i \omega (t-x)} dx \right) d\omega \\
&\overset{1}{=} \int_{-\infty}^{\infty} f(x) \left(\int_{-\sigma}^{\sigma} e^{2\pi i \omega (t-x)} d\omega \right) dx \\
&\overset{2}{=} \int_{-\infty}^{\infty} f(x) \left. \frac{e^{2\pi i \omega (t-x)}}{2i\pi(t-x)} \right|_{-\sigma}^{\sigma} dx \\
&= \int_{-\infty}^{\infty} f(x) \frac{e^{2\pi i \sigma (t-x)} - e^{-2\pi i \sigma (t-x)}}{2i\pi(t-x)} dx \\
&= \int_{-\infty}^{\infty} f(x) \frac{\sin(2\pi\sigma(t-x))}{\pi(t-x)} dx \\
&= 2\sigma \int_{-\infty}^{\infty} f(x) \text{ sinc } (2\sigma(t-x)) dx.
\end{aligned}$$

Two remarks on these steps. First, the equality marked by a 1 is not mathematically rigorous, since the order of integration may not be changed for all f. Second, the antiderivative obtained for the integration with respect to ω (equality marked by a 2) is the right one, except when $t = x$. In this case the antiderivative should be ω and the integral 2σ. But this is precisely the value given to this integral when $t = x$ in the last line, since the value $\text{sinc}(x = 0)$ is defined to be 1.

To relate the last expression to the sampling theorem we need to study the rate of variation of the two functions $f(x)$ and sinc $(2\sigma(t-x))$ in the integrand. For that purpose set $\Delta = \frac{1}{2\sigma}$. Since σ is the maximal frequency (number of oscillations per second), Δ is to be understood as the time in seconds between two extrema of the wave with highest possible frequency in f. If the overall feature of the graph of f varies slowly on the scale of Δ, the two values $f(t)$ and $f(t+\Delta)$ will almost be equal. The function

$$\text{sinc } (2\sigma(t-x)) = \text{sinc } ((t-x)/\Delta),$$

on the other hand, varies more rapidly. Note that increasing x to $x+\Delta$ in this function changes its argument by one unit. As can be seen on the graph of sinc displayed in Figure 10.11, the sign of the function sinc changes each time its argument changes by one unit (except for x in the interval $(-1, 1)$). Therefore the function sinc changes more

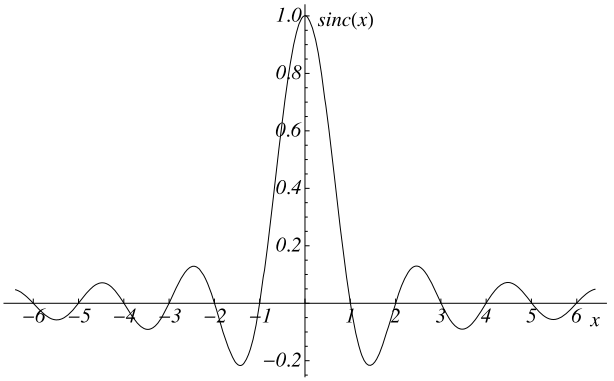


Fig. 10.11. The function sinc.

rapidly than f in the above integral. To approximate the integral by a sum, it is natural to probe the integrand at every change of sign of the function sinc, that is, at every $x = n\Delta, n \in \mathbb{Z}$. Replacing the infinitesimal dx by Δ, we get the following estimate for $f(t)$:

$$\begin{aligned} f(t) &\approx 2\sigma \sum_{n=-\infty}^{\infty} f(n\Delta)\ \mathrm{sinc}\left(\frac{t-n\Delta}{\Delta}\right)\Delta \\ &= \sum_{n=-\infty}^{\infty} f(n\Delta)\ \mathrm{sinc}\left(\frac{t-n\Delta}{\Delta}\right), \end{aligned}$$

that is, the form proposed in (10.18). Note finally that if f varies significantly on an interval of width Δ, it is unlikely that the above approximation will give a good estimate of f. This argument is not a proof. But it underlines the role of sinc and the interplay between Δ and the maximal frequency that may appear in f.

Thus the theorem says that it is sufficient to sample the function f at a regular interval $\Delta \leq \frac{1}{2\sigma}$ in order to reconstruct this function. Alternatively, the sampling rate of a function f must be at least twice the maximum frequency contained in f. Thus, we are again brought back to Nyquist's limit.

This theorem has been attributed to many scientists, since it was independently discovered several times by researchers in very different domains. It is in the domain of telecommunications and signal processing that this result continues to have the greatest impact. Thus, it is not very surprising that we most often associate the names of various electrical engineers (notably Kotelnikov, Nyquist, and Shannon) with this result. However, two mathematicians, E. Borel and J.M. Whittaker, also discovered this result. It is becoming increasingly common for this result to be covered in Fourier analysis courses for mathematicians. (See [2] and [3].)

10.5 Exercises

1. Determine the frequencies of each of the C keys on a piano.

2. Prove the identities in equations (10.4) and (10.5).

3. **(a)** Show that

$$c_k \cos 2\pi kt + s_k \sin 2\pi kt = \sqrt{c_k^2 + s_k^2}\cos(2\pi k(t + t_0))$$

for some $t_0 \in [0, 1]$. The sum $c_k \cos 2\pi kt + s_k \sin 2\pi kt$ therefore corresponds to a pure sound wave of frequency k translated in time. The value t_0 (or more precisely $2\pi kt_0$) is called its phase.
(b) Show that all functions of the form $f(t) = r\cos(2\pi k(t + t_0))$ can be written in the form $f(t) = c_k \cos 2\pi kt + s_k \sin 2\pi kt$. Calculate c_k and s_k as functions of r and t_0.

4. **(a)** How many notes could we add to the high end of a piano such they could still be heard by the average human?
(b) The same question, but for notes added to the low end of the keyboard.
(c) Certain breeds of small dogs can hear frequencies as high as 45,000 Hz. How many octaves would have to be added to a modern piano to completely cover the audible spectrum of such a dog?
(d) How many samples per second should be taken such that a compact disc could faithfully reproduce sound as perceived by a dog?

5. **Alternative temperaments.** Construct the Pythagorean and Zarlino scales. In other words, determine the frequencies of each note between two consecutive A's. You will have to refer to music texts or the Internet in order to discover how these two scales are constructed.

6. Is the function $f : \mathbb{R} \to \mathbb{R}$

$$f(t) = \frac{1}{2}\sin 2\pi t - \frac{1}{3}\sin 6\pi t - \frac{1}{600}\sin 400\pi t$$

periodic? If yes, what is its minimal period? Which of its Fourier coefficients are nonzero?

7. **(a)** Find the Fourier coefficients of the function $f : [0, 1) \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 1, & 0 \le x < \frac{1}{2}, \\ -1, & \frac{1}{2} \le x < 1. \end{cases} \tag{10.19}$$

Hint: the formulas giving c_k and s_k are integrals defined over the interval $[0, 1)$. Partition these integrals into two, the first over the interval $[0, \frac{1}{2})$ and the second over $[\frac{1}{2}, 1)$.
(b) Use mathematical computing software to plot the first few partial sums of the Fourier series (see equation (10.7)) corresponding to this function f. Verify that the partial sums approach the original function.

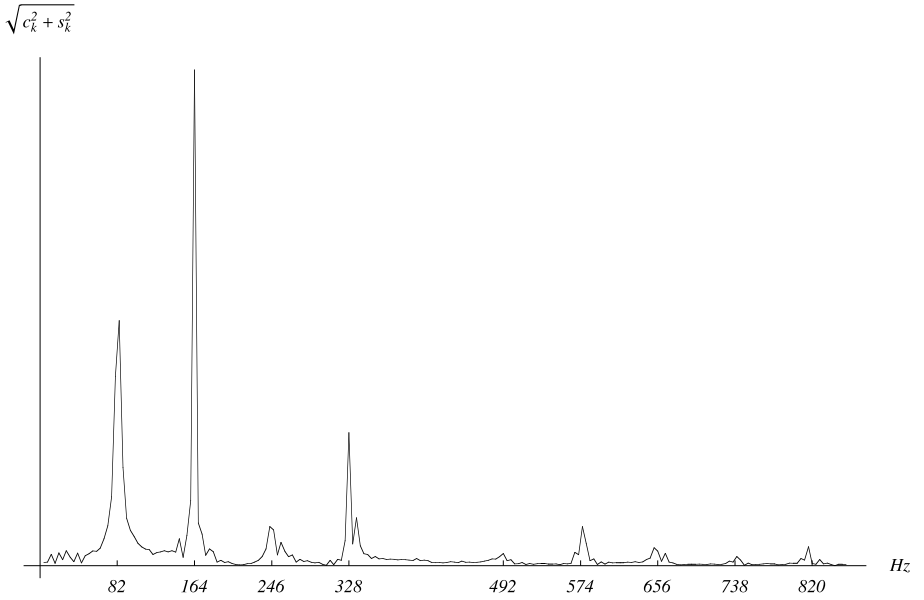


Fig. 10.12. Spectrum of the first note of Brahms's first sonata for cello and piano. The frequencies of the local maxima are indicated. (see Exercise 8.)

8. The first note of Brahms's first sonata for cello and piano is a single note played only by the cello. The graph in Figure 10.12 shows the intensity $\sqrt{c_k^2 + s_k^2}$ of the Fourier coefficients of this note as a function of the frequency k Hz.
(a) On the keyboard of Figure 10.3, identify the note being played by the cello.
(b) One of the following statements is true. Determine which, and justify your response.

1. One of the harmonic frequencies dominates the fundamental frequency.
2. The harmonic frequency with largest amplitude bears a name different from that of the note being played.
3. The peak at 82 Hz cannot be perceived by the human ear.
4. The horizontal axis of the graph covers the entire human audible spectrum.

5. Depending on the phase difference between the fundamental frequency and a given harmonic, the harmonic may or may not be heard.

9. **(a)** The last note of Schubert's first *Impromptu* D. 946 is a chord of four notes, meaning the pianist plays four notes simultaneously. Figure 10.13 shows the spectrum of this chord. Among these four notes, one is rather difficult to identify. Find the three easily identified notes and explain your reasoning.
(b) Why could a note be difficult to identify when a chord is being played? Considering your answer, suggest a few possibilities for the fourth note being played.

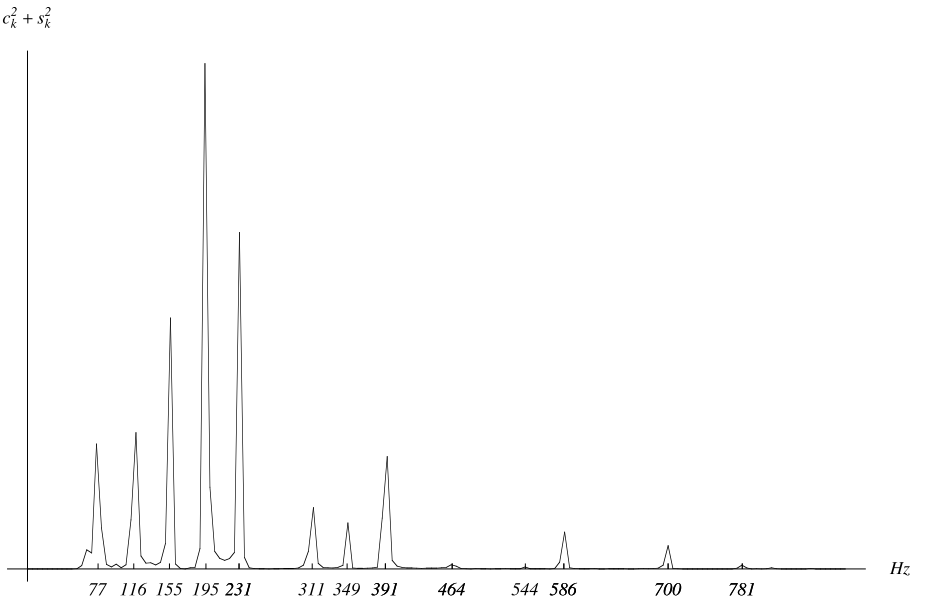


Fig. 10.13. Spectrum of the last chord of Schubert's first *Impromptu* D. 946. The frequencies of local maxima are indicated. (See Exercise 9.)

10. **(a)** G. Gershwin's *Rhapsody in Blue* opens with a clarinet glissando. The clarinet is the only instrument playing at that moment. The spectrum at the beginning of the glissando is shown in Figure 10.14. What note is being played by the clarinet?
(b) The harmonics of a clarinet possess a certain characteristic that may be seen in the spectrum. What is this characteristic? (A little research into the particulars of clarinets might be necessary. A good starting point is Benson's book [1].)

11. **(a)** Using the equations defining d_k in terms of c_k and s_k, show that a periodic function f with period 1 can be written in the form

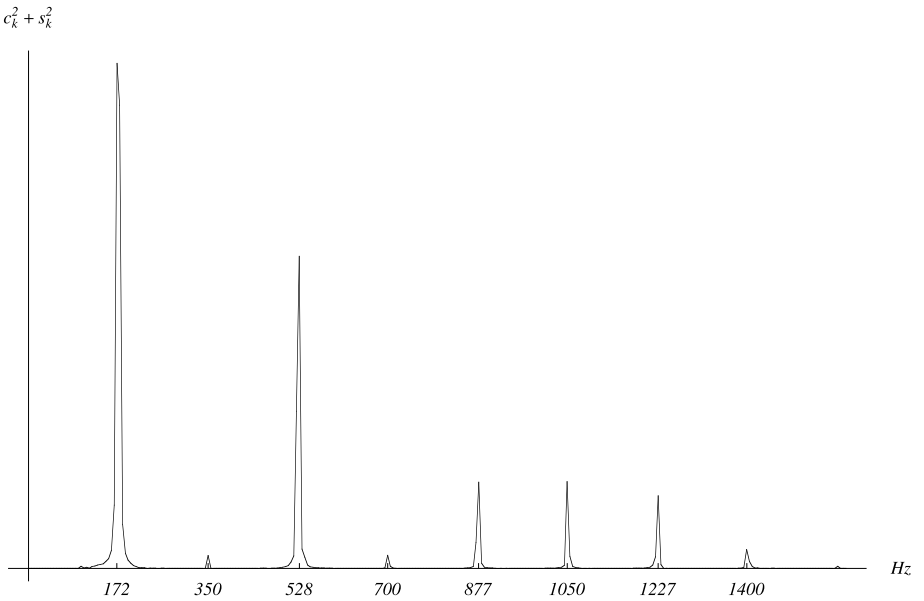


Fig. 10.14. The spectrum of the first note of *Rhapsody in Blue*. The frequencies of local maxima are indicated. (See Exercise 10.)

$$f(t) = \sum_{k \in \mathbb{Z}} d_k e^{2\pi i k t},$$

where the d_k are calculated using

$$d_k = \int_0^1 f(t) e^{-2\pi i k t}\, dt.$$

(b) Suppose that the function $f(t)$ contains only component frequencies with integer frequencies $k \in \{-N, -N+1, \ldots, N-2, N-1\}$ and define the coefficients D_k as those obtained from the sampling of f at interval $\Delta = \frac{1}{2N}$:

$$D_k = \frac{1}{2N} \sum_{l=0}^{2N-1} f(l\Delta) e^{-2\pi i k l / 2N}.$$

Observe that equation (10.14) allows you to conclude that $d_k = D_k$ for such a function f.

12. Another way of showing that the system of equation in (10.11) has a unique solution $\{d_{-N}, \ldots, d_{N-1}\}$ is by showing that the determinant of the matrix is nonzero. Show this

by transforming it into a Vandermonde determinant and using Lemma 6.22 of Chapter 6.

13. Beat patterns. Beat patterns are a well-known musical phenomenon. When two instruments (physically close to each other) play nearly the same note at the same intensity, the perceived sound varies regularly in intensity with time. In other words, the perceived amplitude oscillates periodically. This oscillation may be slow (once every few seconds) or fast (several times per second).

(a) Two flutes emit sound waves f_1 and f_2 with frequencies ω_1 and ω_2 respectively:

$$f_1 = \sin(\omega_1 t) \qquad \text{and} \qquad f_2(t) = \sin(\omega_2 t).$$

(We neglect the harmonics, which we assume are weak.) The resulting sound is $f = f_1 + f_2$. Show that we can write f in the form

$$f(t) = 2 \sin \alpha t \ \cos \beta t$$

and determine α and β in terms of ω_1 and ω_2.

(b) Suppose that ω_1 is a well-tempered E at 659.26 Hz, and that ω_2 is a true E at 660 Hz. Show that the ear would perceive f as a frequency close to these two, but with an amplitude varying with a period of about $\frac{4}{3}$ seconds. This is an example of a beat pattern.

14. Aliasing. This chapter has so far ignored a technical difficulty faced by engineers. We have shown that sampling every $\Delta = \frac{1}{44{,}100}$ seconds allows for all sounds in the (average) human audible spectrum to be reproduced. The problem is that musical instruments often produce harmonic frequencies above our hearing range with $N_{\max} = 20{,}000$ Hz. When the recording is sampled, a frequency $N > N_{\max}$ will be perceived as a sound with frequency between 0 and $N_{\max}$. (See Figure 10.15, where the sampled points appear to describe a sinusoidal curve with a lower frequency than that which actually generated them.) This problem is known as *aliasing*, since certain frequencies are aliased (appear

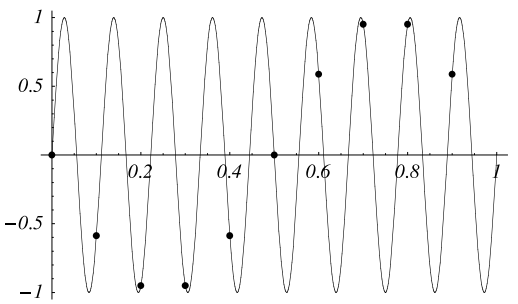


Fig. 10.15. A simple example of aliasing. (See Exercise 14.)

as) other frequencies after sampling. This problem appears in all domains in which signals are digitized. For example, it appears as countour banding or moiré patterns in digital photography. This effect is closely related to another distortion commonly encountered in movies: a spoked wheel rotating quickly in one direction may appear to be rotating in the opposite direction.

Determine the frequency N' that the frequency $N > N_{\max}$ will be "aliased" to after sampling. (Obviously, this aliased frequency must satisfy $0 < N' \leq N_{\max}$.)

15. Sampling theorem This exercise provides an example of reconstructing a continuous signal $f(t)$ using the sampling theorem (Theorem 10.5). Suppose that we wish to reproduce the sound waves of a signal with frequencies constrained to the range $[-\sigma, \sigma]$, where $\sigma = 6$ Hz. As such, we will use a sampling interval of $\Delta = \frac{1}{2\sigma} = \frac{1}{12}$ seconds. We should therefore be able to recover $f(t) = \cos 2\pi\omega_0 t$ using only its sampled values $f(n\Delta), n\mathbb{Z}$, assuming $\omega_0 \in [-\sigma, \sigma]$. Take for example $\omega_0 = 5.5$ Hz.

(a) With the help of software, plot the function $f(t)$ on the interval $t \in [0, 1]$.

(b) Plot the function sinc t over the interval $t \in [-6, 6]$. (The function sinc was defined in equation (10.17).)

(c) Plot the partial sum

$$\sum_{n=M}^{N} f(n\Delta) \text{ sinc } \left(\frac{t - n\Delta}{\Delta}\right)$$

on the interval $t \in [0, 1]$ and compare it with the graph from (a). Start with a small number of terms in the sum (for example $M = 0$ and $N = 11$), and increase the number of terms by lowering M and raising N. Investigate the difference between the function f and its partial sum reconstructions.

(d) This questions is for those with a little more knowledge of the Fourier transform. The function f given here does not satisfy the conditions necessary for Theorem 10.5 to apply. Why? Can you slightly modify this function so that it satisfies these conditions? Will the reconstructions plotted in (c) change significantly after this modification?

References

[1] D.J. Benson. *Music: A Mathematical Offering*. Cambridge University Press, 2006.

[2] D.W. Kammler. *A First Course in Fourier Analysis*. Prentice Hall, NJ, 2000.

[3] T.W. Körner. *Fourier Analysis*. Cambridge University Press, 1988. (See also [4].)

[4] T.W. Körner. *Exercises for Fourier Analysis*. Cambridge University Press, 1993. (The sampling theorem (Theorem 10.5) is discussed here.)

[5] H. Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47:617–644, 1928.

[6] K.C. Pohlmann. *The Compact Disc Handbook*. A-R Editions, Madison, WI, 2nd edition, 1992.

[7] L. van Beethoven. Symphony no. 9, in b minor, opus 125, 1826. (Since then, many other editions have been published (Eulenberg, Breitkopf & Härtel, Kalmus, Bärenreiter, etc.). Affordable reprints of this work are available.)

[8] H. von Hemholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Dover Publications, New York, 1954. (Translated by A.J. Ellis from the 4th German edition of 1877.)

11

Image Compression: Iterated Function Systems

This chapter can be covered in one or two weeks of classes. If only one week is available then you can briefly cover the introduction (Section 11.1) and then explain in detail the concept of an attractor of an iterated function system (Section 11.3) by concentrating on the Sierpiński triangle (Example 11.5). Demonstrate the theorem that constructs affine transformations mapping three points on the plane to three points on the plane and discuss the particular affine transformations that will be used often in iterated function systems (Section 11.2). Explain Banach's fixed-point theorem stressing the point that the proof on $\mathbb{R}$ can be transposed, nearly word by word, to complete metric spaces (Section 11.4). Finally, discuss the intuition behind the Hausdorff distance (beginning of Section 11.5). If you wish to spend a second week, then you can deepen the discussion of the Hausdorff distance and work through a few of the proofs of its various properties (Section 11.5). This leaves sufficient time to discuss fractal dimensions (Section 11.6) and to explain briefly the construction of iterated function systems that allow for the reconstruction of actual photographs (Section 11.7). Sections 11.5, 11.6, and 11.7 are almost independent, so it is possible to treat Section 11.6 or 11.7 without having gone through the more difficult Section 11.5.

Another option for a one-week coverage is to discuss Sections 11.1 to 11.3 and to jump to 11.7, which explains how to adapt the technique to compression of real photographs.

11.1 Introduction

The easiest way to store an image in computer memory is to store the color of each individual pixel. However, a high-resolution photograph (many pixels) with accurate color (many data bits per pixel) would require an enormous of amount of computer memory. And videos, with many such images per second, would required even more.

With widespread adoption of digital cameras and the Internet, people are storing an ever larger number of images on their computers. It is thus critical that these images be stored efficiently so as not to take up an inordinate amount of space. Images on the web can be of lower resolution than digital photographs or large posters. And we are

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_11,

very interested in keeping their sizes small; no doubt you have already experienced slow loading images while browsing the web, even if the images are compressed.

There exist many image compression techniques. The commonly used JPEG (Joint Photographic Experts Group) format makes use of discrete Fourier techniques and is explored in Chapter 12. In this chapter we will concentrate on another technique: image compression using iterated function systems.

There was a great deal of hope and excitement over the possibilities of this technique when it was first introduced in the 1980s, spurring considerable research. Unfortunately, formats based on these techniques have not seen much success because the compression algorithms and the compression ratios are not good enough. However, these techniques continue to be researched and might yet see improvements. We have decided to present these methods for several reasons. First, it is easy to show the underlying mathematics at work, which rely on Banach's powerful fixed-point theorem (the fixed point of the theorem referring to the attractor of an operator). Moreover, the method uses fractals, which we demonstrate how to construct in a very simple manner as fixed points of operators. That such complicated structures can be described through such simple constructions is a striking demonstration of the power and elegance of mathematics; it shows that if we look at an object from just the right point of view, everything is simplified, allowing us to understand its structure.

We stated above that the easiest way to store a picture is simply to store the color associated with each pixel, an approach that is far from efficient. How to do better? Suppose that we were to draw a profile of a city (Figure 11.1). Instead of storing the actual pixels, we could store the underlying geometric constructs, allowing us to reconstruct it:

- all line segments,
- all circular arcs,
- etc.

We have represented the image as a union of known geometric objects.

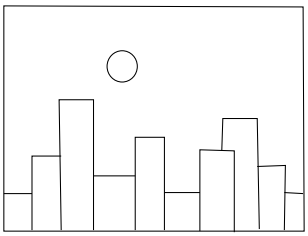


Fig. 11.1. A line drawing of a city.

To store a line segment it is more economical to store only its extremities and to create a program that can draw the line given these two points. Similarly, the arc of

a circle may be specified by its center, radius, and starting and stopping angles. The underlying geometric objects form the *alphabet* with which we can describe an image.

How can we store a more complicated image, for instance, a photograph of a landscape or a forest? It may seem that the previous method cannot work, because our alphabet of geometric objects is too poor. We will discover that we can use the same technique, but with a larger and more advanced alphabet:

- we approximate our image with a finite number of fractal images. For example, consider the fern leaf in Figure 11.2;
- to store the image we create a program that draws the image using the underlying fractals. The fern leaf of Figure 11.2 can be drawn by a program of fewer than 15 lines! (A *Mathematica* program for drawing the fern can be found at the end of Section 11.3.)

In this process the resulting image is the "attractor" of an operator W (defined below) that maps a subset of the plane to a subset of the plane. Beginning from any initial subset B_0 we recursively construct the sequence $B_1 = W(B_0)$, $B_2 = W(B_1)$, ..., $B_{n+1} = W(B_n)$, ... For sufficiently large n (in fact, $n = 10$ suffices if B_0 was carefully chosen), B_n will start to look like the fern leaf.

The technique may sound a little naive: can we really program a computer to approximate any photo using fractals? Indeed, some adaptation of the initial idea will be needed, but we will keep the fundamental idea that the reconstructed image is the attractor of some operator. Since constructing an arbitrary photo is quite advanced, we leave the discussion until the end of the chapter (Section 11.7). To start, we focus on constructing programs that can draw fractals.

11.2 Affine Transformations in the Plane

We start by explaining why we need affine transformations. Consider the fern leaf in Figure 11.2. It is the union of (see Figure 11.2)

- the stalk,
- and three smaller fern leaves: the bottom left branch, the bottom right branch, and the leaf minus the two lowest branches.

Each of these four pieces is the image of the entire fern leaf under an affine transformation. Knowing the four associated transformations allows us to reconstruct the entire image:

- the transformation T_1, which maps the entire leaf to the leaf minus the two lowest branches,
- the transformation T_2, which maps the entire leaf to the bottom left branch (marked L in Figure 11.2),
- the transformation T_3, which maps the entire leaf to the bottom right branch (marked R in Figure 11.2), and

Fig. 11.2. A fern leaf.

- the transformation T_4, which maps the entire leaf to the bottom part of the stalk.

Definition 11.1 *An affine transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ is the composition of a translation with a linear transformation. It can be written as*

$$T(x, y) = (ax + by + e, cx + dy + f). \tag{11.1}$$

This is the composition of the linear transformation

$$S_1(x, y) = (ax + by, cx + dy)$$

and the translation

$$S_2(x, y) = (x + e, y + f).$$

Linear transformations are often represented in matrix notation as

$$S_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

We can also use this notation to represent affine transformations:

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}.$$

We see that the affine transformation is specified by the six parameters a, b, c, d, e, f. Thus, in order to uniquely determine a given affine transformation we require six linear equations.

Theorem 11.2 *There exists a unique affine transformation that maps three distinct noncollinear points P_1, P_2, and P_3 to three points Q_1, Q_2, and Q_3.*

PROOF: Let (x_i, y_i) be the coordinates of P_i and let (X_i, Y_i) be the coordinates of Q_i. The desired transformation is in the form of (11.1), and we must solve for a, b, c, d, e, f, knowing that $T(x_i, y_i) = (X_i, Y_i)$, $i = 1, 2, 3$. This gives us six linear equations in six unknowns a, b, c, d, e, f:

$$\begin{aligned} ax_1 + by_1 + e &= X_1, \\ cx_1 + dy_1 + f &= Y_1, \\ ax_2 + by_2 + e &= X_2, \\ cx_2 + dy_2 + f &= Y_2, \\ ax_3 + by_3 + e &= X_3, \\ cx_3 + dy_3 + f &= Y_3. \end{aligned}$$

The parameters a, b, e are solutions of the system

$$\begin{aligned} ax_1 + by_1 + e &= X_1, \\ ax_2 + by_2 + e &= X_2, \\ ax_3 + by_3 + e &= X_3, \end{aligned} \tag{11.2}$$

while c, d, f are solutions of the system

$$\begin{aligned} cx_1 + dy_1 + f &= Y_1, \\ cx_2 + dy_2 + f &= Y_2, \\ cx_3 + dy_3 + f &= Y_3. \end{aligned} \tag{11.3}$$

Both of these are systems over the same matrix A, whose determinant is

$$\det A = \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}.$$

Note that this determinant is nonzero precisely when the points P_1, P_2, and P_3 are distinct and noncollinear. In fact, the three points are collinear if and only if the vectors $\overrightarrow{P_1P_2} = (x_2 - x_1, y_2 - y_1)$ and $\overrightarrow{P_1P_3} = (x_3 - x_1, y_3 - y_1)$ are collinear, which is the case if and only if the following determinant is zero:

$$\begin{vmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \end{vmatrix} = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

The determinant of a matrix does not change when we add to a row a multiple of another. Subtracting the first row from the second and the third yields

$$\begin{aligned} \det A &= \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 - x_1 & y_2 - y_1 & 0 \\ x_3 - x_1 & y_3 - y_1 & 0 \end{vmatrix} \\ &= (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1). \end{aligned}$$

We see that $\det A = 0$ precisely when the three points are aligned. On the other hand, if $\det A \neq 0$, then each of the systems (11.2) and (11.3) has a unique solution. □

Remark: We must use the technique of Theorem 11.2 to find the four transformations describing the fern leaf. For that we need to specify coordinate axes and measure the coordinates of the points P_i and Q_i. However, in many examples we can guess the affine transformations without having to measure the coordinates of the points P_i and Q_i and solving the associated systems. In these cases we use compositions of simple affine transformations.

Some simple affine transformations.

- Homothety with ratio r: $T(x, y) = (rx, ry)$.
- Reflection about the x axis: $T(x, y) = (x, -y)$.
- Reflection about the y axis: $T(x, y) = (-x, y)$.
- Reflection through the origin: $T(x, y) = (-x, -y)$.
- Rotation through angle θ: $T(x, y) = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$. To find this formula we use the fact that a rotation is a linear transformation. The columns of its matrix are the coordinates of the images of the base vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$ (Figure 11.3). The transformation matrix is therefore

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

- Projection onto the x axis: $T(x, y) = (x, 0)$.
- Projection onto the y axis: $T(x, y) = (0, y)$.
- Translation by a vector (e, f): $T(x, y) = (x + e, y + f)$.

11.3 Iterated Function Systems

Fractals that can be constructed using the technique described above will be *attractors* of *iterated function systems*. We define these terms more clearly.

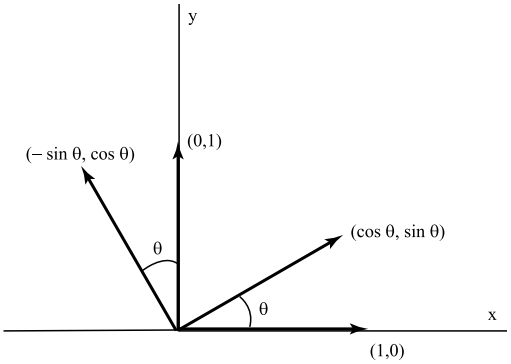


Fig. 11.3. The images of base vectors under a rotation of angle θ.

Definition 11.3 *1. An affine transformation is an* affine contraction *if the image of any segment is a shorter line segment.*
2. An iterated function system *is a set of affine contractions* $\{T_1, \ldots, T_m\}$.
3. The attractor *of an iterated function system* $\{T_1, \ldots, T_m\}$ *will be the unique geometric object* A *such that*

$$A = T_1(A) \cup \cdots \cup T_m(A).$$

Example 11.4 A fern leaf. *We consider the fern leaf from Figure 11.2. It is easy to see that each of the branches of the leaf resembles the entire leaf itself. Thus, the leaf is the union of the stalk and infinitely many smaller copies of the leaf. We want to avoid working with an infinite number of sets of transformations, so a little care is required. Call* A *the subset of the plane consisting of all points of the fern leaf. We introduce a coordinate system. Let* T_1 *be the transformation mapping* P_i *to* Q_i*, as labeled in Figure 11.4. The image* $T_1(A)$ *is a subset of* A*. Now consider* $A \setminus T_1(A)$*. It consists of the bottom portion of the stalk and the bottommost branches on either side, as outlined in Figure 11.2. We can choose points* Q_1'*,* Q_2'*, and* Q_3' *to construct a transformation* T_2 *that maps the entire leaf to the bottommost left branch. (Exercise!) Similarly, we can choose points* Q_1''*,* Q_2''*, and* Q_3'' *describing a transformation* T_3 *that maps to the bottommost right branch. Thus* $A \setminus (T_1(A) \cup T_2(A) \cup T_3(A))$ *is simply the bottommost portion of the stalk. We wish to find another transformation* T_4 *that maps the entire leaf to this portion of the stalk. Such a transformation is simply a projection onto the* y *axis composed with a contraction (homothety with ratio* $r < 1$*) and a translation.*

We have constructed four affine transformations such that

$$A = T_1(A) \cup T_2(A) \cup T_3(A) \cup T_4(A). \tag{11.4}$$

Fig. 11.4. The points P_i and Q_i describing the transformation T_1.

We claim and will prove later that no other set than the fern satisfies (11.4). *The fern leaf will be the attractor of the iterated function system* $\{T_1, T_2, T_3, T_4\}$.

This example is relatively complicated. Thus, we present another easier example to help develop some intuition.

Example 11.5 The Sierpiński triangle. *To simplify the calculations we will consider a Sierpiński triangle with a base and height of* 1 *(see Figure 11.5).*

Here the triangle A *is the union of three smaller copies of itself* $A = T_1(A) \cup T_2(A) \cup T_3(A)$. *In this case we can easily write the explicit equations of the affine contractions. In fact, if we suppose that the origin is situated at the bottom left corner of the triangle, then* T_1 *is the homothety with ratio* $1/2$*:*

$$T_1(x, y) = (x/2, y/2),$$

and T_2 *and* T_3 *are simply compositions of* T_1 *with translations. Since the base and height of the triangle are both* 1*, then* T_2 *is* T_1 *composed with a translation by* $(1/2, 0)$*, while* T_3 *is* T_1 *composed with a translation by* $(1/4, 1/2)$*:*

$$\begin{aligned} T_2(x, y) &= (x/2 + 1/2, y/2), \\ T_3(x, y) &= (x/2 + 1/4, y/2 + 1/2). \end{aligned}$$

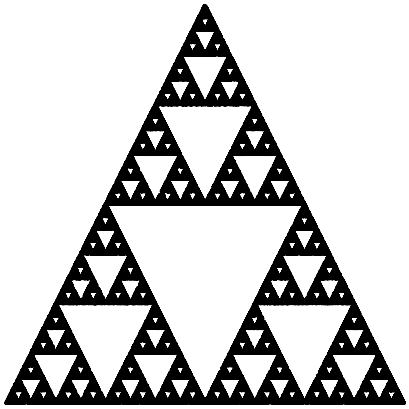


Fig. 11.5. The Sierpiński triangle.

The triangle lies within the square $C_0 = [0,1] \times [0,1]$. *We are interested in the sets*

$$\begin{aligned} C_1 &= T_1(C_0) \cup T_2(C_0) \cup T_3(C_0), \\ C_2 &= T_1(C_1) \cup T_2(C_1) \cup T_3(C_1), \\ &\vdots \\ C_n &= T_1(C_{n-1}) \cup T_2(C_{n-1}) \cup T_3(C_{n-1}), \\ &\vdots \end{aligned}$$

the first few of which are shown in Figure 11.6. Observe that for sufficiently large n *(even at* $n = 10$*), the set* C_n *already begins to resemble* A. *The set*

$$C_n = T_1(C_{n-1}) \cup T_2(C_{n-1}) \cup T_3(C_{n-1})$$

is called the n*th iteration of the initial set* C_0 *under the operator*

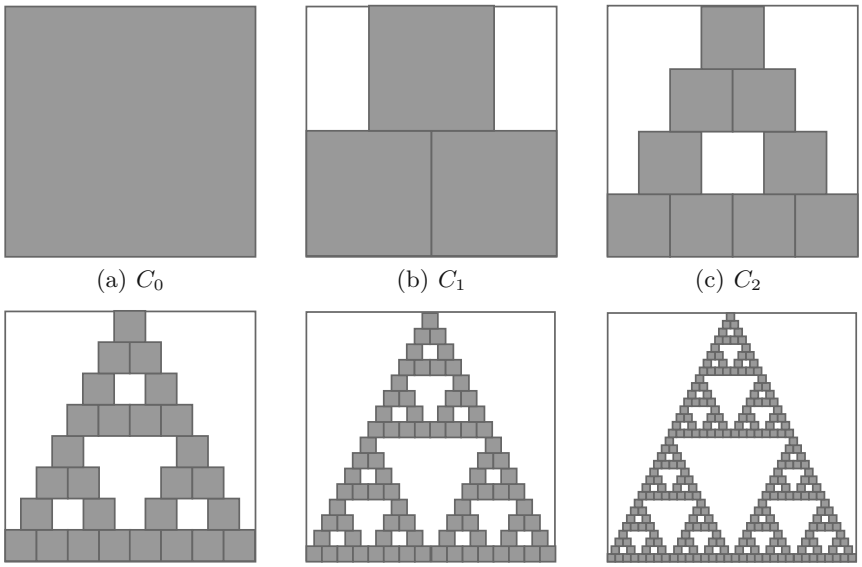
$$C \mapsto W(C) = T_1(C) \cup T_2(C) \cup T_3(C),$$

which maps a subset C *to another subset* $W(C)$.

It is for this reason we say that A *is an* attractor. *The remarkable thing is, had we started with any subset of the plane other than* C_0, *the limit of the process would still be the Sierpiński triangle (see Figure 11.7).*

The general principle. The Sierpiński triangle example allowed us to see the general process at work. Given an iterated function system $\{T_1, \ldots, T_m\}$ of affine contractions, we construct an *operator* W that acts on subsets of the plane. A subset C is mapped to the subset $W(C)$ as follows:

$$W(C) = T_1(C) \cup T_2(C) \cup \cdots \cup T_m(C). \tag{11.5}$$



(a) C_0 (b) C_1 (c) C_2

(d) C_3 (e) C_4 (f) C_5

Fig. 11.6. C_0 and the first five iterations C_1–C_5.

The fractal A that we wish to construct is a subset of the plane satisfying $W(A) = A$. We say that A is a *fixed point* of the operator W.

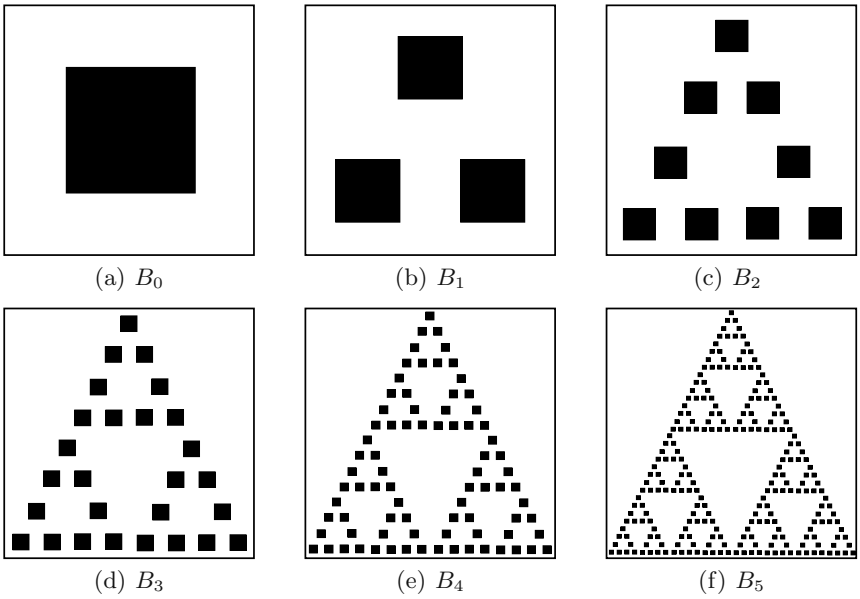
In the next section we will see that for all iterated function systems there exists a unique subset A of the plane that is a fixed point of the operator W. Moreover, we will show that for all nonempty subsets $C_0 \subset \mathbb{R}^2$, the subset A is the *limit* of the sequence $\{C_n\}$ defined by the recurrence

$$C_{n+1} = W(C_n).$$

The subset A is called the attractor of the iterated function system. Thus, if we know of a set B satisfying $B = W(B)$, then we know that B will be the limit of the sequence $\{C_n\}$.

In our Sierpiński triangle example we used the unit square $[0, 1] \times [0, 1]$ as our initial set C_0, and we constructed the sequence $\{C_n\}_{n\geq 0}$ using the recurrence $C_{n+1} = W(C_n)$. The experimental results of Figure 11.6 convinced us that the sequence $\{C_n\}_{n\geq 0}$ "converges" to the set A, the Sierpiński triangle. We could have performed this experiment with any initial set B_0, for example $B_0 = [1/4, 3/4] \times [1/4, 3/4]$. We would have obtained that the sequence $\{B_n\}_{n\geq}0$, where $B_{n+1} = W(B_n)$, again converges to A (Figure 11.7).

We can convince ourselves that we could have taken an initial set B_0 consisting only of a single point of the square C_0. In this case, the set B_n consists of 3^n points. If



(a) B_0 (b) B_1 (c) B_2

(d) B_3 (e) B_4 (f) B_5

Fig. 11.7. B_0 and the first five iterations B_1–B_5.

for each point in B_n we darken the corresponding pixel in a digitized image, then for sufficiently large n the image would resemble the Sierpiński triangle A.

In fact, traditional programs for drawing fractals function in a slightly different way, since it is simpler to draw a single point at each step than subsets of the plane consisting of 3^n points. We start by choosing a point P_0 in the rectangle R. At each step we randomly choose one of the transformations T_i and we calculate $P_n = T_{i_n}(P_{n-1})$, where T_{i_n} is the randomly chosen transformation at step n. If the point P_0 is already in the set A, then drawing the entire set of points from the sequence $\{P_n\}_{n\geq 0}$ will quickly begin to resemble A. If we are unsure whether P_0 is in A, then we discard the first M generated points $P_0, \ldots, P_{M-1}$, and draw the points $\{P_n\}_{n\geq M}$. The following section will show that there always exists a value for M that will ensure that we achieve a good approximation to A. In practice, M is often taken as small as 10, since convergence to the attractor usually occurs quite rapidly.

When drawing the Sierpiński triangle of Figure 11.5, at each step we randomly chose one of the transformations $\{T_1, T_2, T_3\}$. Thus, at step n we randomly chose a number $i_n \in \{1, 2, 3\}$ and applied the transformation T_{i_n}. Each time we generated 1 we applied T_1. If we generated 2 we applied T_2, and if we generated 3 we applied T_3. For the fern leaf this approach is not very efficient: we would spend too much time drawing points on the stalk and the bottom leaves and not enough time in the rest of the leaf. Let T_1 (respectively T_2, T_3, T_4) be the affine contraction that maps the leaf onto the

upper portion (respectively the left bottom branch, the right bottom branch, and the stalk) of the leaf. We will arrange it so that our random-number generator yields 1 with probability 85%, 2 and 3 with probabilities 7% each, and 4 with probability 1%. To accomplish this we actually generate random-numbers $\bar{a}_n$ in the range 1 to 100, choosing T_1 when $\bar{a}_n \in \{1, \ldots, 85\}$, T_2 when $\bar{a}_n \in \{86, \ldots, 92\}$, T_3 when $\bar{a}_n \in \{93, \ldots, 99\}$, and T_4 when $\bar{a}_n \in \{100\}$.

Mathematica program to draw the fern leaf of Figure 11.2 (The coefficients for the transforms T_i are taken from [1].)

```
chooseT := (r = RandomInteger[{1, 100}];
    If[r <= 85, 1,
      If[r <= 92, 2,
        If[r <= 99, 3, 4]]])

t = { (* { linear transformation, translation } *)
      {{{0.85, 0.04}, {-0.04, 0.85}}, {0., 1.6}},
      {{{0.2, -0.26}, {0.23, 0.22}}, {0., 1.6}},
      {{{-0.15, 0.28}, {0.26, 0.24}}, {0., 0.44}},
      {{{0., 0.}, {0., 0.16}}, {0., 0.}}
      };

transfoAff[t_, pt_] := t[[1]].pt + t[[2]]

nIteration = 20000; A = {{0., 0.}};
Do[AppendTo[A, transfoAff[t[[chooseT]], Last[A]]], {nIteration}]

ListPlot[A, AspectRatio -> Automatic, Axes -> False]
```

11.4 Iterated Contractions and Fixed Points

A full reading of this section requires some familiarity with analysis, but the basic concepts can be understood without it.

We noted previously that for all iterated function systems $\{T_1, \ldots, T_m\}$ there exists a unique subset A of the plane that is a fixed point of the operator W defined by

$$W(B) = T_1(B) \cup \cdots \cup T_m(B). \tag{11.6}$$

This set, satisfying $W(A) = A$, is called the attractor of the iterated function system. We will now justify this claim.

The following theorem from real analysis provides the key.

Theorem 11.6 *Let $f : \mathbb{R} \to \mathbb{R}$ be a contraction. In other words, there exists some $0 < r < 1$ such that for all $x, x' \in \mathbb{R}$ we have that*

$$|f(x) - f(x')| \leq r|x - x'|.$$

Then f has a unique fixed point $a \in \mathbb{R}$ such that $f(a) = a$.

We will prove this theorem in order to understand exactly how it works. While working through the proof, note that we can replace $\mathbb{R}$ by any closed interval $[\alpha, \beta]$ and more generally by any complete metric space (an intuitive definition of this follows). However, we are unable to replace $\mathbb{R}$ by $\mathbb{Q}$, nor by any open interval (α, β). When generalizing this theorem we will replace the notion of a point in $\mathbb{R}$ with that of a closed and bounded subset of $\mathbb{R}^2$, and the function f by the operator W defined in (11.6). We will require the notion of a *distance* between two subsets (the equivalent of $|x - x'|$ in the above formulation) and we will need to show that W is a contraction with respect to this distance. We would like to be able to use the same argument as will be used in the proof of Theorem 11.6 in order to prove the existence of a unique attractor A, a closed and bounded subset of $\mathbb{R}^2$ that is the fixed point of W.

Proof of Theorem 11.6: We start by showing that if f has a fixed point, then it must be unique. Suppose that $a_1 \neq a_2$ are two fixed points of f. Then $f(a_2) - f(a_1) = a_2 - a_1$ because they are both fixed points. However, since f is a contraction, we have that $|f(a_2) - f(a_1)| \leq r|a_2 - a_1|$, where $0 < r < 1$, a contradiction.

We must now prove the existence of a. To obtain a we will start with a point $x_0 \in \mathbb{R}$ and construct the sequence of its iterates $x_1 = f(x_0), x_2 = f(x_1), \ldots, x_{n+1} = f(x_n), \ldots$. If $x_1 = x_0$, then x_0 is a fixed point and we are done. Consider the case $x_1 \neq x_0$. Then

$$|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \leq r|x_n - x_{n-1}|.$$

By iterating we obtain

$$|x_{n+1} - x_n| \leq r^n |x_1 - x_0|.$$

We wish to show that the sequence $\{x_n\}$ converges to a point $a \in \mathbb{R}$ and that the limit a is a fixed point of f. A very powerful tool exists that permits us to show that a sequence of real numbers converges without having to guess a candidate for the limit: it suffices to show that it is a Cauchy sequence. (Recall that a sequence $\{x_n\}$ is a Cauchy sequence if $\forall \epsilon > 0\ \exists N \in \mathbb{N}$ such that if $n, m > N$ then $|x_n - x_m| < \epsilon$.) Suppose that $n > m$. Then

$$\begin{aligned} |x_n - x_m| &= |(x_n - x_{n-1}) + (x_{n-1} - x_{n-2}) + \cdots + (x_{m+1} - x_m)| \\ &\leq |x_n - x_{n-1}| + |x_{n-1} - x_{n-2}| + \cdots + |x_{m+1} - x_m| \\ &\leq (r^{n-1} + r^{n-2} + \cdots + r^m)|x_1 - x_0| \\ &\leq r^m(r^{n-m-1} + \cdots + 1)|x_1 - x_0| \\ &\leq \tfrac{r^m}{1-r}|x_1 - x_0|. \end{aligned}$$

For $|x_n - x_m|$ to be smaller than ϵ it suffices to take m sufficiently large, such that

$$\frac{r^m |x_1 - x_0|}{1 - r} < \epsilon,$$

or in other words, $r^m < \frac{\epsilon(1-r)}{|x_1-x_0|}$. Since $0 < r < 1$, we then take N large enough such that $\frac{r^N|x_1-x_0|}{1-r} < \epsilon$. Since $r^m < r^N$ for $N > m$ we have shown that the sequence $\{x_n\}$ is a Cauchy sequence.

Since every Cauchy sequence of real numbers converges to a real number, this yields that the sequence $\{x_n\}$ converges to some number $a \in \mathbb{R}$. We must now show that a is a fixed point of f. To do this we need to show that f is continuous. In fact, f is actually uniformly continuous on $\mathbb{R}$. Consider $\epsilon > 0$ and take $\delta = \epsilon$. Then if $|x - x'| < \delta$ we have that

$$|f(x) - f(x')| \leq r|x - x'| < r\delta = r\epsilon < \epsilon.$$

Since f is continuous, the image of the convergent sequence $\{x_n\}$ with limit a is itself a convergent sequence with limit $f(a)$. Thus

$$f(a) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} x_n = a.$$

□

We can generalize the statement of the previous theorem while maintaining the same proof. We can replace $\mathbb{R}$ by a general space K sharing certain properties with $\mathbb{R}$. In fact, we require only that K be a *complete metric space*. In order to keep the letters x and y for the Cartesian coordinates of a point we will denote points of K by the letters $v, w, \ldots$. Before we can elaborate on such spaces we must precisely define the notion of a *distance* $d(v, w)$ between two elements v, w of a space K. We will construct our definition of a distance so that it mirrors the properties of $|x - x'|$ in $\mathbb{R}$.

Definition 11.7 *1. A distance function $d(\cdot,\cdot)$ on a set K is a function $d : K \times K \to \mathbb{R}^+ \cup \{0\}$ that satisfies:*
(i) $d(v, w) \geq 0$;
(ii) $d(v, w) = d(w, v)$;
(iii) $d(v, w) = 0$ if and only if $v = w$;
(iv) Triangle inequality: for all v, w, z,

$$d(v, w) \leq d(v, z) + d(z, w).$$

2. A set K equipped with a distance function d is called a metric space.
3. A sequence $\{v_n\}$ of elements in K is a Cauchy sequence if $\forall\epsilon > 0$, $\exists N \in \mathbb{N}$ such that for all $m, n > N$, we have that $d(v_n, v_m) < \epsilon$.
4. A sequence $\{v_n\}$ of elements of K converges to an element $w \in K$ if $\forall\epsilon > 0$, $\exists N \in \mathbb{N}$ such that for all $n > N$, we have that $d(v_n, w) < \epsilon$. The element w is called the limit of the sequence $\{v_n\}$.
5. A metric space K is complete *if any Cauchy sequence of elements from K converges to a limit also in K.*

Example 11.8 *1. $\mathbb{R}^n$ with the Euclidean distance is a complete metric space.*

2. *Let K be the set of all closed and bounded subsets of $\mathbb{R}^2$: we call them* compact *subsets of $\mathbb{R}^2$. The distance we will use over this set of subsets is the Hausdorff distance, which will be defined and discussed in Section 11.5. Equipped with this distance, K will be a complete metric space (the proof of this fact can be found in [1]).*
3. *When moving from theory to practice in Section 11.7, we will consider a black and white photo on a rectangle R as a function $f : R \to S$, where S denotes the set of gray tones. We can then define a distance between two such functions f and f' through the use of the following definitions:*

$$d_1(f, f') = \max_{(x,y)\in R} |f(x, y) - f'(x, y)|$$

and

$$d_2(f, f') = \left(\iint_R (f(x, y) - f'(x, y))^2 \, dx \, dy \right)^{1/2}. \tag{11.7}$$

Equipped with these distances, the set of functions $f : R \to S$ is a complete metric space. We can replace the set $R = [a, b] \times [c, d]$ by a discrete set of pixels over the rectangle R by adapting slightly the above definitions. For example, the double integral in the distance function will be replaced by a discrete sum over the individual pixels. If x and y take the values $\{0, \ldots, h-1\}$ and $\{0, \ldots, v-1\}$) respectively, then the distance (11.7) *becomes*

$$d_3(f, f') = \left(\sum_{x=0}^{h-1} \sum_{y=0}^{v-1} (f(x, y) - f'(x, y))^2 \right)^{1/2}. \tag{11.8}$$

We require that the operator W defined in (11.5) be a contraction with respect to the distance function over the space K. This leads us to the famous Banach fixed-point theorem: since we will apply it with the elements of K being compact subsets of $\mathbb{R}^2$, we will use capital letters for the elements of K.

Theorem 11.9 (Banach fixed-point Theorem) *Let K be a complete metric space and $W : K \to K$ a contraction. In other words, let W be a function such that for all $B_1, B_2 \in K$,*

$$d(W(B_1), W(B_2)) \leq r \, d(B_1, B_2) \tag{11.9}$$

with $0 < r < 1$. Then there exists a unique fixed point $A \in K$ of W such that $W(A) = A$.

We will not give a proof of the Banach fixed-point theorem, since it is exactly the same as that of Theorem 11.6. We only need to replace $|x - x'|$ by $d(B, B')$.

The Banach fixed-point theorem is one of the most important theorems in mathematics. It has applications in many diverse areas.

Example 11.10 *We discuss a few applications of the Banach fixed-point theorem:*

1. *A first classical application of this theorem allows us to prove the existence and uniqueness of solutions to ordinary differential equations satisfying a Lipschitz condition. In this example the elements of K are functions. The fixed point is the unique function that is a solution to the differential equation. We will not go further into this example. However, we wish to point out that simple ideas often have important applications in seemingly unrelated fields.*
2. *The second application is of immediate interest. Let K be the set of all closed and bounded subsets of the plane, together with the Hausdorff distance. Equipped with this distance, K will be a complete metric space. Consider a set of affine contractions $T_1, \ldots, T_m$ forming an iterated function system. We define the operator of* (11.6)*, and we will show that it is a contraction, satisfying* (11.9) *for some $0 < r < 1$. Theorem 11.9 immediately proves both the existence and uniqueness of the attractor A of such an iterated function system.*

Remark: The Banach theorem states that the fixed point A of a contraction W must be *unique*. Thus, if we are already aware of a set A satisfying this property (for example, the fern leaf), then we are sure that it is indeed the fixed point of the iterated function system we have constructed.

11.5 The Hausdorff Distance

The definition of this distance function is somewhat difficult. Thus, we will start by discussing the intuitive foundations on which it was built. The proof of the Banach fixed-point theorem uses the distance function only as a tool for discussing convergence and for discussing the closeness of two elements of K. When we talk of the convergence of a sequence of sets B_n in K to some set A, intuitively we wish to show that for sufficiently large n, the sets B_n strongly resemble A.

Thus, we wish to quantify the notion of closeness between two sets B_1 and B_2, such that we can say precisely when two sets are within some distance ϵ of each other. One way of doing this is to consider "inflating" the set B_1 by an amount ϵ. That is, we consider the set of all points within a distance ϵ of some point in B_1. If the distance between B_1 and B_2 is less than ϵ, then B_2 should be entirely contained in the inflated version of B_1. The ϵ-inflated set B_1 is given by

$$B_1(\epsilon) = \{v \in \mathbb{R}^2 | \exists w \in B_1 \text{ such that } d(v, w) < \epsilon\},$$

where $d(v, w)$ is the usual Euclidean distance between v and w, both points of $\mathbb{R}^2$. We require that $B_2 \subset B_1(\epsilon)$. However, this is not sufficient. The set B_2 could have a very different form and be much smaller than B_1. Thus, we also consider inflating B_2,

$$B_2(\epsilon) = \{v \in \mathbb{R}^2 | \exists w \in B_2 \text{ such that } d(v, w) < \epsilon\},$$

and requiring that $B_1 \subset B_2(\epsilon)$. We denote by $d_H(B_1, B_2)$ the *Hausdorff distance* between B_1 and B_2, which remains to be precisely defined. We want that

$$d_H(B_1, B_2) < \epsilon \iff (B_1 \subset B_2(\epsilon) \quad \text{and} \quad B_2 \subset B_1(\epsilon)).$$

This intuitive idea of inflating a set until it subsumes another helps to make sense of the formal definition of the Hausdorff distance.

Definition 11.11 *1. Let B be a compact (closed and bounded) subset of $\mathbb{R}^2$ and let $v \in \mathbb{R}^2$. The distance of v to B, denoted by $d(v, B)$, is*

$$d(v, B) = \min_{w \in B} d(v, w).$$

2. The Hausdorff distance *between two compact sets B_1 and B_2 of $\mathbb{R}^2$ is*

$$d_H(B_1, B_2) = \max\left(\max_{v \in B_1} d(v, B_2), \max_{w \in B_2} d(w, B_1)\right).$$

Remarks:

(i) The condition that B, B_1, and B_2 be compact ensures that the minima and maxima in Definition 11.11 do indeed exist.

(ii) Given the following fact regarding maxima,

$$\max(a, b) < \epsilon \iff (a < \epsilon \quad \text{and} \quad b < \epsilon),$$

we have that

$$d_H(B_1, B_2) < \epsilon$$

if and only if

$$\max_{v \in B_1} d(v, B_2) < \epsilon \quad \text{and} \quad \max_{w \in B_2} d(w, B_1) < \epsilon$$

if and only if

$$B_1 \subset B_2(\epsilon) \quad \text{and} \quad B_2 \subset B_1(\epsilon).$$

Thus, the Hausdorff distance is intimately related to the concept of inflated sets.

We state the following theorem without proof:

Theorem 11.12 *[1] Let K be the set of all compact subsets of the plane. Then the Hausdorff distance over K is a distance function by Definition 11.7. Moreover, K equipped with the Hausdorff distance is a complete metric space.*

Our set K, equipped with the Hausdorff distance, is a complete metric space. We defined the operator $W : K \to K$ in (11.6). In order to apply Banach fixed-point theorem we must now show that W is a contraction.

To do this we first clarify the notion of the contraction factor r in the context of affine transformations.

Definition 11.13 *Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be an affine contraction.*

1. *A real number $r \in (0,1)$ is a* contraction factor *for T if for all $v, w \in \mathbb{R}^2$ we have that*
$$d(T(v), T(w)) \leq rd(v, w).$$

2. *A contraction factor r is an* exact contraction factor *if for all $v, w \in \mathbb{R}^2$ we have that*
$$d(T(v), T(w)) = rd(v, w).$$

Remark: Only affine transformations whose linear part is some composition of a homothety, a rotation, and a reflection with respect to a line have exact contraction factors.

Theorem 11.14 *Let $\{T_1, \ldots, T_m\}$ be an iterated function system such that each T_i has contraction factor $r_i \in (0,1)$. Then the operator W defined in* (11.5) *is a contraction with contraction factor $r = \max(r_1, \ldots, r_m)$.*

The proof of this theorem requires the following lemmas regarding the Hausdorff distance.

Lemma 11.15 *Let $B, C, D, E \in K$. Then*
$$d_H(B \cup C, D \cup E) \leq \max(d_H(B, D), d_H(C, E)).$$

Proof: By our remark following Definition 11.11 it suffices to show that:

(i) for all $v \in B \cup C$ we have that
$$d(v, D \cup E) \leq d_H(B, D) \leq \max(d_H(B, D), d_H(C, E))$$
or
$$d(v, D \cup E) \leq d_H(C, E) \leq \max(d_H(B, D), d_H(C, E));$$

(ii) and for all $w \in D \cup E$ we have that
$$d(w, B \cup C) \leq d_H(B, D) \leq \max(d_H(B, D), d_H(C, E))$$
or
$$d(w, B \cup C) \leq d_H(C, E) \leq \max(d_H(B, D), d_H(C, E)).$$

We will prove only (i), since the proof of (ii) is completely similar. Let $v \in B \cup C$ be a given point. Since D and E are both compact sets, there exists $z \in D \cup E$ such that $d(v, D \cup E) = d(v, z)$. Thus we have that for all $w \in D \cup E$, $d(v, z) \leq d(v, w)$. In particular, for all $u \in D$ we have that $d(v, z) \leq d(v, u)$, or equivalently $d(v, z) \leq d(v, D)$. Additionally, for all $p \in E$, we have that $d(v, z) \leq d(v, p)$, yielding $d(v, z) \leq d(v, E)$. However, $v \in B \cup C$; hence $v \in B$ or $v \in C$. If $v \in B$ we have that

$$d(v, D) \leq d_H(B, D) \leq \max(d_H(B, D), d_H(C, E)).$$

Similarly, if $v \in C$ we see that

$$d(v, E) \leq d_H(C, E) \leq \max(d_H(B, D), d_H(C, E)).$$

The rest of (i) follows from the fact that $d(v, D \cup E) \leq d(v, D)$ and $d(v, D \cup E) \leq d(v, E)$ (Exercise 14). □

Lemma 11.16 *If $T : \mathbb{R}^2 \to \mathbb{R}^2$ is an affine contraction with contraction factor $r \in (0, 1)$, then the mapping $T : K \to K$ (again labeled T through a slight abuse of notation) defined by*

$$T(B) = \{T(v) | v \in B\}$$

is a contraction on K with the same contraction factor r.

Proof: Consider $B_1, B_2 \in K$. We have to show that

$$d_H(T(B_1), T(B_2)) \leq r d_H(B_1, B_2).$$

As before, it suffices to show that

(i) for all $v \in T(B_1)$ we have that $d(v, T(B_2)) \leq r d_H(B_1, B_2)$;
(ii) and for all $w \in T(B_2)$ we have that $d(w, T(B_1)) \leq r d_H(B_1, B_2)$.

Again, we will prove only (i), since the proof of (ii) is analogous. Since $v \in T(B_1)$, we see that $v = T(v')$ for some $v' \in B_1$. Let $w \in T(B_2)$. Then $d(v, T(B_2)) \leq d(v, w)$. Choose $w' \in B_2$ such that $w = T(w')$. Then it follows that

$$d(v, T(B_2)) \leq d(v, w) = d(T(v'), T(w')) \leq r d(v', w').$$

Since this holds true for all $w' \in B_2$, we deduce that

$$d(v, T(B_2)) \leq r d(v', B_2) \leq r d_H(B_1, B_2).$$

□

Proof of Theorem 11.14. The proof proceeds by induction on the number of transformations m defining the operator W. We will show that if T_i, $i = 1, \ldots, m$, are contractions with contraction factors r_i, then W is a contraction with contraction factor $r = \max(r_1, \ldots, r_m)$. The case $m = 1$ follows immediately from Lemma 11.16.

Although it is not necessary to explicitly treat the case for $m = 2$, we will nonetheless do so in order to clearly illustrate the idea behind the proof before applying it to the general case. If $m = 2$, $W(B) = T_1(B) \cup T_2(B)$. We see that

$$\begin{aligned} d_H(W(B), W(C)) &= d_H(T_1(B) \cup T_2(B), T_1(C) \cup T_2(C)) \\ &\leq \max(d_H(T_1(B), T_1(C)), d_H(T_2(B), T_2(C))) \\ &\leq \max(r_1 d_H(B, C), r_2 d_H(B, C)) \\ &= \max(r_1, r_2) d_H(B, C), \end{aligned}$$

by successively applying Lemmas 11.15 and 11.16.

Suppose that the theorem holds for a system of m iterated functions and consider the case of $m+1$ functions. In this case, we have that $W(B) = T_1(B) \cup \cdots \cup T_{m+1}(B)$. It follows that

$$\begin{aligned} d_H(W(B), W(C)) &= d_H(T_1(B) \cup \cdots \cup T_{m+1}(B), T_1(C) \cup \cdots \cup T_{m+1}(C)) \\ &= d_H\left(\left(\bigcup_{i=1}^{m} T_i(B)\right) \cup T_{m+1}(B), \left(\bigcup_{i=1}^{m} T_i(C)\right) \cup T_{m+1}(C)\right) \\ &\leq \max\left(d_H\left(\bigcup_{i=1}^{m} T_i(B), \bigcup_{i=1}^{m} T_i(C)\right), d_H(T_{m+1}(B), T_{m+1}(C))\right) \\ &\leq \max(\max(r_1, \ldots, r_m) d_H(B, C), r_{m+1} d_H(B, C)) \\ &\leq \max(r_1, \ldots, r_{m+1}) d_H(B, C), \end{aligned}$$

by the inductive hypothesis and the application of Lemmas 11.15 and 11.16. □

Theorem 11.14 assures us that regardless of $B \subset \mathbb{R}^2$, the Hausdorff distance between two consecutive iterates $W^n(B)$ and $W^{n+1}(B)$ decreases as n increases, since

$$d_H(W^n(B), W^{n+1}(B)) \leq r d_H(W^{n-1}(B), W^n(B)) \leq \cdots \leq r^n d_H(B, W(B)),$$

where $r \in (0, 1)$. This does not, however, allow us to say anything about the distance between B and the attractor A. This question is addressed in the following result, Barnsley's *collage theorem*.

Theorem 11.17 (Barnsley's collage theorem [1]) *Let $\{T_1, \ldots, T_m\}$ be an iterated function system with contraction factor $r \in (0, 1)$ and attractor A. Let B and $\epsilon > 0$ be chosen such that*

$$d_H(B, T_1(B) \cup \cdots \cup T_m(B)) \leq \epsilon.$$

Then

$$d_H(B, A) \leq \frac{\epsilon}{1 - r}. \tag{11.10}$$

Proof: We will reuse a portion of the proof of Theorem 11.6 in order to bound the distance $d_H(B, W^n(B))$. By the triangle inequality we have that

$$\begin{aligned} d_H(B, W^n(B)) &\leq d_H(B, W(B)) + \cdots + d_H(W^{n-1}(B), W^n(B)) \\ &\leq (1 + r^1 + \cdots + r^{n-1}) d_H(B, W(B)) \\ &\leq \tfrac{1-r^n}{1-r} d_H(B, W(B)) \\ &\leq \tfrac{1}{1-r} d_H(B, W(B)) \leq \tfrac{\epsilon}{1-r}. \end{aligned}$$

Consider an arbitrary $\eta > 0$. Then there exists N such that if $n > N$ then $d_H(W^n(B), A) < \eta$. Thus, if $n > N$ we have that

$$d_H(B, A) \leq d_H(B, W^n(B)) + d_H(W^n(B), A) < \frac{\epsilon}{1-r} + \eta.$$

Since this inequality holds for all $\eta > 0$, we can conclude that $d_H(B, A) \leq \frac{\epsilon}{1-r}$. □

The collage theorem is extremely important for practical applications of iterated function systems. In fact, suppose that rather than the mathematically precise fern leaf of Figure 11.2, we considered a photograph of a real fern leaf; call it B. It is possible (and quite likely) that there does not exist any collection of four affine transformations $T_1, \ldots, T_4$ such that $B = T_1(B) \cup T_2(B) \cup T_3(B) \cup T_4(B)$. We have only that B is approximately equal to

$$C = T_1(B) \cup T_2(B) \cup T_3(B) \cup T_4(B)$$

for four affine transformations $T_1, \ldots, T_4$. If we were now to construct (using a computer, for example) the attractor A of the iterated function system $\{T_1, \ldots, T_4\}$ and if $d_H(B, C) \leq \epsilon$, then the collage theorem assures us that $d_H(A, B) \leq \frac{\epsilon}{1-r}$. In other words, A will resemble B. Thus our method is "robust": it performs well when we approximate arbitrary images.

11.6 The Fractal Dimension of the Attractor of an Iterated Function System

It is not necessary to have seen the entire previous section in order to cover this section. In fact, it suffices to be familiar with the definition of a contraction factor (Definition 11.13).

We have constructed several iterated function systems $\{T_1, \ldots, T_m\}$ (where T_i has contraction factor r_i) and their attractors, for example the Sierpiński triangle and the fern leaf. Given their richly repeating structure, these objects seem in some ways more "dense" than simple curves through the plane. However, somewhat counterintuitively we can actually show that they have zero area assuming that $r_1^2 + \cdots + r_m^2 < 1$, which is the case in both of our examples.

Proposition 11.18 *Consider the attractor A of an iterated function system $\{T_1, \ldots, T_m\}$ with contraction factors $r_1, \ldots, r_m$ (respectively). If*

$$r_1^2 + \cdots + r_m^2 < 1, \tag{11.11}$$

then it follows that A has zero area.

PROOF. Let $S(B)$ be the area of a compact subset B of $\mathbb{R}^2$. Then $S(T_i(B)) \le r_i^2 S(B)$ and therefore $S(W(B_0)) \le (r_1^2+\cdots+r_m^2)S(B_0)$. If $B_{n+1} = W(B_n)$, iterating then yields that

$$S(B_{n+1}) \le (r_1^2 + \cdots + r_m^2)S(B_n) \le \cdots \le (r_1^2 + \cdots + r_m^2)^{n+1} S(B_0).$$

Hence

$$\lim_{n\to\infty} S(B_n) = S(A) = 0.$$

□

Thus we see that the notion of area is not adequate to express that such objects are denser than a simple curve: their area is zero. In some sense, these fractal objects are "more than a curve but something less than an surface." This concept will be formalized by formally defining *dimension*. To be consistent with the usual definition of dimension we require a definition that will evaluate to 1 for simple curves, 2 for surfaces, and 3 for volumes. At the same time, we wish the value to be calculable for the fractals we are considering here. Since the attractors we are considering fall somewhere between a curve and a surface, their dimensions should lie between 1 and 2. *Any coherent theory of dimension must yield noninteger values for certain fractal objects.*

There are several definitions of *dimension*. They all coincide with the usual values for curves, surfaces, and volumes. However, they may differ for fractal objects. We will consider only *fractal dimension*.

Start by considering the line segment $[0,1]$, the square $[0,1]\times[0,1]$, and the cube $[0,1]^3$ and take small segments of length $1/n$, small squares with side length $1/n$, and small cubes with edge length $1/n$.

- The segment $[0,1]$ can be considered in $\mathbb{R}$, $\mathbb{R}^2$, or $\mathbb{R}^3$. In each case we can cover the entire original segment with n small segments of length $1/n$, n small squares with side length $1/n$, or n small cubes with side length $1/n$.
- The square $[0,1]^2$ may be considered in $\mathbb{R}^2$ or $\mathbb{R}^3$. We require n^2 small squares or small cubes to cover it, while it may not be covered by any finite number of small line segments.
- The cube $[0,1]^3$ can be considered only in $\mathbb{R}^3$. In this space it can be covered by n^3 small cubes, while no finite number of small segments or squares will do.
- If we had considered the segment $[0,L]$ instead of $[0,1]$ we would have required roughly nL small segments, squares, or cubes to cover it.
- If we had considered the square $[0,L]^2$ rather than $[0,1]^2$ we would have required roughly n^2L^2 small squares, or cubes to cover it.
- If we had considered the cube $[0,L]^3$ rather than $[0,1]^3$ we would have required roughly n^3L^3 small cubes to cover it.

We try to extract a general rule from the above observations:

(i) If we had a finite differentiable curve through $\mathbb{R}^2$ or $\mathbb{R}^3$, we would require a finite number $N(1/n)$ of small squares or small cubes with side or edge length $\frac{1}{n}$ to cover it such that, provided n is large enough,

$$C_1 n \leq N(1/n) \leq C_2 n.$$

The above statement requires some thought to convince ourselves of its validity. If the curve has length L we can cut it into Ln pieces with length less than or equal to $\frac{1}{n}$, and each such piece can be covered by a small square (cube) of side (edge) length $\frac{1}{n}$. Thus, $N(1/n) \leq C_2 n$ for some C_2. The other inequality is harder to get, and valid only for sufficiently large n. In fact, the curve could be sufficiently winding that a square or cube of side length $\frac{1}{n}$ could actually contain a long length of it. However, since the curve is differentiable (and not fractal), the width of the smallest kink is bounded below. Thus, if we take $\frac{1}{n}$ sufficiently small, then a small square or cube can possibly contain only a portion of the curve of length less than or equal to $C\frac{1}{n}$. The minimum number of squares or cubes will thus be greater than or equal to $C_1 n$, where $C_1 = \frac{L}{C}$.

(ii) Similarly, had we considered a finite smooth surface in the plane or in space, we would require a finite number $N(1/n)$ of small cubes with edge length $\frac{1}{n}$ to cover it such that, provided n is sufficiently large,

$$C_1 n^2 \leq N(1/n) \leq C_2 n^2.$$

(iii) Finally, a volume of space will require a number $N(1/n)$ of small cubes with edge length $\frac{1}{n}$ to cover it such that, when n is large enough,

$$C_1 n^3 \leq N(1/n) \leq C_2 n^3.$$

(iv) The number $N(1/n)$ is of roughly the same size, regardless of the space we are working in! In fact, whether we consider covering a curve with segments, squares, or cubes we will obtain roughly the same value.

Thus we see that the dimension of the object corresponds to the exponent of n in the order of magnitude of $N(1/n)$ and that the constants C_1 and C_2 are unimportant. In each case we can verify that the dimension corresponds to

$$\lim_{n\to\infty} \frac{\ln N(1/n)}{\ln n}.$$

In fact, in the case of a curve we have

$$\frac{\ln(C_1 n)}{\ln n} \leq \frac{\ln N(1/n)}{\ln n} \leq \frac{\ln(C_2 n)}{\ln n}.$$

Since $\ln(C_i n) = \ln C_i + \ln n$, then

$$\lim_{n\to\infty} \frac{\ln(C_i n)}{\ln n} = 1.$$

We can use the same reasoning with surfaces and volumes to obtain dimensions of 2 and 3.

We will now give the formal definition of fractal dimension. Rather than just considering side lengths of $1/n$, we will generalize the above concepts to permit segments, squares, and cubes with side length ϵ for any small $\epsilon > 0$.

Definition 11.19 *We consider a compact subset B of $\mathbb{R}^i$, $i = 1, 2, 3$. Let $N(\epsilon)$ be the minimum number of small segments (respectively squares or cubes) with length (respectively side length, edge length) ϵ necessary to cover B. Then the* fractal dimension $D(B)$ *of B is, provided it exists, the limit*

$$D(B) = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln 1/\epsilon}.$$

Remark:

1. In the previous definition, suppose that B is a subset of a line in $\mathbb{R}^3$. Then the previous definition leads to the same limit whether we cover B using segments, squares, or cubes. A similar observation applies if B is a subset of a line in $\mathbb{R}^2$.
2. The wording of the definition implies that the limit may not always exist. The fractals we have constructed up until now are *self-similar*, which means that at any scale we see the same repeating structure. In this case, we can show that the limit exists. However, the limit may not exist if B is very complicated and not self-similar.

Definition 11.20 *An iterated function system $\{T_1, \ldots, T_m\}$ with attractor A is* totally disconnected *if the sets $T_1(A), \ldots, T_m(A)$ are disjoint.*

We present the following theorem without proof.

Theorem 11.21 *Let A be the attractor of a totally disconnected iterated function system. Then the limit defining its fractal dimension exists.*

Example 11.22 *We calculate the dimension of the Sierpiński triangle A. From Figure 11.6 it is possible to count the number of squares with side length $\frac{1}{2^n}$ required to cover A.*

- *We need one square with side length 1 to cover A: $N(1) = 1$.*
- *We need three squares with side length $\frac{1}{2}$ to cover A: $N(\frac{1}{2}) = 3$.*
- *We need nine squares with side length $\frac{1}{4}$ to cover A: $N(\frac{1}{4}) = 9$.*
- ...
- *We need 3^n squares with side length $\frac{1}{2^n}$ to cover A: $N(\frac{1}{2^n}) = 3^n$.*

Letting $\epsilon = 1/2^n$ we have that $\epsilon_n \to 0$ as $n \to \infty$. Since the limit defining the dimension $D(A)$ of the Sierpiński triangle exists by Theorem 11.21, this limit is equal to

$$D(A) = \lim_{n \to \infty} \frac{\ln N(1/2^n)}{\ln(2^n)} = \lim_{n \to \infty} \frac{n \ln 3}{n \ln 2} = \frac{\ln 3}{\ln 2} \approx 1.58496.$$

Thus $1 < D(A) < 2$. As announced earlier, the dimension of A is therefore greater than that of a curve but less than that of a surface.

The method of Example 11.22 can be quite difficult for complicated attractors. We now present a theorem that allows a direct calculation of the fractal dimension of an attractor without our having to explicitly count covering squares.

Theorem 11.23 *Let $\{T_1, \ldots, T_m\}$ be a totally disconnected iterated function system where T_i has the exact contraction factor $0 < r_i < 1$. Let A be its attractor. Then the fractal dimension $D = D(A)$ of A is the unique solution to the equation*

$$r_1^D + \cdots + r_m^D = 1. \tag{11.12}$$

In the particular case $r_1 = \cdots = r_m = r$, we have that

$$D(A) = \frac{\ln m}{-\ln r} = -\frac{\ln m}{\ln r}. \tag{11.13}$$

(The quotient is positive, since $\ln r < 0$.)

SKETCH OF PROOF: We start by verifying that (11.13) is a consequence of (11.12). In fact, if $r_1 = \cdots = r_m = r$, then (11.12) yields

$$r^D + \cdots + r^D = m r^D = 1.$$

From this it follows that $r^D = 1/m$. Taking the logarithm of both sites yields

$$D \ln r = \ln 1/m = -\ln m,$$

from which the result follows.

We provide an intuitive sketch of the proof for the first equation. Let A be the attractor of the system and let $N(\epsilon)$ be the minimum number of squares with side length ϵ necessary to cover it. Since A is the disjoint union of $T_1(A), \ldots, T_m(A)$, then $N(\epsilon)$ is roughly equal to $N_1(\epsilon) + \cdots + N_m(\epsilon)$, where $N_i(\epsilon)$ is the number of such squares required to cover $T_i(A)$. This approximation becomes better and better as ϵ approaches 0. The set $T_i(A)$ is obtained from A by applying an affine contraction with an exact contraction factor of r_i. Thus T_i is the composition of a homothety of factor r_i and an isometry, preserving angles and distances. It follows that if we require $N_i(\epsilon)$ squares with side length ϵ to cover $T_i(A)$, then applying T_i^{-1} to these squares gives us $N_i(\epsilon)$ squares with side length ϵ/r_i covering A. Hence

$$N(\epsilon/r_i) \approx N_i(\epsilon).$$

We therefore have that

$$N(\epsilon) \approx N(\epsilon/r_1) + \cdots + N(\epsilon/r_m). \tag{11.14}$$

In this form it is difficult to calculate the limit $\lim_{\epsilon \to 0} N(\epsilon)$. Thus we suppose that $N(\epsilon) \approx C\epsilon^{-D}$, where D is the dimension (here we are giving only an intuitive argument!); this is certainly the case for the segments, squares, and cubes considered in our simple examples. With this assumption, equation (11.14) yields

$$C\epsilon^{-D} = C\left(\frac{\epsilon}{r_1}\right)^{-D} + \cdots + C\left(\frac{\epsilon}{r_m}\right)^{-D}.$$

We can simplify $C\epsilon^{-D}$, leaving us with

$$1 = \frac{1}{r_1^{-D}} + \cdots + \frac{1}{r_m^{-D}} = r_1^D + \cdots + r_m^D.$$

□

Example 11.24 *For the Sierpiński triangle we have that $r = 1/2$ and $m = 3$. Thus the theorem gives us a direct way to calculate its dimension as $\frac{\ln 3}{\ln 2} \approx 1.58496$, the same value obtained by directly counting covering squares as shown in Example 11.22.*

Calculating $D(A)$ when the r_i are not all equal and satisfy equation (11.11). Even if it is not simple to give a completely rigorous proof, an inspection of several examples convinces us that the condition of equation (11.11) is often satisfied by totally disconnected iterated function systems. Equation (11.12) cannot be solved exactly, but we can use numerical methods. To begin with, we know that the dimension lies in the range $[0, 2]$. The function

$$f(D) = r_1^D + \cdots + r_m^D - 1$$

is strictly decreasing on $[0, 2]$, since

$$f'(D) = r_1^D \ln r_1 + \cdots + r_m^D \ln r_m < 0.$$

Indeed, the condition $r_i < 1$ implies that $\ln r_i < 0$. Moreover, $f(0) = m - 1 > 1$ and $f(2) = r_1^2 + \cdots + r_m^2 - 1 < 0$ by (11.11). Thus by the intermediate value theorem the function $f(D)$ must have a unique root in $[0, 2]$. We may graph this function or use any numerical root-finding procedure (such as Newton's method) to find the solution to the desired accuracy.

Example 11.25 *Consider a totally disconnected iterated function system $\{T_1, T_2, T_3\}$ with contraction factors $r_1 = 0.5$, $r_2 = 0.4$, and $r_3 = 0.7$. Figure 11.8(a) shows the graph of the function*

$$f(D) = 0.5^D + 0.4^D + 0.7^D - 1$$

for $D \in [0, 2]$. Figure 11.8(b) shows the same function for $D \in [1.75, 1.85]$, allowing us to evaluate the root with higher precision. Inspection shows that $D(A) \approx 1.81$.

11.7 Photographs as Attractors to Iterated Function Systems?

Everything we have seen up until now is elegant from a theoretical point of view, but it does not really help us compress images. We have seen that iterated function systems

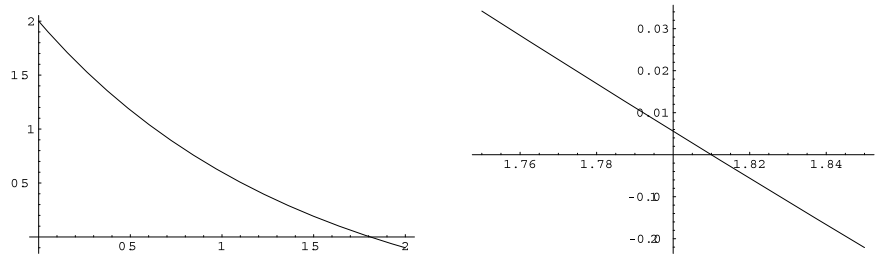


Fig. 11.8. The graph of $f(D)$ for $D \in [0, 2]$ and $D \in [1.75, 1.85]$ for Example 11.25.

allow us to store in memory a fractal image with a very short program. However, to take advantage of this powerful compression we must be able to recognize portions of an image that exhibit strong self-similarity and write short programs constructing them. Are all the parts of an image describable in such a fractal manner? Probably not! Even if a human is able to approximate certain photographs using carefully crafted iterated function system (there are some nice examples in [1]), this is far from providing a systematic algorithm that can operate on hundreds of photographs. If we wish to apply iterated function systems to image compression, we must broaden the ideas we have developed in this chapter.

The concepts of this chapter will thus be applied slightly differently. The common point is that we will still be using a specific type of iterated function system (called *partitioned iterated function system*) whose attractor will approximate the image we wish to compress. The following discussion was inspired by [2]. Research is ongoing to find better-performing alternative methods.

Representing an image as the graph of a function. We discretize a photograph by considering it as a finite set of squares with varying intensity, called *pixels* (for *picture elements*). We associate each pixel in the photo with a number representing its color. To simplify our discussion we will limit ourselves to grayscale images. Thus each point (x, y) of a rectangular photo is associated with a value z that represents its gray tone. Most digital photographs assign integer values in the range $\{0, \ldots, 255\}$ corresponding to black through white, with 0 representing black and 255 representing white. Thus, a photograph may be viewed as a two-dimensional function. If a photograph contains h pixels horizontally and v vertically and we denote by S_N the set $\{0, 1, 2, \ldots, N-1\}$, then a photograph is a function

$$f : S_h \times S_v \longrightarrow S_{255}.$$

In other words, it is a function that associates a gray tone

$$z = f(x, y) \in \{0, \ldots, 255\}$$

to every pixel (x, y) for $0 \le x \le h-1$ and $0 \le y \le v-1$. The iterated functions that we will introduce will transform a photograph f into another photograph f' whose gray tones will not always be integers between 0 and 255. Thus it will be easier for us to work with functions

$$f : S_h \times S_v \longrightarrow \mathbb{R}.$$

Constructing a partitioned iterated function system. A partitioned iterated function system acts on the set $\mathcal{F} = \{f : S_h \times S_v \to \mathbb{R}\}$ of all photographs. Here is how such a system is constructed for an arbitrary photograph. We divide the image into disjoint neighboring tiles of 4×4 pixels. Each such tile C_i is called a *small tile*, and I is the set of all small tiles. We also consider the set of all possible 8×8 tiles, called big tiles. Each small tile C_i is associated with the big tile G_i that "resembles" it the most (see Figure 11.9). (We will precisely define what we mean by "resemble" a little later.)

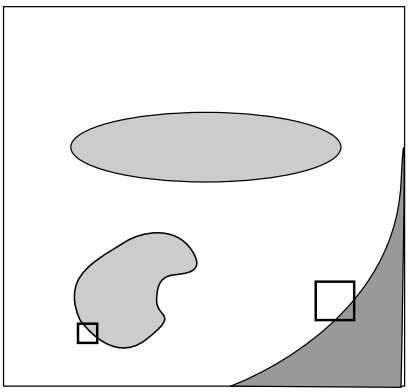


Fig. 11.9. Choosing a big tile that resembles a small tile.

Each point in the image is represented by its coordinates (x, y, z), where z is the gray tone of the pixel at (x, y). An affine transformation T_i will be chosen that maps a big tile G_i onto a small tile C_i, where T_i has the form

$$T_i \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} \alpha_i \\ \beta_i \\ g_i \end{pmatrix}. \tag{11.15}$$

Restricting ourselves to the integer coordinates (x, y), this transformation is a simple affine contraction

$$t_i(x, y) = (a_i x + b_i y + \alpha_i, c_i x + d_i y + \beta_i). \tag{11.16}$$

Consider now the gray tone of the tile. The parameter s_i serves to modify the spread of the gray tones used in the tile: if $s_i < 1$ then the small tile C_i has less contrast than the large tile G_i, while it has more contrast if $s_i > 1$. The parameter g_i corresponds to a translation of the grayscale. If $g_i < 0$ then the large tile is paler than the small tile and vice versa (remember that 0 is black and 255 is white). In practice, since a large tile ($8 \times 8 = 64$) contains four times as many pixels as a small tile ($4 \times 4 = 16$), we start by replacing the color of each 2×2 block of G_i by a uniform color given by the average color of the four pixels originally located there. We compose this operation with the transformation T_i, calling the composition $\overline{T}_i$. Since the sides of a large tile are mapped to those of a small tile, the parameters of the linear part $\left(\begin{smallmatrix} a_i & b_i \\ c_i & d_i \end{smallmatrix}\right)$ of the transformation T_i are greatly limited. In fact, the linear portion of the transformation will be the composition of the homothety of scale $1/2$,

$$(x, y) \mapsto (x/2, y/2),$$

and one of the eight following transformations:

1. the identity transform with matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$;
2. rotation by $\pi/2$ with matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$;
3. rotation by π with matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, also called symmetry with respect to the origin;
4. rotation by $-\pi/2$ with matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$;
5. reflection about the horizontal axis with matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$;
6. reflection about the vertical axis with matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$;
7. reflection about the first diagonal axis with matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$;
8. reflection about the second diagonal axis with matrix $\left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$.

Note that all of the matrices associated with these linear transformations are orthogonal. (Exercise: which of the above transformations will be used in mapping the big tile to the small tile in Figure 11.9?)

To decide whether two tiles resemble each other we will define a distance function d. The partitioned iterated function system we construct will produce iterates approaching a limit with respect to this same distance as applied to the set $\mathcal{F}$ of all photographs. If $f, f' \in \mathcal{F}$, that is, both f and f' are digitized images of the same size, then the distance between them is defined as

$$d_{h\times v}(f, f') = \sqrt{\sum_{x=0}^{h-1} \sum_{y=0}^{v-1} (f(x, y) - f'(x, y))^2},$$

corresponding to the distance d_3 given in (11.8) of Example 11.8. This distance may seem somewhat intimidating when written out, but it is simply the Euclidean distance on the vector space $\mathbb{R}^{h\times v}$. To decide whether a small tile C_i resembles a large tile G_i we define a similar distance between G_i and C_i. In fact, we calculate the distance between f_{C_i} (the image function restricted to the small tile C_i) and $\overline{f}_{C_i} = \overline{T}_i(f_{G_i})$, that is, the image by $\overline{T}_i$ of the photograph f restricted to the large tile G_i. Recall that the

transformation $\overline{T}_i$ is the composition of replacing the gray tones in each 2×2 block by their average, and then applying T_i to map G_i onto C_i. Let H_i be the set of horizontal indices of the pixels in the C_i, and let V_i be the corresponding set of vertical indices. Then

$$d_4(f_{C_i}, \overline{f}_{C_i}) = \sqrt{\sum_{x \in H_i} \sum_{y \in V_i} \left(f_{C_i}(x,y) - \overline{f}_{C_i}(x,y)\right)^2}. \tag{11.17}$$

It is by carefully choosing s_i and g_i that we obtain a partitioned iterated function system that converges with respect to this distance. Let C_i be a small tile. We discuss how to choose the best large tile G_i and the transform T_i between the two. For a given C_i, we repeat the following steps for each potential large square G_j and each of the possible linear transformations L above:

- apply the smoothing transformation replacing 2×2 blocks of G_j by their average;
- apply the transformation L to the 8×8 square, resulting in a 4×4 square whose pixels are functions in the variables s_i and g_i;
- choose s_i and g_i to minimize the distance d_4 between the original and transformed tiles;
- calculate the minimized distance for the chosen s_i and g_i.

We do the above for each G_j and L and keep track of which G_j, L, s_i, and g_i resulted in the smallest distance between C_i and the resulting transformed tile. This will be one of the transformations in the partitioned iterated function system. We then repeat the above steps for each C_i, for each one determining the optimal associated G_i and T_i. If the image contains $h \times v$ pixels, there are $(h \times v)/16$ small tiles. For each of these, the number of large tiles that it must be compared against is enormous! In fact, a large tile is uniquely specified by its upper left corner, for which there are $(h-7) \times (v-7)$ choices. Since this is too large and would result in too slow an algorithm, we artificially limit ourselves to nonoverlapping large tiles, of which there are $(h \times v)/64$. It is thus with this "alphabet" of tiles that we attempt to accurately reconstruct the original image by associating to each small tile C_i a large tile G_i and a transform $\overline{T}_i$. If $h \times v = 640 \times 640$ then we will have to inspect $(\frac{1}{64} h \times v) \times 8 \times (\frac{1}{16} h \times v) \approx 1.3 \times 10^9$ potential transforms. This is still quite a lot! There are other tricks that may be employed to reduce the search space, but despite these optimizations, this method still has a high compression cost.

Method of least squares. This is the method that is employed in the second-to-last step of the above algorithm, which searches for the best values for s_i and g_i. It is likely that you have already seen this technique in a multivariable calculus, linear algebra, or statistics course. We wish to minimize

$$d_4(f_{C_i}, \overline{f}_{C_i}) = \sqrt{\sum_{x \in H_i} \sum_{y \in V_i} \left(f_{C_i}(x,y) - \overline{f}_{C_i}(x,y)\right)^2}. \tag{11.18}$$

Minimizing d_4 is equivalent to minimizing its square d_4^2, which frees us of the square root. So we must derive the expression of $\overline{f}_{C_i}$ as a function of s_i and g_i. Let us look at how we get $\overline{f}_{C_i}$:

- we start by replacing each 2×2 large square of G_i by a uniform square with the mean color;
- we apply the transformation (11.16), which amounts to sending G_i to C_i without any color adjustment;
- we compose with the mapping $(x, y, z) \mapsto (x, y, s_i z + g_i)$, which is just the color adjustment.

The composition of the first two transformations produces an image on C_i that is described by a function $\tilde{f}_{C_i}$, and we have

$$\overline{f}_{C_i} = s_i \tilde{f}_{C_i} + g_i. \tag{11.19}$$

To minimize d_4^2 in (11.18) we replace $\overline{f}_{C_i}$ by its expression in (11.19) and we require that the partial derivatives with respect to both s_i and g_i be equal to zero. The vanishing of the derivative with respect to g_i yields

$$\sum_{x \in H_i} \sum_{y \in V_i} f_{C_i}(x, y) = s_i \sum_{x \in H_i} \sum_{y \in V_i} \tilde{f}_{C_i}(x, y) + 16 g_i,$$

which implies that f_{C_i} and $\overline{f}_{C_i}$ have the same average gray tone. Requiring that the partial derivative with respect to s_i also vanish implies (after a few simplifications) that

$$s_i = \frac{\mathrm{Cov}(f_{C_i}, \tilde{f}_{C_i})}{\mathrm{var}(\tilde{f}_{C_i})},$$

where the covariance, $\mathrm{Cov}(f_{C_i}, \tilde{f}_{C_i})$, of f_{C_i} and $\tilde{f}_{C_i}$ is defined as follows:

$$\begin{aligned} \mathrm{Cov}(f_{C_i}, \tilde{f}_{C_i}) &= \frac{1}{16} \sum_{x \in H_i} \sum_{y \in V_i} f_{C_i}(x, y) \tilde{f}_{C_i}(x, y) \\ &\quad - \frac{1}{16^2} \left(\sum_{x \in H_i} \sum_{y \in V_i} f_{C_i}(x, y) \right) \left(\sum_{x \in H_i} \sum_{y \in V_i} \tilde{f}_{C_i}(x, y) \right), \end{aligned}$$

and the variance $\mathrm{var}(\tilde{f}_{C_i})$ is defined as

$$\mathrm{var}(\tilde{f}_{C_i}) = \mathrm{Cov}(\tilde{f}_{C_i}, \tilde{f}_{C_i}).$$

The operator W associated with a partitioned iterated function system $\{T_i\}_{i \in I}$. Given a gray tone image $f \in \mathcal{F}$, $W(f)$ is the image obtained by replacing

the image f_{C_i} of the tile C_i by the transformed image $\overline{f}_{C_i}$ of the associated big tile G_i. This gives us a transformed image $\overline{f} \in \mathcal{F}$ defined by

$$\overline{f}(x,y) = \overline{f}_{C_i}(x,y) \qquad \text{if} \quad (x,y) \in C_i.$$

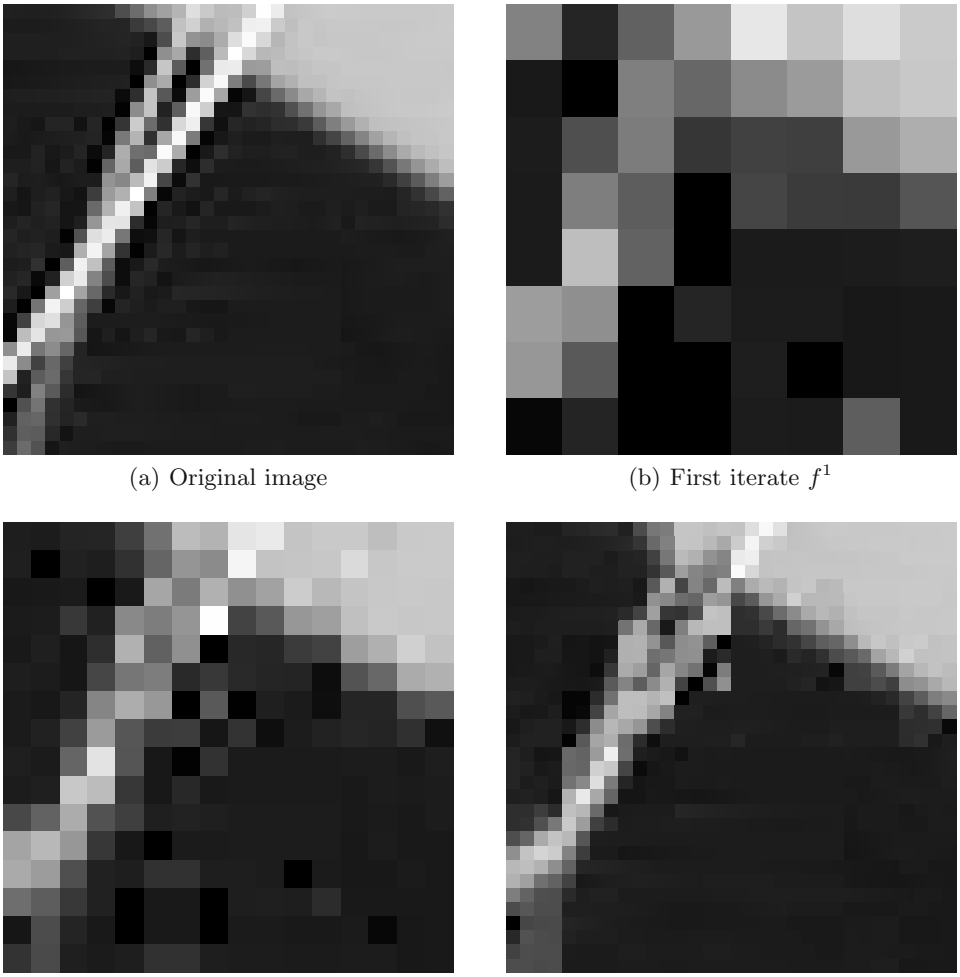
The attractor of this iterated function system should hopefully be something very close to the original image we wished to compress. Thus $W : \mathcal{F} \to \mathcal{F}$ is an operator on the set of all photographs. This technique replaces the *alphabet of geometric objects* we used in our first example with an *alphabet of gray tone tiles*, more specifically the large 8×8 tiles of the photograph to be compressed.

Reconstructing the image. The image can be reconstructed using the following procedure.

- Choose an arbitrary initial function $f^0 \in \mathcal{F}$. A natural choice is the function $f^0(x,y) = 128$ for all x and y, corresponding to a uniformly gray initial image.
- Calculate the iterates $f^j = W(f^{j-1})$. At step $j-1$ the image on each small tile C_i is given by the restriction of f^{j-1} to it. At step j here is how we calculate f^j restricted to C_i: we apply $\overline{T}_i$ to the image given by f^{j-1} on the associated large tile G_i. In practice, we keep track of the distance between successive iterates by calculating $d_{h\times v}(f^j, f^{j-1})$. Once this distance is below a given threshold (the image has largely stabilized), we stop the iteration.
- Replace the real-valued gray tone associated with each pixel by its closest integer value in the range $[0, 255]$.

As it will be shown in the following example, even the iterates f^1 and f^2 give quite good approximations to the original photograph. Furthermore, the distance between successive iterations quickly becomes small, and f^5 is already an excellent approximation to the attractor of the system (and, we hope, of the original image).

Remark: When considered as affine transformations on $\mathbb{R}^3$, the T_i are not always contractions; in fact, T_i is never a contraction if $s_i > 1$! However, most T_i will be contractions, since it is natural to have more contrast across a large tile than across a small one. As far as we know, there is no theorem guaranteeing the convergence of this algorithm for all images. However, in practice we generally see convergence, as if the system $\{T_i\}_{i\in I}$ were in fact a contraction. Benoît Mandelbrot introduced fractal geometry as a way to describe naturally occurring forms, that proved too complicated to be described with traditional geometry. Besides fern leaves and other plants there are many self-similar shapes occurring in nature: rocky coastlines, mountains, river networks, the human capillary system, etc. The technique of compressing images using iterated function systems is particularly well adapted to images having a strong fractal nature, that is, having a strong self-similarity across many scales. For such photos we can generally hope not only for convergence of the resulting system, but for an accurate reproduction of the original image.



(a) Original image

(b) First iterate f^1

(c) Second iterate f^2

(d) Sixth iterate f^6

Fig. 11.10. Reconstructing a 32×32 image (see Example 11.26).

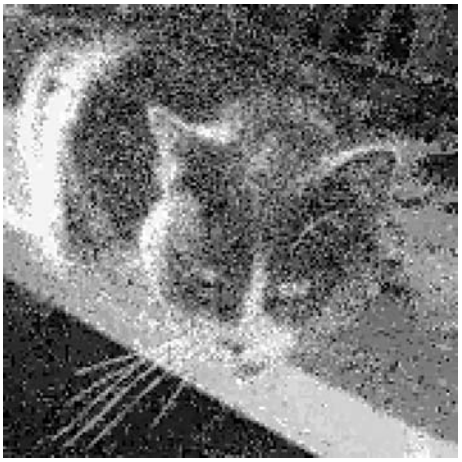
Example 11.26 *An example at last! The above comments may lead one to wonder whether this approach has any chance of accurately reproducing a real photograph. The following example should answer that question. We will use the same photograph used in the discussion of the JPEG image compression standard of Chapter 12, that of Figure 12.1. This photograph contains* $h \times v = 640 \times 640$ *pixels. We will produce two partitioned iterated function systems: the first for reconstructing the* 32×32 *pixel block where two of the cat's whiskers cross (see the zoomed portion of Figure 12.1), and another for the*

entire image. The 32×32 *pixel image is a demanding test of the algorithm. In fact, there are only 16 large tiles to choose from, restricting our chances of finding a good match. We will see, however, that despite this limited "alphabet" the resulting reconstruction is quite accurate!*

For the 32×32 *block there are only* 16 *nonoverlapping* 8×8 *tiles, each of which may be transformed by one of the* 8 *allowed transformations. This creates an "alphabet" of* $16 \times 8 = 128$ *tiles. This is quite limited, but at least it allows the best transformations to be quickly determined. After having found the best tile* G_i *and transformation* T_i *for each of the* $8 \times 8 = 64$ *small tiles* C_i*, we can proceed to the reconstruction. The results are shown in Figure 11.10. Figure 11.10(a) shows the original image to be displayed. For the reconstruction we began with the function* f^0 *associating a constant gray tone of value* 128 *to each of the pixels, halfway between black and white. Figures 11.10(b) through (d) show the reconstruction after* 1, 2, *and* 6 *iterations, respectively. The first surprise is that the first iteration appears to consist of only* 8×8 *pixels. This is easy to explain, since each of the large tiles began as a uniform block and was mapped to a uniform* 4×4 *tile. For the same reason the second iterate appears to consist of only* 16×16 *pixels of width* 2 *each. However, even after only two iterations, the edge of the table and the rough form of the whiskers is clearly visible. The iterates* f^4 *through* f^6 *are very similar to each other, only the last having been shown here. In fact,* f^5 *and* f^6 *are so close that the system is very likely convergent and* f^6 *is quite close to the attractor! In the sixth iterate the two whiskers are nearly completely visible, but with some errors: some pixels are much paler or much darker than in the original image. This is largely due to the limited alphabet of large tiles that we were restricted to working with.*

To obtain the complete partitioned iterated function system of the entire image we made a few concessions. (Recall that the number of individual transformations to be explored is over a billion!) In fact, for each small tile, each large tile, and each of the eight transformations we calculate a pair (s_j, g_j)*. Thus, for each small square we must repeat the calculation eight times the number of large squares. To make this process more efficient we have decide to abandon the search as soon as a large tile* G_i *and associated transform* T_i *are found that are within a distance of* $d_4 = 10$ *to the original small tile. Is this a large distance in the Euclidean space* $\mathbb{R}^{h \times v} = \mathbb{R}^{16}$*? No; in fact, it is quite close! If the distance is* 10*, then the square distance is* 100*. In each small square there are* 16 *pixels; thus we can expect an average squared error of* $\frac{100}{16} \approx 6.3$ *per pixel, corresponding to an expected gray tone error of* $\sqrt{6.3} \approx 2.5$ *per pixel, a relative error of* 1% *on the scale from* 0 *to* 255*. As we will see, the eye is easily able to overlook such a small error. The second compromise we have made is to reject all transformations in which* $|s_i| > 1$*. We have done this to improve the chances that the resulting system is convergent.*

Figure 11.11 presents the first, second, fourth, and sixth iterates of the reconstruction. Again, you can clearly see the 4×4 *uniform blocks in the first iterate and the* 2×2 *uniform blocks in the second iterate. As for* f^4 *and* f^6*, the two are nearly identical and distinguished only by small details. The quality of the sixth iterate is quite good and generally comparable to the original image, the exceptions being areas of fine detail and high contrast, such as the white whiskers against the shadowed background under*



(a) The first iterate f^1

(b) The second iterate f^2

(c) The fourth iterate f^4

(d) The sixth iterate f^6

Fig. 11.11. Reconstructing the entire image of a cat (see Example 11.26).

the table. It should be noted that a majority of the small tiles were approximated by transformations with a distance less than 10 *from the original. However, roughly* 15% *of the tiles were approximated by transformations with a larger error, and the worst offender had a distance of roughly* 280.

Compression ratio. As of 2007, consumer-level digital cameras are commonly available that capture images of up to 8 million pixels (and professional cameras can reach

up to 50 million!). We consider the compression ratio achieved on a 3000×2000 pixel grayscale image with $2^8 = 255$ gray tones. The gray tone of each pixel can be specified using exactly 8 bits, thus one byte,[1] and thus the original image requires $3000 \times 2000 = 6 \times 10^6$ B = 6 MB. Now consider the space required to represent the partitioned iterated function system.

Each small tile has an associated transformation T_i and large tile G_i. Consider:

(i) the number of bits necessary to represent a transformation T_i of the form in (11.15):
- 3 bits to specify one of the $2^3 = 8$ possible affine transformations L;
- 8 bits to specify s_i, the gray tone scaling factor; and
- 9 bits to specify g_i, the gray tone shift (we must permit negative values, requiring another bit).

(ii) the number of bits necessary to identify the associated large tile G_i. If we permit all possible overlapping large tiles, then each of them may be uniquely specified by indicating the upper left corner of the block. However, since we limited ourselves to nonoverlapping blocks, there are only $3000/8 \times 2000/8 = 93{,}750$ possible choices. Since $2^{16} = 65{,}536 < 93{,}750 < 2^{17} = 131{,}072$, we require 17 bits to specify a large tile.

(iii) the number of small tiles in the image: $\frac{3000}{4} \times \frac{2000}{4} = 375{,}000$.

Thus, we require $3 + 8 + 9 + 17 = 37$ bits per small tile, yielding $37 \times 375{,}000$ bits or roughly 1.73 MB, yielding that the compression ratio is roughly 3.46 times. In this approach we see that it is possible to vary the number of candidate large tiles. Had we restricted the search of large tiles to the one-fourth of them immediately neighboring the small tile in question, we could have reduced the number of bits necessary to encode each small tile by 2 (from 37 to 35). The resulting compression ratio would improve to a factor of $\frac{37}{35} \times 1.73 \approx 3.66$.

A more substantial gain is achieved by making small tiles 8×8 and large tiles 16×16. A factor of 4 is immediately gained, but at the expense of reconstructed image quality. Finally, one last improvement is to let the size of both the small and large tiles vary. In areas with little detail we can increase the tile size, while we could correspondingly decrease it in areas of fine detail. Thus, the compression ratio may be smoothly varied according to storage needs or desired quality of reconstruction.

Iterated function systems and JPEG. The method described here is very different from that employed by the JPEG standard. Which image compression technique is the best? This depends greatly on the type of images, the desired compression ratio, and the amount of computational power available. As with the improvements discussed above, the compression ratio of JPEG may be smoothly varied (at the expense of image quality) by changing the quantization tables (see Section 12.5). Digital cameras typically store images in the JPEG format, offering the user two or three resolution settings. The degree of compression actually obtained for a given resolution depends on the

[1] One byte equals eight bits and is abbreviated B. One megabyte is 10^6 bytes and is abbreviated MB.

photograph itself (in contrast to the algorithm presented here), but is typically between 6 and 10 times. These are compression ratios that are comparable to those we have just calculated. Compression using iterated function systems has been studied for quite some time but is not used in practice. Its weak point is the amount of time required to compress an image. (Recall that in our earliest discussion of the algorithm the number of steps was proportional to the square of the number of pixels, $(h \times v)^2$. In comparison, the complexity of the JPEG algorithm grows only linearly with image size, and is proportional to $h \times v$. For a photographer in the field snapping photos one after the other, this is a big advantage. For research images being processed on a high-powered computer, it is less so. Regardless, the domain moves quite fast, and iterated function systems may not have spoken their last words.

11.8 Exercises

Certain of the following fractals have been constructed based on the figures found in [1].

1. **(a)** For the fractals of Figure 11.12, find iterated function systems describing them. In each case clearly specify the coordinate system you have chosen. Afterward, reconstruct each of the figures in software.
(b) Given your chosen coordinate system, find two different iterated function systems describing the fractal (b).

2. For the fractals of Figure 11.13, find iterated function systems describing them. In each case clearly specify the coordinate system you have chosen. Afterward, reconstruct each of the figures in software.

3. For the fractals of Figure 11.14, find iterated function systems describing them. In each case clearly specify the coordinate system you have chosen. Afterward, reconstruct each of the figures in software. Attention: here the triangle in Figure 11.14(b) is equilateral, in contrast to the Sierpiński triangle in our earlier example.

4. For the fractals of Figure 11.15, find iterated function systems describing them. In each case clearly specify the coordinate system you have chosen. Afterward, reconstruct each of the figures in software.

5. Amuse yourself by constructing arbitrary iterated function systems and trying to intuit their attractors. Afterward, confirm or disprove your intuitions by plotting them on a computer.

6. Calculate the fractal dimensions of the fractals in Exercises 1 (except (a)), 2, 3, and 4. (In certain cases you will be required to pursue numeric approaches.)

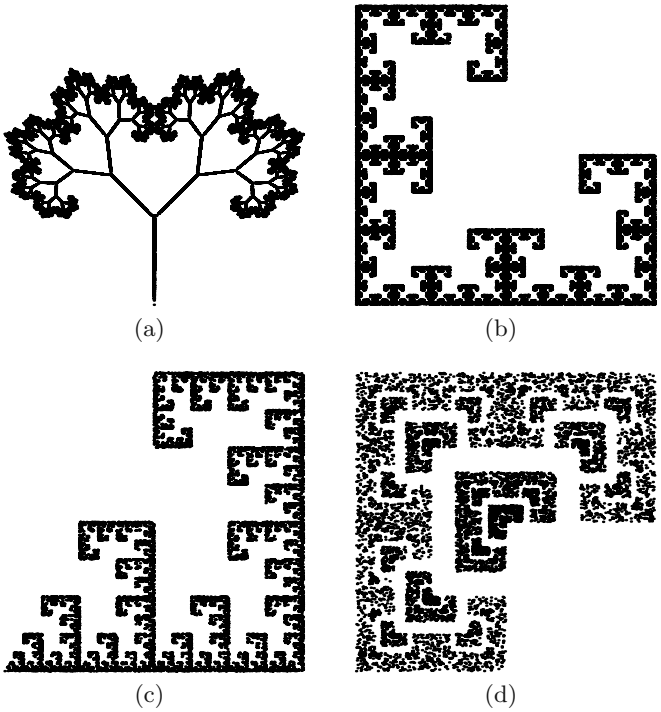


Fig. 11.12. Exercise 1.

7. The Cantor set is a subset of the unit interval $[0, 1]$. It is obtained as the attractor of the iterated function system $\{T_1, T_2\}$, where T_1 and T_2 are the affine contractions defined by $T_1(x) = x/3$ and $T_2(x) = x/3 + 2/3$.
(a) Describe the Cantor set.
(b) Draw the Cantor set. (You may pursue the first few iterations by hand, but it is easiest to use a computer.)
(c) Show that there exists a bijection between the Cantor set and the set of real numbers with base-3 expansions of the form

$$0.a_1a_2 \ldots a_n \ldots,$$

where $a_i \in \{0, 2\}$.
(d) Calculate the fractal dimension of the Cantor set.

8. Show that the fractal dimension of the Cartesian product $A_1 \times A_2$ is the sum of the fractal dimensions of A_1 and A_2:

$$D(A_1 \times A_2) = D(A_1) + D(A_2).$$

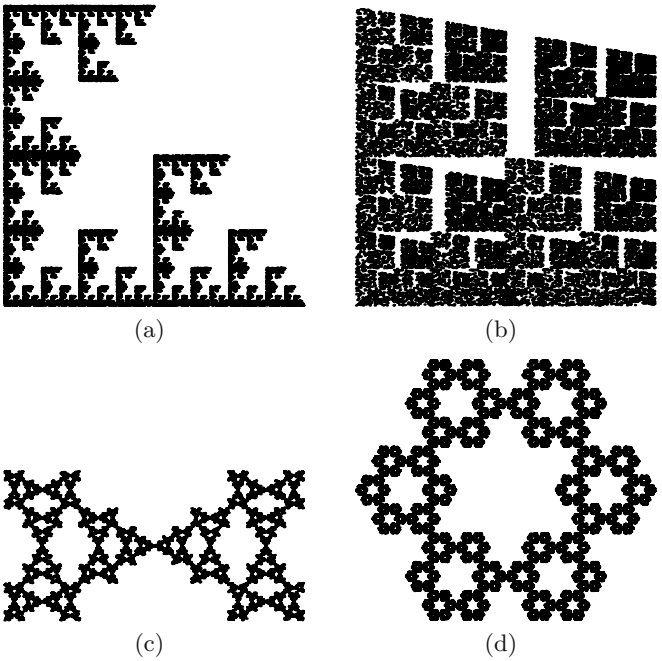


Fig. 11.13. Exercise 2.

9. Let A be the Cantor set, as described in Exercise 7. This is a subset of $\mathbb{R}$. Find an iterated function system on $\mathbb{R}^2$ whose attractor is $A \times A$.

10. The Koch snowflake (or von Koch snowflake) is constructed as the limiting object of the following process (see Figure 11.16):

- Begin with the segment $[0, 1]$.
- Replace the initial segment with four segments, as shown in Figure 11.16(b)).
- Iterate the process, at each step replacing each segment by four smaller segments (see Figure 11.16(c)).

(a) Give an iterated function system that constructs the von Koch snowflake.
(b) Can you give an iterated function system for building the von Koch snowflake that requires just two affine contractions?
(c) Calculate the fractal dimension of the von Koch snowflake.

11. Explain how to modify an iterated function system on $\mathbb{R}^2$

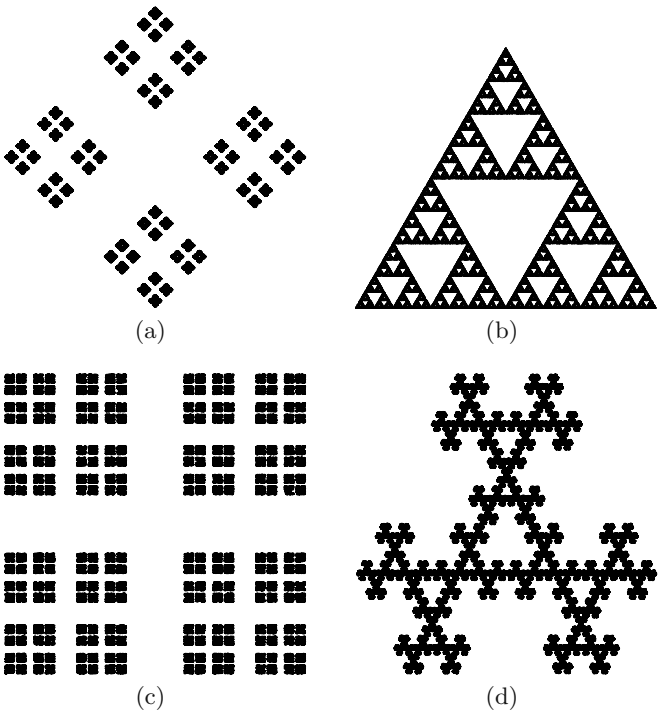


Fig. 11.14. Exercise 3.

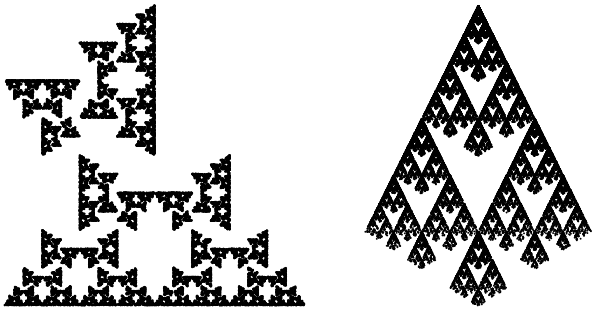
(a) such that its attractor will be twice as large in both dimensions;
(b) to translate the location of its bottom leftmost point.

12. Consider an affine transformation $T(x, y) = (ax + by + e, cx + dy + f)$.
(a) Show that T is an affine contraction if and only if the associated linear transformation $U(x, y) = (ax + by, cx + dy)$ is a contraction.
(b) Show that U contracts distances if

$$\begin{cases} a^2 + c^2 < 1, \\ b^2 + d^2 < 1, \\ a^2 + b^2 + c^2 + d^2 - (ad - bc)^2 < 1. \end{cases}$$

Suggestion: it suffices to show that the square of the length of $U(x, y)$ is less than the square of the length of (x, y) for all nonzero (x, y).

13. Let $P_1, \ldots, P_4$ be four noncoplanar points in $\mathbb{R}^3$. Let $Q_1, \ldots, Q_4$ be four other points of $\mathbb{R}^3$. Show that there exists a unique affine transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ such that $T(P_i) = Q_i$.



(a) (b)

Fig. 11.15. Exercise 4.

(a) The initial segment (b) The first iteration (c) The second iteration (d) The von Koch snowflake

Fig. 11.16. Constructing the von Koch snowflake of Exercise 10.

Remark: We can consider systems of iterated functions in $\mathbb{R}^3$. As an example, we could use an iterated function system in this space to describe a fern leaf bent under its own weight. We could then project this image to the plane in order to display it.

14. Consider $v \in \mathbb{R}^2$ and A, B, two closed and bounded subsets of $\mathbb{R}^2$. Show that $d(v, A \cup B) \leq d(v, A)$ and $d(v, A \cap B) \geq d(v, A)$.

15. Proceeding numerically, find the contraction factors of the individual transforms T_i for the fern leaf. Are any of these exact contraction factors?

16. **(a)** Let B_1 and B_2 be two disks in $\mathbb{R}^2$ with radius r, and whose centers are at a distance of d from each other. Calculate $d_H(B_1, B_2)$.
(b) Let B_1 and B_2 be two concentric disks in the plane with radii r_1 and r_2, respectively. Calculate $d_H(B_1, B_2)$.

References

[1] M. Barnsley. *Fractals Everywhere*. Academic Press, 1988.

[2] J. Kominek. Advances in fractal compression for multimedia applications. *Multimedia Systems Journal*, 5(4):255–270, 1997.

12

Image Compression: The JPEG Standard

Presenting the JPEG standard at the level of detail contained in this chapter will require about four hours. To fit within this amount of time, you will have to skip Section 12.4; this section proves the orthogonality of the matrix C and can be seen as the advanced part of this chapter. It is necessary, however, to discuss the relationship between the matrices f and α and to present the 64 basis elements A_{ij}. The central idea underlying the JPEG standard is a change of basis in a 64-dimensional space; this chapter provides the perfect occasion to review this portion of linear algebra.

12.1 Introduction: Lossless and Lossy Compression

Data compression is at the very heart of computer science, and the Internet has made its use an everyday occurrence for most. Many of us may not even know we are using compression, or at least have little knowledge of how the underlying algorithms work. Even so, many compression algorithms have names that are familiar to general computer users (*WinZip*, *gzip*, and, in the UNIX world, *compress*), to music lovers and Internet users (*GIF*, *JPG*, *PNG*, *MP3*, *AAC*, etc). If not for the common use of compression algorithms, the Internet would be completely paralyzed by the volume of uncompressed data being transferred.

The goal of this chapter is to study a commonly used algorithm for the compression of black-and-white or color still images ("still" as opposed to "moving" images). This method of compression is commonly known as JPEG, the acronym of *Joint Photographic Experts Group*, the consortium of companies and researchers that developed and popularized it. The group started its work in June 1987, and the first draft of the standard was published in 1991. Internet users will no doubt associate this compression method with the "jpg" suffix that is a part of the names of many images and photographs transmitted over the Internet. The JPEG algorithm is the most commonly used compression method in digital cameras.

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_12,

Before diving into the details of this algorithm and the underlying mathematics it is good to have a basic knowledge of data compression in general. There are two broad families of data compression algorithms: those that actually degrade the original information to some extent (called *lossy* algorithms) and those that allow for the reconstruction of the original with perfect accuracy (called *lossless* algorithms). Two simple observations can be made.

The first is that it is impossible to compress *without loss* all files of a given size using the same algorithm. Suppose that such a technique exists for files of exactly N bits in length. Each of these bits can take on 2 different values (0 or 1) and thus there are 2^N distinct N-bit files. If the algorithm compresses each of these files, then each one of them will be represented by some new file containing at most $N-1$ bits. There are 2^{N-1} distinct files of $N-1$ bits, 2^{N-2} distinct files of $N-2$ bits, ..., 2^1 distinct files of 1 bit and a single one with 0 bit. Thus, the number of distinct files containing at most $N-1$ bits is

$$1 + 2^1 + 2^2 + \cdots + 2^{N-2} + 2^{N-1} = \sum_{n=0}^{N-1} 2^i = \frac{2^N - 1}{2 - 1} = 2^N - 1.$$

Thus the algorithms we are using must compress at least two of the original N-bit files to some identical file containing fewer than N bits. These two compressed files will then be indistinguishable, and it is impossible to determine which original file they should decompress to. Again: *it is impossible to losslessly compress all files of a given size!*

The second observation is a consequence of the first: when developing a compression algorithm, the person charged with this task must decide whether the information must be preserved perfectly or whether a slight loss (or transformation) is tolerable. Two examples can help make this choice clear while also demonstrating different approaches once this decision has been made.

Webster's Ninth New Collegiate Dictionary has 1592 pages, most being typeset in two columns, each column having around 100 lines, each line having about 70 characters, spaces, or punctuation marks. This amounts to a total of about 22 million characters. These characters can be represented by an alphabet of 256 characters, each being coded by 8 bits, or 1 byte (see Section 12.2). About 22 MB are therefore needed to hold *Webster's*. If one recalls that compact disks store approximately 750 MB, a single CD can carry 34 copies of the whole of *Webster's* (without the figures and drawings, however). No author of a dictionary, an encyclopedia, or a textbook (or *any* book for that matter!) would tolerate the changing of a single character. Thus, in compressing such material it is extremely important to use a lossless compression algorithm allowing for a perfect reconstruction of the original document.

A simple approach to such an algorithm assigns variable length codes to each letter of the alphabet.[1] The most common characters in English are the "␣" (space) character

[1] This approach is common to text compression. Different algorithms may assign codes to "words" rather than "letters," and more complicated algorithms may change the assigned codes based on context.

and the letter "e" followed by the letters "t", "a", "o", "i", "n", "h", "s", "r" (see Table 12.1). The most uncommonly used letters are "x", "z", "j", and "q". The actual frequencies depend on the author and the text. They may vary significantly if the text is short. It is natural to try to assign short codes to more frequently occurring characters (such as "␣" and "e") and longer codes to less frequently occurring ones (such as "j" and "q"). In this manner, characters are represented by a variable number of bits rather than always requiring a single byte. Does this approach violate our first observation? No, since in order for each assigned code to be uniquely decodable the codes for rarely occurring letters will be *longer* than 8 bits. Thus, files containing an unusually high percentage of such characters will actually be longer than the original uncompressed file. The idea of assigning variable length codes to individual symbols as a function of their frequency of use is the main idea underlying Huffman codes.

letter	frequency
e	0.125
t	0.088
a	0.080
o	0.077
i	0.069
n	0.068
h	0.066
s	0.060
r	0.059

letter	frequency
d	0.047
l	0.041
u	0.027
m	0.026
w	0.025
c	0.023
g	0.022
f	0.021
y	0.021

letter	frequency
p	0.018
b	0.016
v	0.010
k	0.0090
j	0.0014
x	0.0014
q	0.0010
z	0.0002

Table 12.1. Frequencies of letters in Dickens's *Oliver Twist*. (Spaces and punctuation marks have been ignored. Capital letters have been mapped to the corresponding lowercase letters. *Oliver Twist* contains a little over 680,000 letters.)

Our second example lies a little closer to the subject of this chapter. All computer screens have a finite resolution. Usually, this is measured by counting the number of pixels that it can display. Each pixel may be illuminated to take on any color and intensity.[2] Early screens could display $640 \times 480 = 307{,}200$ pixels.[3] (Resolution is

[2]This is not exactly true. Computer screens are able to reproduce only a portion of the visible color gamut, broken down into a finite set of discrete colors that are roughly uniformly close to each other. As such, they can generally reproduce a large number of colors but not the entire visible spectrum.

[3]It is now common to have displays capable of displaying many millions of pixels, with the largest surpassing four million.

normally reported as "number of pixels per horizontal line × number of lines.") Suppose that the Louvre decided to digitize its entire collection of painted works. The museum would ideally like to do this with sufficient quality so as to please art experts. However, at the same time they would like to have lower-quality versions for transmission over the Internet and display on typical computer screens. In this case, it doesn't make any sense for the image to be of a higher resolution than a typical computer monitor. Thus, the image satisfying art experts and that for display on a typical computer monitor are going to be of very different resolutions and sizes. The latter will contain significantly less detail but will be entirely satisfactory for displaying on a monitor. In fact, transmitting the higher-quality image would be a complete waste of time given the limited resolution of the display! The decision about the number of pixels to send is then a fairly obvious one. But suppose that Louvre technical people want to further reduce the size of the transmitted files. They argue that mathematicians often approximate functions around a given point by a straight line, and if one looks at the graph of the function and the approximating line they usually agree fairly well, at least locally. If we imagine the pale tones of a picture as the peaks and ridges of a function graph and the dark ones as its valleys, could we use the mathematical idea of approximation to this "function"?

This last question is more physiological than mathematical: can one fool the user by sending a picture that has been "mathematically approximated"? If the answer is yes, it will mean that a certain loss of quality is acceptable depending on the use of the data. Other criteria (such as human physiology) therefore play an equally important role in deciding how to compress. For example, in digitizing music it is useful to know that the (average) human ear is unable to perceive sounds above 20,000 Hz. In fact, the standard used for recording compact discs ignores frequencies over 22,000 Hz and is capable of accurately reproducing only those frequencies below this threshold, a loss that would bother only dogs, bats, or other animals with a keener sense of hearing than our own. For images are there limits to the variations in colors and intensities of light that may be perceived by the human eye? Are our eyes and mind content with receiving less than an exact reproduction of an image? Should photographic images and cartoons be compressed in the same manner? The JPEG compression standard, through its successes and its limits, answers these questions.

12.2 Zooming in on a JPEG Compressed Digital Image

A photograph can be digitized in a variety of ways. In the JPEG method the photograph is first divided into very small elements, called *pixels*, each one associated with a uniform color or gray tone. The photograph of a cat in Figure 12.1 has been subdivided into 640×640 pixels. Each of these $640 \times 640 = 409{,}600$ pixels has been associated with a uniform tone of gray between black and white. This particular photograph has been digitized using a scale of 256 gray tones where 0 represents black and 255 represents white. Since $256 = 2^8$, each of these values may be stored using 8 bits (a single byte).

Without compression we would require 409,600 bytes to store the photo of the cat, which equates to roughly 410 KB. (Here we are using the metric convention: a KB represents 1000 bytes, a MB represents 10^6 bytes, etc.) To encode a color image, each pixel is associated with three color values (red, green, and blue) each encoded using an 8-bit value between 0 and 255. An image of this size would require over 1.2 MB to store uncompressed. However, as frequent users of the Internet will know, large color JPEG-compressed images (files with a "jpg" suffix) rarely exceed 100 KB. The JPEG method is thus able to efficiently store the information in the image. The JPEG algorithm's utility is not strictly confined to the Internet. It is the principal standard used in digital photography. Nearly all digital cameras will compress images to JPEG format by default; the compression occurs at the instant the photo is taken, and therefore a part of the information is lost forever. As we will see in this chapter, this loss is usually acceptable, but sometimes it is not. Depending on the specific use of the camera, it is up to the photographer to decide. (Exercise: As of 2006, many digital cameras offer resolutions exceeding 10 million pixels (megapixels). What is the space that would be required by such a color image in an uncompressed form?)

Rather than processing the entire photograph at once, the JPEG standard divides the image into little tiles of 8×8 pixels. Figure 12.1 shows two closeups of the image of the cat. In the bottom left, a 32×32 pixel region has been shown. The bottom right shows a further closeup of an 8×8 region of this closeup. The closeups focus on a small region depicting the intersection of two of the cat's whiskers close to the edge of the table. This particular block of the image is unique in that it contains fine details and high contrast. This is not typical of most 8×8 tiles! In most of the image we see that the changes in color and texture are quite gradual. The surface under the table, the table itself, and even the cat's fur consist largely of smooth gradients when looked at as 8×8 blocks. This is the case with most photographs; just think of any landscape photo containing open regions of land, water or sky. The JPEG standard was built on this uniformity; it tries to represent a nearly uniform 8×8 block using as little information as possible. When such a block contains significant detail (such as is the case in our closeup), the use of more space is accepted.

12.3 The Case of 2×2 Blocks

It is simpler to characterize 2×2 blocks than 8×8 blocks, so we will start with that.

We have seen that gray tones are typically represented using a scale with 256 increments. We could equally imagine a scale with infinitely fine increments that covers all of $[-1, 1]$ or any interval $[-L, L]$ of $\mathbb{R}$. In this case, we may associate negative values with dark grays tending to black and positive values to lighter grays tending to white. The origin would then correspond to a gray between levels 127 and 128 on the scale with 256 levels. Even though this change of scale and origin may be perfectly natural in some ways, it is not necessary for our discussion. We will, however, ignore the fact that our gray tones are integers between 0 and 255 and instead treat them as real numbers in

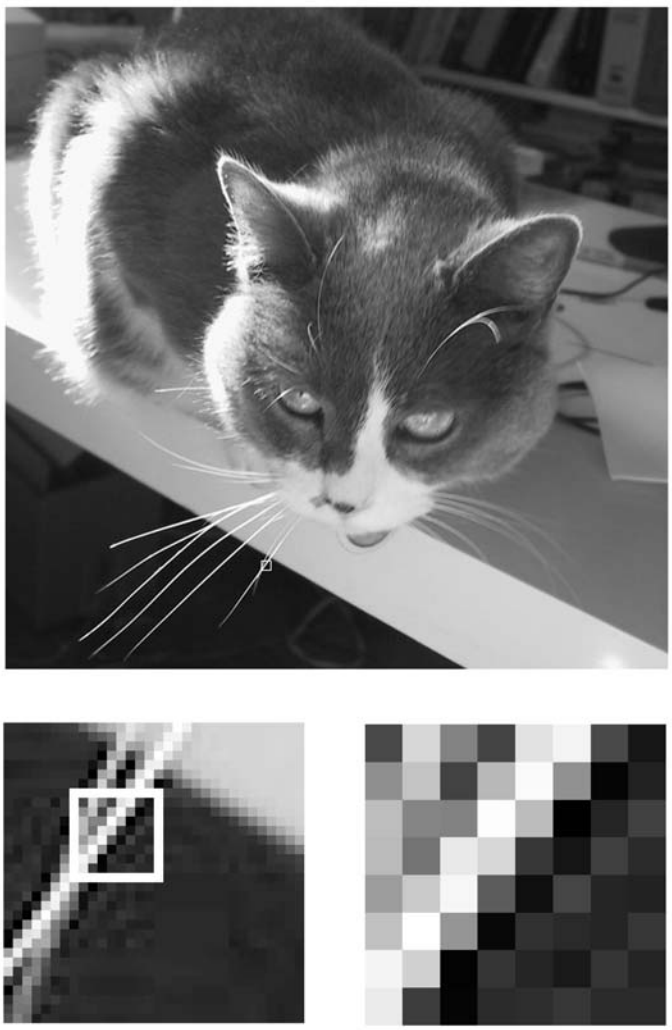


Fig. 12.1. Two successive closeups are made of the original photo (top), which contains 640×640 pixels. The first closeup (bottom left) contains 32×32 pixels. The second closeup (bottom right) contains 8×8 pixels. The white frames on the first and second images denotes the boundaries of the 8×8 closeup in the last image.

this same range. The tone of each pixel will therefore be represented by a real number, and a 2×2 block will require four such values, or equivalently, a point in $\mathbb{R}^4$. (When we are dealing with an $N \times N$ block, we can consider it as a vector in $\mathbb{R}^{N^2}$.)

Given that we perceive the blocks in two dimensions, it is more natural to number the individual pixels using two indices i and j from the set $\{0,1\}$ (or the set $\{0,1,\ldots,N-1\}$ when we are dealing with $N \times N$ blocks). The first index will indicate the row, while the second will indicate the column, as is typical in linear algebra. For example, the values of the function f giving the gray tones on the 2×2 square of Figure 12.2 are

$$f = \begin{pmatrix} f_{00} & f_{01} \\ f_{10} & f_{11} \end{pmatrix} = \begin{pmatrix} 191 & 207 \\ 191 & 175 \end{pmatrix}.$$

Many of the functions that we will study naturally take their values in the range $[-1,1]$. When representing them as gray tones we will use the obvious affine transformation to map them to the range $[0,255]$. This transformation can be

$$\mathrm{aff}_1(x) = 255(x+1)/2 \tag{12.1}$$

or

$$\mathrm{aff}_2(x) = [255(x+1)/2], \tag{12.2}$$

where $[x]$ denotes the integer part of x. (This last transformation will be used when the values need to be constrained to integers in the range $[0,255]$. See Exercise 1.) We will use f to denote a function defined in the range $[0,255]$ and g to denote functions defined in the range $[-1,1]$. The following box summarizes this notation and specifies the translation we will use. Using this method, the function g associated with the above function f is

$$g = \begin{pmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{5}{8} \\ \frac{1}{2} & \frac{3}{8} \end{pmatrix}:$$

$$\begin{array}{ccc} f_{ij} \in [0,255] \subset \mathbb{Z} & \longleftrightarrow & g_{ij} \in [-1,1] \subset \mathbb{R} \\ f_{ij} = \mathrm{aff}_2(g_{ij}), \quad \text{where} & & \mathrm{aff}_2(x) = \left[\tfrac{255}{2}(x+1)\right]. \end{array}$$

We will graphically represent a 2×2 block in two different manners. The first will be simply to draw it using the associated gray tones that would appear in a photograph. The second is to interpret the values g_{ij} as a two-dimensional function of the variables i and j, $i,j \in \{0,1\}$. Figure 12.2 represents the function $g = (g_{00}, g_{01}, g_{10}, g_{11}) = (\frac{1}{2}, \frac{5}{8}, \frac{1}{2}, \frac{3}{8})$ in these two manners. The coefficients giving the gray values for both the top left g_{00} and bottom left g_{10} pixels are identical. Those of the right column are g_{01} (the paler of the two) and g_{11}. In other words, if we use the matrix notation

$$g = \begin{pmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{pmatrix},$$

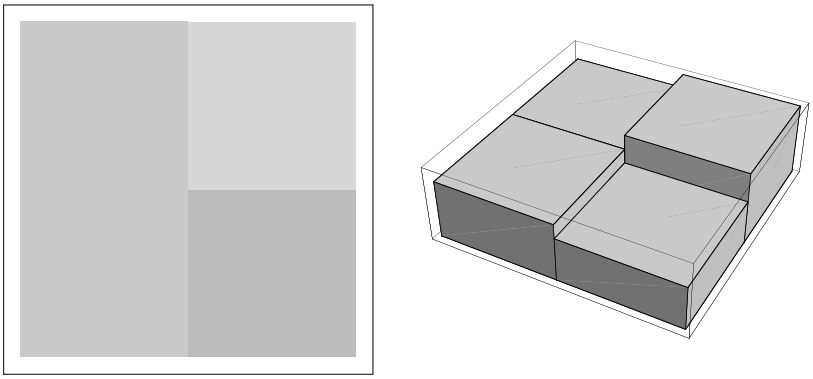


Fig. 12.2. Two graphical representations of the function $g = (g_{00}, g_{01}, g_{10}, g_{11}) = (\frac{1}{2}, \frac{5}{8}, \frac{1}{2}, \frac{3}{8})$.

then the elements of the matrix g are in the same positions as the pixels of Figure 12.2. The second image interprets these same values but displays them as a histogram in two variables i and j, with darker colors being associated to lesser heights. This particular 2×2 block was chosen because all of the pixels are closely related gray tones, as is typical of most 2×2 blocks in a photograph. (In fact, the higher the resolution of the photo, the gentler the gradients become.)

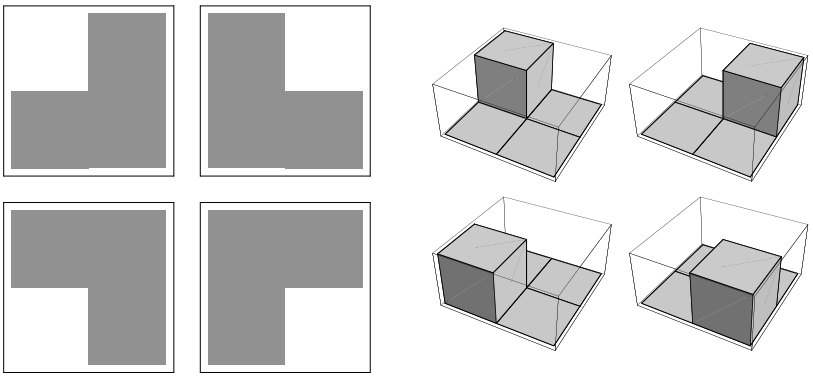


Fig. 12.3. The four elements of the usual basis $\mathcal{B}$ of $\mathbb{R}^4$ represented graphically.

The coordinates $(g_{00}, g_{01}, g_{10}, g_{11})$ (or equivalently $(f_{00}, f_{01}, f_{10}, f_{11})$) represent the small 2×2 block without any loss. (In other words, no compression has yet been done.) These coordinates are expressed in the usual basis $\mathcal{B}$ of $\mathbb{R}^4$, where each element of the basis contains a single nonzero entry with value 1. This basis is depicted graphically in

Figure 12.3. If we were to apply a change of basis

$$[g]_{\mathcal{B}} = \begin{pmatrix} g_{00} \\ g_{01} \\ g_{10} \\ g_{11} \end{pmatrix} \mapsto [g]_{\mathcal{B}'} = \begin{pmatrix} \beta_{00} \\ \beta_{01} \\ \beta_{10} \\ \beta_{11} \end{pmatrix} = [P]_{\mathcal{B}'\mathcal{B}}[g]_{\mathcal{B}},$$

the new coordinates β_{ij} would also accurately represent the contents of the block. The coordinates g_{ij} are not appropriate to our end goal. In fact, we would like to easily recognize blocks where all of the pixels are nearly the same color or gray tone. To do this, it is useful to construct a basis in which completely uniform blocks are represented by a single nonzero coefficient. Similarly, we would like a cursory inspection of the coordinates to reveal when the block is far from being uniform.

The JPEG standard proposes using another basis $\mathcal{B}' = \{A_{00}, A_{01}, A_{10}, A_{11}\}$. Each element A_{ij} of this basis can be expressed using the standard basis shown in Figure 12.3. In the standard basis $\mathcal{B}$ their coefficients are

$$[A_{00}]_{\mathcal{B}} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \quad [A_{01}]_{\mathcal{B}} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \quad [A_{10}]_{\mathcal{B}} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \quad [A_{11}]_{\mathcal{B}} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}. \tag{12.3}$$

The elements of this new basis are represented graphically in Figure 12.4. The first element A_{00} represents a uniform block. If the 2×2 block is completely uniform, only the coefficient of A_{00} will be nonzero. The two elements A_{01} and A_{10} represent left/right and top/bottom contrasts, respectively. The last element A_{11} represents a mixture of these two, where each pixel is in contrast with its neighbor along both directions, much like a checkerboard.

Knowing the A_{ij} in the standard basis, it is easy to obtain the change of basis matrix $[P]_{\mathcal{B}\mathcal{B}'}$ from $\mathcal{B}'$ to $\mathcal{B}$. In fact, its columns are given by the coordinates of the elements of $\mathcal{B}'$ expressed in the basis $\mathcal{B}$. It is therefore given by

$$[P]_{\mathcal{B}\mathcal{B}'} = [P]^{-1}_{\mathcal{B}'\mathcal{B}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{12.4}$$

To calculate $[g]_{\mathcal{B}'}$ we will need to use $[P]_{\mathcal{B}'\mathcal{B}}$, that is, the inverse of $[P]_{\mathcal{B}\mathcal{B}'}$. Here the matrix $[P]_{\mathcal{B}\mathcal{B}'}$ is orthogonal. (Exercise: A matrix A is orthogonal if $A^tA = AA^t = I$. Verify that $P_{\mathcal{B}\mathcal{B}'}$ is orthogonal.) The computation is therefore easy:

$$[P]_{\mathcal{B}'\mathcal{B}} = [P]^{-1}_{\mathcal{B}\mathcal{B}'} = [P]^{t}_{\mathcal{B}\mathcal{B}'} = [P]_{\mathcal{B}\mathcal{B}'}.$$

The last equality comes from the fact that the matrix $[P]_{\mathcal{B}\mathcal{B}'}$ is symmetric. The coefficients of g in this basis are simply

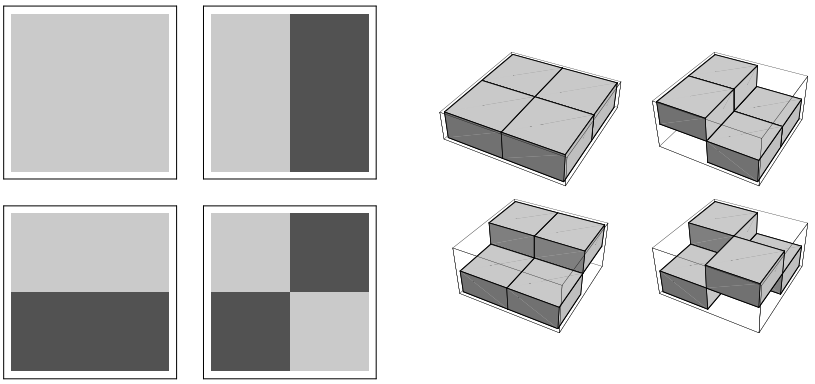


Fig. 12.4. The four elements of the proposed basis $\mathcal{B}'$. (Element A_{00} is at the upper left and element A_{01} is at the upper right.)

$$[g]_{\mathcal{B}'} = \begin{pmatrix} \beta_{00} \\ \beta_{01} \\ \beta_{10} \\ \beta_{11} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ \frac{5}{8} \\ \frac{1}{2} \\ \frac{3}{8} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \frac{1}{8} \\ -\frac{1}{8} \end{pmatrix}.$$

In this basis the largest coefficient is $\beta_{00} = 1$. This is the weight of the element A_{00} that gives an equal importance to each of the four pixels; in other words, this element of the new basis assigns them all the same gray tone. The two remaining nonzero coefficients, both much smaller in magnitude ($\beta_{10} = -\beta_{11} = \frac{1}{8}$), contain information regarding the small amount of contrast between the left and the right columns, and between the two pixels in the right column. The careful choice of the basis highlights spatial contrast information rather than giving individual pixel information. This is the heart of the JPEG standard. To make this technique lossy, one needs only to decide what coefficients correspond to visible contrasts for each of the elements of the basis. The rest of the coefficients may simply be thrown away.

12.4 The Case of $N \times N$ Blocks

The JPEG standard divides the image into 8×8 blocks. The definition of the basis that puts the focus on contrast information rather than individual pixels can equally be defined for arbitrary $N \times N$ blocks. The basis $\mathcal{B}'$ that we introduced in the previous section ($N = 2$) and that used in the JPEG standard ($N = 8$) are particular cases.

The *discrete cosine transform*[4] replaces the function $\{f_{ij},\ i, j = 0, 1, 2, \ldots, N-1\}$ defined over an $N \times N$ square grid by a set of coefficients $\alpha_{kl},\ k, l = 0, 1, \ldots, N-1$.

[4]The discrete cosine transform is a particular instance of a more general mathematical technique called *Fourier analysis*. Introduced at the beginning of the nineteenth century by

The coefficients α_{kl} are given by

$$\alpha_{kl} = \sum_{i,j=0}^{N-1} c_{ki} c_{lj} f_{ij}, \qquad 0 \le k, l \le N-1, \tag{12.5}$$

where the c_{ij} are defined as

$$c_{ij} = \frac{\delta_i}{\sqrt{N}} \cos \frac{i(2j+1)\pi}{2N}, \qquad i, j = 0, 1, \ldots, N-1, \tag{12.6}$$

with

$$\delta_i = \begin{cases} 1, & \text{if } i = 0, \\ \sqrt{2}, & \text{otherwise.} \end{cases} \tag{12.7}$$

(Exercise: For the case $N = 2$, show that the coefficients c_{ij} are given by

$$C = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Is it possible for the transformation (12.5) to be equivalent to the change of basis embodied by the matrix $[P]_{\mathcal{B}\mathcal{B}'}$ of (12.4)? Explain.)

The transformation in (12.5) from the $\{f_{ij}\}$ to the $\{\alpha_{kl}\}$ is clearly linear. By writing

$$\alpha = \begin{pmatrix} \alpha_{00} & \alpha_{01} & \cdots & \alpha_{0,N-1} \\ \alpha_{10} & \alpha_{11} & \cdots & \alpha_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N-1,0} & \alpha_{N-1,1} & \cdots & \alpha_{N-1,N-1} \end{pmatrix}, \quad f = \begin{pmatrix} f_{00} & f_{01} & \cdots & f_{0,N-1} \\ f_{10} & f_{11} & \cdots & f_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N-1,0} & f_{N-1,1} & \cdots & f_{N-1,N-1} \end{pmatrix},$$

and

$$C = \begin{pmatrix} \sqrt{\frac{1}{N}} & \sqrt{\frac{1}{N}} & \cdots & \sqrt{\frac{1}{N}} \\ \sqrt{\frac{2}{N}} \cos \frac{\pi}{2N} & \sqrt{\frac{2}{N}} \cos \frac{3\pi}{2N} & \cdots & \sqrt{\frac{2}{N}} \cos \frac{(2N-1)\pi}{2N} \\ \sqrt{\frac{2}{N}} \cos \frac{2\pi}{2N} & \sqrt{\frac{2}{N}} \cos \frac{6\pi}{2N} & \cdots & \sqrt{\frac{2}{N}} \cos \frac{2(2N-1)\pi}{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{2}{N}} \cos \frac{(N-1)\pi}{2N} & \sqrt{\frac{2}{N}} \cos \frac{3(N-1)\pi}{2N} & \cdots & \sqrt{\frac{2}{N}} \cos \frac{(2N-1)(N-1)\pi}{2N} \end{pmatrix},$$

we see that the transformation of (12.5) takes on the matrix form

$$\alpha = C f C^t, \tag{12.8}$$

Jean Baptiste Joseph Fourier for studying the propagation of heat, this technique has since invaded the world of engineering. It also plays an important role in Chapter 10.

where C^t denotes the transpose of the matrix C. In fact,

$$\alpha_{kl} = [\alpha]_{kl} = [CfC^t]_{kl} = \sum_{i,j=0}^{N-1} [C]_{ki}[f]_{ij}[C^t]_{jl} = \sum_{i,j=0}^{N-1} c_{ki} f_{ij} c_{lj},$$

which is the same as (12.5).

This transformation is an isomorphism if the matrix C is invertible. (That this is the case will be shown later.) If it is so, we are able to write

$$f = C^{-1}\alpha(C^t)^{-1}$$

and recover the values $f_{ij}, i, j = 0, 1, \ldots, N-1$, from the $\alpha_{kl}, k, l = 0, 1, \ldots, N-1$. The transformation $f \mapsto \alpha$ given by (12.8) is also a linear transformation. Indeed, suppose that f and g are related to α and β through (12.8) (namely $\alpha = CfC^t$ and $\beta = CgC^t$). Then

$$C(f+g)C^t = CfC^t + CgC^t = \alpha + \beta$$

follows from the distributivity of matrix multiplication. And if $c \in \mathbb{R}$ then

$$C(cf)C^t = c(CfC^t) = c\alpha.$$

The two previous identities are the defining properties of linear transformations. Since this linear transformation is an isomorphism, it is a *change of basis*! Note that the passage from f to α is not expressed through a matrix $[P]_{\mathcal{B}'\mathcal{B}}$ as in the previous section. But linear algebra assures us that the transformation $f \mapsto \alpha$ could be written with such a matrix. (If the two indices of f run through $\{0, 1, \ldots, N-1\}$, then there are N^2 coordinates f_{ij}, and the matrix $[P]_{\mathcal{B}'\mathcal{B}}$ doing the change of basis is of size $N^2 \times N^2$. The form (12.8) has the advantage of using only $N \times N$ matrices.)

The proof of the invertibility of C rests on the observation that C is orthogonal:

$$C^t = C^{-1}. \tag{12.9}$$

This observation simplifies the calculations because the above expression for f becomes

$$f = C^t \alpha C. \tag{12.10}$$

We will give a proof of this property at the end of the section.

For the moment we will accept this fact and give an example of the transformation $f \mapsto \alpha$. To do this we will use the gray tones defined over the 8×8 block of Figure 12.1. The $f_{ij}, 0 \le i, j \le 7$, are given in Table 12.2. The positions of pixels in the picture correspond to positions of entries in the table, and the entries are the gray intensities with 0 = black and 255 = white. The large numbers (> 150) correspond to the two white whiskers. The principal characteristic of this 8×8 block is the presence of diagonal stripes with high contrast. We will see how this contrast influences the coefficients α of this function.

The α_{kl} of the function f from Table 12.2 are given in Table 12.3. They are presented in the same order as previously, with α_{00} in the upper left and α_{07} in the upper right. None of the entries are exactly zero-valued, but we see that the largest coefficients (in terms of absolute value) are α_{00}, α_{01}, α_{12}, α_{23}, ... To interpret these numbers we need to have a better "visual" understanding of the elements of the basis $\mathcal{B}'$.

Consider once again the change of basis expressions

$$\alpha = CfC^t \qquad \text{and} \qquad f = C^t \alpha C.$$

In terms of the coefficients themselves, the relationship giving f from α is

$$f_{ij} = \sum_{k,l=0}^{N-1} \alpha_{kl}(c_{ki}c_{lj}).$$

Let A_{kl} be the $N \times N$ matrix whose elements are $[A_{kl}]_{ij} = c_{ki}c_{lj}$. We see that f is a linear combination of the matrices A_{kl} with weights α_{kl}. The set of N^2 matrices $\{A_{kl}, 0 \le k, l \le N-1\}$ forms a basis in terms of which the function f is described. The 64 basis matrices A_{kl} of this example ($N = 8$) are shown in Figure 12.5. Matrix

40	193	89	37	209	236	41	14
102	165	36	150	247	104	7	19
157	92	88	251	156	3	20	35
153	75	220	193	29	13	34	22
116	173	240	54	11	38	20	19
162	255	109	9	26	22	20	29
237	182	5	28	20	15	28	20
222	33	8	23	24	29	23	23

Table 12.2. The 64 values of the function f.

681.63	351.77	−8.671	54.194	27.63	−55.11	−23.87	−15.74
144.58	−94.65	−264.52	5.864	7.660	−89.93	−24.28	−12.13
−31.78	−109.77	9.861	216.16	29.88	−108.14	−36.07	−24.40
23.34	12.04	53.83	21.91	−203.72	−167.39	0.197	0.389
−18.13	−40.35	−19.88	−35.83	−96.63	47.27	119.58	36.12
11.26	9.743	24.22	−0.618	0.0879	47.44	−0.0967	−23.99
0.0393	−12.14	0.182	−11.78	−0.0625	0.540	0.139	0.197
0.572	−0.361	0.138	−0.547	−0.520	−0.268	−0.565	0.305

Table 12.3. The 64 coefficients α_{kl} of the function f.

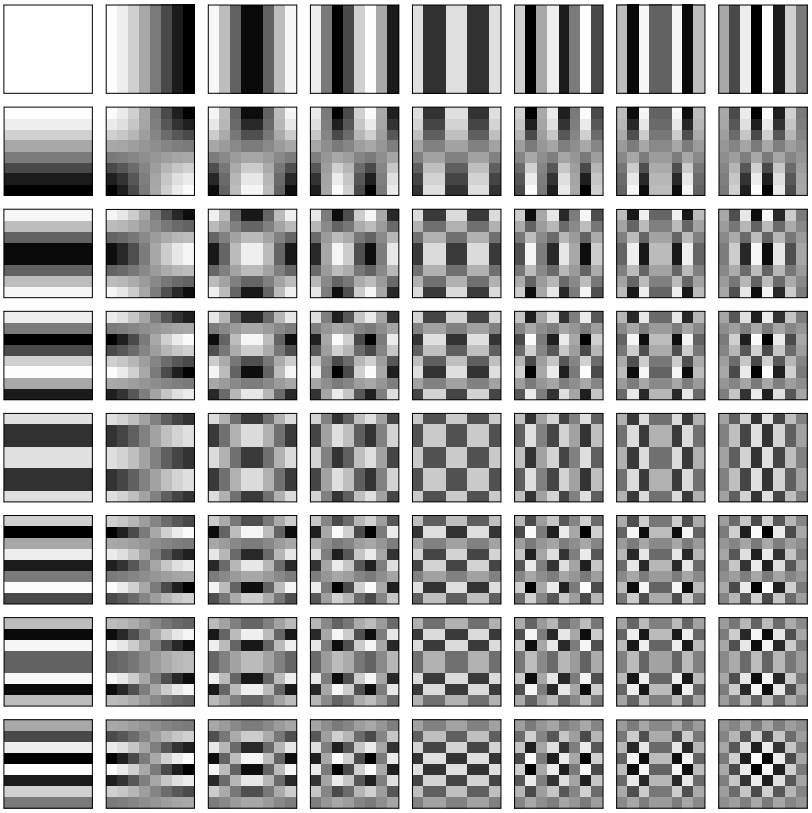


Fig. 12.5. The 64 elements A_{kl} of the basis $\mathcal{B}'$. Element A_{00} is at the upper left and element A_{07} is at the upper right.

A_{00} is in the upper left corner of the image, while A_{07} is found in the upper right. To graphically represent each basis matrix we needed to have their coefficients mapped to gray tones in the range 0 to 255. This was done by first replacing the $[A_{kl}]_{ij}$ by

$$[\tilde{A}_{kl}]_{ij} = \frac{N}{\delta_k \delta_l}[A_{kl}]_{ij},$$

where δ_k and δ_l are given by (12.7). This transformation ensures that $[\tilde{A}_{kl}]_{ij} \in [-1, 1]$. Next, the transformation aff_2 of (12.2) was applied to each scaled coefficient to obtain

$$[B_{kl}]_{ij} = \mathrm{aff}_2([\tilde{A}_{kl}]_{ij}) = \left\lfloor \frac{255}{2}([\tilde{A}_{kl}]_{ij} + 1) \right\rfloor.$$

The $[B_{kl}]_{ij}$ can be directly interpreted as gray tones, since $0 \leq [B_{kl}]_{ij} \leq 255$. These are the values represented in Figure 12.5.

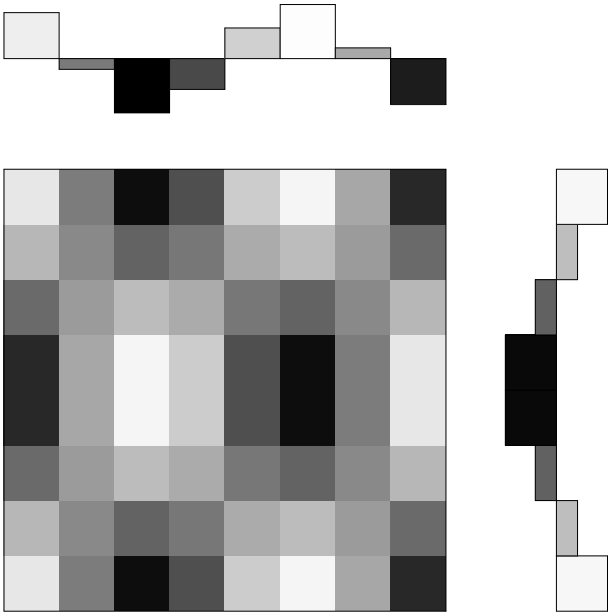


Fig. 12.6. Constructing the graphic representation of A_{23}.

It is possible to understand the graphic representations of the A_{kl} directly from their definitions. Here we consider the details of the construction of the element A_{23}, given by

$$[A_{23}]_{ij} = \frac{2}{N} \cos \frac{2(2i+1)\pi}{2N} \cos \frac{3(2j+1)\pi}{2N}.$$

The upper portion of Figure 12.6 shows the function

$$\cos \frac{3(2j+1)\pi}{16},$$

and at right, vertically, the function

$$\cos \frac{2(2i+1)\pi}{16}$$

has been shown. Since j varies from 0 to $N-1=7$, the argument of the cosine of the first function passes from $3\pi/16$ to $3 \cdot 15\pi/16 = 45\pi/16 = 2\pi + 13\pi/16$ and the figure therefore shows roughly one and one-half cycles of the cosine. Each rectangle of the histogram has been assigned the gray tone corresponding to

$$\frac{255}{2}\left(\cos\left(\frac{3(2j+1)\pi}{16}\right)+1\right).$$

The same process has been repeated for the second function, $\cos 2(2i+1)\pi/16$, and the results of this shown vertically at the right of the figure. The function A_{23} is obtained by multiplying these two functions. This multiplication is between two cosine functions, thus between values in the range $[-1, 1]$. The result of this multiplication can be interpreted visually from the image. Multiplying two very light rectangles (corresponding to values near $+1$) or two very dark rectangles (corresponding to values near -1) results in light values. The 8×8 "product" of the two histograms is the matrix of basis element A_{23}.

We return to the 8×8 block depicting the two cat whiskers. What coefficients α_{kl} will be the most important? A coefficient α_{kl} will have larger magnitude if the extrema of the basis matrix correspond roughly to those of f. For example, the basis A_{77} (bottom right corner of Figure 12.5) alternates rapidly between black and white in both directions. It has many extrema, while f depicts only a diagonal pattern. As can be predicted, the associated coefficient is quite small at $\alpha_{77} = 0.305$. On the other hand, the coefficient α_{01} will be quite large. The basis matrix A_{01} (second from the left in the top row of Figure 12.5) contains a bright left half and a dark right half. Even though the two white whiskers of f extend into the right half of the 8×8 block, the left half is significantly paler than the right one. The actual coefficient is $\alpha_{01} = 351.77$.

How should we interpret a negative coefficient α_{kl}? The coefficient $\alpha_{12} = -264.52$ is negative, and a closer inspection yields an answer. The basis matrix A_{12} is roughly divided into six contrasting bright and dark regions, three at the top and three at the bottom. Observe that two of the dark regions are roughly aligned with the brightest region of f, the whiskers. Multiplying this basis matrix by -1 would make these dark regions light, indicating that $-A_{12}$ describes the contrast between the whiskers and the background relatively well, thus the importance of this (negative) coefficient. We can easily repeat this "visual calculation" for each of the basis matrices, but it quickly becomes tedious. In fact, it is faster to program a computer to perform the calculations of (12.5). Regardless, this discussion has demonstrated the following intuitive rule: *the coefficient α_{kl} associated with a function f will have a significant magnitude if the extrema of A_{kl} are similar to those of f. A negative coefficient indicates that the bright spots of f matched dark spots of the basis element and vice versa.* As such, the nearly constant basis matrices A_{00}, A_{01}, and A_{10} are likely to have large factors α_{kl} for nearly constant functions f. At the other extreme, the basis matrices A_{67}, A_{76}, and A_{77} will be important for representing rapidly varying functions.

Proof of the orthogonality of C (12.11): To show this somewhat surprising fact, we rewrite the identify $C^t C = I$ in terms of its coefficients:

$$[C^t C]_{jk} = \sum_{i=0}^{N-1} [C^t]_{ji}[C]_{ik} = \sum_{i=0}^{N-1} [C]_{ij}[C]_{ik} = \delta_{jk} = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise,} \end{cases}$$

or equivalently

$$[C^tC]_{jk} = \sum_{i=0}^{N-1} \frac{\delta_i^2}{N} \cos\frac{i(2j+1)\pi}{2N} \cos\frac{i(2k+1)\pi}{2N} = \delta_{jk}. \tag{12.11}$$

Proving (12.11) is equivalent to proving (12.9), the orthogonality of C, which implies the invertibility of (12.5). The proof that follows is not that difficult, but it contains several cases and subcases that must be carefully considered.

We expand the product of cosines from (12.11) using the trigonometric identity

$$\cos\alpha\cos\beta = \frac{1}{2}\cos(\alpha+\beta) + \frac{1}{2}\cos(\alpha-\beta).$$

Let $S_{jk} = [C^tC]_{jk}$. Then we have that

$$\begin{aligned} S_{jk} &= \sum_{i=0}^{N-1} \frac{\delta_i^2}{N} \cos\frac{i(2j+1)\pi}{2N} \cos\frac{i(2k+1)\pi}{2N} \\ &= \sum_{i=0}^{N-1} \frac{\delta_i^2}{2N} \left(\cos\frac{i(2j+2k+2)\pi}{2N} + \cos\frac{i(2j-2k)\pi}{2N} \right) \\ &= \sum_{i=0}^{N-1} \frac{\delta_i^2}{2N} \left(\cos\frac{2\pi i(j+k+1)}{2N} + \cos\frac{2\pi i(j-k)}{2N} \right). \end{aligned}$$

Since $\delta_i^2 = 1$ if $i = 0$ and $\delta_i^2 = 2$ otherwise, we can add the $i = 0$ term and subtract it to obtain

$$S_{jk} = \frac{1}{N} \sum_{i=0}^{N-1} \left(\cos\frac{2\pi i(j+k+1)}{2N} + \cos\frac{2\pi i(j-k)}{2N} \right) - \frac{1}{N}.$$

We split the proof into the following three cases: $j = k$, $j - k$ is even but nonzero, $j - k$ is odd. Observe that exactly one of $(j-k)$ and $(j+k+1)$ is even, while the other is odd. We consider each of these cases by separating the sum and the term $-\frac{1}{N}$ as follows:

$\boxed{j = k}$ We write $S_{jk} = S_1 + S_2$ with

$$S_1 = -\frac{1}{N} + \frac{1}{N} \sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N}, \qquad S_2 = \frac{1}{N} \sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N},$$

where $l = j + k + 1$ is odd, where $l = j - k = 0$.

$\boxed{j - k \text{ even and } j \neq k}$ Write $S_{jk} = S_1 + S_2$ with

$$S_1 = -\frac{1}{N} + \frac{1}{N} \sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N}, \qquad S_2 = \frac{1}{N} \sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N},$$

where $l = j + k + 1$ is odd, where $l = j - k$ is even and nonzero.

$\boxed{j-k \text{ odd}}$ Write $S_{jk} = S_1 + S_2$ with

$$S_1 = \frac{1}{N}\sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N}, \qquad S_2 = -\frac{1}{N} + \frac{1}{N}\sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N},$$

where $l = j + k + 1$ is even, nonzero, and $< 2N$, where $l = j - k$ is odd.

There are three distinct sums to be studied:

$$\frac{1}{N}\sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N}, \qquad \text{where } l = 0, \tag{12.12}$$

$$\frac{1}{N}\sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N}, \qquad \text{where } l \text{ even, nonzero, and } < 2N, \tag{12.13}$$

$$-\frac{1}{N} + \frac{1}{N}\sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N}, \qquad \text{where } l \text{ odd}. \tag{12.14}$$

The first case is simple, since if $l = 0$ it follows that

$$\frac{1}{N}\sum_{i=0}^{N-1} \cos\frac{2\pi il}{2N} = \frac{1}{N}\sum_{i=0}^{N-1} 1 = \frac{N}{N} = 1.$$

Since we wish to show that S_{jk} is zero unless $j = k$ (otherwise, $S_{jj} = 1$), the proof is finished if we can show that (12.13) and (12.14) are both zero. For (12.13) recall that

$$\sum_{i=0}^{2N-1} e^{2\pi il\sqrt{-1}/2N} = \frac{e^{2\pi l\cdot 2N\sqrt{-1}/2N} - 1}{e^{2\pi l\sqrt{-1}/2N} - 1} = 0 \tag{12.15}$$

if $e^{2\pi l\sqrt{-1}/2N} \neq 1$. If $l < 2N$ this inequality is always satisfied. By taking the real part of (12.15) we find that

$$\sum_{i=0}^{2N-1} \cos\frac{2\pi il}{2N} = 0.$$

The sum contains twice as many terms as (12.13). However, we can rewrite it as

$$\begin{aligned}
0 &= \sum_{i=0}^{2N-1} \cos \frac{2\pi il}{2N} \\
&= \sum_{i=0}^{N-1} \cos \frac{2\pi il}{2N} + \sum_{i=N}^{2N-1} \cos \frac{2\pi il}{2N} \\
&= \sum_{i=0}^{N-1} \cos \frac{2\pi il}{2N} + \sum_{j=0}^{N-1} \cos \frac{2\pi (j+N)l}{2N}, \qquad \text{for } i = j+N, \\
&= \sum_{i=0}^{N-1} \cos \frac{2\pi il}{2N} + \sum_{j=0}^{N-1} \cos \left(\frac{2\pi jl}{2N} + \frac{2\pi Nl}{2N} \right).
\end{aligned}$$

If l is even, the phase $\frac{2\pi Nl}{2N} = \pi l$ is an even multiple of π and can therefore be dropped, since the cosine is periodic with period 2π. Thus

$$0 = \sum_{i=0}^{N-1} \cos \frac{2\pi il}{2N} + \sum_{j=0}^{N-1} \cos \frac{2\pi jl}{2N} = 2 \sum_{i=0}^{N-1} \cos \frac{2\pi il}{2N},$$

and hence the sum of (12.13) is zero-valued.

Observe that the first term $i = 0$ of the sum from (12.14) is

$$\frac{1}{N} \cos \frac{2\pi \cdot 0 \cdot l}{2N} = \frac{1}{N},$$

which cancels the term $-\frac{1}{N}$. As such, the sum from (12.14) simplifies to

$$\sum_{i=1}^{N-1} \cos \frac{2\pi il}{2N}.$$

We must now divide case (12.14) into two subcases, N even and N odd. We divide the sum $\sum_{i=1}^{N-1} \cos \frac{2\pi il}{2N}$ as follows:

N odd

$$\sum_{i=1}^{\frac{N-1}{2}} \cos \frac{2\pi il}{2N} \qquad \text{and} \qquad \sum_{i=\frac{N-1}{2}+1}^{N-1} \cos \frac{2\pi il}{2N}$$

and

N even

$$\text{the term } i = \frac{N}{2}, \qquad \sum_{i=1}^{\frac{N}{2}-1} \cos \frac{2\pi il}{2N}, \qquad \text{and} \qquad \sum_{i=\frac{N}{2}+1}^{N-1} \cos \frac{2\pi il}{2N}.$$

We start with this last subcase. If N is even, then for $i = N/2$ we have

$$\cos\frac{2\pi}{2N}\cdot\frac{N}{2}\cdot l = \cos\frac{\pi}{2}l = 0,$$

since l is odd. Rewrite the second sum by letting $j = N - i$; since $\frac{N}{2}+1 \le i \le N-1$, the domain of j is $1 \le j \le \frac{N}{2}-1$:

$$\sum_{i=\frac{N}{2}+1}^{N-1}\cos\frac{2\pi il}{2N} = \sum_{j=1}^{\frac{N}{2}-1}\cos\frac{2\pi(N-j)l}{2N} = \sum_{j=1}^{\frac{N}{2}-1}\cos\left(\pi l - \frac{2\pi jl}{2N}\right).$$

And since l is odd, the phase πl is always an odd multiple of π, and

$$\sum_{i=\frac{N}{2}+1}^{N-1}\cos\frac{2\pi il}{2N} = \sum_{j=1}^{\frac{N}{2}-1} -\cos\left(-\frac{2\pi jl}{2N}\right).$$

Since the cosine function is even, we have finally that

$$\sum_{i=\frac{N}{2}+1}^{N-1}\cos\frac{2\pi il}{2N} = -\sum_{j=1}^{\frac{N}{2}-1}\cos\frac{2\pi jl}{2N},$$

and the two sums of the subcase cancel each other. The subcase of (12.14) where N is odd is left as an exercise to the reader. □

12.5 The JPEG Standard

As discussed in the introduction, a good compression method will be tailored to the specific use and type of the object being compressed. The JPEG standard is intended for use in compressing images, more specifically photorealistic ones. As such, the compression technique is based on the fact that most photographs consist primarily of gentle gradients and transitions, while rapid variations are relatively rare. With what we have just learned about the discrete cosine transform and the coefficients α_{kl}, it seems natural to let the low-frequency components (with small l and k) play a large role, while letting high-frequency components (with l and k near N) play a small role. The following rule serves as a guide: all loss of information that is imperceptible to the human visual system (eyes and brain) is acceptable.

The compression algorithm can be broken down into the following major steps:

- translation of the image function,
- application of the discrete cosine transform to each 8×8 block,
- quantization of the transformed coefficients,

- zigzag ordering and encoding of the quantized coefficients.

We will describe each of these steps as applied to the image of a cat from Figure 12.1. This photo was taken by a digital camera that natively compressed the image in JPEG format. A 640×640 crop of the image was taken and subsequently converted to grayscale, with each pixel taking an integer value between 0 and 255. Recall that each pixel requires one byte of raw storage and therefore that the image requires 409,600 B = 409.6 KB = 0.4096 MB to store uncompressed.

Translation of the image function. The first step is the *translation* of the values of f by the quantity 2^{b-1}, where b is the number of bits (or *bit depth*) used to represent each pixel. In our case we are using $b = 8$, and we therefore subtract $2^{b-1} = 2^7 = 128$ from each pixel. This first step produces a function $\tilde{f}$ whose values are in the interval $[-2^{b-1}, 2^{b-1} - 1]$, which is (nearly) symmetric with respect to the origin, like the range of the cosine functions that form the basis matrices A_{kl}. We will follow the details of the algorithm on the 8×8 block shown in Table 12.2. The values of the translated function $\tilde{f}_{ij} = f_{ij} - 128$ are shown in Table 12.4, while the original values of the function f may be found in Table 12.2.

-88	65	-39	-91	81	108	-87	-114
-26	37	-92	22	119	-24	-121	-109
29	-36	-40	123	28	-125	-108	-93
25	-53	92	65	-99	-115	-94	-106
-12	45	112	-74	-117	-90	-108	-109
34	127	-19	-119	-102	-106	-108	-99
109	54	-123	-100	-108	-113	-100	-108
94	-95	-120	-105	-104	-99	-105	-105

Table 12.4. The 64 values of the function $\tilde{f}_{ij} = f_{ij} - 128$.

Discrete cosine transformation of each 8×8 block. The second step consists in partitioning the image into nonoverlapping blocks of 8×8 pixels. (If the image width is not a multiple of 8, then columns are added to the right until it is. The pixels in these additional columns are assigned the same gray tone as the rightmost pixel in each row of the original image. A similar treatment is applied to the bottom of the picture if the height is not a multiple of 8.) After *partitioning* the image into 8×8 blocks the *discrete cosine transform* is applied to each block. The result of this second step as applied to $\tilde{f}$ is given in Table 12.5. If we compare these coefficients to the α_{kl} of f shown in Table 12.3, we see that only the coefficient α_{00} has changed. This is no coincidence and is a direct result of the fact that $\tilde{f}$ is obtained from f by a translation. Exercise 11 (b) investigates why this happens.

−342.38	351.77	−8.671	54.194	27.63	−55.11	−23.87	−15.74
144.58	−94.65	−264.52	5.864	7.660	−89.93	−24.28	−12.13
−31.78	−109.77	9.861	216.16	29.88	−108.14	−36.07	−24.40
23.34	12.04	53.83	21.91	−203.72	−167.39	0.197	0.389
−18.13	−40.35	−19.88	−35.83	−96.63	47.27	119.58	36.12
11.26	9.743	24.22	−0.618	0.0879	47.44	−0.0967	−23.99
0.0393	−12.14	0.182	−11.78	−0.0625	0.540	0.139	0.197
0.572	−0.361	0.138	−0.547	−0.520	−0.268	−0.565	0.305

Table 12.5. The 64 coefficients α_{kl} of the function $\tilde{f}$.

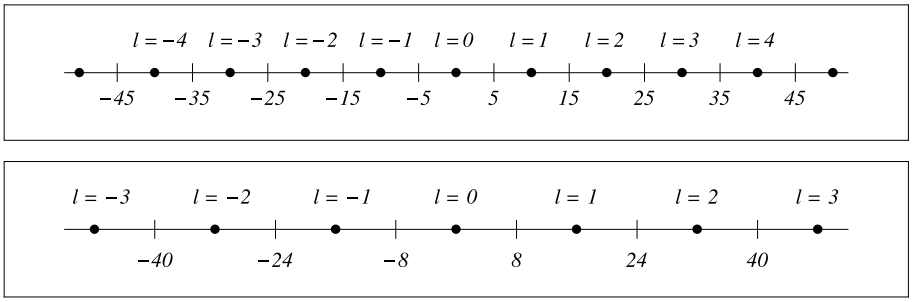


Fig. 12.7. The discrete scales used to measure α_{00} (top) and both α_{01} and α_{10} (bottom).

Quantization. The third step is called *quantization*: it consists in transforming the real-valued coefficients α_{kl} into integers ℓ_{kl}. The integer ℓ_{kl} is obtained from α_{kl} and q_{kl} by the formula

$$\ell_{kl} = \left[\frac{\alpha_{kl}}{q_{kl}} + \frac{1}{2} \right], \tag{12.16}$$

where $[x]$ is the integer part of x.

We explain the origins of this formula. Since the set of real numbers that can be represented on a computer is finite, the mathematical concept of the real line is not natural on computers. These numbers must be discretized, but must it be to the full precision that the computer is capable of representing? Could we not discretize them at a coarser scale? The JPEG standard gives a large amount of flexibility at this step: each coefficient α_{kl} is discretized with an individually chosen quantization step. The size of the step is encoded in the *quantization table*, which is fixed across all 8×8 blocks in a single image. The quantization table that we will use is shown in Table 12.6. For this table the step size for α_{00} will be 10, while already for α_{01} and α_{10} it will be 16. Figure 12.7 shows the effects of these step sizes for these three coefficients. Observe that all α_{00} from 5 up to but not including 15 will be mapped to the value $\ell_{00} = 1$; in

10	16	22	28	34	40	46	52
16	22	28	34	40	46	52	58
22	28	34	40	46	52	58	64
28	34	40	46	52	58	64	70
34	40	46	52	58	64	70	76
40	46	52	58	64	70	76	82
46	52	58	64	70	76	82	88
52	58	64	70	76	82	88	94

Table 12.6. The quantization table q_{kl} used in this example.

fact, from

$$\ell_{00}(5) = \left\lfloor \frac{5}{10} + \frac{1}{2} \right\rfloor = [1] = 1$$

and

$$\ell_{00}(15 - \epsilon) = \left\lfloor \frac{15 - \epsilon}{10} + \frac{1}{2} \right\rfloor = \left\lfloor 2 - \frac{\epsilon}{10} \right\rfloor = 1$$

for an arbitrarily small positive number ϵ. Figure 12.7 shows the window of values that are mapped to the same quantized coefficient, each delimited by a small vertical bar. Any values of α_{kl} between two numbers below the axis will share the same ℓ at the moment of reconstruction, the ℓ noted above the central dot. These dots indicate the middle of each region, and the value $\ell_{kl} \times q_{kl}$ will be assigned to the coefficient when they are uncompressed. The fraction $\frac{1}{2}$ in (12.16) ensures that $\ell_{kl} \times q_{kl}$ falls in the middle of each window. The second axis of Figure 12.7 depicts the situation for α_{01} and α_{10}, whose quantification factor is larger, namely $q_{01} = q_{10} = 16$. More values of α_{01} (and α_{10}) will be identified to the same ℓ_{01} (and ℓ_{10}) due to this wider window. As can be seen, the larger the value of q_{kl}, the rougher the approximation of the reconstructed α_{kl} and the more information that is lost. The largest step size in our quantification table is $q_{77} = 94$. All coefficients α_{77} whose values lie in the range $[-47, 47)$ will map to the value $\ell_{77} = 0$. The precise value of the original coefficient in this interval will be irrevocably lost during the compression process.

Having chosen the quantization table shown in Table 12.6, we can quantifiy the transform coefficients of the original block f; they are shown in Table 12.7.

Most digital cameras offer a way to save images at various quality levels (basic, normal, and fine, for example). Most software packages for manipulating digital images offer similar functionality. Once a given quality level has been chosen, the image is compressed using a quantization table that has been predetermined by the makers of the hardware or software. The same quantization table is used for *all* 8×8 blocks

-34	22	0	2	1	-1	-1	0
9	-4	-9	0	0	-2	0	0
-1	-4	0	5	1	-2	-1	0
1	0	1	0	-4	-3	0	0
-1	-1	0	-1	-2	1	2	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Table 12.7. The quantization ℓ_{kl} of the transformed coefficients α_{kl}.

in the image. It is transmitted once from the header of the JPEG file, followed by the transformed, quantized, and compressed block coefficients. Even though the JPEG standard suggests a family of quantization tables, any one may be used. As such, the quantization table offers a large amount of flexibility to the end user.

Zigzag ordering and encoding. The last step of the compression algorithm is the *encoding* of the table of quantized coefficients ℓ_{kl}. We will not delve too far into the details of this step. We will say only that the coefficient ℓ_{00} is encoded slightly differently from the rest and that the encoding uses the ideas discussed in the introduction: the values of ℓ_{kl} occurring more frequently are assigned shorter code words and vice versa. What are the most likely values? The JPEG standard prefers coefficients with a small absolute value: the smaller $|\ell_{kl}|$, the smaller the code word for ℓ_{kl}. Is it surprising that many coefficients ℓ_{kl} are nearly zero-valued? No, it is not if we recall that the α_{kl} (and hence the ℓ_{kl}) typically measure changes that are relatively small in scope with respect to the actual size of the image.

Thanks to the quantization step, many ℓ_{kl} with large k and l are zero-valued. The encoding makes use of this fact by ordering the coefficients such that long strings of zero-valued coefficients are more likely. The precise ordering defined by the JPEG standard is shown in Figure 12.8: $\ell_{01}, \ell_{10}, \ell_{20}, \ell_{11}, \ell_{02}, \ell_{03}, \ldots$. Given that most of the nonzero coefficients tend to be clustered in the upper left corner, it often happens that the coefficients ordered in this manner are terminated by a long run of zero values. Rather than encoding each of these zero values, the encoder sends a single special code word indicating the "end of block." When the decoder encounters this symbol it knows that the rest of the 64 symbols are to filled in with zeros. Looking at Table 12.7, note that $\ell_{46} = 2$ is the last nonzero coefficient in the proposed zigzag ordering. The eleven remaining coefficients $(\ell_{37}, \ell_{47}, \ell_{56}, \ell_{65}, \ell_{74}, \ell_{75}, \ell_{66}, \ell_{57}, \ell_{67}, \ell_{76}, \ell_{77})$ are all zero-valued and will not be explicitly transmitted. As we will see in the example of the image of the cat, this provides an enormous gain to the compression ratio.

Reconstruction. A computer can quickly reconstruct a photo from the information in a JPEG file. The quantification table is first read from the file header. Then the

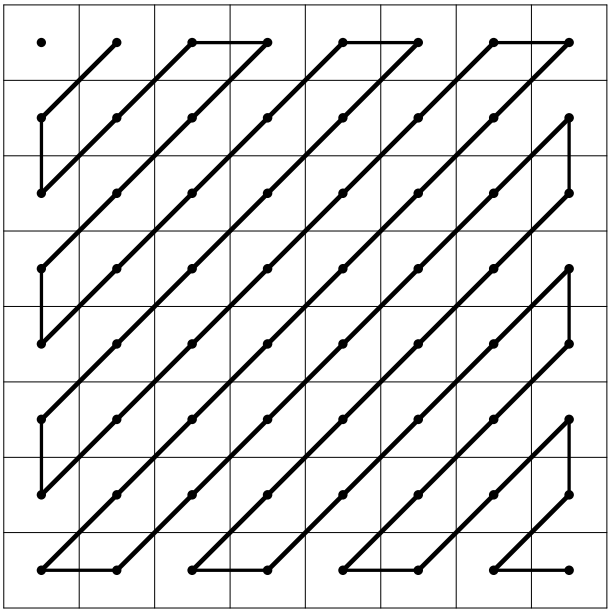


Fig. 12.8. The order in which the coefficients ℓ_{kl} are transmitted: $\ell_{01}, \ell_{1,0}, \ldots, \ell_{77}$.

following steps are performed for each 8×8 block: the information for a block is read until the "end of block" signal is encountered. If fewer than 64 coefficents were read, the missing ones are set to zero. The computer then multiplies each ℓ_{kl} by the corresponding q_{kl}. The coefficient $\beta_{kl} = \ell_{kl} \times q_{kl}$ is therefore chosen in the middle of the quantification window where the original α_{kl} lay. The inverse of the discrete cosine transformation (12.10) is then applied to the β's to get the new gray tones $\bar{f}$:

$$\bar{f} = C^t \beta C.$$

After correcting for the translation of the original image, the gray tones for this 8×8 block are ready to be shown on screen.

Figure 12.9 shows the visual results of JPEG compression, applied to the entire image as described in this section. Recall that the original photo contains 640×640 pixels and therefore $80 \times 80 = 6400$ blocks of 8×8 pixels. The four steps (translation, transformation, quantization, and encoding) are thus performed 6400 times. The left column of Figure 12.9 contains the original image plus two successive closeups.[5] The right column contains the same image after being JPEG compressed and decompressed using the quantization table of Table 12.6.

[5]Recall that the original photo was obtained from a digital camera that itself stores the image in JPEG form.

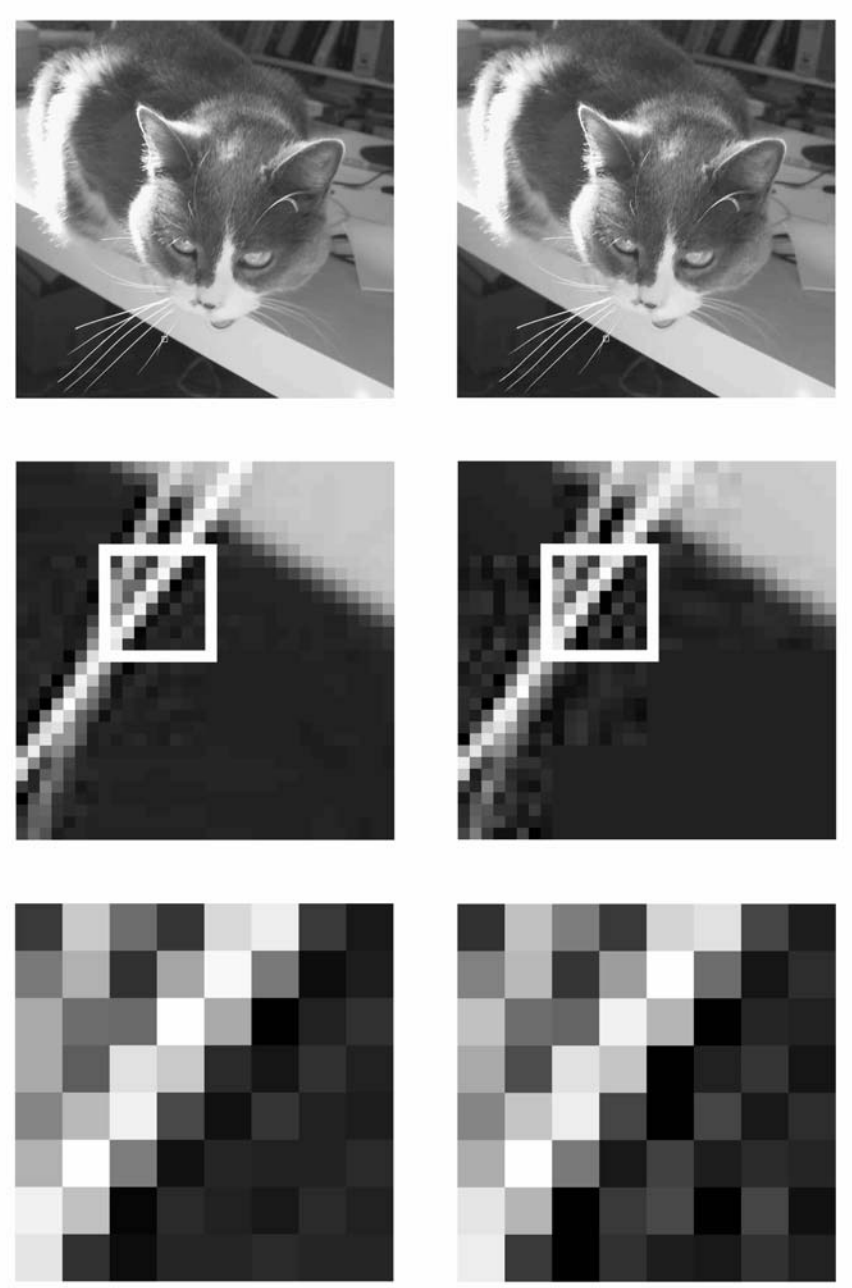


Fig. 12.9. The three images at the left are the same as those of Figure 12.1. Those at the right have been obtained from this image after being heavily JPEG compressed. The middle blocks are 32×32 pixels, while those at the bottom are 8×8 pixels.

The 8×8 block containing the crossing of two whiskers has been chosen because it is a block with high contrast. These are the types of blocks that are the least well compressed by the JPEG standard. By comparing the closeups we can see the effect of the aggressive compression. Close to the border between the highly contrasting regions the effect is most noticeable. Since this block contains high-contrast quickly varying data, we would have had to store the coefficients α_{kl} with more precision in order to reproduce them clearly. The aggressive zeroing of many of these coefficients in the quantization step has introduced a certain "noise" close to the whiskers. Note that a certain amount of noise was already present in this region in the original photograph, a clear sign that the camera was using JPEG compression. Another clear sign that JPEG compression has been used is the often visible boundaries of 8×8 blocks, specifically blocks containing high contrast next to smooth blocks, as is the case in the region of the whiskers. Notice the 8×8 block second from the bottom and third from the left of the 32×32 blocks in Figure 12.9. This block is completely "under the table" and has been compressed to a uniform gray. As such, it is not surprising that after quantization it contains only two nonzero coefficients (ℓ_{00} and ℓ_{10}). The encoding of this block omits 62 coefficients, and the compression is very good!

Is this block the rule or the exception? There are $640 \times 640 = 409{,}600$ pixels in the entire image. After transforming and quantizing these coefficients, the image is encoded by a series of 409,600 coefficients ℓ_{kl}. By ordering them in zigzag order and omitting the trailing runs of zeros, we are able to avoid storing over 352,000 zeros, roughly $\frac{7}{8}$ of the coefficients! It is not surprising that the compression achieved by the JPEG standard is so good.[6]

The ultimate test is the comparison of the two images with the naked eye. It is up to the user to judge whether the compression (in this case, the zeroing of roughly $\frac{7}{8}$ of the Fourier coefficients α_{kl}) has damaged the photograph. It is important to note that this comparison should be performed under the same conditions in which the compressed photograph will be used. Recall the example of the digitized works from the Louvre. If the image is going to be looked at using a low-resolution screen, then the compression can be relatively aggressive. However, if the image is to be closely studied by art historians, is to be printed at high resolution, or is to be viewed through software that allows zooming in, then a higher resolution and a less-aggressive compression should be used.

The JPEG standard offers an enormous amount of flexibility through its quantization tables. In certain cases we can imagine that using even higher values in this table will lead to better compression and acceptable quality. However, the weaknesses of the JPEG standard are made apparent in areas of high contrast and detail, especially when the quantization table contains overly large values. This is why the JPEG standard performs so poorly at compressing line art and cartoons, which consist largely of black lines on a white background. These lines become marred (with a characteristic JPEG

[6]Through careful choice of the quantization table this photo can be compressed to less than 30 KB in size (compared to 410 KB uncompressed) without the degradation being intolerable.

"speckle") after aggressive compression. It would be equally inappropriate to take a picture of a page of text and compress it using the JPEG standard; the letters are in high contrast with the page and would become blurred. The JPEG standard was created with the goal of compressing photographs and photorealistic images and it excels at this task.

What about color images? It is well known that colors can be described using three dimensions. For example, the color of a pixel on a computer screen is normally described as a ratio of the three (additive) primary colors: red, green, and blue. The JPEG standard uses a different set of coordinates (or *color space*). It is based on recommendations made by the *Commission internationale de l'éclairage* (International Commission on Illumination), which in the 1930s developed the first standards in this domain. The three dimensions of this color space are separated, leading to three independent images. These images, each corresponding to one coordinate, are then individually treated in the same manner as discussed in this chapter for gray tones. (For those who want to learn more, the book [2] contains a self-contained description of the standard with enough information to fully implement the standard, a discussion of the science underlying the various mathematical tools used in it, and the necessary knowledge on the human visual system. References [3, 4] are good entry points in the field of data compression.)

12.6 Exercises

1. **(a)** Verify that if $x \in [-1, 1] \subset \mathbb{R}$, then $\text{aff}_1(x) = 255(x+1)/2$ is an element of $[0, 255]$.
(b) Is aff_1 the ideal transformation? For which x will $\text{aff}_1(x) = 255$? Can you propose a function aff' such that all integers in $\{0, 1, 2, \ldots, 255\}$ will be images of equal-length subintervals of $[-1, 1]$?
(c) Give the inverse of aff_1. The function aff' cannot have an inverse. Why? Despite this, can you propose a rule that would allow you to construct a function g starting from a function f as in Section 12.3?

2. **(a)** Verify that the four vectors A_{00}, A_{01}, A_{10}, and A_{11} of (12.3) (expressed in the usual basis $\mathcal{B}$) are orthonormal, that is, they have length 1 and are pairwise orthogonal.
(b) Let v be the vector whose coefficients in the basis $\mathcal{B}$ are

$$[v]_{\mathcal{B}} = \begin{pmatrix} -\frac{3}{8} \\ \frac{5}{8} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}.$$

Give the coefficients of this vector in the basis $\mathcal{B}' = \{A_{00}, A_{01}, A_{10}, A_{11}\}$. What is the largest coefficient of $[v]_{\mathcal{B}'}$ in terms of absolute value? Could you have guessed which one it was going to be without explicitly calculating them? How?

3. **(a)** Show that the $N \times N$ matrix C used in the discrete cosine transform for $N = 4$ is given by

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \gamma & \delta & -\delta & -\gamma \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \delta & -\gamma & \gamma & -\delta \end{pmatrix}.$$

Express the two unknowns γ and δ in terms of the cosine function.
(b) Using the trigonometric identity $\cos 2\theta = 2\cos^2\theta - 1$, explicitly give the numbers γ and δ. (Here "explicitly" means as an algebraic expression with integer numbers and radicals *but* without the cosine function.) Using these expressions, show that the second line of C represents a vector with unit norm as is required by the orthogonality of C.

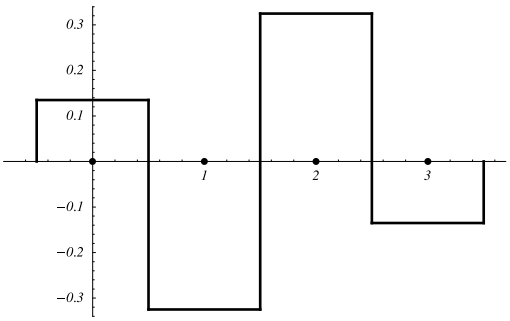


Fig. 12.10. The discrete function g of Exercise 4 (b).

4. **(a)** The discrete cosine transformation allows the expression of discrete functions $g : \{0, \ldots, N-1\} \to \mathbb{R}$ (given by $g(i) = g_i$) as linear combinations of the N discrete basis vectors C_k, where $C_k(i) = (C_k)_i = c_{ki}, k = 0, 1, 2, \ldots, N-1$. This transformation expresses g in the form $g = \sum_{k=0}^{N-1} \beta_k C_k$, which yields

$$g_i = \sum_{k=0}^{N-1} \beta_k (C_k)_i.$$

For $N = 4$, represent the function $(C_2)_i$ by a histogram. (This exercise reuses results from Exercise 3, but the reader is not required to have completed that exercise.)
(b) Knowing that the numeric values of γ and δ of the previous exercise are roughly 0.65 and 0.27 respectively, what will be the coefficient β_k with the largest magnitude for the function g represented in Figure 12.10?

5. Complete the calculation of (12.14) for the subcase in which N is odd.

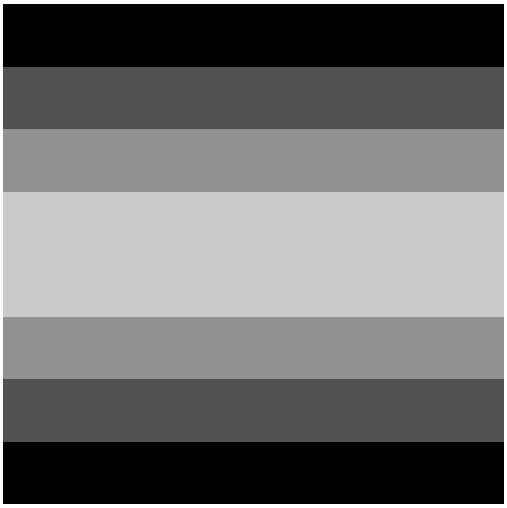


Fig. 12.11. The function f of Exercise 6.

6. A function

$$f : \{0,1,2,3,4,5,6,7\} \times \{0,1,2,3,4,5,6,7\} \to \{0,1,2,\ldots,255\}$$

is represented graphically by the gray tones of Figure 12.11. The values f_{ij} are constant along a given row; in other words, $f_{ij} = f_{ik}$ for all $j,k \in \{0,1,2,\ldots,7\}$.
(a) If $f_{0j} = 0, f_{1j} = 64, f_{2j} = 128, f_{3j} = 192, f_{4j} = 192, f_{5j} = 128, f_{6j} = 64, f_{7j} = 0$ for all j, calculate α_{00} as defined by the JPEG standard, but without doing the translation of f as described in the first step of Section 12.5.
(b) If the discrete cosine transform is carried out as suggested by the JPEG standard, several of the coefficients α_{kl} will be zero-valued. Determine which elements of α_{kl} will be zero-valued and explain why.

7. Let C be the matrix representing the discrete cosine transform. Its elements $[C]_{ij} = c_{ij}, 0 \le i,j \le N-1$, are given by (12.6). Let N be even. Show that each of the elements of rows i of C where i is odd is one of the following N values:

$$\pm\sqrt{\frac{2}{N}}\cos\frac{k\pi}{2N}, \qquad \text{with } k \in \{1,3,5,\ldots,N-1\}.$$

8. Figure 12.12 displays an 8×8 block of gray tones. Which coefficient α_{ij} will have the largest magnitude (ignoring α_{00})? What will its sign be?

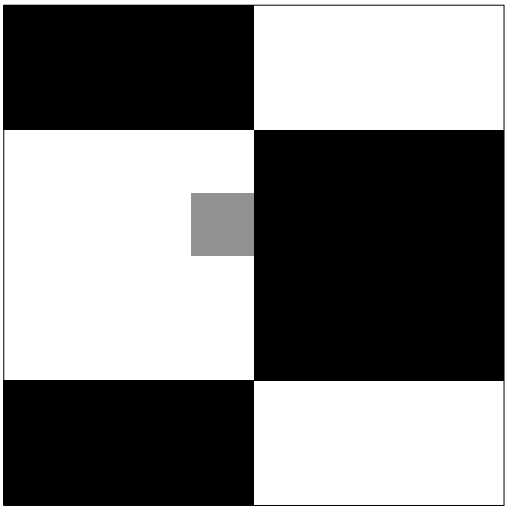


Fig. 12.12. An 8×8 block of gray tones for Exercise 8.

9. With the rising popularity of digital photography, programs allowing for the manipulation and retouching of photographs have become increasingly popular. Among other things, they allow images to be reframed (or *cropped*) by removing rows or columns from the outer edges. If an image is JPEG compressed, explain why it is better to remove groups of rows or columns that are multiples of 8.

10. **(a)** Two copies of the same photograph are independently compressed using distinct quantization tables q_{ij} and q'_{ij}. If $q_{ij} > q'_{ij}$ for all i and j, what will be, in general, the larger file, the second or the first? Which quantization table will lead to a larger loss of quality in the photograph?

(b) If the quantization table from Table 12.6 is used and if $\alpha_{34} = 87.2$, what will be the value of ℓ_{34}? What if $\alpha_{34} = -87.2$?

(c) What is the smallest value of q_{34} that will lead to a zero-valued ℓ_{34} for the values of α_{34} in the preceding question?

(d) Does $\ell_{kj}(-\alpha_{kj}) = -\ell_{kj}(\alpha_{kj})$? Explain.

Note: Another slightly different problem is raised by technology. Suppose a photo is already in the JPEG format and is available through the Internet. If the file remains large, it could be useful to recompress the file using a more aggressive quantification table for users having slower Internet connections. The choice of the new quantification table would then depend on the speed of the connection and perhaps on the use of the photo. It turns out that the choice of this second table is delicate, since the degradation of the picture does not increase monotonically with the size of its coefficients. See, for example, [1].

11. **(a)** Calculate the difference between the α_{00} of the function f given in Table 12.2 and that of the function $\tilde{f}$ obtained through translation.
(b) Show that a translation of f by any constant (for example 128) changes only the coefficient α_{00}.
(c) Using the definition of the discrete cosine transform, predict the difference between the two coefficients α_{00} calculated in (a).
(d) Show that α_{00} is N times the average gray tone of the block.

12. Let g be a step function representing a checkerboard: the upper left corner $(0, 0)$ has value $+1$, and the rest of the squares are filled in such a way that they have the opposite sign to their horizontal and vertical neighbors.
(a) Show that the step function g_{ij} can be described by the formula

$$g_{ij} = \sin(i + \tfrac{1}{2})\pi \cdot \sin(j + \tfrac{1}{2})\pi.$$

(b) Calculate the eight numbers

$$\lambda_i = \sum_{j=0}^{7} c_{ij} \sin(j + \tfrac{1}{2})\pi, \qquad \text{for } i = 0, \ldots, 7,$$

where c_{ij} is given by (12.6). (If this exercise is taking too long to perform by hand, consider using a computer!)
(c) Calculate the coefficients β_{kl} of the checkboard function g given by $\beta_{kl} = \sum_{i,j=0}^{N-1} c_{ki} c_{lj} g_{ij}$ (calculating the values λ_i is helpful). Could you have guessed exactly which coefficients would be zero-valued? Is the position of the largest nonzero coefficient β_{kl} surprising?

References

[1] H.H. Bauschke, C.H. Hamilton, M.S. Macklem, J.S. McMichael, and N.R. Swart. Recompression of JPEG images by requantization. *IEEE Transactions on Image Processing*, 12:843–849, 2003.

[2] W.B. Pennebaker and J.L. Mitchell. *JPEG Still Image Data Compression Standard*. Springer, New York, 1996.

[3] D. Salomon. *Data Compression: The Complete Reference*. Springer, New York, 2nd edition, 2000.

[4] K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann, San Francisco, 1996.

13

The DNA Computer[1]

Covering this entire chapter could easily consume two full weeks of course time. However, it is equally possible to condense the core material into one week. In the latter case, provided the students have sufficient mathematical maturity, we construct the theory of recursive functions starting from simple functions and the operations of composition, recurrence, and minimization. We explain the mechanics of a Turing machine and examine a few Turing machines that calculate simple functions (Section 13.3). We state without proof Theorem 13.40, which shows that all recursive functions are Turing-calculable. At this point, we have a choice: we can decide to discuss parts of the proof in further detail, or we can skip directly to discussing DNA computers. In the latter case we have sufficient time only to discuss biological operations that can be performed on DNA, and walk through the example of Adleman's technique for solving the Hamiltonian path problem using DNA (Section 13.2).

For students with more of a computer science background it is worthwhile to spend a solid two weeks on this chapter. We spend more time describing Turing machines and we discuss at least one step in the proof that recursive functions are Turing-calculable (Theorems 13.32 and 13.40). We introduce insertion–deletion systems (Section 13.4) and we explain how enzymes are able to perform insertions and deletions on DNA. We state without proof Theorem 13.44, showing that for each Turing machine there exists an insertion–deletion system that executes the same program, and we stress the significance of this result. We discuss at least one of the cases of the proof, and if time is too short, we skip Adleman's technique.

13.1 Introduction

The subject of this chapter is an area of active research. Even though they have been used to solve an actual mathematical problem, DNA computers are still a thing of

[1]This chapter was written by Hélène Antaya and Isabelle Ascah-Coallier while supported by an NSERC Undergraduate Student Research Award.

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_13,

science fiction. Research is ongoing and requires multidisciplinary teams with expertise in computing and biochemistry.

Compare this to the development of classic computers. Their development was spurred once somebody realized that electric circuits were capable of performing logical operations. (Simple examples of this are explored in Section 15.7 of Chapter 15.) Modern computers are constructed by connecting an enormous number of transistors. In the time of the first computers, programming required an implicit knowledge of the inner workings of the computer in order to decompose the program into a sequence of operations that the computer was able to perform. Advances in several directions were made quickly, with computers becoming more and more sophisticated on one side and programming languages being developed on the other side. With this progress, it became less and less important to know the inner workings of a computer in order to use one.

Somewhere along the way we asked ourselves, what questions may be solved by a computer? In order to respond to this question we must first define exactly what we mean by an "algorithm" and a "computer." The two questions are rather difficult and push the limits of philosophy. Rather than talking about algorithms we often talk about "calculable functions." All approaches to calculability have led to equivalent definitions. In particular, if we limit ourselves to functions $f : \mathbb{N}^n \to \mathbb{N}$, then calculable functions are the recursive functions we will discuss in Section 13.3.2. In order to analyze the power of computers, rather than thinking about the most complex computers the future will bring, scientists instead focused on the simplest computer imaginable: a Turing machine, described in Section 13.3. The central theorem on this topic shows that a function $f : \mathbb{N}^n \to \mathbb{N}$ is recursive if and only if it is calculable by a Turing machine (see Theorem 13.41 for one of the two directions). This led Church to formulate his famous thesis, which states that a function is "calculable" if and only if it is calculable by a Turing machine.

The above theory yields a method for programmers to calculate all recursive functions. However, such solutions are often far from being the most elegant or the most efficient. When we are interested in numeric solutions, theoretical algorithms offer little utility, and the algorithms used in practice bear little resemblance to them. Many of the most simply stated problems are effectively unsolvable by traditional computers in reasonable time. This is the case for the problem of large integer factorization discussed in Chapter 7 and the Hamiltonian path problem discussed in this chapter. Given a set of cities and oriented paths between them, the Hamiltonian path problem asks whether there exists a path that starts in the first city, goes through each city exactly once, and ends in the last city. When the number of cities is sufficiently large (more than a hundred or so), the number of possible paths becomes so large that even the most powerful computers are unable to explore them all. There are two ways to improve performance for these types of problems: find better algorithms, or build faster computers. A simple way of building faster computers is to increase the number of processors and to connect many modern computers in parallel, allowing them to work on the same problem simultaneously. In 2005 the largest computer on the planet had 131,072 paral-

lel processors. Parallel computers are not an ideal solution, however, since the largest ones are expensive to build and they are quickly out of date.

The concept of the DNA computer was born in 1994. Leonard Adleman, a computer scientist and one of the creators of the well known RSA cryptographic system (see Chapter 7), observed that the biological operations performed on strands of DNA inside cells could be used to perform logical operations. DNA is a very large molecule arranged in a double helix, which is able to be separated into two single strands in the same way as we open a zipper. Each strand consists of a simple sequence of bases, each one of four types: A (adenine), C (cytosine), G (guanine), and T (thymine). Two single strands can be assembled into a double strand if they are complementary: A bases can pair only with T bases, while C bases can pair only with G bases. Certain enzymes are able to cut a strand of DNA at specific locations, called "loci." A snippet of DNA may be removed from a strand if it lies between two loci (deletion), and snippets may be added in a similar manner (insertion). DNA polymerase (another enzyme) allows for the duplication of DNA molecules and hence the cloning of entire DNA strands. Adleman saw these operations and was reminded of the basic operations being performed by electrical circuits and transistors in a computer (see Section 15.7 of Chapter 15). In order to demonstrate the potential computing power of DNA, Adleman used DNA manipulation to construct the solution to a Hamiltonian path problem involving seven cities. This initial demonstration quickly spurred further research on the subject. As with conventional computers, research has gone in many directions. On the theoretical side things are quite advanced. Kari and Thierrin [3] showed that all Turing-calculable functions are able to be calculated on DNA strands using insertion and deletion operations. We will show this result in Section 13.4. As is the case with conventional computers, theoretical algorithms used in proofs are not necessarily the most efficient or practical for solving actual problems. Thus, much research has focused on the more practical aspects. Adleman required seven days in the lab to find the Hamiltonian path over a set of seven cities, while most anyone would be able to find the solution in a few minutes using pencil and paper. It is not known whether large problems could be efficiently tackled with DNA computing. The technique used by Adleman is known to be practical only for small numbers of cities. However, as noted above, parallelism in conventional computers is somewhat limited by its cost. Many researchers are therefore interested in the potential parallelism of DNA computers. It is known that DNA strands can be efficiently cloned in very large numbers. Mixing them all together with the appropriate enzymes, a large number of insertions and deletions may be performed in parallel. Can this property be used to construct hugely parallel DNA computers? The research continues.

13.2 Adleman's Hamiltonian Path Problem

Even if we are not yet able to build a practical DNA computer, several simple calculations have already been performed using DNA operations. As just said, Leonard

Adleman demonstrated in 1994 the potential of DNA computing by solving an actual (albeit small) problem.

The problem starts with a directed graph, as shown in Figure 13.1. A *directed graph* is a set of nodes (here labeled by the numbers 0 through 6) and a collection of directed edges connecting pairs of nodes (here represented by arrows between nodes).

The Hamiltonian path problem consists in finding a path starting at the first node (node 0) and finishing at the last node (node 6) while passing through all other nodes exactly once, while satisfying the directions imposed on connections between nodes. This is a classic problem in mathematics.

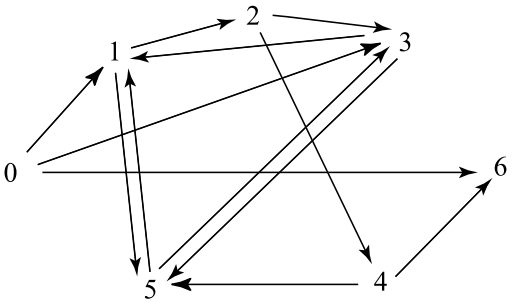


Fig. 13.1. The directed graph investigated by Adleman.

Adleman's solution: Adleman started by encoding each node using a small DNA strand consisting of eight bases. For example, node 0 may be represented by the strand

$$AGTTAGCA$$

and node 1 by

$$GAAACTAG.$$

We will use the word "prename" to refer to the first four bases in a node label, and the word "name" to refer to the last four. Directed edges are encoded as strands of eight bases, consisting of the complementary bases of the name of the departure node, followed by the complementary bases of the prename of the destination node. Recall that A is complementary to T, and C to G. For example, the arrow from 0 to 1 would be encoded by the strand $TCGTCTTT$, since $TCGT$ is the complement of the last four bases (the name) of the encoding for 0, $AGTT\underline{AGCA}$, and $CTTT$ is the complement of the first four bases (the prename) of the encoding for 1, $\underline{GAAA}CTAG$.

Adleman then placed a large number (roughly 10^{14}) of copies of each strand of DNA encoding for nodes and edges into a single test tube. DNA strands have a strong tendency to join themselves with complementary strands. For example, if the strands corresponding to nodes 0 and 1 were to come into proximity to a strand encoding for the directed edge from 0 to 1, they would likely join to create the following double strand:

$$\begin{array}{cccccccccccccccc} & & & & T & C & G & T & C & T & T & T & & & & \\ & & & & | & | & | & | & | & | & | & | & & & & \\ A & G & T & T & A & G & C & A & G & A & A & A & C & T & A & G \end{array}$$

where the vertical line represents a stable chemical bond between complementary bases. The bottom strand still contains unpaired bases. These bases can now attract the ends of other directed edges, which in turn will attract nodes. Thus the molecules in the test tube perform a large parallel computation by constructing a large number of possible paths through the graph. Any finite path through the graph of length $\leq N$ for some N could possibly be generated. This level of parallelism is simply not possible with a conventional computer or even a large cluster of conventional computers.

If the mixture is heated, the double strands of DNA separate into single strands, thus producing single strands encoding sequences of nodes and others encoding sequences of directed edges. Adleman focused on the strands encoding node sequences, since these encode the actual path walked through the graph.

The approach effectively assumes that all possible paths through the graph will be generated. If the problem has a solution, we are nearly guaranteed that this path will exist somewhere in the test tube. The problem now becomes to isolate and read this solution. How to recognize which chain is the right one among the billions of others? To succeed at this task, Adleman had to use several sophisticated biological and chemical techniques. In fact, this was by far the most difficult and onerous part of the solution. The basic approach is relatively simple to understand from a theoretical point of view. In fact, Adleman used a brute-force method, which involved making an exhaustive search through the paths and finally selecting the correct one.

To isolate the solution strand, Adleman proceeded in five steps:

Step 1. We must first select only those paths that start at node 0 and finish at node 6. The idea is to duplicate these chains until they completely dominate all others. The details of this step require a certain familiarity with chemistry, and we will discuss it in more detail in Section 13.6.3.

Step 2. Among the chains selected in step 1 we must now select those that contain exactly seven nodes (hence six directed edges). These chains will be 56 bases long, as opposed to the 48 base chains encoding directed edges (see Figure 13.2).

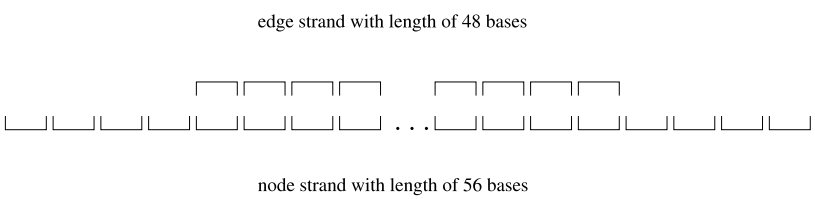


Fig. 13.2. Length of chain encoding paths.

To accomplish this, Adleman used electrophoresis, a well-known technique from biology. The basic idea is to induce a negative charge on the strands of DNA, and to place them along one edge of a plate covered in gel. Next, a voltage difference is applied across the plate, as shown in Figure 13.3.

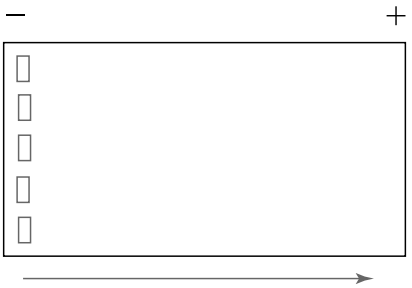


Fig. 13.3. A schematic plate for electrophoresis. (The first lane contains a DNA ladder for sizing.)

Attracted by the positive end of the plate, the strands of DNA slowly travel through the gel. As the first negatively charged molecules reach the positive end of the plate, the plate is deactivated, halting the motion. The speed of travel through the gel depends on the length of the strand of DNA, with shorter strands traveling faster than longer strands. Thus, we can estimate the position of strands on the platter as a function of their length. In order to calculate this precisely, the process is calibrated by also applying electrophoresis to a sample of molecules of known length. Thus, this technique allowed Adleman to extract only those strands of DNA with lengths of 48, 52, and 56 bases, while discarding the rest. Why did Adleman choose strands with these three lengths, rather than just those of length 56? This is due to limitations in the chemical methods being used, and will be explained further in Section 13.6.3.

Step 3. The next step is to select only those strands of DNA that also contain the five other nodes. To do this, Adleman used the principle of complementarity of bases. The basic idea is to isolate those strands of DNA that contain a particular intermediate node, one node at a time. Suppose we wish to isolate all strands that contain node 1. To start, we heat the solution so as to separate double strands into simple strands, and we mix into the solution microscopic particles of iron, attached to which are complementary strands encoding for node 1. Once mixed, all of the strands of DNA containing node 1 will attach to the complementary strands, and thus they will all have iron particles attached to them. Next, the strands of interest are separated from the others by attracting them to one side of the test tube with a magnet, and pouring out the others. The strands of interest are then put back into a solution, and heated to separate the paths from the complementary node strands and iron particles. The iron particles can now be removed

using a magnet, and the process repeated for each of the other intermediate nodes: $2, 3, 4, 5$.

Step 4. We check to see whether there are any DNA molecules left in the test tube. If there are, then we have found one or more solutions; if not, then the problem more than likely does not have a solution.

Step 5. If we found any chains in the previous step, then they must be analyzed in order to determine the exact sequence(s) they encode.

Adleman spent seven days in the laboratory to come up with the simple solution above for the graph of Figure 13.1!

13.3 Turing Machines and Recursive Functions

As mentioned in the introduction, in studying the theoretical capabilities of a computer, the most commonly used model is that of Turing machines. This approach was invented by Alan Turing in 1936 [7] with the goal of clearly defining the concept of an algorithm.

In this section we will discuss the operation of a standard Turing machine. Afterward, we will establish the connection to recursive functions. We will conclude this section with a discussion of Church's thesis, which is often considered as the formal definition of an algorithm.

13.3.1 Turing Machines

It is interesting to compare a Turing machine to a computer program. A Turing machine consists in an infinitely long tape, which may be considered as the computer memory (which is finite in the real world). The tape is divided into individual cells, each capable of storing a single symbol from a finite alphabet. At any point in time, only a finite number of cells contain symbols other than the blank symbol. The machine operates on one cell at a time, with the current cell being indicated by a pointer to it. The operation to be performed on the cell depends on a function φ, which effectively describes the program being run on the machine. The function φ takes as input the symbol in the cell being pointed to and the state of the pointer. As in normal computer programming, the function φ must obey several rules of syntax, and the function rule depends on the problem to be solved.

As an example, we will start this section with a discussion of a Turing machine built to solve a particular problem. Afterward, we will formalize the theory of Turing machines.

Example 13.1 *Consider a tape that extends infinitely to the right and that is separated into individual cells as shown in Figure 13.4. The first cell is initialized with the blank symbol* B*. It is followed by a series of cells containing* 1 *and* 0 *symbols and is terminated by another blank cell. The set of symbols* $\{0, 1, \mathrm{B}\}$ *forms the alphabet of the machine. There is a pointer in an initial state (from a finite set of states) that is pointing to the*

first cell on the tape. Our task is to change all 1 *symbols into* 0 *symbols, and vice versa, terminating with the pointer on the first cell.*

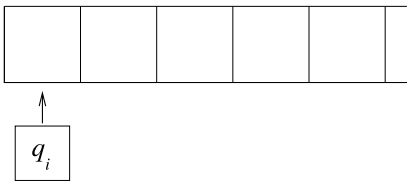


Fig. 13.4. A semi-infinite tape.

The actions to be followed by the machine depend on the state of the pointer and the symbol to which it is pointing. There are three actions:

1. *change the symbol in the cell;*
2. *change the state of the pointer;*
3. *move the pointer left or right by one cell.*

We now describe the algorithm that will complete our task. When the pointer is on the first blank cell, we move the pointer to the right. From then on, each time a 1 *is encountered it is exchanged for a* 0 *and the pointer moves to the right. Similarly, each time a* 0 *is encountered it is exchanged for a* 1 *and the pointer moves to the right. When the pointer encounters a second blank cell, it reverses direction and continues until it returns to the first blank cell. This algorithm is represented graphically in Figure 13.5.*

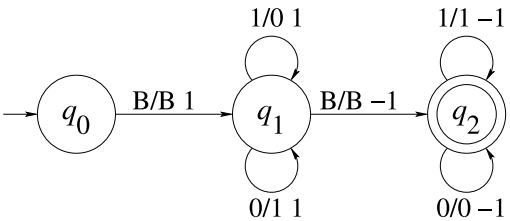


Fig. 13.5. The algorithm for Example 13.1.

We will discuss this diagram in further detail, since others of its type will be used throughout this chapter. The circles represent the possible pointer states, while arrows indicate possible actions. The arrow pointing to state q_0 *indicates that this is the initial state, while the double circle indicates that* q_2 *is the final, or halting, state. An arrow from state* q_i *to state* q_j *is labeled with a label of the form "*x_k/x_l c*" (where* $c \in \{-1, 0, 1\}$*) and is interpreted as follows: if the machine is in state* q_i *and points to a cell containing*

symbol x_k, then the symbol x_k in the cell is replaced with the symbol x_l, the pointer moves c cells (with positive entries meaning go right), and the machine transitions to state q_j.

Walk through the steps performed by the machine with an initial tape containing B10011B. *At the beginning the pointer is in state q_0 and points at the first blank cell. We will represent the state of the machine as*

$$q_0\text{B10011B}$$

Note that the pointer has been written immediately to the left of the cell it points to. This string signifies that the machine is in state q_0, that the pointer points to the first cell containing a B*, and that the tape contains the symbols* B10011B. *The machine transitions to state q_1 and the pointer is moved one cell to the right. The machine will then toggle* 1 *and* 0 *symbols, each time moving one cell to the right. Since the machine performs the same action at each of these steps it does not need to change states. This sequence of configurations is represented by*

$$\text{B}q_1\text{10011B}$$

$$\text{B0}q_1\text{0011B}$$

$$\text{B01}q_1\text{011B}$$

$$\text{B011}q_1\text{11B}$$

$$\text{B0110}q_1\text{1B}$$

$$\text{B01100}q_1\text{B}$$

Now that the pointer encounters a second B *it transitions to state q_2 and begins moving back to the left. This continues until the machine encounters the first cell containing a* B*:*

$$\text{B0110}q_2\text{0B}$$

$$\text{B011}q_2\text{00B}$$

$$\text{B01}q_2\text{100B}$$

$$\text{B0}q_2\text{1100B}$$

$$\text{B}q_2\text{01100B}$$

$$q_2B\text{01100B}$$

The machine now terminates with the task completed. In fact, the algorithm does not define what to do when the machine encounters a B *while in state q_2; thus it halts operation.*

The utility of states is now clear: they allow the machine to react differently when encountering the same symbol. We also see why we should not change state when we repeat the same operation. This allows the machine, which has a finite number of instructions, to perform the program on arbitrarily long inputs between the two B *symbols.*

We are now ready to rigorously define Turing machines.

Definition 13.2 *A standard Turing machine M is a triplet*

$$M = (Q, X, \varphi),$$

where Q is a finite set called the state alphabet, X is a finite set called the tape alphabet, and $\varphi : D \to Q \times X \times \{-1, 0, 1\}$ is a function with domain $D \subset Q \times X$. As in our example, the last item returned by the function indicates how the pointer is moved, where $-1, 0$, and 1 mean move to left, do not move, and move the right, respectively. Note that Q and X are generally chosen to be disjoint alphabets, that is, $Q \cap X = \emptyset$. Moreover, the state $q_0 \in Q$ is the initial state, $\mathrm{B} \in X$ is the blank symbol, and $Q_f \subset Q$ is the set of possible halting states.

End of Example 13.1. *Using this notation the Turing machine from Example 13.1 is described as $Q = \{q_0, q_1, q_2\}$, $X = \{1, 0, \mathrm{B}\}$, and $Q_f = \{q_2\}$, with φ being defined in Table 13.1. In this table the input states are labeled in the top row, while the input symbols (which are elements of the alphabet X) are labeled in the left column. The action of the machine on encountering a given state and symbol is defined at the intersection of the row and column containing these two labels, and contains a triplet in $Q \times X \times \{-1, 0, 1\}$.*

	q_0	q_1	q_2
B	$(q_1, \mathrm{B}, 1)$	$(q_2, \mathrm{B}, -1)$	
0		$(q_1, 1, 1)$	$(q_2, 0, -1)$
1		$(q_1, 0, 1)$	$(q_2, 1, -1)$

Table 13.1. The function φ from Example 13.1.

Remark: The tape in a standard Turing machine is unlimited in one direction. There are alternative forms of Turing machines using tapes that are unlimited in both directions, as well as Turing machines using multiple tapes. However, it can be proved that all of these machines are fundamentally equivalent to standard Turing machines [6], which is why we focus our discussion on the simplest device. Note that at any moment, even if the tape is infinite, only a finite number of cells may be nonblank. This is a direct result of the restriction that input tapes may have only a finite number of nonblank cells and that at each step of operation at most one more cell may be filled.

It is important to clearly define the class of functions that are calculable using a Turing machine, which we will call T-calculable functions. First, we define the concept of "words" over an alphabet X, which will be used often.

Definition 13.3 *Let X be an alphabet and λ the null word containing no characters. The set X^* of all words over the alphabet X, is defined as follows:*

1. $\lambda \in X^*$;
2. *If* $a \in X$ *and* $c \in X^*$, *then* $ca \in X^*$, *where* ca *represents the word constructed by appending the symbol* a *to the word* c.
3. $\omega \in X^*$ *only if it can be obtained starting with* λ *and through a finite number of applications of (ii).*

Often we will find it convenient to use the *concatenation* of two words. We formalize this operation in the following definition.

Definition 13.4 *Let* b *and* c *be two words from* X^*. *The concatenation of* b *and* c *is the word* $bc \in X^*$, *obtained by appending the characters from* c *to those of* b.

Definition 13.5 *A Turing machine* $M = (Q, X, \varphi)$ *can calculate a function* $f : U \subset X^* \to X^*$ *if*

1. *there exists a unique transition from* q_0 *of the form* $\varphi(q_0, B) = (q_i, \mathrm{B}, 1)$, *where* $q_i \neq q_0$;
2. *there does not exist a transition of the form* $\varphi(q_i, x) = (q_0, y, c)$, *where* $i \neq 0$, $x, y \in X$, *and* $c \in \{-1, 0, 1\}$;
3. *there does not exist a transition of the form* $\varphi(q_f, \mathrm{B})$, *where* $q_f \in Q_f$;
4. *for all* $\mu \in U$, *the operation performed by* M *on* μ *with an initial configuration of* $q_0\mathrm{B}\mu\mathrm{B}$ *stops in the final configuration* $q_f\mathrm{B}\nu\mathrm{B}$ *with* $\nu \in X^*$ *after a finite number of steps if* $f(\mu) = \nu$ *(we say that a Turing machine stops in the configuration* $q_i x_1 \ldots x_n$ *if* $\varphi(q_i, x_1)$ *is not defined);*
5. *the calculation performed by* M *continues indefinitely if the input is the word* $\mu \in X^*$ *and* $f(\mu)$ *is undefined (in other words, where* $\mu \in X^* \setminus U$*).*

If these properties are satisfied we say that f *is T-calculable.*

At first sight it may seem difficult to imagine performing numeric calculations using Turing machines. However, they are perfectly capable of dealing with functions defined over natural numbers. We will use the unary representation of natural numbers.

Definition 13.6 *A number* $x \in \mathbb{N}$ *has a* unary representation *of* 1^{x+1}, *where* 1^{x+1} *is interpreted as the concatenations of* $x + 1$ *consecutive* 1 *symbols. Thus, the unary representation of* 0 *is* 1, *that of* 1 *is* 11, *and that of* 2 *is* 111, *etc. We will use* $\overline{x}$ *to denote the unary representation of an integer* x.

Example 13.7 The successor function. *It is rather straightforward to construct a Turing machine that calculates the successor function* s, *defined as follows:* $s(x) = x+1$. *The tape alphabet is* $X = \{1, \mathrm{B}\}$, *the state alphabet is* $Q = \{q_0, q_1, q_2\}$, $Q_f = \{q_2\}$, $U = \{\mathrm{B1B}, \mathrm{B11B}, \mathrm{B111B}, \ldots\}$, *and the state transition function* φ *is shown in Figure 13.6. Note that the tape will contain a number in unary representation preceded by a single blank. All other cells in the tape will also be blank.*

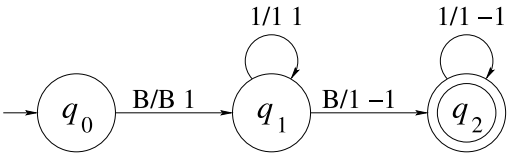


Fig. 13.6. The successor function.

The pointer initially encounters a blank cell; it changes state and moves to the right until it encounters another blank. This blank is replaced by a 1 *and the pointer starts moving to the left until it returns to the initial blank cell. At this point, computation halts, since* $\varphi(q_2, \mathrm{B})$ *is not defined.*

Example 13.8 The zero function. *We consider constructing a machine that implements the zero function* z*, defined as* $z(x) = 0$*. We must erase all the* 1 *symbols except the first, and then return to the initial blank cell. The tape alphabet will be the same as in the preceding example and the state alphabet will be* $Q = \{q_0, q_1, q_2, q_3, q_4\}$*. The initial configuration of the tape is* $q_0\mathrm{B}\overline{x}\mathrm{B}$*, and the final configuration will be* $q_f\mathrm{B1B}$ *(here* $q_f = q_4$*). The function* φ *is shown in Figure 13.7.*

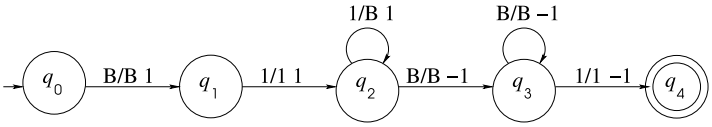


Fig. 13.7. The zero function.

Example 13.9 Addition. *We will now construct a Turing machine that performs addition. The tape will contain the entries* $\mathrm{B}\overline{x}\mathrm{B}\overline{y}\mathrm{B}$*, where* x *and* y *are the two numbers to be added (in their unary representation). The machine will replace the blank symbol between the two numbers with a* 1*, and then erase the final two* 1 *symbols. Thus, the final configuration will be* $q_f\mathrm{B}\overline{x+y}\mathrm{B}$*, where* $q_f = q_5$*. The state alphabet is* $Q = \{q_i : i = 0, \ldots, 5\}$*, with the tape alphabet remaining the same as in the previous examples. The function* φ *is shown in Figure 13.8.*

Example 13.10 Projection functions. *We construct one final machine for a type of function that will be important later: projection functions. We define the projection function* ${p_i}^{(n)}$ *as follows:*

$$ {p_i}^{(n)}(x_1, x_2, \ldots, x_n) = x_i, \quad 1 \leq i \leq n. $$

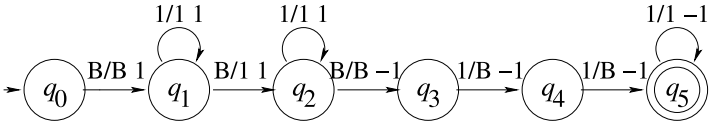


Fig. 13.8. The addition function.

In order to implement this function on a Turing machine we want to erase the first $i-1$ numbers on the tape, preserve the ith number, and erase the $n-i$ remaining numbers. The tape alphabet remains the same as before, while the state alphabet is $\{q_i : i = 0, \ldots, n+2\}$. The function φ is shown in Figure 13.9. Note that the tape will have an initial configuration of $q_0 \mathrm{B}\overline{x_1}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}$ and a final configuration of $q_f \mathrm{B}\overline{x_i}\mathrm{B}$.

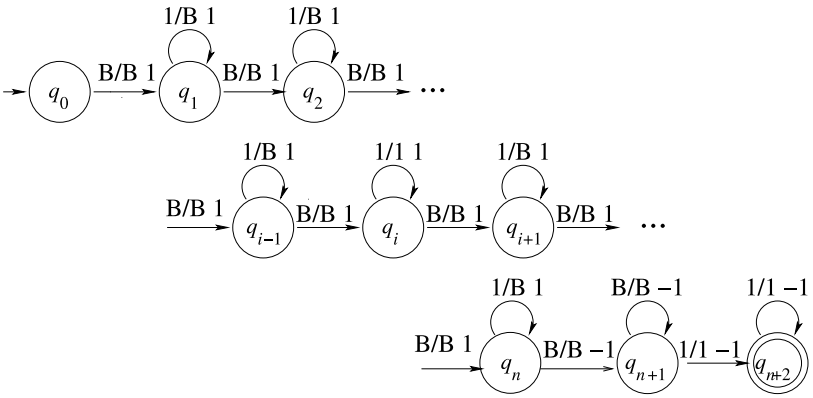


Fig. 13.9. The projection function.

Figure 13.9 shows the steps taken by the machine. After the initial state, the first $i-1$ states direct the machine to erase the first $i-1$ numbers, replacing them with blanks. The machine finally reaches state q_i, which instructs it to skip the ith number without changing it. States q_{i+1} through q_n instruct the machine to erase the remaining numbers, while state q_{n+1} returns the machine to the right of the ith number. Finally, state q_{n+2} ensures that the pointer returns to the blank cell preceding the ith number, where it will halt, since $\varphi(q_{n+2}, B)$ is undefined. Note that the machine does not return to the initial cell, that is, the leftmost cell of the half-infinite tape.

We could have added additional instructions directing the machine to translate the ith number back to the beginning of the tape, preceded by a single blank (see Exercise 3) and to halt with the pointer at the initial cell with the result immediately to its right, as in our other examples. However, this is not strictly necessary based on the definition of calculable functions (Definition 13.5).

13.3.2 Primitive Recursive Functions and Recursive Functions

The previous section showed that there exist numeric functions that are calculable using a Turing machine. This leads to the more general question of exactly what functions are T-calculable. The primitive recursive functions and recursive functions we discuss in this section are examples of such functions.

Before discussing primitive recursive functions we need a few preliminary definitions. In all this chapter we will have

$$\mathbb{N} = \{0, 1, 2, \dots\}.$$

Definition 13.11 *An arithmetic function is a function of the form*

$$f : \mathbb{N} \times \mathbb{N} \times \cdots \times \mathbb{N} \to \mathbb{N}$$

Example 13.12 *The successor function*

$$s : \mathbb{N} \to \mathbb{N}, \qquad x \mapsto x + 1,$$

and projection function

$$p_i^{(n)} : \mathbb{N} \times \mathbb{N} \times \cdots \times \mathbb{N} \to \mathbb{N}, \qquad (x_1, x_2, \dots, x_n) \mapsto x_i,$$

are examples of arithmetic functions.

We can represent a function $f : \mathrm{X} \to \mathrm{Y}$ using the pairs of all its inputs and corresponding outputs, as a subset of $\mathrm{X} \times \mathrm{Y}$. Thus, $(x, y) \in f$ is equivalent to saying that $y = f(x)$.

Definition 13.13 *A function* $f : \mathrm{X} \to \mathrm{Y}$ *is called a* total function *if it satisfies the following two conditions:*

1. $\forall x \in X$, $\exists y \in Y$ *such that* $(x, y) \in f$;
2. *if* $(x, y_1) \in f$ *and* $(x, y_2) \in f$, *then* $y_1 = y_2$.

This definition is the one that is usually used for a function whose domain is X. However, we have formalized it here to allow us to distinguish between total functions and partial functions, which will be defined a little later.

The primitive recursive functions are generated from the following base functions.

Base primitive recursive functions:

1. the successor function s: $s(x) = x + 1$;
2. the zero function z: $z(x) = 0$;
3. the projection functions ${p_i}^{(n)}$: ${p_i}^{(n)}(x_1, x_2, \dots, x_n) = x_i, 1 \le i \le n$.

Note in particular that the identity function is a base function, since it is equal to the projection function $p_1^{(1)}$.

Primitive recursive functions are constructed using two operations that may be iterated, starting from the base functions listed above. As will be shown later, these operations (*composition* and *recurrence*) preserve the T-calculability of the starting functions.

Definition 13.14 *Let $g_1, g_2, \ldots, g_k$ be arithmetic functions in n variables, and let h be an arithmetic function in k variables. Let f be the function defined by*

$$f(x_1, x_2, \ldots, x_n) = h(g_1(x_1, x_2, \ldots, x_n), \ldots, g_k(x_1, x_2, \ldots, x_n)).$$

The function f is called the composition *of h with $g_1, g_2, \ldots, g_k$, denoted by $f = h \circ (g_1, g_2, \ldots, g_k)$.*

Example 13.15 *Let $h(x_1, x_2) = s(x_1) + x_2$, $g_1(x) = x^3$ and $g_2(x) = x^2 + 9$. Define $f(x) = h \circ (g_1, g_2)(x)$ for $x \geq 0$. Then the composite function f simplifies to*

$$f(x) = x^3 + x^2 + 10.$$

Example 13.16 The constant functions. *Let $c_n(x) = n$ be the constant function taking the value n. It is primitive recursive. Indeed the function $c_1(x) = 1$ is defined as $c_1(x) = s \circ z(x)$. If c_n has been shown to be primitive recursive, then $c_{n+1} = s \circ c_n$ is primitive recursive.*

We are now ready to define the operation of recurrence.

Definition 13.17 *Let g and h be total arithmetic functions of n and $n + 2$ variables respectively. Define the function f of $n + 1$ variables as follows:*

1. $f(x_1, x_2, \ldots, x_n, 0) = g(x_1, x_2, \ldots, x_n)$;
2. $f(x_1, x_2, \ldots, x_n, y + 1) = h(x_1, x_2, \ldots, x_n, y, f(x_1, x_2, \ldots, x_n, y))$.

We say that f has been constructed by recurrence *with base g and step h. We allow $n = 0$, with the convention that a function g of zero variables is a constant.*

We now have the necessary tools to define primitive recursive functions.

Definition 13.18 *A function is called* primitive recursive *if it may be constructed using the successor function, the zero function, the projection functions, and through a finite number of composition and recurrence operations.*

Example 13.19 The addition function. *We can define addition,* $\text{add}(m,n) = m + n$*, using the successor function, two projection functions and a recurrence operation with base* $g(x) = p_1^{(1)}(x) = x$ *and step* $h(x,y,z) = s \circ p_3^{(3)}(x,y,z) = s(p_3^{(3)}(x,y,z)) = s(z)$*:*

$$\begin{cases} \text{add}(m,0) = g(m) = m, \\ \text{add}(m,n+1) = h(m,n,\text{add}(m,n)) = s(\text{add}(m,n)). \end{cases}$$

Example 13.20 The multiplication function. *Using the addition function we just defined, we can define multiplication using the recurrence operation with base* $g(x) = 0$ *and step* $h(x,y,z) = \text{add}(p_1^{(3)}(x,y,z), p_3^{(3)}(x,y,z)) = \text{add}(x,z)$*:*

$$\begin{cases} \text{mult}(m,0) = g(m) = 0, \\ \text{mult}(m,n+1) = h(m,n,\text{mult}(m,n)) = \text{add}(m,\text{mult}(m,n)). \end{cases}$$

Example 13.21 The exponential function. *In a similar manner we can define the exponential function* $\exp(m,n) = m^n$*, by taking* $g(x) = 1$ *and* $h(x,y,z) = \text{mult}(x,z)$*:*

$$\begin{cases} \exp(m,0) = 1, \\ \exp(m,n+1) = \text{mult}(m,\exp(m,n)). \end{cases}$$

Note that we have dropped the projection functions, in an effort to make the notation a little lighter and more readable.

Example 13.22 *To define the addition function* $\text{add}(m,n+1)$ *we used the successor function. To define multiplication* $\text{mult}(m,n+1)$ *we used* $\text{add}(\ldots)$ *and to define* $\exp(m,n+1)$ *we used* $\text{mult}(\ldots)$*. Continuing this process, the next function in the chain is the* power tower *or* tetration *function. Let* $\text{add}(m,n) = f_1(m,n)$*,* $\text{mult}(m,n) = f_2(m,n)$*, and* $\exp(m,n) = f_3(m,n)$*. We define* f_4 *by*

$$\begin{cases} f_4(m,0) = 1 \\ f_4(m,n+1) = f_3(m, f_4(m,n)). \end{cases}$$

Thus we have that

$$f_4(m,n) = \underbrace{m^{m^{m^{\cdots^m}}}}_{n \text{ times}}.$$

Similarly, we can continue this process by defining $f_i(m,n)$ *as*

$$\begin{cases} f_i(m,0) = 1, \\ f_i(m,n+1) = f_{i-1}(m, f_i(m,n)), \end{cases}$$

for $i > 4$. This generates the sequence of hyperoperators, each one a function that grows unimaginably faster than the previous one. (Exercise: what are the functions g, h used to define f_i according to Definition 13.17?)

Example 13.23 *The factorial function is a primitive recursive function. We define the factorial function as*

$$\begin{cases} \text{fact}(0) = 1, \\ \text{fact}(n+1) = \text{mult}(n+1, \text{fact}(n)). \end{cases}$$

After having seen that addition is a primitive recursive function, it is natural to ask whether subtraction is as well. However, our normal notion of subtraction is not a total function. In fact, if we define $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that $f(x, y) = x - y$, we observe that among others, $f(3, 5)$ is not defined. Thus we have to define another type of subtraction in order to have a total function on $\mathbb{N} \times \mathbb{N}$. We will call this function *proper subtraction*.

Definition 13.24

$$\begin{cases} \text{sub}(x, y) = x - y & \text{if} \quad x \geq y, \\ \text{sub}(x, y) = 0 & \text{if} \quad x < y. \end{cases}$$

Example 13.25 *Proper subtraction is a primitive recursive function. Showing this requires two steps. We start by showing that the predecessor function is a primitive recursive function and then we construct the proper subtraction function from it.*

Definition 13.26 *The predecessor function is defined by the recurrence*

$$\begin{cases} \text{pred}(0) = 0, \\ \text{pred}(y+1) = y. \end{cases}$$

As with addition, we can now construct the proper subtraction function using the operations of recurrence and composition:

$$\begin{cases} \text{sub}(m, 0) = m, \\ \text{sub}(m, n+1) = \text{pred}(\text{sub}(m, n)). \end{cases}$$

Primitive recursive functions also allow us to construct Boolean operators, which are necessary for constructing logical propositions. The three basic operators are NOT ($\neg$), AND ($\wedge$), and OR ($\vee$) (see also Section 15.7 of Chapter 15). Before we can do this we must first define the functions *sgn* and *cosgn*, which correspond to the "sign" of a natural number. These functions are primitive recursive (see Exercise 11):

1.
$$\begin{cases} \text{sgn}(0) = 0 \\ \text{sgn}(y+1) = 1; \end{cases}$$

2.
$$\begin{cases} \text{cosgn}(0) = 1 \\ \text{cosgn}(y+1) = 0. \end{cases}$$

Definition 13.27 *An n variable* predicate*, or an* open proposition*, is a proposition that will take a value of true or false depending on the values assigned to its variables $x_1, \ldots, x_n$. We will use $P(x_1, \ldots, x_n)$ to denote such a predicate.*

Example 13.28 *Let $P_1(x, y)$, $P_2(x, y)$, and $P_3(x, y)$ be respectively the three statements $x < y$, $x > y$, and $x = y$, respectively. Then P_1, P_2 are P_3 binary predicates.*

Once evaluated, a predicate can return the truth value of TRUE or FALSE. Since we are interested in working with numeric values, we will associate the number 1 with the value TRUE, and the number 0 with the value FALSE.

Definition 13.29 *Let P be a predicate on n variables. Its value function, which we denote by $|P|$, is the function that given numbers $x_1, \ldots, x_n$ returns the truth value of $P(x_1, \ldots, x_n)$ in $\{0, 1\}$.*

We can now define the value functions of the binary predicates from the previous example as primitive recursive functions which we call $\text{lt}(x, y)$, $\text{gt}(x, y)$, and $\text{eq}(x, y)$:

$$\begin{array}{lcll} |x < y| & = & \text{lt}(x, y) & = \text{sgn}(\text{sub}(y, x)) \\ |x > y| & = & \text{gt}(x, y) & = \text{sgn}(\text{sub}(x, y)) \\ |x = y| & = & \text{eq}(x, y) & = \text{cosgn}(\text{lt}(x, y) + \text{gt}(x, y)), \end{array} \tag{13.1}$$

where, by an abuse of notation, we have written $\text{lt}(x, y) + \text{gt}(x, y)$ to represent $\text{add}(\text{lt}(x, y), \text{gt}(x, y))$.

We are now ready to define the Boolean operators. Let P_1 and P_2 be two predicates such that $|P_1| = p_1$ and $|P_2| = p_2$. The following equations define the Boolean operators using the functions *sgn* and *cosgn* and other known primitive recursive functions

$$\begin{aligned} |\neg P_1| &= \text{cosgn}(p_1), \\ |P_1 \vee P_2| &= \text{sgn}(p_1 + p_2), \\ |P_1 \wedge P_2| &= p_1 * p_2, \end{aligned}$$

where by another abuse of notation, we have written $p_1 * p_2$ for $\text{mult}(p_1, p_2)$. In Exercise 6, the reader is asked to verify that these three functions do in fact correspond to the Boolean operators.

Definition 13.30 *A predicate is called* primitive recursive *if its value function is a primitive recursive function.*

Example 13.31 *The predicates $x < y$, $x > y$, and $x = y$ from Example 13.28 are primitive recursive. In fact, we have already constructed their value functions as compositions of primitive recursive functions.*

Now that we have introduced primitive recursive functions, we can make the link between them and Turing machines.

Theorem 13.32 *All primitive recursive functions are T-calculable.*

PROOF: Since we have already constructed Turing machines that calculate the successor, zero, and projection functions, it remains only to show that the set of T-calculable functions is closed under the operations of composition and recurrence.

We start by showing closure under composition. Let

$$f(x_1, \ldots, x_n) = h \circ (g_1(x_1, \ldots, x_n), \ldots, g_k(x_1, \ldots, x_n)),$$

where $g_i, i = 1, \ldots, k$, and h are total arithmetic functions that are T-calculable. We use H and G_i to denote the Turing machines that are capable of calculating the functions h and g_i, respectively. We will use these Turing machines to construct a Turing machine that is able to calculate the function $f(x_1, \ldots, x_n)$.

1. The calculation of $f(x_1, \ldots, x_n)$ begins with initial tape configuration of

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}.$$

2. We construct a copy of the information on the tape immediately to its right, such that the tape now reads

$$\underbrace{\mathrm{B}\overline{x_1}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}}\underbrace{\overline{x_1}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}}.$$

   (The Turing machine that performs this copying is constructed in Exercise 2.)
3. We use machine G_1 to obtain

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{g_1(x_1, \ldots, x_n)}\mathrm{B}.$$

   We can now copy $\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}$ to the end of the tape to obtain the configuration

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{g_1(x_1, \ldots, x_n)}\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}.$$

   It is now possible to use G_2 on the last n numbers. We will do these steps k times, yielding the configuration

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{g_1(x_1, \ldots, x_n)}\mathrm{B}\ldots\mathrm{B}\overline{g_k(x_1, \ldots, x_n)}\mathrm{B}.$$

4. We now erase the first n numbers by replacing them with blanks and we translate the remaining numbers to the left (as shown in Exercise 3), yielding the configuration

$$\mathrm{B}\overline{g_1(x_1,\ldots,x_n)}\mathrm{B}\ldots\mathrm{B}\overline{g_k(x_1,\ldots,x_n)}\mathrm{B}.$$

5. Machine H is used to perform the final operation, yielding a final configuration of

$$\mathrm{B}\overline{h(y_1,\ldots,y_k)}\mathrm{B},$$

where $y_i = g_i(x_1,\ldots,x_n)$, which is equivalent to the desired final configuration of

$$\mathrm{B}\overline{f(x_1,\ldots,x_n)}\mathrm{B}.$$

We now show closure under recurrence. Let g and h be T-calculable arithmetic functions and let f be the function

$$\begin{cases} f(x_1,\ldots,x_n,0) = g(x_1,\ldots,x_n), \\ f(x_1,\ldots,x_n,y+1) = h(x_1,\ldots,x_n,y,f(x_1,\ldots,x_n,y)), \end{cases}$$

defined using recurrence with base g and step h. Let G and H be the Turing machines calculating g and h respectively.

1. The calculation of $f(x_1,\ldots,x_n,y)$ starts with an initial tape configuration of

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}.$$

2. A counter with an initial value of zero is placed to the right of the above configuration. This counter is used to keep track of the recursive variable during the calculation. The numbers $x_1,\ldots,x_n$ are repeated to the right of the counter, producing a configuration of

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}\overline{0}\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}.$$

3. Machine G is used to calculate g on the last n values of the tape, producing a configuration of

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}\overline{0}\mathrm{B}\overline{g(x_1,\ldots,x_n)}\mathrm{B}.$$

Note that the last value on the tape, $g(x_1,\ldots,x_n)$, corresponds to $f(x_1,\ldots,x_n,0)$.

4. The tape is now in the configuration

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}\overline{i}\mathrm{B}\overline{f(x_1,\ldots,x_n,i)}\mathrm{B},$$

where $i = 0$. The operation performed at this point will simply be iterated for other values of i, so we describe the general case.

5. If $i < y$ (equivalently if $\mathrm{lt}(i, y) = 1$), then the machine makes a copy of the variables and the counter i found to the left of $f(x_1, \ldots, x_n, i)$. (Exercise 10 shows how to build a Turing machine that calculates $\mathrm{lt}(i, y)$. Thus, it is possible to build a Turing machine that places itself in one state if $\mathrm{lt}(i, y) = 1$ and in another state if not.) The tape now has the configuration

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}\overline{i}\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{i}\mathrm{B}\overline{f(x_1, \ldots, x_n, i)}\mathrm{B}.$$

The successor function is applied to the counter, yielding the configuration

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}\overline{i+1}\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{i}\mathrm{B}\overline{f(x_1, \ldots, x_n, i)}\mathrm{B}.$$

Machine H is applied to the last $n + 2$ variables of the tape, producing

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{y}\mathrm{B}\overline{i+1}\mathrm{B}\overline{h(x_1, \ldots, x_n, i, f(x_1, \ldots, x_n, i))}\mathrm{B}.$$

Note that $h(x_1, \ldots, x_n, i, f(x_1, \ldots, x_n, i)) = f(x_1, \ldots, x_n, i + 1)$. If the counter is such that $i = y$ (equivalently $\mathrm{lt}(i, y) = 0$), then the calculation is completed by erasing the first $n+2$ numbers on the tape. Otherwise the calculation continues by returning to step 5. □

It is natural to ask whether all T-calculable functions are primitive recursive functions. As it turns out, the answer is no. This is demonstrated in the following theorem and example.

Theorem 13.33 *The set of primitive recursive functions is a proper subset of the set of T-calculable functions. In other words, there exists a function f that is T-calculable but that is not primitive recursive.*

Example 13.34 *The Ackermann function A, defined as*

1. $A(0, y) = y + 1$,
2. $A(x + 1, 0) = A(x, 1)$,
3. $A(x + 1, y + 1) = A(x, A(x + 1, y))$,

is T-calculable but is not primitive recursive. The Ackermann function has the property that it "grows faster" than all primitive recursive functions, which explains why it is fascinating. But since it grows faster than all primitive recursive functions, it cannot actually be one. The proof of this fact is difficult and will not be presented here. However, you may find it, for instance, in [8].

To define a new family of functions that contains the primitive recursive functions we will make use of Boolean and relational operators. They permit us to define a new operation: *minimization*.

Definition 13.35 *Let P be a predicate on $n+1$ variables and $p = |P|$ its associated value function. The expression $\mu z[p(x_1, \ldots, x_n, z)]$ represents the smallest natural number z, if it exists, such that $p(x_1, \ldots, x_n, z) = 1$. Otherwise, it is undefined. In other words, z is the smallest natural number such that $P(x_1, \ldots, x_n, z)$ is true. This construction is called the* minimization *of p, and μ is the* minimization operator.

An $(n+1)$-variable predicate allows us to define an n variable function f,

$$f(x_1, \ldots, x_n) = \mu z[p(x_1, \ldots, x_n, z)],$$

whose domain is the set of $(x_1, \ldots, x_n)$ for which there exists a z such that $P(x_1, \ldots, x_n, z)$ is true.

Example 13.36 *We consider the "function"*

$$f: \mathbb{N} \to \mathbb{N},$$
$$x \mapsto \sqrt{x}.$$

As such, this is not a function in the usual sense, but if we look at Definition 13.5 we could imagine creating a Turing machine that calculates the square roots of perfect squares and that otherwise does not stop:

$$f : \{0, 1, 4, 9, \ldots\} = U \to \mathbb{N}.$$

In this example it is relatively easy to identify the domain U, but this is not always the case. Thus, Definition 13.37 introduces the notion of partial functions. *The function f can be written with the minimization operator μ as*

$$f(x) = \mu z[\mathrm{eq}(x, z * z)].$$

This function can be imagined as a type of search procedure. Starting at $z = 0$, we verify whether the equality is satisfied. If this is the case, then the appropriate value z has been found. If not, then we increment z and continue the search. For values of x that are not perfect squares, equality will never be attained; thus the calculation will continue indefinitely.

Definition 13.37 *A* partial function *$f : X \to Y$ is a subset of $X \times Y$ such that if $(x, y_1) \in f$ and $(x, y_2) \in F$, then $y_1 = y_2$. We say that f is defined for x if there exists $y \in Y$ such that $(x, y) \in f$. Otherwise, we say that f is undefined for x.*

We can be certain that the function f of Example 13.36 is not a primitive recursive function because all functions of this type are total functions. This shows that even if the value function p of a predicate is primitive recursive, the function constructed with the minimization of p is not necessarily primitive recursive. Such functions are part of the set of *recursive functions*, which is defined below.

Definition 13.38 *The families of* recursive functions *and* recursive predicates *are defined as follows:*

1. *The successor, zero, and projection functions are recursive.*
2. *Let $g_1, g_2, \ldots, g_k$, and h be recursive functions. Let f be the composition of h with $g_1, g_2, \ldots, g_k$. Then f is a recursive function.*
3. *Let g and h be two recursive functions. Let f be defined by the recurrence with base g and step h. Then f is a recursive function.*
4. *A predicate is called recursive if its value function is recursive. Similarly, it is called total if its value function is a total function.*
5. *Let P be a total recursive predicate over $n+1$ variables. The function f obtained by the minimization of $|P|$ is a recursive function.*
6. *A function is recursive if it can be constructed using a finite number of composition, recurrence, and minimization operations, starting from the successor, zero, and projection functions.*

The first three items in the above definition imply that all primitive recursive functions are themselves recursive. Example 13.36 shows formally that the set of primitive recursive functions is a proper subset of the set of recursive functions. We state without proof the following result.

Proposition 13.39 *The Ackermann function defined in Example 13.34 is a recursive function.*

Theorem 13.40 *All recursive functions are T-calculable.*

Proof: We have already shown that the successor, zero, and projection functions are T-calculable. Moreover, Theorem 13.32 has already shown the closure of T-calculability with respect to the operations of composition and recurrence. It remains to show that the set of T-calculable functions contains the minimization of recursive predicates.

Let $f(x_1, \ldots, x_n) = \mu z[p(x_1, \ldots, x_n, z)]$, where $p(x_1, \ldots, x_n, z)$ is the value function of a total T-calculable predicate, calculated with Turing machine Π.

1. The tape has an initial configuration of

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}.$$

2. We append the number 0 to the right end of the tape, obtaining

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{0}\mathrm{B}.$$

   We call this value the *minimization index*, denoted by j.
3. We duplicate the entries of the tape, appending them to the right end of the tape, resulting in the following configuration:

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{j}\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{j}\mathrm{B}.$$

4. Machine Π is applied to the copies of the initial entries, yielding

$$\mathrm{B}\overline{x_1}\mathrm{B}\overline{x_2}\mathrm{B}\ldots\mathrm{B}\overline{x_n}\mathrm{B}\overline{j}\mathrm{B}\overline{p(x_1,\ldots,x_n,j)}\mathrm{B}.$$

5. If $p(x_1,\ldots,x_n,j)=1$, then $f(x_1,\ldots,x_n)=j$, and the rest of the entries are erased. If not, the value $p(x_1,\ldots,x_n,j)$ is erased and the minimization index j incremented using the successor function. The algorithm continues by returning to step 3.

If $f(x_1,\ldots,x_n)$ is defined, then the algorithm will eventually find the correct value. If it is not defined, then the machine will continue calculating indefinitely, as specified in Definition 13.5. □

This theorem shows that a large number of functions are calculable with Turing machines. In fact, the relationship between Turing machines and recursive functions is even tighter, as shown by the following theorem (which will not be proved here).

Theorem 13.41 *[6] A function is T-calculable if and only if it is recursive.*

We will now introduce Church's thesis, which makes the connection between our intuitive notion of "calculability" and T-calculability.

This thesis is stated in many forms, but all forms being proven equivalent, we have chosen to present the form that complements the previous theorem.

Church's thesis *A partial function is "calculable" if and only if it is recursive.*

Thus, if we accept this thesis, then all "calculable" functions are T-calculable. This leads to the following definition: a function is "calculable" if and only if there exists a Turing machine that can calculate it.

The problem with this thesis is that it is impossible to prove, since we have no formal definition of "calculable." It would be possible to disprove it, however, by finding a function that is calculable with a precise algorithm but for which no equivalent Turing machine exists. However, there is no rigorous definition of an "algorithm" either. It is interesting to note that all attempts to formalize the notion of an algorithm have validated Church's thesis; despite taking a variety of approaches, all such formalizations have led to equivalent definitions of T-calculability.

13.4 Turing Machines versus Insertion–Deletion Systems and the DNA Computer

We have seen Turing machines that can execute programs. Let us now construct similarly a "DNA computer." As with Turing machines, we start with a finite alphabet X of symbols. In DNA computers this alphabet is naturally

$$X=\{A,C,G,T\}.$$

Such a small alphabet may seem restrictive, but recall that ordinary computers use only the binary alphabet $\{0, 1\}$.

We can construct strands (or words) with the symbols of this alphabet, and we define X^* to be the set of finite strands that can be constructed by the method of Definition 13.3. In the case of DNA, X^* represents the set of all strands of DNA that could possibly be constructed using the four bases A, C, G, and T, including the "null" strand.

In a Turing machine the words are the entries on the tape. The Turing machine has a finite set of instructions that transform an entry on the tape into another entry on the tape.

Here the instructions will transform strands of DNA into other strands of DNA. The best-known model used in analyzing DNA computers is the *insertion–deletion* model. The idea is to use enzymes to perform two basic operations:

- the deletion operation: remove a prescribed substrand of DNA at a precise location;
- the insertion operation: insert a prescribed substrand of DNA at a precise location.

We now formalize this model by rigorously defining the operations of insertion and deletion, often called *production rules*.

Definition 13.42 *1. Insertion. If $x = x_1x_2$ is a portion of a word $z = v_1xv_2 \in X^*$, we may insert a sequence $u \in X^*$ between x_1 and x_2, yielding the word $w = v_1yv_2$, where $y = x_1ux_2$. We describe this operation using the following simplified notation:*

$$x \Longrightarrow_I y.$$

We say that y is derived *from x by the insertion production rule. (It is understood that x and y may be portions of larger words.) This rule is represented by a triplet $(x_1, u, x_2)_I$.*

2. Deletion. If $x = x_1ux_2$ is a portion of a word $z = v_1xv_2 \in X^*$, we may delete the sequence u, yielding the word $w = v_1yv_2$ where $y = x_1x_2$. We use the notation

$$x \Longrightarrow_D y$$

and say that y is derived from x by the deletion production rule. This rule is represented by a triplet $(x_1, u, x_2)_D$.

As such, each rule of insertion and deletion can be seen as an element of $(X^*)^3$.

We introduce the general notation $x \Longrightarrow y$ to say that y was derived from x using one of the production rules. If y was derived from x through the application of several production rules applied one after the other, we use the notation

$$x \Longrightarrow^* y.$$

Definition 13.43 *An insertion–deletion system is a* 3*-tuple*

$$\mathrm{ID} = (X, I, D)$$

where X *is an alphabet,* I *is the set of insertion rules, and* D *is the set of deletion rules. In the case of DNA, the alphabet* $X = \{A, G, T, C\}$ *is formed of the four bases. Both* I *and* D *are subsets of* $(X^*)^3$*.*

Theoretically, this model is very efficient. In fact, we will prove that any recursive problem is able to be calculated using insertion and deletion operations. However, it is often very difficult to find a practical algorithm that will solve a given problem using only insertions and deletions.

Theorem 13.44 *[3] For each Turing machine there exists an insertion–deletion system that executes the same program.*

Remark: This statement is rather vague and difficult to understand. Stating it formally would require introducing a number of difficult notions such as formal languages and grammars. In simple words, the theorem means that for each Turing machine (that we can identify to a program), we can construct an insertion–deletion system that executes the program, that is, the different instructions of the Turing machine. For a Turing machine, to carry out one operation, a tape input is needed as well as the state of the machine, and the position of the pointer.

We associate a DNA strand to each 3-tuple formed by a tape input, the state of the machine, and the position of the pointer. A portion of the strand corresponds to the tape input, another one contains the information on the state of the machine, and yet another stores the position of the pointer. The proof discussed below gives, for each instruction of the Turing machine, a set of insertions and deletions transforming the strand into a strand corresponding to the new 3-tuple. The corresponding sequence of insertions and deletions must transform the first portion of the chain so that it corresponds to the new tape input. It must also cut the portions containing the information on the old state and the old position of the pointer and replace them by new portions of strand corresponding to the new state and the new position of the pointer.

Sketch of proof of Theorem 13.44: We want to show that all of the actions performed on a tape by a Turing machine can also be performed on words by an insertion–deletion system. To do this, for each transition that may be performed on the tape by a Turing machine we will construct an insertion–deletion system that performs the same action on a word of symbols representing the input on the tape.

Let $M = (Q, X, \varphi)$ be a Turing machine. If $\varphi(q_i, x_i) = (q_j, x_j, c)$, we define $(q_i, x_i) \to (q_j, x_j, c)$, where $c \in \{-1, 0, 1\}$. We will consider each case $c = 0$, $c = 1$ and $c = -1$. We must verify that for each of these transition rules there exists an equivalent set of production rules of an insertion–deletion system $\mathrm{ID} = (N, I, D)$, where

$$N = X \cup Q \cup \{L, R, O\} \cup \{q_i' : q_i \in Q\}.$$

The role of the sets $\{q_i' : q_i \in Q\}$ and $\{L, R, O\}$ will be made clear in the proof. For each transition rule of the Turing machine, the goal is to construct a sequence of insertions and deletions that when applied in the *prescribed order*, have the same effect as the transition rule. However, a word of a warning: we must ensure that these insertions and deletions cannot occur in any order other than the prescribed one, thus producing a different result from the one prescribed by the Turing machine.

We will use many sequences of symbols throughout this proof. To help make things clear, keep in mind that $\mu, \mu_1, \mu_2, \nu, x_i, x_j, \rho, \sigma, \tau \in X$ and that $q_i, q_j \in Q$.

1. For each rule of the form $(q_i, x_i) \rightarrow (q_j, x_j, 0)$ we will add to ID the following three rules: $(q_i x_i, q_j O x_j, \nu)_I$, $(\mu, q_i x_i, q_j O x_j)_D$, and $(\rho\sigma q_j, O, x_j)_D$, for all $\mu, \nu, \rho, \sigma \in X$. In fact, for each character $\nu \in X$ we must add a rule to ID of the form $(q_i x_i, q_j O x_j, \nu)_I$, and likewise for the two other rules. Since X is finite, we will therefore add only a finite number of rules to ID.
   Thus, if we process a word of the form $\mu q_i x_i \nu$, we will perform the following sequence of operations:

$$\mu q_i x_i \nu \Longrightarrow_I \mu q_i x_i q_j O x_j \nu \Longrightarrow_D \mu q_j O x_j \nu \Longrightarrow_D \mu q_j x_j \nu.$$

   To begin with, we had the word $\mu q_i x_i \nu$. First, we inserted $q_j O x_j$ between $q_i x_i$ and ν. This operation was followed by two deletions: the first removed $q_i x_i$, while the second removed the remaining O between q_j and x_j. The final result is $\mu q_j x_j \nu$, which is exactly the word we wanted. Recall that the symbol q_j represents the pointer state, and is found immediately before the symbol being pointed to. Thus, the previous operations have allowed us to proceed from configuration $\mu \underline{x_i} \nu$ in state q_i to configuration $\mu \underline{x_j} \nu$ in state q_j.
   Why did we use this O symbol? Could we not just have executed

$$\mu q_i x_i \nu \Longrightarrow_I \mu q_i x_i q_j x_j \nu \Longrightarrow_D \mu q_j x_j \nu?$$

   The next transition to be executed is $(q_j, x_j) \rightarrow (q_k, x_k, c)$. We need to ensure that the system does not start this operation before finishing the present one. That is, $q_i x_i$ needs to be erased before $q_j x_j$ is modified. The presence of the O between q_j and x_j ensures that the pattern $q_j x_j$ cannot be matched until after the O is removed.
2. For each rule of the form $(q_i, x_i) \rightarrow (q_j, x_j, 1)$ we will add to ID the following six rules: $(q_i x_i, q_i' O x_j, \nu)_I$, $(\mu, q_i x_i, q_i' O x_j)_D$, $(\rho\sigma q_i', O, x_j)_D$, $(q_i' x_j, q_j R, \nu)_I$, $(\mu, q_i', x_j q_j R)_D$, and $(\tau x_j q_j, R, \nu)_D$, for all $\mu, \nu, \rho, \sigma, \tau$ in X.
   Thus, if we process a word of the form $\mu q_i x_i \nu$, we will perform the following sequence of operations:

$$\begin{aligned}\mu q_i x_i \nu &\Longrightarrow_I \mu q_i x_i q_i' O x_j \nu \Longrightarrow_D \mu q_i' O x_j \nu \Longrightarrow_D \mu q_i' x_j \nu \\ &\Longrightarrow_I \mu q_i' x_j q_j R \nu \Longrightarrow_D \mu x_j q_j R \nu \Longrightarrow_D \mu x_j q_j \nu.\end{aligned}$$

   Here we see that the first three operations simply repeat those that were performed for the rule $(q_i, x_i) \rightarrow (q_j, x_j, 0)$. These three productions, one insertion and two

deletions, allow us to exchange x_i for x_j without moving the pointer. We use an artificial state q_i' to signify that the operation is not yet finished, thus preventing other transition rules of the Turing machine from starting. The three following operations move the pointer to the right and finally replace q_i' with the actual state q_j. The machine is now ready to execute the command $(q_j, \nu) \to (q_k, x_k, c)$ with $c \in \{-1, 0, 1\}$, if such a command exists.
Here we again used artificial symbols (O, R, and q_i') to force production rules to be applied in the exact order we specify. For example, the rule $(\rho\sigma q_i', O, x_j)_D$ is used to ensure that we cannot remove the O from $\mu q_i x_i q_i' O x_j \nu$ before removing $q_i x_i$. In fact, in configuration $\mu q_i x_i q_i' O x_j \nu$, the artificial state q_i' is preceded by a unique symbol $x_i \in X$, itself preceded by a state symbol. This works because we can remove O only when the symbol q_i' is preceded by two symbols from X (one of which could be B). We let the reader convince himself (herself) of the necessity of the remaining production rules.

3. For each rule of the form $(q_i, x_i) \to (q_j, x_j, -1)$ we will add to ID the following six rules: $(q_i x_i, q_i' O x_j, \nu)_I$, $(\mu_2, q_i x_i, q_i' O x_j)_D$, $(\rho\sigma q_i', O, x_j)_D$, $(\mu_1, q_j L, \mu_2 q_i' x_j)_I$, $(q_j L \mu_2, q_i', x_j)_D$, and $(q_j, L, \mu_2 x_j)_D$ for all $\mu_1, \mu_2, \nu, \rho, \sigma \in X$.
Thus, if we process a word of the form $\mu_1\mu_2 q_i x_i \nu$, we will perform the following sequence of operations:

$$\mu_1\mu_2 q_i x_i \nu \Longrightarrow_I \mu_1\mu_2 q_i x_i q_i' O x_j \nu \Longrightarrow_D \mu_1\mu_2 q_i' O x_j \nu$$
$$\Longrightarrow_D \mu_1\mu_2 q_i' x_j \nu \Longrightarrow_I \mu_1 q_j L \mu_2 q_i' x_j \nu \Longrightarrow_D \mu_1 q_j L \mu_2 x_j \nu \Longrightarrow_D \mu_1 q_j \mu_2 x_j \nu.$$

Therefore, all of the commands $(q_j, \nu) \to (q_k, x_k, c)$ with $c \in \{-1, 0, 1\}$ may be performed by an insertion–deletion system. □

This theorem shows that an insertion–deletion system has at least the computational power of a Turing machine: all functions that can be calculated on a Turing machine can also be calculated on a DNA computer using insertions and deletions. This illustrates how powerful a DNA computer is in theory.

13.5 NP-Complete Problems

This section will be relatively light in theory, and concentrate more on examples.

NP-complete problems are a very important class of problems in computer science. These are problems that are simple to describe, often extremely important in their respective applications, yet difficult to solve using a computer. The precise definition of NP-completeness can be found in [6].

13.5.1 The Hamiltonian Path Problem

An example of an NP-complete problem is the Hamiltonian path problem, as discussed earlier in Section 13.2. Recall that the problem consists in finding a path through a

directed graph that passes through each node exactly once. It is easy to imagine real-world applications that need to solve such a problem, for example in the domain of transportation.

Looking at the simple example from Figure 13.1, the solution may be found easily by hand. In fact, the solution is to pass through the seven nodes in the following order: $0, 3, 5, 1, 2, 4, 6$. Finding the solution is even easier with a conventional computer: even with a rudimentary and inefficient algorithm the calculation takes only a fraction of a second.

What makes a problem "complex"? It is related to the amount of time necessary to find a solution as a function of the input size. For example, classic algorithms for solving the Hamiltonian path problem require time exponential in the number of nodes in the graph. Beyond a certain number of nodes it becomes effectively impossible for a computer to find the solution. Even with graphs containing only 100 nodes, modern computers require an inordinate amount of time to find a solution. This comes from the fact that conventional computers are sequential: each operation is performed one after the other. This is the reason why computer scientists are interested in parallelism.

We already mentioned that Adleman spent seven days in the laboratory to come up with the simple solution above. So what exactly is the advantage of a DNA computer? A DNA computer is able to perform billions of operations in parallel, and this ability is what fascinates researchers. The slowest steps in performing calculations with a DNA computer are those that must be performed by humans in a laboratory. With the method proposed by Adleman, the number of such steps is linear in the number of nodes in the graph. However, it should be remarked that Adleman's method will not scale particularly well for another reason. Although the number of steps to be performed in the laboratory is linear with the number of nodes, the number of potential paths through these nodes remains exponential. With a billion snippets of DNA in a test tube the millions of generated paths will cover all paths in a small graph with very high probability. However, when there are billions of possible paths to consider, it becomes very probable that not all of them will be generated. Other practical problems can occur in isolating such a small fraction of all generated DNA strands. Thus, much work remains to be done before DNA computers can have their parallelism fully exploited.

13.5.2 Satisfiability

Another example of a classic NP-complete problem is that of satisfiability. This problem can be efficiently solved using a DNA computer in a method similar to that used by Adleman for the Hamiltonian path problem. This shows that the general approach taken by Adleman is not completely specific to the Hamiltonian path problem.

The problem of satisfiability concerns itself with logical statements built using the Boolean operators $\vee$ (OR), $\wedge$ (AND), and $\neg$ (NOT) and the Boolean variables $x_1, \ldots, x_n$, which may each take a value of TRUE or FALSE. We consider two examples.

Example 13.45 *Consider the statement α, defined as*

$$\alpha = (x_1 \vee x_2) \wedge \neg x_3.$$

The value of α is the value of the logical statement when values for x_1, x_2, and x_3 have been substituted. As such, α will be either TRUE *or* FALSE*, depending on the values of the variables. For example, if x_1, x_2, and x_3 are all* TRUE*, then α is* FALSE*.*

The problem of satisfiability asks the following question: can we assign truth values to the variables x_1, x_2, and x_3 such that α is TRUE*? In this case, it is simple to see that we can. In fact, we could simply let $x_1 = x_2 =$ TRUE and $x_3 =$ FALSE. We say that we are able to verify the logical equation $\alpha =$ TRUE and that α is satisfiable.*

Example 13.46 *Now consider the logical statement*

$$\beta = (x_1 \vee x_2) \wedge (\neg x_1 \vee x_2) \wedge (\neg x_2).$$

In this case we can easily convince ourselves that there is no configuration of truth values for the variables such that $\beta =$ TRUE. Hence, β is not satisfiable.

Definition 13.47 *A logical statement built using the Boolean operators $\vee$ (*OR*), $\wedge$ (*AND*), and $\neg$ (*NOT*) and Boolean variables $x_1, \ldots, x_n$ is satisfiable if there exists an assignment of truth values to the variables such that the statement becomes true.*

Example 13.45 is simple to visualize, and equally simple for a computer to analyze, even using the most inefficient of algorithms. In fact, a computer may simply enumerate the 2^3 assignments of truth values (there are 3 variables, and each of them can take one of 2 values), and evaluate the statement for each of them. However, such an approach quickly breaks down when we are dealing with a large number of variables. With 100 variables there are already 2^{100} possible configurations to be tested. In the general case there is no known algorithm that is more efficient than the exhaustive approach.

This large search space is one of the reasons why DNA computers seem suitable for solving this problem. In fact, like any algorithm based on an "exhaustive search" (which requires a computer to check all possible solutions, one after the other), this algorithm benefits greatly from the massive parallelism inherent in DNA computing. In effect, a DNA computer is able to test all solutions at the same time. The problem then becomes to extract the correct solution, if it exists.

To start, we need to find a method that will construct all possible solutions as strands of DNA. If we have 3 variables, we need a method that allows us to uniquely encode each of the $2^3 = 8$ possibilities. This is possible with the help of a little graph theory. We model each of the possible assignments of truth values as a maximal path through the graph shown in Figure 13.10. There is a bijection between the maximal paths in the graph and the sequences of truth value assignments to all variables. We denote FALSE by 0 and TRUE by 1. Node a_j^0 represents assigning a value of 0 to x_j, node a_j^1 represents assigning a value of 1 to x_j, and the nodes v_i are simply spacers. For example, the path

$a_1^0 v_1 a_2^0 v_2 a_3^0 v_3$ represents assigning FALSE to each of the three variables x_i. It is easy to see that all of the possible paths of length 5 (they are the maximal ones) enumerate exactly the 8 possible assignments of truth values.

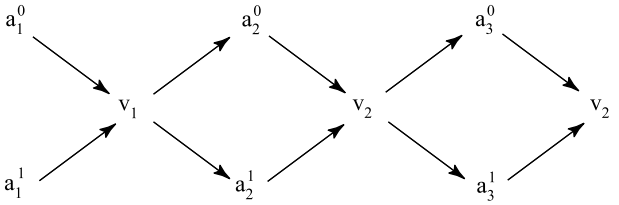


Fig. 13.10. Truth variable assignment graph for logical statement α or any logical statement with three variables.

This is useful because the first step of the DNA computing algorithm is to generate many copies of each possible path. To do this, we will use the same technique used by Adleman: encoding nodes as unique strands of DNA and directed edges as complementary strands that will join two nodes. More specifically, each node will be encoded by a strand of length $2N$, and each directed edge as the complement of the last N bases from the departure node followed by the complement of the first N bases from the destination node. The exact value of N depends on the size of the problem to be solved; it must be large enough that each node and edge can be uniquely labeled.

As before, we assemble a large quantity of each DNA strand encoding for nodes and directed edges. After a given amount of time, these strands of DNA will join to create longer strands representing the possible paths through the graph. With high probability, all of the possible paths will be enumerated. It remains to extract those paths that correspond to possible solutions of the logical statement, if such paths exist.

The first step is to transform the statement into *conjunctive normal form*, such that

$$\alpha = C_1 \wedge C_2 \wedge C_3 \wedge \cdots \wedge C_m,$$

where the C_i are logical statements using only $\vee$ and $\neg$. A theorem from logic ensures that such a transformation is always possible. It is done using the following rules:

1. For all x_1, x_2, x_3,
$$x_1 \wedge (x_2 \vee x_3) = (x_1 \wedge x_2) \vee (x_1 \wedge x_3).$$
2. For all x_1, x_2, x_3,
$$x_1 \vee (x_2 \wedge x_3) = (x_1 \vee x_2) \wedge (x_1 \vee x_3).$$
3. For all x_1, x_2,
$$\neg(x_1 \vee x_2) = \neg x_1 \wedge \neg x_2.$$

4. For all x_1, x_2,

$$\neg(x_1 \wedge x_2) = \neg x_1 \vee \neg x_2.$$

Note that although the conversion always exists, it is not always easy to find it. In fact, the known algorithms for converting to conjunctive normal form are quite complex and sometimes require a relatively long time to run. However, in many cases the logical statement is already given in the appropriate form or is easy to convert. The statement in Example 13.45 is already in conjunctive normal form using $C_1 = x_1 \vee x_2$ and $C_2 = \neg x_3$.

To satisfy a logical statement of the form $C_1 \wedge \cdots \wedge C_m$ we must satisfy C_1, **and** we must satisfy C_2, $\cdots$, **and** we must satisfy C_m.

The conversion to conjunctive normal form is used to guide the following procedure. We start by extracting all strands that satisfy statement C_1. In our case, $C_1 = x_1 \vee x_2$, so we want to extract all strands that encode x_1 or x_2 as 1.

This can be done by first extracting all solutions that encode x_1 as 1. To do this, we again borrow from Adleman's technique. We place in test tube A strands of DNA encoding the complement of edge $a_1^1 v_1$, each of these being attached to a small particle of iron. These attract all strands containing $a_1^1 v_1$, while the other chains remain in the solution. We then attract these using a magnet to the border of test tube A. We pour the rest of the solution in test tube B. We put back some liquid free of DNA in test tube A and separate the strands from the iron particles.

In order for $x_1 \vee x_2$ to be true, it could also be that $x_2 = 1$. Thus, we repeat the procedure on the remaining strands rejected in the first step and now in test tube B, this time selecting strands containing the directed edge $a_2^1 v_2$. The strands retained at this step are placed back into the solution of test tube A containing the strands selected in the previous step. We now have a single test tube containing all strands encoding for $x_1 = 1$ or $x_2 = 1$. So the remaining strands in test tube B can be discarded.

Thus, we now have all strands that satisfy statement C_1. We can now repeat the same procedure to extract all strands that satisfy statement C_2, with the surviving strands therefore satisfying both C_1 and C_2, hence satisfying $C_1 \wedge C_2$. In our example, $C_2 = \neg x_3$. Thus we need to extract all strands encoding $x_3 = 0$. In the general case this procedure is repeated for each C_i.

We may ask ourselves whether a DNA computer is required to solve a problem in conjunctive normal form, or whether other algorithms would be more efficient. Suppose that $\alpha = C_1 \wedge C_2 \wedge C_3 \wedge \cdots \wedge C_m$ and that all C_i are formed from n distinct variables x_j and their negations $\neg x_j$ (it could happen that not all variables appear in each C_i). Then there are 2^n paths in the graph. However, as we have seen, there are at most n verifications to do for each C^i, so at most a total of mn verifications. Hence the method is an improvement compared to the systematic exploration of all paths, unless m is very large compared to n.

13.6 More on DNA Computers

13.6.1 The Hamiltonian Path Problem and Insertion–Deletion Systems

Section 13.4 showed that all recursive functions can be calculated using a DNA computer, performing only insertion and deletion operations. However, Adleman's solution to the Hamiltonian path problem did not use any insertions or deletions. As discussed in the introduction, this is because theoretical algorithms and the best algorithms in practice are often far from each other. This is equally true for Turing machines. Consider the function $\text{add}(m, n) = m + n$. Being a primitive recursive function, the proof of Theorem 13.32 provides an algorithm on a Turing machine to calculate it. This algorithm is recursive, applying the successor function n times. However, the algorithm depicted in Figure 13.8 (constructed in Example 13.9) calculates it in a much simpler manner!

The DNA computer algorithm solving the Hamiltonian path problem using only insertions and deletions is no doubt much more complex than that presented by Adleman. However, given that there are so few algorithms conceived for DNA computers it is extremely hard to judge which biological operations will be used the most, provided that one day, the gap between theory and practicality is bridged.

13.6.2 Current Limits of DNA Computers

Up until now we have painted a rather rosy picture of DNA computers. We have shown how to use to DNA computers to solve a few difficult mathematical problems. Both of these algorithms have played off of the biggest strength of DNA computers, their massive parallelism, which lets us test effectively all possible configurations simultaneously instead of sequentially. Moreover, we have seen that DNA computers are also fully capable of computing anything that may be computed using Turing machines, and thus they are potentially very powerful.

However, one must keep in mind that all of our theoretical models have made one rather presumptuous hypothesis: that nature is ideal, and we can manipulate DNA strands with perfect precision. In reality, this is far from being the case. In fact, in nature it happens often that DNA strands in solution break (hydrolize) spontaneously. Similarly, there are often errors when two complementary strands unite. For example, the strand

$$AAGTACCA$$

with complement

$$TTCATGGT$$

could pair up with a "false complement" that matches very closely its true complement. Thus, we could find ourselves with the double strand

$$\begin{array}{cccccccc} A & A & \mathbf{G} & T & A & C & C & A \\ T & T & \mathbf{T} & A & T & G & G & T \end{array}$$

where a single G has been paired with a T instead of a C. Such an error could be a problem for any algorithm, such as that of Adleman, which relies implicitly on the perfect pairing of complements. Research is under way to counter these problems. Certain researchers have proposed performing the calculations inside of a living cell (*in vivo*) rather than simply in a solution. In fact, living cells already have several advanced control mechanisms for dealing with such errors.

It should be noted that the Hamiltonian path experiment that was performed by Adleman in 1994 was repeated (without success!) by Kaplan, Cecci and Libchaber in 1995. Their experiment produced poor results at the electrophoresis step. The location on the plate that should have contained only paths of length 7 contained many contaminants (paths with fewer or more than seven nodes). The gel used in the electrophoresis had many imperfections, but more importantly the strands of DNA where often folded over themselves too much and did not travel through the gel with the expected velocity. Adleman himself admitted repeating the electrophoresis step several times before obtaining satisfactory results.

Using Adleman's approach there is always a risk that the solution path will not actually be generated. Let us look at the graph of Figure 13.1. There are paths, called cycles, that have the same first and last node, for instance the path 12351. Nothing prevents the existence of an infinite path always repeating this loop. So the number of possible paths is infinite, while the quantity of DNA material in the test tube is finite. We must manage that the quantity of DNA in the test tube be sufficient to ensure, with very high probability, that all paths with length $\leq N$ are generated, where N is larger than the number of nodes. Of course, there is no 100% guarantee that they will all be present. It may happen that the solution, even if it exists, is not in the test tube.

Using this type of algorithm, if we have found a solution, then we know with certainty that it is a solution. On the other hand, if we do not find a solution, we are not completely certain that a solution does not exist. All that we are able to say is that there is a very high probability that no solution exists. Thus, such algorithms are inherently probabilistic.

There is also a problem with the theoretical model of insertion–deletion systems. We assumed that it is possible to perform an arbitrary insertion and an arbitrary deletion. Since these operations are actually performed by enzymes, we have implicitly assumed that there are effectively an infinite number of enzymes able to perform any insertions and deletions we desire, and moreover that we can place a large number of them together in a test tube, where they will work as intended without any interference. In reality, we have not yet mastered biochemistry to this point, and we do not have a great enough understanding of enzymes to be able to effectuate arbitrary insertions or deletions.

Can we program a DNA computer? Conventional computers are not built to perform a single calculation. Rather, we are able to program them explicitly, allowing them to run any number of algorithms. From what we have seen, it is tempting to assume that DNA computers are effectively impossible to program in such a manner. Indeed, the method used by Adleman is adapted to the special problem (or type of

problem) to be solved. But we have also seen that DNA computers are able to act as Turing machines. There exists a universal Turing machine [6], which when provided as input the instructions of another Turing machine M and also a problem instance ω is able to produce the same output as would have been produced by feeding ω to M. Such a Turing machine is therefore programmable, and thus DNA computers are similarly theoretically programmable.

The challenges that must be overcome in order to make this technology a reality are enormous. But the concepts are very seductive and attract a great deal of research.

13.6.3 A Few Biological Explanations Concerning Adleman's Experiment

Section 13.2 presented an overview of the method applied by Adleman in order to solve an instance of the Hamiltonian path problem. The algorithm consisted of five steps:

- select all paths that start at node 0 and finish at node 6;
- from among these, select all paths that have the desired length of seven nodes;
- from the remaining paths, select those that contain all nodes;
- test to see whether any paths (solutions) remain;
- analyze the solution(s) in order to determine the paths they encode.

We briefly discussed each of these steps, without delving too far into the chemistry. Here we will discuss in detail the method used by Adleman to perform the first step of the algorithm.

Adleman used a gene amplification technique known as PCR (*polymerase chain reaction*). The idea is to replicate only those chains containing the correct starting and finishing nodes until they completely dominate all others.

In this example we want to replicate the chains starting with node 0 (encoded by $AGTTAGCA$) and terminating with node 6 (encoded by $CCGAGCAA$). Looking at Figure 13.1, we observe that it is impossible to arrive at node 0 from any other node, and that it is similarly impossible to leave node 6. Thus, if node 0 is encoded in a strand it must be at the beginning of the strand. Similarly, the encoding for node 6 is always found at the end of a chain. Thus, a chain that satisfies both of these properties resembles

$$\begin{array}{cccccccccccccccc} & & & & T & C & G & T & \ldots & G & G & C & T & & & \\ & & & & | & | & | & | & \ldots & | & | & | & | & & & \\ A & G & T & T & A & G & C & A & \ldots & C & C & G & A & G & C & A & A \end{array}$$

Nature has devised a very powerful mechanism for replicating strands of DNA: DNA polymerase. This enzyme is able to complete the complementary strand of a paired DNA molecule, provided that the first few bases (a *primer*) are already in place.

Polymerase can work in only one direction. If we apply a primer, the reaction is able to proceed in only one direction from that primer. We will rewrite our chains using the notation of biochemists, marking one end of the strand as the $3'$ end, and the other

as the $5'$ end (the origin of this notation will be explained shortly). It is important to note that polymerase is able to proceed only in the $5'$ to $3'$ direction, and that a $5'-3'$ oriented chain is able to pair only with a $3'-5'$ oriented complement. Using this notation, a chain with the desired starting and ending nodes is of the form

$$\begin{array}{cccccccccccccccc} & & & & 3' & & & & & & & & 5' & & & \\ & & & & T & C & G & T & ... & G & G & C & T & & & \\ & & & & | & | & | & | & ... & | & | & | & | & & & \\ A & G & T & T & A & G & C & A & ... & C & C & G & A & G & C & A & A \\ 5' & & & & & & & & & & & & & & & & 3' \end{array}$$

How can we use polymerase to multiply the desired strands of DNA? The first step is to separate the double strands of DNA into single strands. This is done by simply heating the solution to an appropriate temperature. The double strands thus separate into two types of strands: "node strands" and "edge strands." For example, the double strand

$$\begin{array}{cccccccccccccccc} & & & & 3' & & & & & & & & 5' & & & \\ & & & & T & C & G & T & ... & G & G & C & T & & & \\ & & & & | & | & | & | & ... & | & | & | & | & & & \\ A & G & T & T & A & G & C & A & ... & C & C & G & A & G & C & A & A \\ 5' & & & & & & & & & & & & & & & & 3' \end{array}$$

produces an edge strand of

$$\begin{array}{cccccccccc} 3' & & & & & & & & & 5' \\ T & C & G & T & ... & G & G & C & T & \end{array}$$

and a node strand of

$$\begin{array}{cccccccccccccccc} 5' & & & & & & & & & & & & & & & 3' \\ A & G & T & T & A & G & C & A & ... & C & C & G & A & G & C & A & A \end{array}$$

Explaining the $5'-3'$ notation. Let us first consider a simple DNA strand. Its backbone is formed from a chain of sugars linked together. Each sugar contains five carbon atoms, numbered from $1'$ to $5'$. Each base is attached to one sugar. It is bonded to carbon $1'$ of its associated sugar, while a hydroxyl group (OH) is attached to carbon $3'$ on one side and a phosphate to carbon $5'$ on the other side. When two sugars corresponding to two neighboring bases connect, the hydroxyl group ($3'$) of one attaches to the phosphate ($5'$) of the other.

Thus if we imagine a strand of DNA, the sugar of its first base being connected by its hydroxyl group to the phosphate of a second base, then its own phosphate group is unattached and exposed. Consequently, this first base is labeled the $5'$ end of the molecule. Similarly, the last base in the strand has an exposed hydroxyl group, and is therefore labeled the $3'$ end. This DNA single strand therefore has a $5'-3'$ orientation.

A single strand with $5'-3'$ orientation can link only with one with $3'-5'$ orientation to form a double strand. The chains of sugar are located on the outer side of the double helix and form its backbone. The pairing of bases is by means of hydrogen bonds.

Replication in Adleman's experiment. Adleman introduced into the solution a large number of two types of primers. The first encodes the name of node 0 ($AGCA$) and is called the *node primer*. The second encodes the complement of the prename of node 6 ($GGCT$) and is called the *edge primer*. The primers will pair up with their complements. For example, the edge strand

$$\begin{array}{cccccccccc} 3' & & & & & & & & 5' \\ T & C & G & T & ... & G & G & C & T \end{array}$$

will pair up with the node primer to form the partial double strand

$$\begin{array}{cccccccccc} 3' & & & & & & & & 5' \\ T & C & G & T & ... & G & G & C & T \\ A & G & C & A & & & & & \\ 5' & & & 3' & & & & & \end{array}$$

Similarly, the node strand

$$\begin{array}{cccccccccccccccc} 5' & & & & & & & & & & & & & & & 3' \\ A & G & T & T & A & G & C & A & ... & C & C & G & A & G & C & A & A \end{array}$$

will pair up with the edge primer to form the partial double strand

$$\begin{array}{cccccccccccccccc} & & & & & & & & & 3' & & & 5' & & & \\ & & & & & & & & & G & G & C & T & & & \\ A & G & T & T & A & G & C & A & ... & C & C & G & A & G & C & A & A \\ 5' & & & & & & & & & & & & & & & 3' \end{array}$$

The DNA polymerase will attach itself to the $3'$ ends of the primers and fabricate the rest of the complement strand, leaving almost complete double strands. For that purpose it uses free bases that have been added to the solution. Thus, each original double strand starting with a 0 and ending with a 6 has now been replicated, doubling the number of such strands. This process can be repeated several times, and after a few repetitions these strands will dominate all others.

We will walk through an example that clearly demonstrates the entire replication process.

Example 13.48 *Consider a double strand consisting only of nodes* 0 *and* 6 *and the single edge connecting them:*

$$\begin{array}{cccccccccccccccc} & & & & 3' & & & & & & & & 5' & & & \\ & & & & T & C & G & T & G & G & C & T & & & & \\ A & G & T & T & A & G & C & A & C & C & G & A & G & C & A & A \\ 5' & & & & & & & & & & & & & & & 3' \end{array}$$

By heating this double strand we obtain the two single strands $TCGTGGCT$ *and* $AGTTAGCACCGAGCAA$*. The primers will attach themselves to these single strands, forming two partial double strands,*

$$\begin{array}{cccccccccccccccc} & & & & & & & & 3' & & & 5' & & & & \\ & & & & & & & & G & G & C & T & & & & \\ A & G & T & T & A & G & C & A & C & C & G & A & G & C & A & A \\ 5' & & & & & & & & & & & & & & & 3' \end{array}$$

and

$$\begin{array}{cccccccc} 3' & & & & & & & 5' \\ T & C & G & T & G & G & C & T \\ A & G & C & A & & & & \\ 5' & & & 3' & & & & \end{array}$$

DNA polymerase will attach itself to the $3'$ *ends of the primers and complete the replication, yielding the following two double strands:*

$$\begin{array}{cccccccccccccccc} 3' & & & & & & & & & & & 5' & & & & \\ T & C & A & A & T & C & G & T & G & G & C & T & & & & \\ A & G & T & T & A & G & C & A & C & C & G & A & G & C & A & A \\ 5' & & & & & & & & & & & & & & & 3' \end{array}$$

and

$$\begin{array}{cccccccc} 3' & & & & & & & 5' \\ T & C & G & T & G & G & C & T \\ A & G & C & A & C & C & G & A \\ 5' & & & & & & & 3' \end{array}$$

The solution is again heated and cooled, separating the newly formed double strands into single strands. Notice that from the two initial strands, one cycle of this process leaves us with four strands with the same properties as the originals, the edge strand being slightly longer than initially, while the node strand is slightly shorter.

Consider now a double strand encoding for nodes 0 *and* 1*:*

$$\begin{array}{cccccccccccccccc} & & & & 3' & & & & & & & 5' & & & & \\ & & & & T & C & G & T & C & T & T & T & & & & \\ A & G & T & T & A & G & C & A & G & A & A & A & C & T & A & G \\ 5' & & & & & & & & & & & & & & & 3' \end{array}$$

After heating we obtain the two single strands

$$TCGTCTTT \qquad \textit{and} \qquad AGTTAGCAGAAACTAG.$$

Only the node primer $AGCA$ *can attach to one of these single strands, yielding*

$$
\begin{array}{cccccccc}
3' & & & & & & & 5' \\
T & C & G & T & C & T & T & T \\
A & G & C & A & & & & \\
5' & & & 3'. & & & &
\end{array}
$$

This strand can be doubled using DNA polymerase. Thus, we see that strands encoding for a starting node of 0 *or an ending node of* 6 *will also be replicated. However, each round of replication will produce only one additional strand instead of two, thus the strands starting with node* 0 *and ending with node* 6 *will eventually dominate them.*

These operations are repeated several times, in a continuous cycle of heating (whereby strands are separated) and cooling (whereby strands bond with primers and are replicated). Thus, the number of strands with the correct starting and ending nodes will grow exponentially, doubling at each cycle. Meanwhile, the strands satisfying neither the right starting nor ending node will never be replicated, and remain the same in number. Strands with either the right starting node or the right ending node will be replicated, but at a much smaller rate than the interesting ones, as shown in Example 13.48.

Thus, after n cycles there are 2^n truncated node and edge strands for each of the strands starting at node 0 and ending at node 6. Among this multitude of strands, we hope that if n is sufficiently large, the number of strands starting at node 0 and ending at node 6 becomes sufficiently important that we can hope to find them when using the other steps of Adleman's technique.

13.7 Exercises

Turing machines

1. Let Figure 13.11 represent the function φ of a Turing machine M, and consider the initial configuration

B111111B11111B11111B11B.

At the beginning of operation the pointer points to the leftmost B. Describe the action of the machine and calculate the final position of the pointer when the machine terminates.

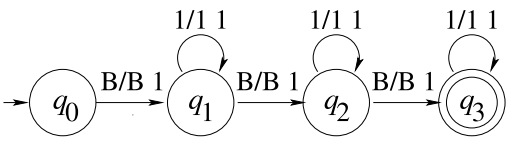


Fig. 13.11. The function φ for Exercise 1.

2. **(a)** Construct a Turing machine that duplicates a unary number to the right of an existing one, with a blank between them. At the end of the calculation the pointer should be returned to the blank preceding the first number.
(b) Construct a Turing machine that permits the copying of a number k times. Use induction.

3. **(a)** Construct a Turing machine that is able to translate a sequence of symbols (containing no blank symbols) by n cells.
(b) Construct a Turing machine that is able to translate k sequences of symbols (each separated by a blank symbol) by n cells.
(c) Consider a sequence of symbols preceded by an arbitrary number of blanks. Construct a Turing machine that will translate the sequence of nonblank symbols to the left until it is preceded by only one blank. For example, the machine will transform BBBBBB$\overline{x}$B to B$\overline{x}$B.

4. Construct a Turing machine that calculates the predecessor function.

5. Construct a Turing machine that calculates the function $cosgn : \mathbb{N} \to \mathbb{N}$ defined by

$$\text{cosgn}(n) = \begin{cases} 1, & n = 0, \\ 0, & n \geq 1. \end{cases}$$

6. Verify that

$$\begin{aligned} |\neg P_1| &= \text{cosgn}(p_1), \\ |P_1 \vee P_2| &= \text{sgn}(p_1 + p_2), \\ |P_1 \wedge P_2| &= p_1 * p_2, \end{aligned}$$

correspond to the value functions of the Boolean operators AND, OR, and NOT. The truth tables for these operators are given in Section 15.7 of Chapter 15.

7. **(a)** Explain how to construct a Turing machine that calculates the function

$$f : \mathbb{N} \times \mathbb{N} \to \{0, 1\} \subset \mathbb{N}$$

defined by

$$f(x, y) = \begin{cases} 1, & x = y, \\ 0, & \text{otherwise.} \end{cases}$$

(b) Describe how to construct a Turing machine that calculates the function

$$f : \mathbb{N} \times \mathbb{N} \to \{0, 1\} \subset \mathbb{N}$$

defined by

$$f(x, y) = \begin{cases} 1, & x \geq y, \\ 0, & \text{otherwise.} \end{cases}$$

8. Construct a Turing machine that exchanges two numbers on the tape. For example, starting with configuration B$\overline{x}$B$\overline{y}$B the machine will terminate with the configuration B$\overline{y}$B$\overline{x}$B. The question is easier if we use the alphabet $\{\mathrm{B}, 1, A\}$, where A will be used as a marker on the ribbon. Note that it is not necessary that the B to the left of y be the first entry on the ribbon (in other words, do not worry about translating the result).

9. Given a Turing machine M that calculates multiplication with two numbers, describe how to construct a Turing machine that calculates the factorial function.

10. The functions $\mathrm{lt}(x, y)$, $\mathrm{gt}(x, y)$, and $\mathrm{eq}(x, y)$ were defined in (13.1).
(a) Explain how to construct a Turing machine that calculates $\mathrm{lt}(x, y)$.
(b) Explain how to construct a Turing machine that calculates $\mathrm{gt}(x, y)$.
(c) Explain how to construct a Turing machine that calculates $\mathrm{eq}(x, y)$.

Recursive functions

11. Show that the functions *sgn* and *cosgn* defined by

$$\begin{cases} \mathrm{sgn}(0) = 0, \\ \mathrm{sgn}(y + 1) = 1, \end{cases} \qquad \begin{cases} \mathrm{cosgn}(0) = 1, \\ \mathrm{cosgn}(y + 1) = 0, \end{cases}$$

are primitive recursive functions.

12. Show that the function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ defined by $f(m, n) = mn + 3n^2 + 1$ is primitive recursive.

13. Show that the following functions are recursive:
(a)

$$abs(x, y) = |x - y|.$$

(b)

$$max(x, y) = \begin{cases} x, & x \geq y, \\ y, & x < y. \end{cases}$$

(c)

$$f(x) = \lfloor \log_2(x) \rfloor.$$

Here $f(x)$ is a total function that maps x to the integer part of $\log_2 x$.

(d)

$$div(x, y) = \lfloor x/y \rfloor.$$

Here $div(x, y)$ is the integer part of the quotient x/y. For example, $div(7, 3) = 2$.

(e)

$$rem(x, y) = x \pmod{y}.$$

This function is the remainder after integer division. For example, $rem(7, 3) = 1$.

(f)

$$f(x) = \begin{cases} 5, & x = 0, \\ 2, & x = 1, \\ 4, & x = 2, \\ 3x, & x > 3. \end{cases}$$

14. Show that if g is a primitive recursive function of $n+1$ variables, then $f(x_1, \ldots, x_n, y) = \sum_{i=0}^{y} g(x_1, \ldots, x_n, i)$ is a primitive recursive function.

Insertion–deletion systems

15. Develop an algorithm that performs addition of two numbers using insertions and deletions. Use the alphabet $X = \{0, 1\}$.

Satisfiability

16. Give the variable assignment graph (like that of Figure 13.10) associated with the statement of Example 13.46.

17. **(a)** Consider the logical statement

$$\gamma = (x_1 \wedge x_2) \vee (\neg x_3 \wedge x_4),$$

where x_1, x_2, x_3, and x_4 are Boolean variables. Express γ in conjunctive normal form.

(b) Repeat the same question with the statement

$$\delta = (\neg(x_1 \vee x_2)) \vee (\neg(x_3 \vee \neg x_4)).$$

(c) Give the variable assignment graph associated with statement γ.

References

[1] L. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(11):1021–1024, November 1994.

[2] A. Church. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, pages 345–363, 1936.

[3] L. Kari and G. Thierrin. Contextual insertions/deletions and computability. *Information and Computation*, 131(1):47–61, 1996.

[4] G. Paun, G. Rozenberg, and A. Salomaa. *DNA Computing: New Computing Paradigms*. Springer, 1998.

[5] M. Sipser. *Introduction to the Theory of Computation*. Course Technology, Boston, 2nd edition, 2006.

[6] T.A. Sudkamp. *Languages and Machine: An Introduction to the Theory of Computer Science*. Addison-Wesley, Boston, 3rd edition, 2006.

[7] A.M. Turing. On computable numbers with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–265, 1937.

[8] A. Yasuhara. *Recursive Function Theory and Logic*. Academic Press, New York and London, 1971.

14

Calculus of Variations and Applications[1]

This chapter is a little more "classic" than the others. It introduces calculus of variations, an elegant field not often covered in modern math curricula. A knowledge of multivariable calculus will suffice, but it helps to also have a familiarity with differential equations.

This chapter covers more material than can be covered in a week of classes. If you want to dedicate only a week of time to this chapter, you could start by motivating the material with a few examples that require minimizing a functional (Section 14.1). Afterward, you may move on to the Euler–Lagrange equation and the Beltrami identity (Section 14.2). Finally, finish the week by solving the problems listed in Section 14.1, including the classic brachistochrone problem (Section 14.4). Covering the rest of the material in this chapter will easily require a second and maybe even a third week. However, the level of difficulty remains constant through the chapter, there being no advanced sections.

Several sections study the properties of cycloids, the solutions to the brachistochrone problem: the tautochrone property is detailed in Section 14.6, and Huygens's isochronous pendulum is studied in Section 14.7. These two sections do not specifically use calculus of variations, but are examples of modeling having given hope, in their time, of technological applications.

All other sections discuss specific problems with solutions in calculus of variations: the fastest tunnel (Section 14.5), soap bubbles (Section 14.8), and isoperimetric problems such as suspended cables, self-supporting arches (both in Section 14.10), and liquid telescopes (Section 14.11).

Section 14.9 discusses Hamilton's principle for classical mechanics, which reformulates the field using the principles of calculus of variations. Less technological than the others, this section offers a cultural enrichment to math students who have been introduced to Newtonian classical mechanics but who have not had the chance to further their studies in physics.

[1] The first version of this chapter was written by Hélène Antaya as an undergraduate math student.

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_14,

14.1 The Fundamental Problem of Calculus of Variations

Calculus of variations is a branch of mathematics dealing with the optimization of physical quantities (such as time, area, or distance). It finds applications in many diverse fields, such as aeronautics (maximizing the lift of an airplane wing), sporting equipment design (minimizing air resistance on a bicycle helmet, optimizing the shape of a ski), mechanical engineering (maximizing the strength of a column, a dam, or an arch), boat design (optimizing the shape of a boat hull), physics (calculating trajectories and geodesics in both classical mechanics and general relativity).

We begin with two examples illustrating the types of problems that may be solved using calculus of variations.

Example 14.1 *This example is very simple and we already know the answer. However, formalizing it will be of help later. The problem consists in finding the shortest path between two points in the plane,* $A = (x_1, y_1)$ *and* $B = (x_2, y_2)$*. We already know that the answer is simply the straight line connecting the two points, but we will go through this solution using the framework of calculus of variations. Suppose that* $x_1 \neq x_2$ *and that it is possible to write the second coordinate as a function of the first. Then the path is parameterized by* $(x, y(x))$ *for* $x \in [x_1, x_2]$*, where* $y(x_1) = y_1$ *and* $y(x_2) = y_2$*. The quantity* I *that we wish to minimize is the length of the path between* A *and* B*. This length depends on the specific trajectory being followed, and is thus a function of* y*,* $I(y)$*. This "function of a function" is called a* functional.

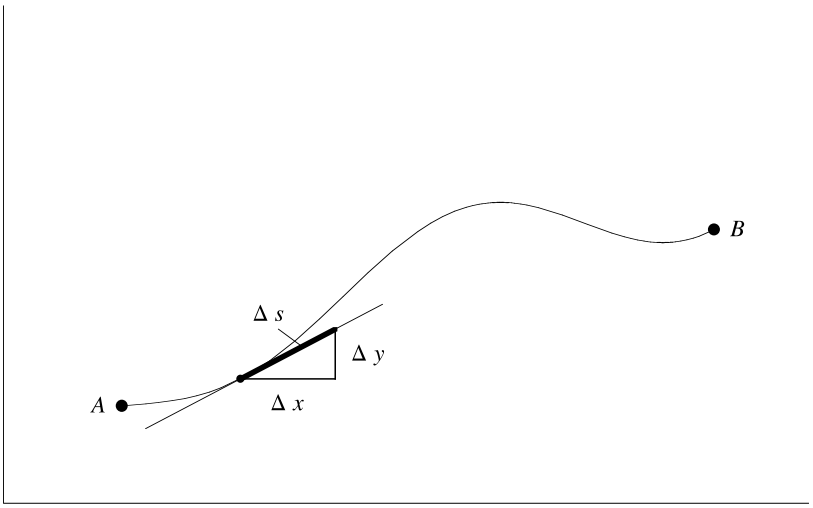


Fig. 14.1. A trajectory between the two points A and B.

Each step Δx corresponds to a step along the trajectory whose length Δs depends on x. The total length of the trajectory is given by

$$I(y) = \sum \Delta s(x).$$

Using the Pythagorean theorem, the length of Δs can be approximated (provided Δx is sufficiently small) as $\Delta s(x) = \sqrt{(\Delta x)^2 + (\Delta y)^2}$, as shown in Figure 14.1. Thus

$$\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2} = \sqrt{1 + \left(\frac{\Delta y}{\Delta x}\right)^2}\,\Delta x.$$

As Δx tends to zero the fraction $\frac{\Delta y}{\Delta x}$ becomes the derivative $\frac{dy}{dx}$, and the integral I may be rewritten as

$$I(y) = \int_{x_1}^{x_2} \sqrt{1 + (y')^2}\,dx. \tag{14.1}$$

Finding the shortest path between the points A and B may be stated, using the language of calculus of variations, as follows: what trajectory $(x, y(x))$ between the points A and B minimizes the functional I? We will return to this problem in Section 14.3.

This first example is not likely to convince anyone of the utility of calculus of variations. The problem posed (find the path $(x, y(x))$ minimizing the integral I) seems way too difficult a method for finding the solution to a problem whose answer is known to be simple. This is why we provide a second example, whose solution is decidedly less obvious.

Example 14.2 *What is the best shape for a skateboard ramp? Half-pipes are very popular in skateboarding and also in snowboarding, a sport that became an Olympic discipline at the 1998 Nagano Olympics. They have a lightly rounded bowl shape. The athlete, either on a skateboard or a snowboard, travels from one side to the other and performs acrobatic stunts at the summits. Three possible profiles for a half-pipe are shown in Figure 14.2. The three shapes all have the same extreme points (A and C) and the same base (B). The bottommost profile requires a small explanation: one must imagine adding a small quarter of a circle in each corner, thus allowing the vertical speed to be transformed into horizontal speed, and then to take the limit as the radius of the circles go to zero. This profile would be fairly dangerous because it contains right angles; however, it allows the athlete to pick up a great deal of speed very quickly, since the path starts with a vertical drop starting at A. The topmost path consists in the two straight line segments AB and BC, and is therefore the shortest possible path going from A through B to C.*

What exactly do we mean by "the best shape"? This formulation is hardly mathematical. We will refine it as follows: what shape will permit the athlete to travel between points A and B in the least amount of time? With this precise definition, what is *the best shape? Should the path giving the greatest speed (at the expense of a longer overall*

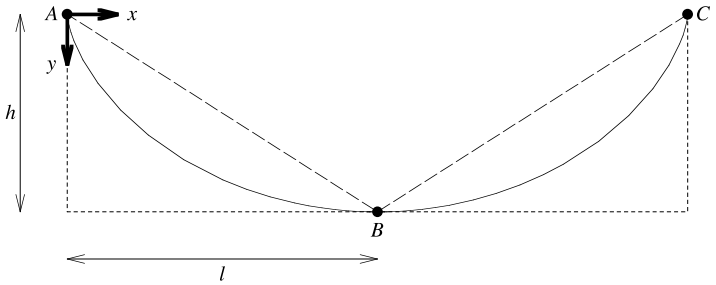


Fig. 14.2. Three candidate profiles for the best half-pipe.

distance) be taken? Should the path covering the shortest distance be taken? Or should it be something between these two extremes, such as the smooth profile in Figure 14.2?

It is relatively easy to calculate the time taken to travel the two extreme profiles. But we will show that the best profile is actually a smooth curve between these two extremes. To this end, we show how to calculate the travel time for a smooth curve described by $(x, y(x))$.

Lemma 14.3 *We choose our coordinate system such that the* y *axis is oriented downward and the* x *axis proceeds from point* A *to* B *and we choose a profile described by a curve* $y(x)$*, where* $A = (x_1, y(x_1))$ *and* $B = (x_2, y(x_2))$*. We consider the time taken for a point mass, propelled only by the force of gravity, to travel from point* A *to point* B*. The time is given by the integral*

$$I(y) = \frac{1}{\sqrt{2g}} \int_{x_1}^{x_2} \frac{\sqrt{1 + (y')^2}}{\sqrt{y}} dx. \tag{14.2}$$

Proof. The key to calculating the travel time is the physical principle of conservation of energy. The total energy E of a point mass is the sum of its kinetic energy ($T = \frac{1}{2}mv^2$) and its potential energy ($V = -mgy$). (Warning: the negative sign in our potential energy term comes from us using an inverted y axis.) In these equations m is the mass of the point, v its speed, and g the acceleration due to gravity. The constant g is approximately $g = 9.8$ m/s^2 on the surface of the Earth. The total energy $E = T + V = \frac{1}{2}mv^2 - mgy$ of the point mass is constant throughout its trip along the curve. If its speed is zero at A, then E is initially zero, and remains so along the entire trajectory. Thus the speed of the point mass is related strictly to its height through the equation $E = 0$, which simplifies to $\frac{1}{2}mv^2 = mgy$ and finally

$$v = \sqrt{2gy}. \tag{14.3}$$

The time taken to travel the path is the sum over all the infinitesimally small dx of the time dt taken to travel the corresponding distance ds. The time is the quotient of the distance ds divided by its speed at the moment. Thus

$$I(y) = \int_A^B dt = \int_A^B \frac{ds}{v}.$$

Example 14.1 showed that for infinitesimal dx, then $ds = \sqrt{1 + (y')^2}dx$, where y' is the derivative of y with respect to x. The travel time is thus given by the integral (14.2). □

A return to Example 14.2. By Lemma 14.3, the integral to minimize is (14.2), where we have the boundary conditions $A = (x_1, 0)$ and $B = (x_2, y_2)$. The problem of finding the best shape for a half-pipe is thus equivalent to finding the function $y(x)$ that minimizes the integral I. This problem seems much harder than the one of our first example!

The two problems shown in Examples 14.1 and 14.2 both belong to the domain of *calculus of variations*. It is possible that they remind you of optimization problems as encountered in calculus. These problems require you to find the extrema of a function $f : [a, b] \to \mathbb{R}$, which can be found at precisely those points where the derivative vanishes or at the extreme points of the interval. Calculus provides us with an extremely powerful tool for solving these problems. However, the problems of Examples 14.1 and 14.2 are of a different breed. In calculus the quantity that varies as we search for the extrema of $f(x)$ is a simple variable x; in calculus of variations, the quantity that varies is itself a function, $y(x)$. We will show that the familiar tools of calculus are sufficiently powerful to allow us to resolve the problems of Examples 14.1 and 14.2.

We now state the fundamental problem of calculus of variations:

Fundamental problem of calculus of variations. Given a function $f = f(x, y, y')$, find the functions $y(x)$ corresponding to the extremal points of the integral

$$I = \int_{x_1}^{x_2} f(x, y, y')dx,$$

subject to the boundary conditions

$$\begin{cases} y(x_1) = y_1, \\ y(x_2) = y_2. \end{cases}$$

How do we identify the functions $y(x)$ that maximize or minimize the integral I? Like the vanishing derivative for variables, the Euler–Lagrange condition characterizes precisely these functions.

14.2 Euler–Lagrange Equation

Theorem 14.4 *A necessary condition for the integral*

$$I = \int_{x_1}^{x_2} f(x, y, y')\, dx \tag{14.4}$$

to attain an extremum subject to the boundary conditions

$$\begin{cases} y(x_1) = y_1, \\ y(x_2) = y_2, \end{cases} \tag{14.5}$$

is that the function $y = y(x)$ satisfy the Euler–Lagrange equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0. \tag{14.6}$$

Proof. We consider only the case of a minimum, but a maximum may be treated similarly.

Suppose that the integral I attains a minimum for a particular function y_* that satisfies $y_*(x_1) = y_1$ and $y_*(x_2) = y_2$. If we deform y_* by applying certain variations, while maintaining the boundary conditions of (14.5), the integral I must increase, since it was minimized by y_*. We consider deformations of a particular type, described by a family of functions $Y(\epsilon, x)$ representing curves between the points (x_1, y_1) and (x_2, y_2):

$$Y(\epsilon, x) = y_*(x) + \epsilon g(x). \tag{14.7}$$

Here ϵ is a real number and $g(x)$ is an arbitrary but fixed differentiable function. The function $g(x)$ must satisfy the condition $g(x_1) = g(x_2) = 0$, which in turn guarantees that $Y(\epsilon, x_1) = y_1$ and $Y(\epsilon, x_2) = y_2$ for all ϵ. The term $\epsilon g(x)$ is called a *variation* of the minimizing function, from which comes the name calculus of variations.

Using this family of deformations, the integral I becomes a function $I(\epsilon)$ of a real variable:

$$I(\epsilon) = \int_{x_1}^{x_2} f(x, Y, Y')\, dx.$$

The problem of finding the extrema of $I(\epsilon)$ for this family of deformations is thus an ordinary optimization problem in calculus. We thus calculate the derivative $\frac{dI}{d\epsilon}$ in order to find the critical points of $I(\epsilon)$:

$$I'(\epsilon) = \frac{d}{d\epsilon}\int_{x_1}^{x_2} f(x, Y, Y')\, dx = \int_{x_1}^{x_2} \frac{d}{d\epsilon} f(x, Y, Y')\, dx.$$

By the chain rule we obtain

$$I'(\epsilon) = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial x}\frac{\partial x}{\partial \epsilon} + \frac{\partial f}{\partial y}\frac{\partial Y}{\partial \epsilon} + \frac{\partial f}{\partial y'}\frac{\partial Y'}{\partial \epsilon}\right) dx. \tag{14.8}$$

But in (14.8), $\frac{\partial x}{\partial \epsilon} = 0$, $\frac{\partial Y}{\partial \epsilon} = g(x)$, and $\frac{\partial Y'}{\partial \epsilon} = g'(x)$. We have therefore that

$$I'(\epsilon) = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} g + \frac{\partial f}{\partial y'} g'\right) dx. \tag{14.9}$$

The second term of (14.9) may be integrated by parts:

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'} g' \, dx = \left[\frac{\partial f}{\partial y'} g \right]_{x_1}^{x_2} - \int_{x_1}^{x_2} g \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \, dx,$$

where the term between brackets on the left disappears, since $g(x_1) = g(x_2) = 0$. Thus, we have that

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'} g' \, dx = -\int_{x_1}^{x_2} g \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \, dx, \tag{14.10}$$

and the derivative $I'(\epsilon)$ becomes

$$I'(\epsilon) = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \right] g \, dx.$$

By our hypothesis the minimum of $I(\epsilon)$ is found at $\epsilon = 0$, since that is precisely when $Y(x) = y_*(x)$. The derivative $I'(\epsilon)$ must therefore be zero when $\epsilon = 0$:

$$I'(0) = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \right] \Bigg|_{y=y_*} g \, dx = 0.$$

The notation $|_{y=y_*}$ indicates that the quantity is evaluated when the function Y is the particular function y_*. Recall that the function g is arbitrary. Thus, in order for $I'(0)$ to remain zero regardless of g, it must be that

$$\left(\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \right) \Bigg|_{y=y_*} = 0,$$

which is precisely the Euler–Lagrange equation. □

In certain cases we can use simplified forms of the Euler–Lagrange equation that allow us to find solutions with ease. One of these "shortcuts" is the Beltrami identity.

Theorem 14.5 *In the case that the function $f(x, y, y')$ in the interior of the integral* (14.4) *is explicitly independent of x, a necessary condition for the integral to have an extremum is given by the Beltrami identity, a particular form of the Euler–Lagrange equation:*

$$y' \frac{\partial f}{\partial y'} - f = C, \tag{14.11}$$

where C is a constant.

Proof. Calculate $\frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right)$ in the Euler–Lagrange equation. By the chain rule and the fact that f is independent of x we obtain

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = \frac{\partial^2 f}{\partial y \partial y'}y' + \frac{\partial^2 f}{\partial y'^2}y''.$$

Thus the Euler–Lagrange equation becomes

$$\frac{\partial^2 f}{\partial y \partial y'}y' + \frac{\partial^2 f}{\partial y'^2}y'' = \frac{\partial f}{\partial y}. \tag{14.12}$$

To obtain Beltrami's identity we need to show that the derivative with respect to x of the function $h = y'\frac{\partial f}{\partial y'} - f$ is zero. Calculating this derivative yields

$$\begin{aligned}\frac{dh}{dx} &= \left(\frac{\partial f}{\partial y'}y'' + \frac{\partial^2 f}{\partial y \partial y'}y'^2 + \frac{\partial^2 f}{\partial y'^2}y'y''\right) - \left(\frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial y'}y''\right)\\ &= y'\left(\frac{\partial^2 f}{\partial y \partial y'}y' + \frac{\partial^2 f}{\partial y'^2}y'' - \frac{\partial f}{\partial y}\right)\\ &= 0,\end{aligned}$$

where the last equality comes from (14.12). □

Before giving examples of the use of the Euler–Lagrange equation it is worthwhile to make a few comments.

The Euler–Lagrange and Beltrami equations are *differential equations* for the function $y(x)$. In other words, they are equations that relate the function y to its derivatives. Solving differential equations is one of the most important applications of differential and integral calculus with many applications in science and engineering.

An easy example of a differential equation is $y'(x) = y(x)$ or simply $y' = y$. "Reading" this differential equation gives a hint of its solution: which function y is equal to its derivative y'? Most people will remember that the exponential function has this property. If $y(x) = e^x$, then $y'(x) = e^x$. Actually, the most general solution of $y' = y$ is $y(x) = ce^x$, where c is a constant. This constant can be determined using a boundary condition like (14.5). There are no systematic methods for finding solutions to differential equations. This in itself is not terribly surprising: a simple differential equation such as $y' = f(x)$ has the following solution $y = \int f(x)dx$. However, there does not always exist a closed form even if it is known that a solution exists and the integral $\int_a^b f(x)dx$ can be numerically integrated. As with integration techniques, there exist a number of ad hoc and special-case methods that may be used to solve common and relatively simple differential equations. We will see some of these techniques in some of the solutions presented in this chapter. Where one cannot find closed-form solutions, it is possible to use theoretical techniques to prove the existence and uniqueness of the solutions, and numerical techniques for calculating them approximately. Such methods are beyond the scope of this chapter, but are discussed in [2], for example.

Much as in the optimization of a single-variable function, the Euler–Lagrange equation sometimes returns several solutions, and further tests are required to determine

which are minima, which are maxima, and which are neither a maximum nor a minimum. Moreover, these extrema may be only local extrema rather than global ones. What is a critical point? For a function of a single real variable, a critical point is a point where the derivative of the function vanishes. Such a point may be an extremum or an inflection point. And for a real function of two variables, critical points can also be saddle points. In the framework of calculus of variations we will say that a function $y(x)$ is a critical point if it is a solution to the associated Euler–Lagrange equation.

One last warning. If we reread the proof of the Euler–Lagrange equation we will see that it makes sense only if the function y is twice differentiable. But it is entirely possible for a real solution to an optimization problem to be a function that is not everywhere differentiable on its domain. An example of a such a situation is found in the following problem: for a specified volume and height, find the profile that should be given to a column of revolution such that it can support the most weight from above. We will not go into the equations describing this problem, but its history is interesting. Lagrange thought he had proved that the best shape was simply a cylinder, but in 1992, Cox and Overton [3] proved that the best shape is that shown in Figure 14.3. Strictly speaking, Lagrange's computations did not contain any errors. He obtained the best solution among the set of differentiable functions, but Cox and Overton's optimal solution is not differentiable.

Fig. 14.3. Cox and Overton's optimal load-bearing column.

The column profile problem is not an isolated example. As it turns out, soap bubbles (Section 14.8) can also contain angles. In fact, problems in calculus of variations (also called variational problems) often have nondifferentiable solutions. In order to solve these problems we must first generalize our notion of the derivative, a subject falling under the heading of nonsmooth analysis.

14.3 Fermat's Principle

We are now ready to solve the two examples introduced in Section 14.1.

Example 14.6 A return to Example 14.1. *As stated earlier, the answer to the first problem is intuitively obvious. What is the shortest path between the points* $A = (x_1, y_1)$ *and* $B = (x_2, y_2)$ *in the plane? Using the Euler–Lagrange equation to solve this problem leads us to another simple example of a differential equation. We have already posed this problem as a variational one: what is the function* $y(x)$ *that minimizes the integral*

$$I(y) = \int_{x_1}^{x_2} \sqrt{1 + (y')^2} dx$$

subject to the boundary conditions

$$\begin{cases} y(x_1) = y_1, \\ y(x_2) = y_2. \end{cases}$$

The function $f(x, y, y')$ *is therefore* $\sqrt{1 + (y')^2}$*. Since the three variables* x*,* y*, and* y' *are independent, this function depends on neither* x *nor* y*. So we only need to calculate the second term of the Euler–Lagrange equation:*

$$\frac{\partial f}{\partial y'} = \frac{y'}{\sqrt{1 + (y')^2}}$$

and

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = \frac{y''}{(1 + (y')^2)^{\frac{3}{2}}}.$$

The shortest path is described by the function y *that satisfies the Euler–Lagrange equation. In other words, it is the one that satisfies the differential equation*

$$\frac{y''}{(1 + (y')^2)^{\frac{3}{2}}} = 0.$$

Since the denominator is always positive, we can multiply both sides of the equation by this quantity, leaving us with

$$y'' = 0.$$

Even if you have not yet taken a course on differential equations you can likely identify the function y *that satisfies the above relation. Solving the differential equation amounts to answering the following question: what function has the function that is everywhere* 0 *as its second derivative? The simple answer is that all first-order polynomials* $y(x) = ax + b$ *have this property. These polynomials depend on two parameters* a *and* b *that must be determined so as to satisfy the boundary conditions* $y(x_1) = y_1$ *and* $y(x_2) = y_2$*. (Exercise!) Thus, calculus of variations has assured us that the shortest path between two points is indeed the straight line through these points!*

This exercise has shown us how to apply the Euler–Lagrange equation. Despite its simplicity, this example can quickly be generalized into much more difficult problems.

We know that light travels in a straight line while it is in material with a constant density, and that it refracts when passing between materials with different densities. Moreover, we know that light reflects from a mirror with an angle of reflection equal to its angle of incidence. *Fermat's principle* summarizes these rules as a statement that leads immediately to variational problems: light follows the trajectory that takes the shortest time to travel (see Section 15.1 of Chapter 15).

The speed of light in a vacuum, denoted by c, is fundamental physical constant (approximately equal to 3.00×10^8 m/s). However, the speed of light is not the same in gas or other materials such as glass. The speed of light through such materials, v, is often expressed with the help of the material's index of refraction n as $v = \frac{c}{n}$. If the material is homogeneous, we have that n and therefore v are constant. Otherwise, n depends on (x, y). A simple example to consider is the index of refraction of the atmosphere, which varies as a function of the density and therefore the altitude (the situation is actually slightly more complex than that, since the speed of light can also depend on the wavelength of the particular beam). If we limit ourselves to motion in a plane, integral (14.1) from the above example must be changed to take into account this variable speed:

$$I = \int_{x_1}^{x_2} dt = \int_{x_1}^{x_2} n(x,y)\frac{ds}{c} = \int_{x_1}^{x_2} n(x,y)\frac{\sqrt{1+(y')^2}}{c}\,dx.$$

Here dt represents an infinitesimally small interval of time and ds a correspondingly small length along the trajectory $(x, y(x))$ described by $\sqrt{1+(y')^2}dx$. If n is constant then n and c can be factored out of the integral and we are again left with the problem of Example 14.1.

However, if the material is not homogeneous then the speed of light varies as it travels through the material, and the quickest path is no longer a straight line. The light is therefore refracted, meaning that its path will deviate from a straight line. Engineers must take this fact into account when designing telecommunications systems (in particular when dealing with short wavelengths).

14.4 The Best Half-Pipe.

We are now ready to tackle the more difficult problem of finding the best shape for a half-pipe. This is actually a much older problem in modern guise. In fact, its first formulation precedes the invention of the skateboard by nearly three centuries! In the seventeenth century, Johann Bernoulli announced a contest that occupied the greatest minds of the time. He published the following problem in Leipzig's *Acta Eruditorum*: *"Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, that starts at A and reaches B in the shortest time?"* The

problem was referred to as the *brachistochrone* problem, which literally means "the shortest time." It is known that five mathematicians proposed solutions to this problem: Leibniz, L'Hôpital, Newton, and both Johann and Jacob Bernoulli [7].

The integral to minimize was shown in (14.2) as

$$I(y) = \frac{1}{\sqrt{2g}} \int_{x_1}^{x_2} \frac{\sqrt{1 + (y')^2}}{\sqrt{y}}\, dx,$$

and the function $f = f(x, y, y')$ is therefore

$$f(x, y, y') = \frac{\sqrt{1 + (y')^2}}{\sqrt{y}}.$$

Since x does not explicitly appear in f, we can apply the Beltrami identity (see Theorem 14.5). The best half-pipe is therefore described by the function y satisfying

$$y' \frac{\partial f}{\partial y'} - f = C.$$

Expanding this yields

$$\frac{(y')^2}{\sqrt{1 + (y')^2}\sqrt{y}} - \frac{\sqrt{1 + (y')^2}}{\sqrt{y}} = C.$$

We can simplify this expression by putting the two terms over a common denominator:

$$\frac{-1}{\sqrt{1 + (y')^2}\sqrt{y}} = C.$$

Solving for y', we obtain the differential equation

$$\frac{dy}{dx} = \sqrt{\frac{k - y}{y}}, \tag{14.13}$$

where k is a constant equal to $\frac{1}{C^2}$.

This differential equation is difficult even for someone who has taken a course in differential equations. In fact, it is impossible to express y as a simple function of x. The following trigonometric substitution will allow us to integrate the equation:

$$\sqrt{\frac{y}{k - y}} = \tan \phi.$$

The function ϕ is a new function of x. Isolating y, we obtain

$$y = k \sin^2(\phi).$$

The derivative of $\phi(x)$ can be calculated using the chain rule, yielding

$$\frac{d\phi}{dx} = \frac{d\phi}{dy} \cdot \frac{dy}{dx} = \frac{1}{2k(\sin\phi)(\cos\phi)} \cdot \frac{1}{(\tan\phi)} = \frac{1}{2k\sin^2\phi}.$$

A typical method for resolving this equation involves rewriting it in the form

$$dx = 2k\sin^2\phi\, d\phi,$$

which indicates the relationship between the two infinitesimal values dx and $d\phi$. Integrating both sides yields

$$x = 2k\int \sin^2\phi\, d\phi = 2k\int \frac{1-\cos 2\phi}{2}\, d\phi = 2k\left(\frac{\phi}{2} - \frac{\sin 2\phi}{4}\right) + C_1.$$

We have chosen the initial point A of the trajectory as the origin of the coordinate system (see Figure 14.2). This choice permits us to fix the constant of integration C_1. At A, the two coordinates x and y are both zero. Thus, the equation $y = k\sin^2\phi$ forces $\phi = 0$ (or an integer multiple of π). Substituting this into the above equation for x yields $x = C_1$, which therefore forces $C_1 = 0$. Finally, by substituting $\frac{k}{2} = a$ and $2\phi = \theta$ we obtain

$$\begin{cases} x = a(\theta - \sin\theta), \\ y = a(1 - \cos\theta). \end{cases} \tag{14.14}$$

These are the parametric equations describing a *cycloid*. The cycloid is the curve traced out by a fixed point on the edge of a circle of radius a rolling in a straight line (see Figure 14.4).

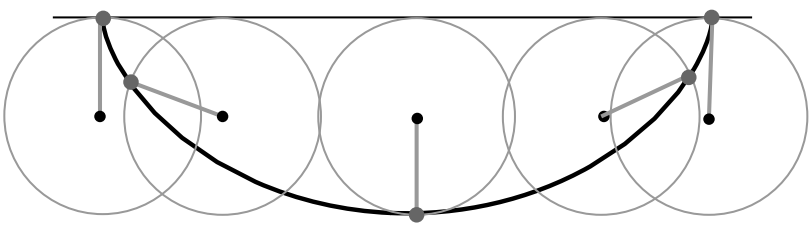


Fig. 14.4. Constructing a cycloid.

Thus, this is the best shape for a half-pipe. More specifically, this is the shape that allows an athlete, powered only by gravity, to travel from point A to point B in the least amount of time. The smooth curve drawn between the two extreme profiles of Figure 14.2 is a cycloid.

Cycloids are very well known by geometers, since they possess a few other interesting properties. For example, Christiaan Huygens discovered that the period of oscillation of

a ball along a cycloid is constant, regardless of its amplitude. In other words, if we place an object anywhere along the side wall of a cycloid, then accelerated only by gravity, it will take exactly the same amount of time to reach the bottom. This independence of the period of oscillation from the amplitude is called the *tautochrone* property. We will prove this in Section 14.6.

14.5 The Fastest Tunnel

We will now discuss a generalization of the brachistochrone that has the potential (in theory) to completely revolutionize transportation. Suppose that we could build a tunnel through the Earth's crust connecting any city A to any other city B in the world. If we neglect friction, a train departing A with zero speed would accelerate as the tunnel gets closer to the center of the Earth and then decelerate as it gets further, finally arriving at B with exactly zero speed! There would be no need for engines, fuel, or brakes! We will push the limits of this fantasy further yet: *we will determine the profile of the tunnel that will be traversed in the shortest time.*

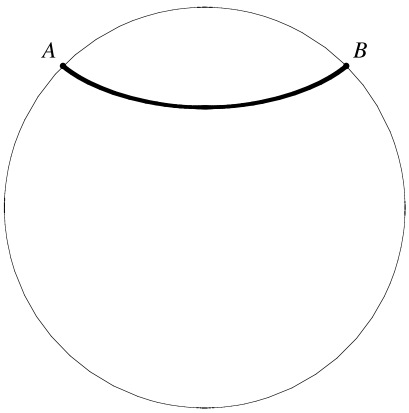


Fig. 14.5. A tunnel between two cities A and B.

Exercise 13 will show that the transit time of such a tunnel between New York and Los Angeles is a little less than half an hour, compared to roughly five hours by air (the great circle route between New York and Los Angeles is roughly 3940 km long). But do not try to buy your tickets yet. This revolutionary transit system has a few difficult problems to overcome. If the two cities being considered are sufficiently far apart, the optimal tunnel between them goes deeper than the Earth's crust and has to travel through its liquid core! What materials can resist the high temperatures and pressures encountered at such depths? Even if we were to overcome such engineering difficulties

there would remain the very real problem of cost. Only the largest of cities (those with many millions of inhabitants) are able to afford building subway lines; the net length of these tracks rarely exceeds a few hundred kilometers (1160 km for the New York subway system). The tunnel running under the English channel is only 50 km long. Opened in 1994, it cost 16 billion euros to build. And there are others: Japan's Seikan rail tunnel is 53.85 km long, and the Swiss are in the middle of building (to be finished in 2015) the Gothard tunnel, whose final length will be 57 km. (Exercise: estimate the size of the hill with 30-degree slopes formed by the Earth removed from the construction of any of these tunnels.) Despite the utopian nature of the following discussion, it remains an elegant exercise.

We can model this situation using physics. We model the Earth as a uniform solid sphere of material with constant density, and the two cities A and B as points on its surface. We will draw the tunnel in the plane defined by the two cities and the center of the sphere, and parameterize it with the curve $(x, y(x))$. The goal of this exercise is again to find the curve $(x, y(x))$ that will be traversed in the shortest amount of time when powered by gravity alone. What is the difference between this problem and the brachistochrone? The main difference is that the strength and the direction of the force of gravity changes as a function of our position along the path.

As with the brachistochrone, the problem is to minimize the integral

$$T = \int \frac{ds}{v}, \tag{14.15}$$

where v designates the speed of the object at point $(x, y(x))$ along its path and ds is an infinitesimally small piece of the trajectory with length

$$ds = \sqrt{1 + (y')^2}\, dx. \tag{14.16}$$

The speed v will be slightly more difficult to express, since the force of gravity is variable.

Proposition 14.7 *The gravitational force at a point a distance $r = \sqrt{x^2 + y^2}$ from the center of the solid sphere of radius $R > r$ and constant density is oriented toward the center of the sphere and has a magnitude of*

$$|F| = \frac{GMm}{R^3} r,$$

where M is the mass of the sphere and G is Newton's gravitational constant.

For now, we will take this classical result on faith and continue our discussion. However, a full proof can be found at the end of the section.

The speed v at point $(x, y(x))$ will again be calculated using the principle of the conservation of energy. This principle says that in the absence of friction, the total energy of an object in motion (that is, the sum of its potential and kinetic energies) remains constant. At the beginning of the trip the speed is assumed to be zero, thus

the object has zero kinetic energy. And since the trajectory starts at the surface of the Earth, the potential energy will be evaluated using $r = R$. The relationship between gravitational force and potential energy is given by $F = -\nabla V$. Since F depends only on the distance r from the center of the sphere, this is easily calculated as

$$V = \frac{GMmr^2}{2R^3}.$$

The potential energy is determined only up to some additive constant, which we choose to be $V(r) = 0$ at $r = 0$. The total energy of the object at the beginning of its trip is therefore given by

$$E = \frac{1}{2}mv^2 + V(r) = 0 + \left.\frac{GMmr^2}{2R^3}\right|_{r=R} = \frac{GMm}{2R}.$$

We are now in a position to calculate the speed v of the object as a function of its position $(x, y(x))$. By the conservation of energy it follows that

$$\frac{GMm}{2R} = \frac{mv^2}{2} + \frac{GMm}{2R^3}r^2$$

and therefore

$$v = \sqrt{\frac{GM(R^2 - r^2)}{R^3}}.$$

Letting $g = \frac{GM}{R^2}$, which corresponds to the force of gravity at the surface of the Earth, we can simplify the speed to

$$v = \sqrt{\frac{g}{R}}\sqrt{R^2 - r^2} = \sqrt{\frac{g}{R}}\sqrt{R^2 - x^2 - y^2}. \tag{14.17}$$

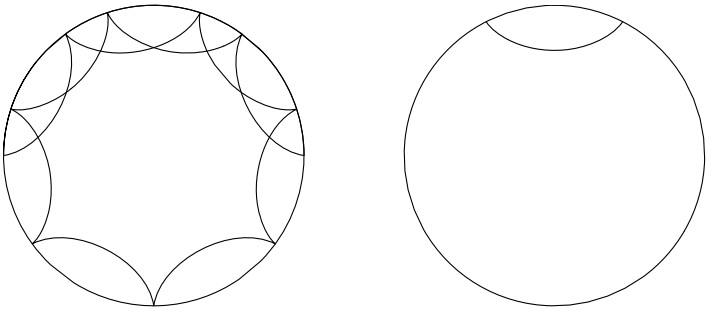
Using (14.15), (14.16), and (14.17), the travel time of the object can be expressed as

$$t = \sqrt{\frac{R}{g}}\int_{x_A}^{x_B} \frac{\sqrt{1 + (y')^2}}{\sqrt{R^2 - x^2 - y^2}}\, dx.$$

We thus end up with an expression very similar to that describing the brachistochrone. Using the Euler–Lagrange equation leads to the curve shown in Figure 14.6, whose parametric equations are

$$\begin{aligned} x(\theta) &= R\left[(1-b)\cos\theta + b\cos\left(\frac{1-b}{b}\theta\right)\right], \\ y(\theta) &= R\left[(1-b)\sin\theta - b\sin\left(\frac{1-b}{b}\theta\right)\right], \end{aligned} \tag{14.18}$$

with $b \in [0, 1]$. This curve is called a *hypocycloid*. We will not step through the details of this solution here. The reader is encouraged to verify that 14.18 is in fact a solution,



(a) $\theta \in [0, 3\pi]$

(b) A tunnel following a hypocycloid trajectory

Fig. 14.6. A hypocyloid with $b = 0.15$.

but the calculation is a little tedious, and mathematical software might be of use. In the particular case $b = \frac{1}{2}$, the hypocycloid is in fact a straight line segment, since $x \in [-R, R]$ and $y = 0$. We showed that the cycloid is drawn by a point on the edge of a circle rolling in a straight line. Similarly, the hypocycloid is drawn by a point on the edge of a circle of radius a rolling along the inside of another circle of radius R (the parameter b of (14.18) is $b = \frac{a}{R}$). Some of you may remember Hasbro's SpiroGraph toy, which involved placing a pencil inside a disk that rolled along the interior of a large ring (one of the many configurations of this toy). In order to draw a hypocycloid with the SpiroGraph, the pencil would have to be placed exactly at the periphery of the disc. It is interesting to note the strong similarities between this problem and the earlier brachistochrone problem.

Proof of Proposition 14.7. We consider a uniform sphere and we study the gravitational force induced by this sphere on a point mass P somewhere inside the sphere. Without loss of generality we may assume that the point mass P is placed along the x axis at a distance $r \leq R$ from the origin (see Figure 14.7). We use spherical coordinates centered at P:

$$\begin{cases} x = \rho \sin\theta, \\ y = \rho \cos\theta \cos\phi, \\ z = \rho \cos\theta \sin\phi, \end{cases}$$

where $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\rho \geq 0$, and $\phi \in [0, 2\pi]$. The Jacobian of this change of coordinates is $\rho^2 \cos\theta \geq 0$, and therefore the infinitesimal volumes of integration are related by $dx\, dy\, dz = \rho^2 \cos\theta\, d\rho\, d\theta\, d\phi$.

Due to symmetry, the sphere with center P and radius $b = R - r$ has a net attraction of zero on the point P. Thus, the net gravitational force exerted on P depends on the remaining volume of the larger sphere, as indicated by the shaded region in Figure 14.7.

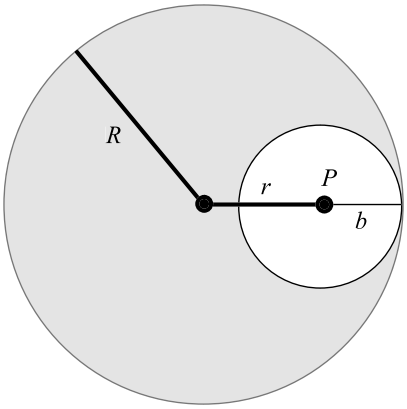


Fig. 14.7. The variables characterizing the interior point P.

The gravitational force exerted by a small element with volume $dx\,dy\,dz$ and centered at (x, y, z) is proportional to the vector $\frac{(x,y,z)}{(x^2+y^2+z^2)^{\frac{3}{2}}}dx\,dy\,dz$. The total gravitational force is the sum of all of these small contributions. For reasons of symmetry it follows that the y and z components of this force are zero.

The (amplitude of the) total force is therefore given by the following triple integral:

$$F = mG\mu \iiint \frac{x}{(x^2 + y^2 + z^2)^{\frac{3}{2}}}\, dx\, dy\, dz,$$

where μ is the density of the sphere, G is Newton's gravitational constant, and m is the mass of the point mass P. The domain of integration is the volume described by the shaded part of Figure 14.7, which is the interior of the large sphere minus the smaller sphere of radius b centered at P. To calculate this integral we first transform it to spherical coordinates:

$$F = mG\mu \iiint \left(\frac{\rho \sin\theta}{\rho^3}\rho^2 \cos\theta\right) d\phi\, d\rho\, d\theta.$$

We must now express the limits of this integral in terms of these new coordinates. The coordinates of a point on the inner sphere satisfy $x^2+y^2+z^2 = \rho^2$, where $\rho = b = R-r$. The coordinates of points on the surface of the outer sphere satisfy $(x+r)^2+y^2+z^2 = R^2$, or equivalently

$$(\rho \sin\theta + r)^2 + \rho^2 \cos^2\theta \cos^2\phi + \rho^2 \cos^2\theta \sin^2\phi = R^2,$$

which simplifies to

$$\rho^2 + r^2 + 2r\rho \sin\theta = R^2.$$

This equation has two roots. We take

$$\rho = -r\sin\theta + \sqrt{r^2\sin^2\theta - r^2 + R^2}$$

so that $\rho \geq 0$. Since we have expressed the limits in spherical coordinates, we can now evaluate the triple integral F:

$$\begin{aligned}
F &= mG\mu \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{R-r}^{-r\sin\theta+\sqrt{R^2-r^2\cos^2\theta}} \int_0^{2\pi} \left(\frac{\rho\sin\theta}{\rho^3}\right) \rho^2\cos\theta\, d\phi\, d\rho\, d\theta \\
&= 2\pi mG\mu \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{R-r}^{-r\sin\theta+\sqrt{R^2-r^2\cos^2\theta}} \sin\theta\cos\theta\, d\rho\, d\theta \\
&= 2\pi mG\mu \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin\theta\cos\theta(-r\sin\theta + \sqrt{R^2 - r^2\cos^2\theta} + r - R)\, d\theta \\
&= 2\pi mG\mu \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(-r\sin^2\theta\cos\theta + \sin\theta\cos\theta\sqrt{R^2 - r^2\cos^2\theta} + (r-R)\frac{\sin 2\theta}{2}\right) d\theta \\
&= 2\pi mG\mu \left(\frac{-r\sin^3\theta}{3}\bigg|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + \frac{1}{3r^2}(R^2 - r^2\cos^2\theta)^{\frac{3}{2}}\bigg|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} - \frac{(r-R)\cos 2\theta}{4}\bigg|_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\right).
\end{aligned}$$

The last two terms are equal to 0. Thus we have that

$$F = -\frac{4\pi}{3} rmG\mu.$$

The negative sign indicates that the force is directed toward the center of the Earth. Finally, if M is the mass of the Earth, we have that $\mu = \frac{M}{4\pi R^3/3}$ and

$$|F| = \frac{GMm}{R^3} r.$$

□

14.6 The Tautochrone Property of the Cycloid

Recall that the cycloid is parameterized by

$$\begin{cases} x(\theta) = a(\theta - \sin\theta), \\ y(\theta) = a(1 - \cos\theta), \end{cases} \tag{14.19}$$

as a function of the variable $\theta \in [0, 2\pi]$. (Figure 14.8 shows such a cycloid; the y axis is oriented downward.) The peaks of the cycloid are at the points $\theta = 0$ and 2π, while the lowest point is at $\theta = \pi$. Consider placing a ball with mass m at the point $(x(\theta_0), y(\theta_0))$

for some $\theta_0 < \pi$ and letting it go with zero initial velocity. If friction is negligible, then the ball will oscillate between the point $(x(\theta_0), y(\theta_0))$ and its corresponding point $(x(2\pi - \theta_0), y(2\pi - \theta_0))$ on the opposite side of the bottom. One trip back and forth is a single period of this oscillation. The goal of this section is to prove that the time taken to complete a period is independent of θ_0.

Proposition 14.8 *Let $T(\theta_0)$ be the period of oscillation for a ball released at $(x(\theta_0), y(\theta_0))$. Then*

$$T(\theta_0) = 4\pi\sqrt{\frac{a}{g}}. \tag{14.20}$$

The period is therefore independent of θ_0.

PROOF. The period is equal to $4\tau(\theta_0)$, where $\tau(\theta_0)$ is the time taken for the ball to roll from its starting point to the lowest point of the cycloid, $(x(\pi), y(\pi))$. We will show that $\tau(\theta_0) = \pi\sqrt{\frac{a}{g}}$.

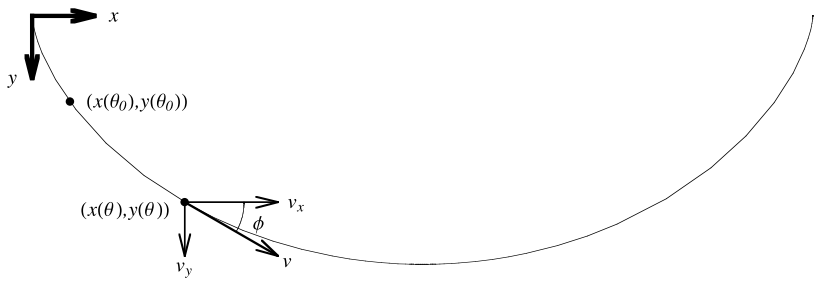


Fig. 14.8. The starting position $(x(\theta_0), y(\theta_0))$ of the ball and the components of its velocity at a later time.

Let $v_y(\theta)$ be the vertical component of the velocity of the ball at position θ. Then we have that

$$\tau(\theta_0) = \int_0^{\tau(\theta_0)} dt = \int_{y(\theta_0)}^{y(\pi)} \frac{dy}{v_y(\theta)} = \int_{\theta_0}^{\pi} \frac{1}{v_y(\theta)} \frac{dy}{d\theta} d\theta. \tag{14.21}$$

By (14.19) we see that

$$\frac{dy}{d\theta} = a \sin\theta.$$

We must calculate $v_y(\theta)$. Again, we may use the conservation of energy. As with (14.3), the total speed $v(\theta)$ of the ball at points $(x(\theta), y(\theta))$ depends on the vertical distance traveled,

$$h(\theta) = y(\theta) - y(\theta_0) = a(\cos\theta_0 - \cos\theta),$$

and therefore

$$v(\theta) = \sqrt{2gh(\theta)} = \sqrt{2ga}\sqrt{\cos\theta_0 - \cos\theta}.$$

The vertical component of this velocity may be computed as

$$v_y(\theta) = v(\theta)\sin\phi, \tag{14.22}$$

where ϕ is the angle between the direction of the ball and the horizontal. Since

$$\tan\phi = \frac{dy}{dx} = \frac{dy}{d\theta}\Big/\frac{dx}{d\theta} = \frac{\sin\theta}{1-\cos\theta},$$

we have

$$1 + \tan^2\phi = \frac{2}{1-\cos\theta}$$

and therefore

$$\sin\phi = \sqrt{1-\cos^2\phi} = \sqrt{1 - \frac{1}{1+\tan^2\phi}} = \sqrt{\frac{1+\cos\theta}{2}}. \tag{14.23}$$

(Careful! Since the y axis is oriented downward, the angle ϕ increases in the clockwise direction rather than counterclockwise. Thus, the angle ϕ indicated in Figure 14.8 is positive.) Thus we get

$$v_y(\theta) = \sqrt{ga}\sqrt{\cos\theta_0 - \cos\theta}\sqrt{1+\cos\theta}. \tag{14.24}$$

The integral in (14.21) is now explicit in terms of θ_0 and θ. Since $\sin\theta$ is positive for $0 \le \theta \le \pi$, then $\sin\theta = \sqrt{1-\cos^2\theta}$ and we obtain

$$\begin{aligned}\frac{1}{v_y(\theta)}\frac{dy}{d\theta} &= \frac{a\sin\theta}{\sqrt{ga}\sqrt{\cos\theta_0 - \cos\theta}\sqrt{1+\cos\theta}} \\ &= \sqrt{\frac{a}{g}}\frac{\sqrt{(1-\cos\theta)(1+\cos\theta)}}{\sqrt{\cos\theta_0 - \cos\theta}\sqrt{1+\cos\theta}} \\ &= \sqrt{\frac{a}{g}}\frac{\sqrt{1-\cos\theta}}{\sqrt{\cos\theta_0 - \cos\theta}}.\end{aligned} \tag{14.25}$$

Thus

$$\tau(\theta_0) = \sqrt{\frac{a}{g}}I(\theta_0), \qquad \text{where} \qquad I(\theta_0) = \int_{\theta_0}^{\pi}\frac{\sqrt{1-\cos\theta}}{\sqrt{\cos\theta_0 - \cos\theta}}d\theta.$$

It remains only to evaluate the integral $I(\theta_0)$. The first step is to rewrite it as

$$I(\theta_0) = \int_{\theta_0}^{\pi}\frac{\sin\frac{\theta}{2}}{\sqrt{\cos^2\frac{\theta_0}{2} - \cos^2\frac{\theta}{2}}}d\theta,$$

using the fact that $\sqrt{1-\cos\theta} = \sqrt{2}\sin\frac{\theta}{2}$ and $\cos\theta = 2\cos^2\frac{\theta}{2} - 1$. In order to evaluate the integral we use a change of variables:

$$u = \frac{\cos\frac{\theta}{2}}{\cos\frac{\theta_0}{2}} \qquad \text{with} \qquad du = -\frac{\sin\frac{\theta}{2}}{2\cos\frac{\theta_0}{2}}d\theta.$$

Under this change of variables $\theta = \theta_0$ and $\theta = \pi$ correspond to $u = 1$ and $u = 0$, respectively. Thus the integral becomes

$$I(\theta_0) = -\int_1^0 \frac{2}{\sqrt{1-u^2}}du = -2\arcsin(u)\Big|_1^0 = \pi,$$

which completes the proof. □

Note that the proof of this section also allows us to calculate the time taken for a ball to travel between $(0,0)$ and $(x(\theta), y(\theta))$; integral (14.21) remains valid, requiring only a change in the limits.

Corollary 14.9 *The time taken for a ball, acted upon only by gravity, to travel along a cycloid from point $\theta = 0$ to θ is given by*

$$T(\theta) = \sqrt{\frac{a}{g}}\theta.$$

In particular, $T(\pi) = \pi\sqrt{\frac{a}{g}}$ (this is the same as $\tau(\theta_0)$ calculated above) and $T(2\pi) = 2\pi\sqrt{\frac{a}{g}}$ (the shortest time taken to travel from $(0,0)$ to $(2\pi a, 0)$ using only gravity).

Proof. The integrand is the same as that of (14.25). Substituting 0 as the lower limit and θ as the upper limit yields

$$T(\theta) = \int_0^{T(\theta)} dt = \sqrt{\frac{a}{g}}\int_0^\theta \frac{\sin\frac{\theta}{2}}{\sqrt{1-\cos^2\frac{\theta}{2}}}d\theta = \sqrt{\frac{a}{g}}\int_0^\theta d\theta = \sqrt{\frac{a}{g}}\theta.$$

□

14.7 An Isochronous Device

When first discovered, the tautochrone property of the cycloid created quite a stir among clockmakers. If we can force a particle to travel without friction along a cycloidal path under the effect of gravity, then it will oscillate with a period of $\left(4\pi\sqrt{\frac{a}{g}}\right)$, regardless of the amplitude of the motion. This is not the case for classic pendulums that swing along a circular arc. For such pendulums the period increases as the angle of maximum displacement increases. Thus in order for such clocks to run true, the pendulum must be precisely positioned when started, and the amplitude must remain constant over

days. In practice, the difference in the period can be neglected if the amplitude of the pendulum is sufficiently small, but the clock will never be precise.[2]

Having discovered the tautochrone property of the cycloid, Huygens had the idea of building a clock whose pendulum would be forced to travel a cycloidal path. At the time, any improvement in the accuracy of clocks implied a corresponding improvement in the accuracy of astronomy and navigation. In fact, having accurate clocks was nearly a question of life or death for maritime navigators. In order to accurately determine their longitude they needed to know the time of day to high precision. However, the imprecise clocks of the era accrued error relatively quickly. Such imprecision could be dangerous, for it could lead navigators to calculate their position as being in safe waters when in reality they were not.

We will describe the device imagined by Huygens, which forced the mass of a pendulum to follow a cycloidal path. The problem with this device is that the friction involved slows down the pendulum much more rapidly than a traditional pendulum.

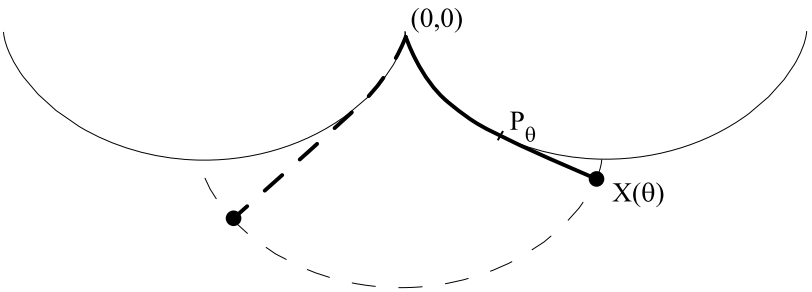


Fig. 14.9. Huygens's device and two positions of the pendulum.

Huygens imagined two "bumpers" with a cycloidal profile of parameter a, and a pendulum of length $4a$ suspended between the two of them (see Figure 14.9). As the pendulum swings, its string is pressed against the cycloidal bumpers for a length $l(\theta)$, running flat with the bumper between the points $(0,0)$ and P_θ. The loose part of the string is a line segment that is tangent to the cycloid at the point P_θ.

Proposition 14.10 *In the absence of friction, Huygens's pendulum (as shown in Figure 14.9) is* isochronous *(in other words, it has a constant period of oscillation regardless of the amplitude of the motion).*

[2] You may already have studied the motion of pendulums in a physics course. The differential equation describing their motion is $\frac{d^2}{dt^2}\theta = -\frac{g}{l}\sin\theta$, which may be approximated by $\frac{d^2}{dt^2}\theta = -\frac{g}{l}\theta$ under the hypothesis that θ remains close to 0. (l is the length of the pendulum's cord.) This approximation yields the solution $\theta(t) = \theta_0\cos(\sqrt{\frac{g}{l}}(t-t_0))$, which has a period independent of the amplitude θ_0. However, this approximation is invalid for sufficiently large θ_0.

PROOF. The position of the end of the pendulum is given by the equation

$$P_\theta + (L - l(\theta))T(\theta) = X(\theta), \tag{14.26}$$

where P_θ is the point of tangency, $T(\theta)$ is the unit tangent vector at P_θ, and $(L-l(\theta))$ is the length of the string that remains free. The quantity $X(\theta)$ represents the position of the end of the pendulum as a function of the parameter θ. (Careful: θ is the parameter that traces out the cycloid, and *not* the angle that the pendulum makes with the vertical axis.)

We begin by finding the components of the vector P_θ. This is straightforward, since P_θ parameterizes the cycloid; thus

$$P_\theta = (a(\theta - \sin\theta), a(1 - \cos\theta)).$$

In order to find the tangent vector to the cycloid at the point θ, it suffices to differentiate the components of P_θ individually:

$$V(\theta) = (a(1 - \cos\theta), a\sin\theta).$$

To make this a *unit* tangent vector, we simply renormalize it by its length,

$$|V(\theta)| = \sqrt{a^2(1-\cos\theta)^2 + a^2\sin^2\theta} = \sqrt{2}a\sqrt{1-\cos\theta},$$

yielding

$$T(\theta) = \frac{V(t)}{|V(t)|} = \left(\frac{\sqrt{1-\cos\theta}}{\sqrt{2}}, \frac{\sin\theta}{\sqrt{2}\sqrt{1-\cos\theta}}\right).$$

The length of the cable has been set to $L = 4a$. Thus it remains only to calculate the value $l(\theta)$, corresponding to the length of the perimeter of the cycloid between the points $(0,0)$ and P_θ (see Figure 14.9). This can be accomplished by evaluating the following integral:

$$l(\theta) = \int_0^\theta \sqrt{(x')^2 + (y')^2}\, d\theta = \int_0^\theta a\sqrt{2}\sqrt{1-\cos\theta}\, d\theta. \tag{14.27}$$

This integral can be simplified by recalling that $\sqrt{1-\cos\theta} = \sqrt{2}\sin\frac{\theta}{2}$, yielding

$$l(\theta) = \int_0^\theta a\sqrt{2}\sqrt{2}\sin\frac{\theta}{2}\, d\theta = \left[-4a\cos\frac{\theta}{2}\right]_0^\theta = -4a\cos\frac{\theta}{2} + 4a.$$

We now have all the tools necessary to describe the trajectory $X(\theta)$. Before we proceed, we simplify the expression for the vector between the point of tangency P_θ and the end $X(\theta)$ of the pendulum:

$$\begin{aligned}\overrightarrow{P_\theta X(\theta)} &= (L - l(\theta))T(\theta) \\ &= 4a\cos\tfrac{\theta}{2}\left(\frac{\sqrt{1-\cos\theta}}{\sqrt{2}}, \frac{\sin\theta}{\sqrt{2}\sqrt{1-\cos\theta}}\right) \\ &= 4a\left(\frac{\sqrt{1-\cos\theta}\sqrt{1+\cos\theta}}{2}, \frac{(\cos\frac{\theta}{2})(2\sin\frac{\theta}{2}\cos\frac{\theta}{2})}{\sqrt{2}\sqrt{2}\sin\frac{\theta}{2}}\right) \\ &= 2a(\sqrt{1-\cos^2\theta}, 2\cos^2\frac{\theta}{2}) \\ &= 2a(\sin\theta, 1+\cos\theta).\end{aligned}$$

Adding the coordinates for the point of tangency P_θ, we finally obtain

$$\begin{aligned}X(\theta) &= (a\theta - a\sin\theta + 2a\sin\theta, a - a\cos\theta + 2a + 2a\cos\theta) \\ &= (a(\theta+\sin\theta), a(1+\cos\theta) + 2a) \\ &= (a(\phi - \sin\phi) - a\pi, a(1-\cos\phi) + 2a),\end{aligned}$$

where we have applied the substitution $\phi = \theta + \pi$ and the two identities $\sin\theta = -\sin(\theta + \pi)$ and $\cos\theta = -\cos(\theta+\pi)$. This curve is thus a cycloid translated by $(-\pi a, 2a)$. Thus, Huygens's device forces the extremity $X(\theta)$ of the pendulum to follow a cycloidal path. □

14.8 Soap Bubbles

What is the form that an elastic sheet will take when it is attached to the edges of a rigid frame? This question has a simple and intuitive answer when the entire perimeter of the frame lies in a plane: the sheet will also lie in the plane of the frame. For example, the skin of a drum is flat, lying within the plane defined by the perimeter of the drum. Calculus of variations is hardly necessary in this case, but what about when the frame does not lie in a plane? As you may have guessed, the answer is much less evident! Nonetheless, finding the answer to this problem is little more than child's play. Armed with nothing more than a little soapy water and a piece of wire that can be bent into any shape, anyone can find the solution. When dipped into the soapy water, the film formed inside the frame will give the experimental answer to the question we have just posed.

In the last half century, architecture has distanced itself from the world of vertical walls and flat roofs. Many large projects have chosen to incorporate nonplanar surfaces, particularly roofs. Although the materials used are far from being elastic and supple, the shapes they take often resemble those of elastic sheets attached to exotic frames.

Calculus of variations allows us to solve this question by noting that the ideal surface is that with minimum surface area. (To convince yourself, recall that the tension in an elastic is at its minimum when it is not stretched. Minimizing the length of an elastic

band and the area of an elastic sheet both serve to minimize the tension of the material.) Thus, answering our question amounts to minimizing the integral

$$I = \iint_D \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dx\, dy, \tag{14.28}$$

which represents the surface area of a function $f = f(x, y)$ situated above a domain D whose perimeter is a closed curve $\mathcal{C}$ (the image of the frame). Under this formulation, the question is equivalent to that of *minimal surfaces* in classical geometry.

Finding the function f that minimizes integral (14.28) requires deriving a form of the Euler–Lagrange equation for functionals defined by two-dimensional integrals. This is not too difficult, and is left to the reader in Exercise 16. For the present discussion we limit ourselves to surfaces of revolution that may be cast as one-dimensional problems.

Example 14.11 *We consider a frame consisting of two parallel circles $y^2 + z^2 = R^2$ situated in the planes $x = -a$ and $x = a$. Consider a curve $z = f(x)$ such that $f(-a) = R$ and $f(a) = R$. The surface of revolution created by rotating this curve around the x axis is a surface that is attached to the two circular frames. We will leave it as an exercise to the reader (Exercise 15) to show that the area of this surface is given by the formula*

$$I = 2\pi \int_{-a}^{a} f\sqrt{1 + f'^2}dx. \tag{14.29}$$

Minimizing this integral amounts to solving the associated Beltrami identity

$$\frac{f'^2 f}{\sqrt{1 + f'^2}} - f\sqrt{1 + f'^2} = C,$$

which may be rewritten as

$$\frac{f}{\sqrt{1 + f'^2}} = C.$$

Thus we have that

$$f' = \pm\frac{1}{C}\sqrt{f^2 - C^2}.$$

In order to solve this differential equation we rewrite it as

$$\frac{df}{\sqrt{f^2 - C^2}} = \pm\frac{1}{C}dx$$

and integrate both sides, yielding

$$\operatorname{arccosh}(f/C) = \pm\frac{x}{C} + K_{\pm}.$$

There are two constants of integration ($K_\pm$) because the solution is given as the union of two functions, $x = g_\pm(z)$, one for each side of $x = 0$. Applying cosh *to both sides leaves*

$$f = C\cosh\left(\frac{x}{C} \pm K_\pm\right).$$

Here we have made use of the hyperbolic cosine (defined using the exponential function as $\cosh x = \frac{1}{2}(e^x + e^{-x})$*) and its inverse* arccosh*. Since we want these two functions to agree for $x = 0$, we define $K_+ = -K_- = K$. It is a good exercise to verify that the derivative of* arccosh x *is $1/\sqrt{x^2-1}$, and in doing so justify the above integration.*

Since $f(-a) = f(a) = R$, we must have that

$$\begin{cases} K = 0, \\ C\cosh(\frac{a}{C}) = R. \end{cases}$$

The second equation fixes C, but only implicitly.

The curve $y = C\cosh\left(\frac{x}{C} + K\right)$ is called a catenary*, and the surface obtained by rotating its graph about the x axis is called the* catenoid*. (See Figure 14.10.) We will discuss it in further detail later.*

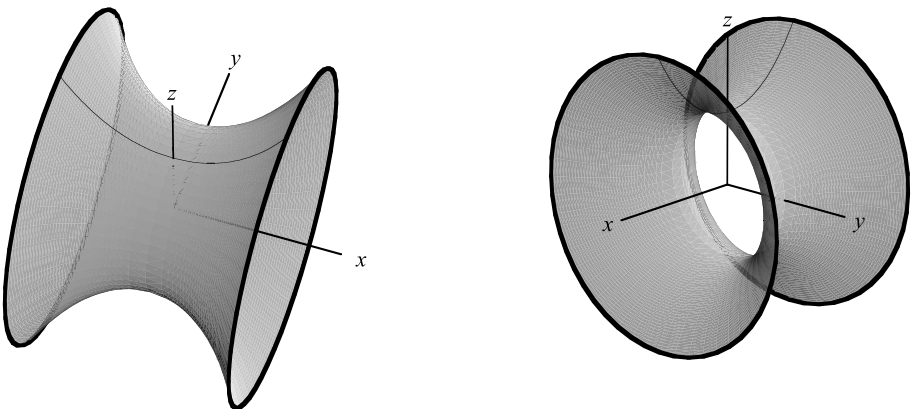


Fig. 14.10. Two points of view of the elastic sheet joining two rings with equal diameter.

It is rare in mathematics that solutions to analytic problems can be constructed and verified, at least approximately, with a toy. As discussed in the introduction to this section, some flexible wire and soapy water is all that is needed to do exactly that for this particular problem. Experimentation also allows us to explore the limitations of calculus of variations, some of which were mentioned in Section 14.2 (see the discussion regarding the optimal column). We encourage the reader to find a "good" recipe for

soapy water on the Internet, and to experiment with diverse shapes. We recommend that you try using the skeleton of a cube as a frame!

Soap bubbles give a simple way to answer several other questions. Here is one:

Example 14.12 The three cities and a soapy film. *Suppose that we have three cities located on a perfectly flat surface. We wish to join these three cities using the shortest possible route. How do we proceed?*

We begin by identifying the cities as three points A, B, and C. Next we construct a model consisting of two parallel plates made of transparent material, joined by perpendicular bars attached between the points corresponding to A, B, and C on each plate. The entire model is then dipped in soapy water and removed. The film joining the three bars will be a minimal surface. Its profile (when viewed through one of the transparent plates) describes the shortest network of roads between the three cities.

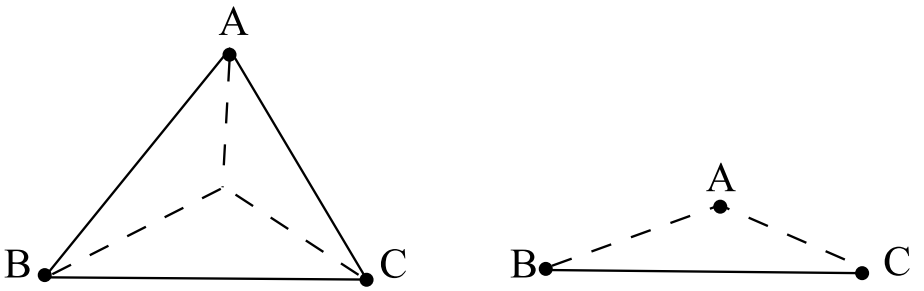


Fig. 14.11. The dotted lines indicate the shortest road network connecting the three cities at the corners of the triangle.

It is somewhat surprising to note that the shape of the soap film does not always correspond to the two shortest edges of the triangle. In fact, if the angles of the triangle ABC are all smaller than $\frac{2\pi}{3}$, we obtain a shorter network by passing through an intermediate point somewhere between the three cities, as shown at the left in Figure 14.11. In contrast, if one of the angles is greater than or equal to $\frac{2\pi}{3}$ then the two incident edges form the shortest network of roads, as shown at the right in Figure 14.11.

The intermediate point between the three cities that minimizes the net distance to all of the cities is called a Fermat point. *The position of the Fermat point can be found by inscribing an equilateral triangle along each side of the triangle, with its peak away from the interior of the triangle. Then, each corner of the triangle is joined with the peak of the equilateral triangle associated with the opposite face. The three lines will intersect at the Fermat point. It will be located inside the triangle only when the three angles of the triangle are all less than $\frac{2\pi}{3}$ (see Figure 14.12).*

Exercise 18 will show that the path constructed in this manner is indeed the shortest.

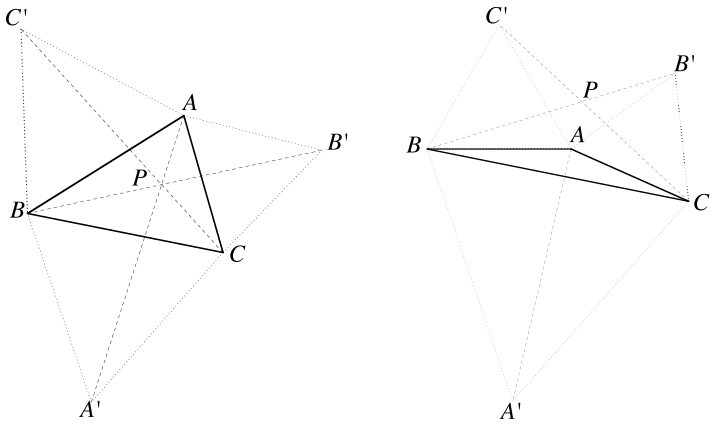


Fig. 14.12. Constructing a Fermat point.

This technique generalizes to networks of more than three cities. It may be used to find the shortest network of roads connecting them. The generalized problem is in fact quite old, and is known as the *minimum Steiner tree* problem.

The minimum Steiner tree problem. The problem can be stated as follows: given n points in the plane, find the shortest network connecting all of the points. It is relatively simple to convince yourself that such a network consists only of line segments (any curve can be replaced by a shorter polygonal line). Moreover, we can convince ourselves that the network will contain no closed triangles, since the above example showed how most efficiently to connect the corners of a triangle. A similar argument will show that the network can contain no closed polygons, and hence no cycles. In graph theory such a network is called a tree.

Minimal surfaces play a natural role in numerous applications. If you keep your eyes open, you will likely encounter a few of them in your studies.

14.9 Hamilton's Principle

Hamilton's principle is one of the greatest successes of calculus of variations. It allows problems from classical mechanics and several other domains of physics to be recast as variational problems.

According to Hamilton's principle, a system in motion will always follow the trajectory that optimizes the following integral:

$$A = \int_{t_1}^{t_2} L\,dt = \int_{t_1}^{t_2} (T - V)\,dt, \tag{14.30}$$

where L, called the Lagrangian, is the difference between the kinetic energy T of the system and its potential energy V. For historic reasons, this integral is called the *action* integral. Thus Hamilton's principle is also referred to as the *principle of least action*.[3]

In many systems, the kinetic energy depends only on the speed of an object (in the case of a moving object, the kinetic energy is given by $\frac{1}{2}mv^2$, where v is the speed of the object and m its mass), and the potential energy depends only on its position. In such systems the Lagrangian L is in fact a function $L = L(t, \mathbf{y}, \mathbf{y}')$, where $\mathbf{y} = \mathbf{y}(t)$ is the position vector and $\mathbf{y}' = \frac{d\mathbf{y}}{dt}$ the corresponding velocity vector. Thus we have an action integral of the form

$$A = \int_{t_1}^{t_2} L(t, \mathbf{y}, \mathbf{y}')\, dt,$$

where the time t now plays the role of the space variable x in Theorem 14.4.

The vector $\mathbf{y}$ describes the position of the entire system. Thus, the number of coordinates required depends on the details of the particular system being considered. If we are describing the motion of a particle in a plane or space, then we would have $\mathbf{y} \in \mathbb{R}^2$ or $\mathbf{y} \in \mathbb{R}^3$, respectively. It the system contains two particles moving in the plane we would have $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and therefore $\mathbf{y} \in \mathbb{R}^4$, where $\mathbf{y}_1$ represents the position of the first particle and $\mathbf{y}_2$ the position of the second. In general, a system whose position is fully described by a vector $\mathbf{y} \in R^n$ is said to have n *degrees of freedom*. (See Chapter 3 for a discussion of degrees of freedom in another context.)

If $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the Lagrangian takes the form $L = L(t, y_1, \ldots, y_n, y_1', \ldots, y_n')$. The Euler–Lagrange equations can be generalized to describe problems with n degrees of freedom. For example, the form discussed below describes a system with two degrees of freedom.

Theorem 14.13 *Consider the integral*

$$I(x, y) = \int_{t_1}^{t_2} f(t, x, y, x', y')\, dt. \tag{14.31}$$

The pair (x^*, y^*) minimizes this integral only if (x^*, y^*) is a solution to the following system of Euler–Lagrange equations:

$$\frac{\partial f}{\partial x} - \frac{d}{dt}\left(\frac{\partial f}{\partial x'}\right) = 0, \qquad \frac{\partial f}{\partial y} - \frac{d}{dt}\left(\frac{\partial f}{\partial y'}\right) = 0.$$

[3] It is difficult to understand exactly why nature behaves in such a manner as to minimize the difference between kinetic and potential energies. Why this difference rather than any of the many other possible differences? Most physics texts are surprisingly silent on this point. In his introductory physics courses, Feynman devotes an entire chapter to the principle of least action. His amazement with the subject stems not from the fact that nature minimizes the difference between kinetic and potential energies, but rather from the existence of such a simple formula that describes physical interactions. For those who wish to explore the connection between calculus of variations and physics further, Feynman's course is an excellent starting point [5].

In our previous examples the behavior of the solution was fixed by the boundary conditions of the function y. For example, the constants of integration that arise in finding the cycloid are determined by knowing that it starts at (x_1, y_1) and ends at (x_2, y_2). In physics, rather than defining the starting and ending points of a particle, it is more common to describe the initial conditions of the system by defining both the position and velocity of the particle. We demonstrate this approach in the following example.

Example 14.14 ***Projectile motion.*** *As an example of Hamilton's principle we consider the trajectory of a projectile of mass m. We suppose that air friction is negligible. The projectile is launched at time $t_1 = 0$ from an initial position $(x(0), y(0)) = (0, h)$ with an initial velocity $\mathbf{v}_0$ at an angle θ above the horizontal. Using the angle of the velocity vector, the components will be $(v_{0x}, v_{0y}) = |\mathbf{v}_0|(\cos\theta, \sin\theta)$.*

The action of such a projectile (see (14.30)*) is described by*

$$A = \int_{t_1}^{t_2} L(t, x, y, x', y')dt = \int_{t_1}^{t_2} (T - V)dt,$$

where $'$ denotes the time derivative. The kinetic energy of the projectile is $T = \frac{1}{2}m|\mathbf{v}|^2$ and the potential energy is $V = mgy$. Since the square of the magnitude of the velocity vector is given by $|\mathbf{v}|^2 = (x')^2 + (y')^2$, the integral may be rewritten in terms of the variables x, y, x', and y' as

$$A = \int_{t_1}^{t_2} m\left(\tfrac{1}{2}(x')^2 + \tfrac{1}{2}(y')^2 - gy\right) dt.$$

The equations describing the motion of the projectile are found with the help of the two-dimensional Euler–Lagrange equations described in Theorem 14.13, where the Lagrangian $L = m\left(\frac{1}{2}(x')^2 + \frac{1}{2}(y')^2 - gy\right)$ is the function whose integral is to be optimized. We use equivalently $f = \frac{L}{m}$. The first equation yields

$$0 = \frac{\partial f}{\partial x} - \frac{d}{dt}\left(\frac{\partial f}{\partial x'}\right) = -\frac{d}{dt}(x') = -x'', \tag{14.32}$$

where the second equality follows from the fact that L is independent of x. Since the second derivative of x is zero, its first derivative must be a constant. We already know the value of this constant: it is the horizontal component of the initial velocity of the particle, v_{0x}. Thus

$$x' = v_{0x} = |\mathbf{v}_0|\cos\theta.$$

Thus we have demonstrated a well-known physical fact: in the absence of friction, a thrown object has a constant horizontal speed. A second integration gives the x coordinate of the particle as a function of time: $x = v_{0x}t + a$. The constant of integration a can also be determined using the initial conditions. Given that $x(0) = 0$, it follows that $a = 0$ and therefore

$$x = v_{0x}t = |\mathbf{v}_0|t\cos\theta.$$

The second Euler–Lagrange equation leads to

$$0 = \frac{\partial f}{\partial y} - \frac{d}{dt}\left(\frac{\partial f}{\partial y'}\right) = -g - \frac{d}{dt}y' = -g - y'',$$

which simplifies to

$$y'' = -g. \tag{14.33}$$

Thus, in the vertical direction the particle is subject to a constant downward force due to gravity. Integrating this once yields

$$y' = -gt + b,$$

where the constant of integration b is fixed by the initial vertical velocity v_{0y} of the particle. Indeed, at $t_1 = 0$, the vertical velocity is $y' = |\mathbf{v}_0|\sin\theta$. Thus it follows that

$$y' = -gt + |\mathbf{v}_0|\sin\theta.$$

Integrating again yields the vertical position of the particle as a function of time, yielding

$$y = \frac{-gt^2}{2} + |\mathbf{v}_0|t\sin\theta + c.$$

The constant c is equal to the initial y coordinate of the particle, and therefore $c = h$. Thus the complete trajectory of the particle is given by

$$x = v_{0x}t = |\mathbf{v}_0|t\cos\theta \qquad \text{and} \qquad y = \frac{-gt^2}{2} + |\mathbf{v}_0|t\sin\theta + h. \tag{14.34}$$

As we will now show, these equations parameterize a parabola when $\theta \neq \pm\frac{\pi}{2}$. Indeed, if $\cos\theta \neq 0$, then $t = x/(|\mathbf{v}_0|\cos\theta)$. This allows the coordinate y to be rewritten as a function of x, yielding

$$y = \frac{-gx^2}{2|\mathbf{v}_0|^2 cos^2\theta} + x\tan\theta + h,$$

the anticipated parabola. The case $\cos\theta = 0$ corresponds to a vertical launch (either upward or downward), and the corresponding trajectory is simply a vertical line.

Note that both (14.32) *and* (14.33) *are the equations that we would have arrived at had we applied Newton's laws. Here they appeared naturally as a consequence of Hamilton's principle.*

Example 14.15 ***Spring motion.*** *This simple example is explored in Exercise 14.*

Example 14.16 ***Systems in equilibrium.*** *Systems in equilibrium can be easily simplified. The configuration of such systems remains constant for all time, and thus the Lagrangian is a constant as a function of time. If we want the action integral* $\int_{t_1}^{t_2} L\,dt$ *to attain an extremum, then the underlying Lagrangian must itself have some extremum. We will see several examples of this in Section 14.10: suspended cables, self-supporting arches, and liquid mirrors.*

The reformulation of physical laws into variational problems using Hamilton's principle is not limited to classical mechanics. In fact, the principle of least action plays an important role in quantum mechanics, electromagnetism, general relativity, and in both classic and quantum field theory.

14.10 Isoperimetric Problems

Isoperimetric problems are an important class of variational problems. They represent problems in which the optimization is subject to one or more constraints.

The term "isoperimetric problems" likely does not make you think of optimization with constraints. However, they have been given this name due to their origin, a problem from antiquity. Given a fixed perimeter, the problem asked to find the geometric figure that encloses the largest possible area. The answer is, perhaps intuitively, the circle. The techniques developed in this section show how to use calculus of variations to answer this and other similar questions. We begin by presenting a variant of this problem.

Example 14.17 *We wish to maximize the integral*

$$I = \int_{x_1}^{x_2} y\,dx$$

under the constraint that

$$J = \int_{x_1}^{x_2} \sqrt{1 + (y')^2}\,dx = L,$$

where L *is a constant that represents the length of the curve. The perimeter is therefore* $L + (x_2 - x_1)$*. The first integral computes the area under the curve* $y(x)$ *between the points* x_1 *and* x_2*, while the second computes its length.*

A review of Lagrange multipliers. For functions with real variables, the problem of optimization with constraints it solved using the classic method of Lagrange multipliers. We discuss the broad strokes of the technique. We wish to find the extrema of a two-variable function $F = F(x, y)$ under the constraint $G(x, y) = C$. We can imagine walking along the contour of points where $G(x, y) = C$. Since the contours of F and G are generally distinct, walking along the $G = C$ contour crosses many contours of F. Thus, we can increase or decrease the value of F by walking along this contour.

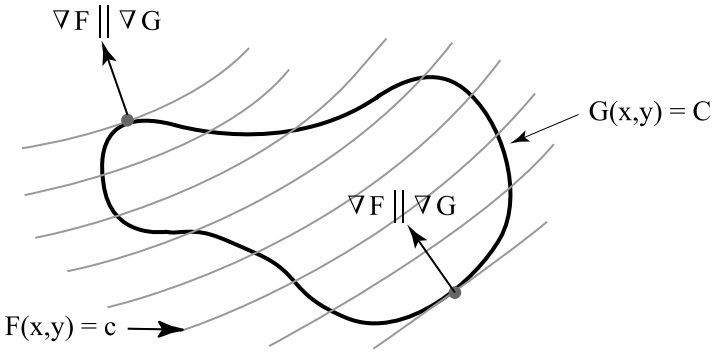


Fig. 14.13. Explaining the role of Lagrange multipliers.

When the contour $G = C$ touches tangentially a contour of F, then movements in both directions along the $G = C$ contour change the value of F in the same direction. Thus, such a point corresponds to a local extremum of the constrained optimization. More precisely, extrema occur where the gradients ∇F and ∇G are parallel; in other words, where $\nabla F \parallel \nabla G$ and therefore $\nabla F = \lambda \nabla G$ for some real λ. This λ is known as a Lagrange multiplier. Figure 14.13 shows a graphical depiction of the intuition behind this technique. The constraint $G = C$ is shown as a black closed curve, while several contours of F are shown in gray. Two constrained extrema can be found at the indicated points, both occurring where the contours are tangential. Thus, for functions of real variables, optimization with a constraint amounts to solving

$$\begin{cases} \nabla F = \lambda \nabla G, \\ G(x, y) = C. \end{cases}$$

This technique can be generalized to handle multiple constraints. As shown without proof in the following theorem, the technique may also be extended to constrained variational problems.

Theorem 14.18 *A function $y(x)$ which is an extremum of the integral $I = \int_{x_1}^{x_2} f(x, y, y')\, dx$ under the constraint $J = \int_{x_1}^{x_2} g(x, y, y')\, dx = C$ is a solution to the Euler–Lagrange differential equation associated with the functional*

$$M = \int_{x_1}^{x_2} (f - \lambda g)(x, y, y') dx.$$

Thus we must resolve the following system:

$$\begin{cases} \dfrac{d}{dx}\left(\dfrac{\partial(f - \lambda g)}{\partial y'}\right) = \dfrac{\partial(f - \lambda g)}{\partial y}, \\ J = \displaystyle\int_{x_1}^{x_2} g(x, y, y')\, dx = C. \end{cases} \tag{14.35}$$

If f and g are independent of x we can again appeal to Beltrami's identity and instead solve the following system:

$$\begin{cases} y' \dfrac{\partial(f-\lambda g)}{\partial y'} - (f-\lambda g) = K, \\ J = \displaystyle\int_{x_1}^{x_2} g(x,y,y')\,dx = C. \end{cases} \tag{14.36}$$

Example 14.19 ***A suspended cable.*** *Suppose that we have a cable suspended between two points, for example a high-voltage power line suspended between two poles (Figure 14.14). Intuitively, we know that if the cable is longer than the distance between the two points it will sag and form a curve. The constrained Euler–Lagrange equations will allow us to deduce that this curve is a catenary and gives its exact equation. The functional to minimize will be that of the potential energy of the cable. Since the cable is stationary and has no kinetic energy, this is another example of Hamilton's principle at work (see Example 14.16).*

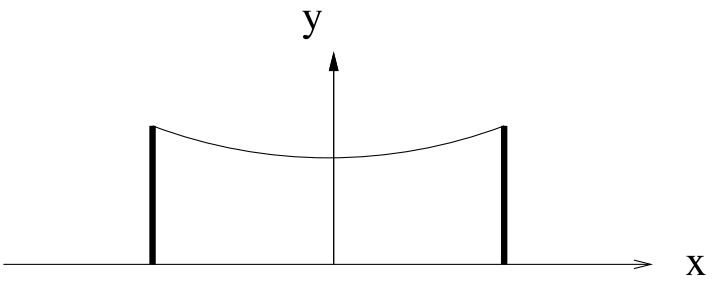


Fig. 14.14. What equation describes the shape of this suspended cable?.

Suppose that the cable has linear density σ (where linear density is mass per unit of length) and that L is its length. Since the potential energy of a mass m at height y is mgy, the potential energy of an infinitesimal piece of cable of length ds at height y is therefore $\sigma g y\,ds$. Thus, the potential energy of the entire cable is given by

$$I = \sigma g \int_0^L y\,ds,$$

or equivalently,

$$I = \sigma g \int_{x_1}^{x_2} y\sqrt{1+(y')^2}\,dx. \tag{14.37}$$

The constraint to be satisfied is that of the length L of the cable. Thus, we must have that

$$J = \int_{x_1}^{x_2} \sqrt{1 + (y')^2}\, dx = L.$$

This problem is therefore an isoperimetric problem.

Since neither $f = y\sqrt{1 + (y')^2}$ *nor* $g = \sqrt{1 + (y')^2}$ *depends on* x*, we can use the Beltrami identity from Theorem 14.18 and apply it to the function*

$$F = \sigma g y\sqrt{1 + (y')^2} - \lambda\sqrt{1 + (y')^2} = (\sigma g y - \lambda)\sqrt{1 + (y')^2}.$$

Substituting the above function into the Beltrami identity

$$y'\frac{\partial F}{\partial y'} - F = C$$

yields

$$\frac{(y')^2(\sigma g y - \lambda)}{\sqrt{1 + (y')^2}} - (\sigma g y - \lambda)\sqrt{1 + (y')^2} = C,$$

which may be simplified to

$$-\frac{\sigma g y - \lambda}{\sqrt{1 + (y')^2}} = C.$$

Solving for y' *yields*

$$\frac{dy}{dx} = \pm\sqrt{\left(\frac{\sigma g y - \lambda}{C}\right)^2 - 1}. \tag{14.38}$$

Like that of the brachistochrone, this differential equation is separable, meaning that the parts depending on x *and* y *may be moved to opposite sides of the relation:*

$$dx = \pm\frac{dy}{\sqrt{\left(\frac{\sigma g y - \lambda}{C}\right)^2 - 1}}.$$

This method allows us to find x *as a function of* y*. However, knowing the rough form of the solution (Figure 14.14), we see that we will need two functions to describe it in this manner, one for the left half and another for the right.*

As before, this approach allows us to integrate the two sides of the differential equation, leading to

$$x = \pm\frac{C}{\sigma g}\operatorname{arccosh}\left(\frac{\sigma g y - \lambda}{C}\right) + a_{\pm},$$

where $a_{\pm}$ *is a constant of integration. Thus*

$$x - a_{\pm} = \pm\frac{C}{\sigma g}\operatorname{arccosh}\left(\frac{\sigma g y - \lambda}{C}\right).$$

Since the function cosh *is even (*$\cosh x = \cosh(-x)$*), it follows that*

$$\frac{\sigma g y - \lambda}{C} = \cosh \frac{\sigma g}{C}(x - a_{\pm}).$$

Finally, we arrive at

$$y = \frac{C}{\sigma g} \cosh \frac{\sigma g}{C}(x - a_{\pm}) + \frac{\lambda}{\sigma g}.$$

As in our earlier discussion in Example 14.11, it follows that $a_+ = a_- = a$ *in order for the two equations to meet smoothly in the middle.*

Thus we see that a suspended chain (assumed to be perfectly uniform and flexible) will naturally take the form of a catenary as in Example 14.11. In order to find the values of C, a, *and* λ *we must solve the system of three equations implied by the boundary conditions:*

$$\begin{cases} J = L, \\ y(x_1) = y_1, \\ y(x_2) = y_2. \end{cases}$$

Note that in some cases it is very difficult to express the values of C, a, *and* λ *in terms of* L, x_1, y_1, x_2, *and* y_2. *In these cases it is necessary to use numerical methods.*

Like the cycloid, the catenary is a shape found throughout nature. In fact, it is even the name given to the system of electric cables suspended above railroad tracks. We also find inverted catenaries: this is the optimal form for a self-supporting arch. Additionally, in Section 14.8 we saw that a soap bubble stretched between two rings is a catenoid, that is, the surface of revolution with a catenary as generatrix.

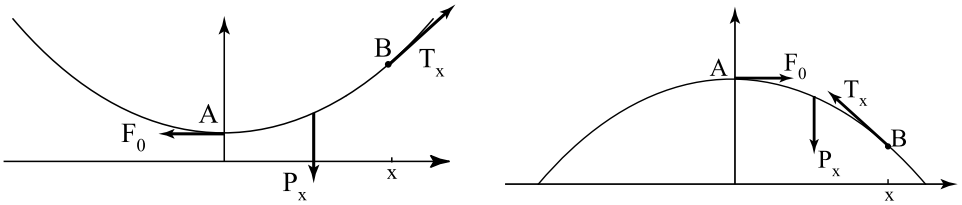
Example 14.20 Self-supporting arch. *The use of arches as a weight-bearing architectural structure dates back probably to Mesopotamia. Almost all civilizations and epochs have left examples of this long-lasting structure. Many forms exist, but one can be singled out for its properties: it is the catenary arch. We will say that an arch is self-supporting if the forces responsible for its equilibrium originate from its own weight and are transmitted tangentially to the curve defined by the arch and if other stress forces in the building material can be neglected.*[4] *An example of such an arch is shown*

[4]This is certainly not the case for all arches. Let us imagine an extreme case in which two (vertical) walls are separated by exactly the width of three bricks. This allows to squeeze in three bricks and, if the pressure on them is sufficient (that is, if the fit is extremely tight), the bricks could stand in the void, without falling. These three bricks form a horizontal arch. The middle brick should fall due to gravity (a vertical force) but is held there by the other two bricks. The latter are in contact with the walls and are subjected only to horizontal forces (from the wall) and one vertical force (gravity). The internal structure of the material must transform the horizontal forces into vertical ones on the middle brick. These forces due to (minute) molecular deformation of the material are known as stress forces. They give rise to compression, shear, and torsion in the material. Many construction materials, including stone and concrete, resist well under compression, but not under shear and torsion. An arch minimizing stress within its components can therefore be useful.

in Figure14.15(b). We will not use calculus of variations in the example, but rather we will use an indirect method to show that the inverted catenary does in fact maximize the potential energy of the arch under the constraint that the length is fixed.

Rather than approaching the problem as in Example 14.19, we will work backward. We will compute the shape of a self-supporting arch and show that it satisfies the Euler–Lagrange equation associated with (14.37) *under the constraint that the length is fixed.*

We will use nearly the same model as that of the suspended cable. As shown in Figure 14.15, they are effectively the same and agree up to symmetry. Consider a



(a) A suspended cable (b) A self-supporting arch

Fig. 14.15. Modelling a suspended cable and a self-supporting arch.

section of a chain or an arch that is above the segment $[0, x]$ *of the* x *axis. Since the section is in equilibrium, then the net sum of forces acting on it must be zero. For the suspended chain, there are three forces at work: the weight* P_x*, the tension* F_0 *at the point* $(0, y(0))$*, and the tension* T_x *at the point* $(x, y(x))$*. In the case of the arch, there are three similar forces in play except that the forces* F_0 *and* T_x *are inverted. The force* $F_0 = (f_0, 0)$ *is constant, but both* P_x *and* T_x *are dependent on* x*. Gravity acts in the vertical direction; thus* $P_x = (0, p_x)$*. Let* $T_x = (T_{x,h}, T_{x,v})$*. Saying that the sum of forces must be zero yields the following equations:*

$$\begin{cases} T_{x,h} = -f_0, \\ T_{x,v} = -p_x. \end{cases} \tag{14.39}$$

Let θ *be the angle between the tangent of the curve at* B *and the horizontal. Then it follows that*

$$\begin{cases} T_{x,h} = |T_x| \cos\theta, \\ T_{x,v} = |T_x| \sin\theta, \end{cases}$$

and

$$y'(x) = \tan\theta.$$

Let σ *be the linear density,* g *the gravitational constant, and* $L(x)$ *the length of the section of curve we are considering. Then* $p_x = -L(x)g\sigma$*. Putting these data into* (14.39) *yields*

$$\begin{cases} |T_x| \cos\theta = -f_0, \\ |T_x| \sin\theta = L(x)\sigma g. \end{cases}$$

Dividing the second equation by the first leaves

$$\tan\theta = y' = -\frac{\sigma g}{f_0} L(x).$$

We take the derivative, arriving at

$$y'' = -\frac{\sigma g}{f_0} L'(x) = -\frac{\sigma g}{f_0}\sqrt{1+y'^2}, \tag{14.40}$$

using the fact that $L'(x) = \sqrt{1+y'^2}$. *(Recall that in Example 14.1 the infinitisemal increase in the length of a curve was computed to be* $ds = \sqrt{1+y'^2}dx$. *This means that the derivative of this length is* $L' = \frac{ds}{dx}$.*)*

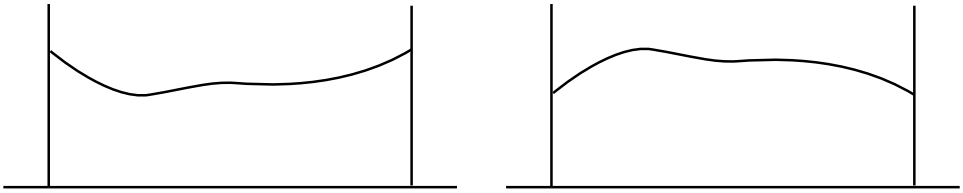
It is an easy exercise in differential calculus to check that

$$y(x) = -\frac{f_0}{\sigma g}\cosh\left(\frac{\sigma g}{f_0}(x - x_0)\right) + y_0$$

satisfies the equation (14.40) *above. To get the maximum in* $x = 0$, *one has to set* $x_0 = 0$. *The curve then intercepts the* x *axis in* $\pm x_1$, *where* x_1 *depends on* y_0. *This constant* y_0 *is determined by the requirement that the length of the curve between* $-x_1$ *and* x_1 *be equal to* L. *The remarkable property of* $y(x)$ *is that it is* also *a solution of the Beltrami equation* (14.38) *used for the cable if the constant* C *is set to* f_0 *and the Lagrange multiplier* λ *to* $\sigma g y_0$. *(Again checking this is a straightforward exercise in calculus!) The solution* $y(x)$ *is therefore a critical point of the functional potential energy* (14.37) *under the constraint of fixed length. Or in other words, the self-supporting arch is a critical point of the potential energy, under the constraint of a given arch length!*

We are sure that it is not a minimum. Is it a maximum under the constraint that the arch length is fixed? It is easy to convince ourselves that this is the case. Here again we will make use of the earlier solution to the suspended cable. In that case, all other solutions (for example, that shown in Figure 14.16(a)) had a higher potential energy than the catenary. By symmetry, all forms other than the inverted catenary (for example that of Figure 14.16(b)) must have a lower potential energy.

Example 14.20 shows that the catenary arch has the lowest possible internal stress forces. This is in contrast to a circular arch, where portions of the arch nearer the peak endure higher stresses than those at the base. It is not surprising that this shape is used in architecture. Perhaps the most famous example is the "Gateway Arch" of St. Louis, Missouri. Similarly, the arches of many buildings have a catenary shape. Each winter in Jukkasjärvi, Sweden, sees the construction of the Icehotel, built entirely of ice. Since ice is brittle, it becomes important to minimize stresses. It is for this reason that the builders of the Icehotel have chosen to construct most arches in the form of a catenary.



(a) A suspended cable (b) A self-supporting arch

Fig. 14.16. Another possible form for a suspended cable and a self-supporting arch.

For the same reason, the optimal profile for constructing an igloo is a catenary. One may wonder whether the Inuits knew this intuitively long before the rest of us?

The famous Catalan architect Antoni Gaudí knew not only of the properties of the catenary arch, but also of its intimate ties with the shape taken by cables under their own weight. To study complex system of arches where, for example, the feet of some rest on the heads of others, he devised the following system. He would attach to the ceiling small chains tied to each other the way the arches were meant to be. He would then look at the resulting structure through a mirror on the floor in order to "read" the form to give to the arches he had in mind.

14.11 Liquid Mirrors

In order to focus light onto a single point, the mirrors in telescopes must have the shape of a paraboloid of revolution (see section 15.2.1). The precise construction of such mirrors is therefore very important in astronomy. The difficulties in constructing such mirrors are enormous, since they are sometimes very large (the Hale telescope on Mount Palomar is more than 5 m in diameter, and it is not even the largest!).

As a way of getting around these difficulties, some physicists had the idea of building liquid mirrors, obtained by rotating a round container of fluid at a constant speed. The first to describe this idea was the Italian Ernesto Capocci in 1850. In 1909 the American Robert Wood built the first liquid telescopes with mercury. Since the quality of the image was low, the idea was not seriously pursued until 1982, when the team of Ermanno F. Borra, at Laval University (Quebec), started working actively on the project. Now several teams worked on the project, including that of Paul Hickson, at the University of British Columbia. The different technical difficulties were mastered, one after the other, and the liquid telescope was here to stay. The paper [6] gives a history of the subject.

Before going further, let us start by explaining the principle. When a liquid contained in a cylinder rotates at constant speed, its shape is a paraboloid of revolution, so the exact shape of a telescope mirror! We will prove this fact with the help of calculus of variations. Such mirrors can be constructed using any reflective liquid, such as mercury.

There are many advantages to this technology: these mirrors are much cheaper than traditional mirrors and they nonetheless have an extremely high quality surface finish. As such, it is possible to construct very large liquid mirrors. Moreover, it is very easy to change the focal length of these mirrors, simply by adjusting the speed of rotation. The largest problem with these mirrors is that it is impossible to orient them in any direction other than vertical. Thus, telescopes using such mirrors are able to observe only the portion of the sky directly above them, unless we use additional mirrors.

Among the problems solved by the researchers we find elimination of vibrations; control of the rotation speed, which must be perfectly constant; and elimination of atmospheric turbulence near the surface of the mirror. Since we cannot orient the telescope to counter the rotation of the Earth (see Exercise 18 of Chapter 3), the observed celestial objects leave traces of light, similar to what you see on night photos. Borra's team solved the problem by replacing the traditional film by a CCD (Charge Couple Device, which, for instance, replaces film in digital cameras), and the technique is called the sweeping technique. This same team also built liquid mirrors in the 1990s with diameter up to 3.7 m that produced images of excellent optic quality.

Near Vancouver, Canada, Hickson's team built a telescope equipped with a liquid mirror with a diameter of six meters, the Large Zenith Telescope (LZT). Even if we cannot orient them, these telescopes are useful. Indeed, when one wants to study the density of far-away galaxies, the zenith is a direction as interesting as any other. During the time the telescope with a liquid mirror is being used, the other more-expensive telescopes can be used for other purposes.

Now that the images produced by liquid mirror telescopes are very satisfactory, there are numerous new ambitious projects. Among these let us mention the ALPACA project

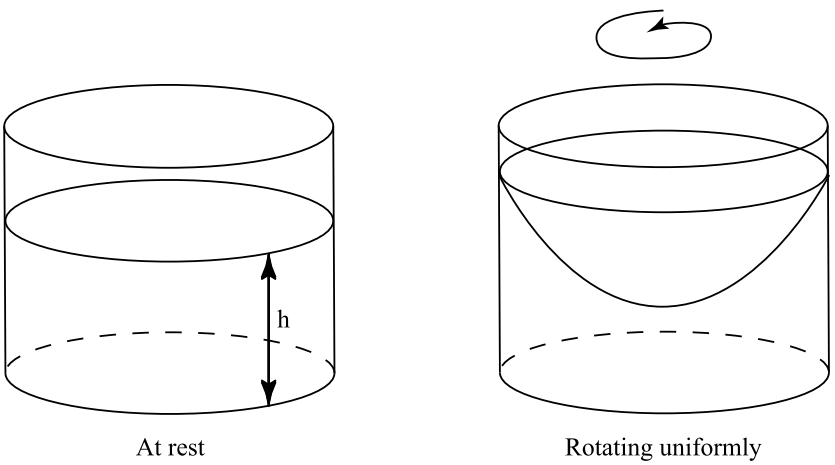


Fig. 14.17. A liquid mirror.

(Advanced Liquid-Mirror Probe for Astrophysics, Cosmology and Asteroids) concerned with the installation of a telescope with a liquid mirror of diameter 8 m on the summit of a Chilean mountain. Exercise 5 of Chapter 15 describes the disposition of the mirrors of this future telescope: only the primary mirror is liquid, while the secondary and tertiary mirrors are glass. And Roger Angel, from the University of Arizona, is the manager of an international team that with the support of NASA (National Aeronautics and Space Administration) is developing plans for a telescope with a liquid mirror that could be installed on the moon! Indeed, telescopes with liquid mirrors are much easier to transport than large glass mirrors. Also, a telescope on the moon would profit from the absence of atmosphere, which on Earth, produces fuzzy images. Moreover, due to the low gravity and the absence of air, which eliminates turbulence close to the surface of the mirror, a project for a mirror of 100 m diameter is being considered! Borra's team has already made progress in replacing mercury, which freezes at $-39°$ C by an ionic liquid that does not evaporate and stays liquid above $-98°$ C.

Borra's team is also working on techniques to deform liquid mirrors so that they can observe in directions other than straight up. Since mercury is very heavy, efforts are being made to replace it with a magnetic liquid (called a *ferrofluid*) that can easily be deformed by an external magnetic field. Unfortunately, ferrofluids are not reflective. The team at Laval University resolved this problem through the use of a thin film of silver nanoparticles called MELLF (MEtal Liquid Like Film), which is very reflective and conforms to the surface of the underlying ferrofluid. Research into these mirrors continues.

Using Hamilton's principle it is possible to prove that the surface of a liquid mirror is a paraboloid of revolution.

Proposition 14.21 *We consider a vertical cylinder of radius R that is full of liquid up to a height h. If the liquid in the cylinder is rotated at a constant angular velocity ω about its axis, then the surface of the liquid will be a paraboloid of revolution whose axis is the axis of the cylinder. The form of the paraboloid is independent of the density of the liquid.*

Proof. We will use the cylindrical coordinates (r, θ, z), where $(x, y) = (r\cos\theta, r\sin\theta)$. The liquid is in a cylinder of radius R. We assume that the surface of the liquid is a surface of revolution described by $z = f(r) = f(\sqrt{x^2 + y^2})$. Identifying the shape of this surface amounts to finding the function f. In order to do this, we apply Hamilton's principle. Since the system is in equilibrium, this is done by finding the extremum of the Lagrangian $L = T - V$ (see Example 14.16).

Calculating the potential energy V. We divide the liquid into infinitesimally small elements of volume centered at (r, θ, z) with side lengths dr, $d\theta$, and dz. Thus the volume of such an element is $dv \approx r\,dr\,d\theta\,dz$. Suppose that the density of the liquid is σ. Then the mass of such an element is given by $dm \approx \sigma r\,dr\,d\theta\,dz$. Since the height of the element is z, its potential energy is given by $dV = \sigma g r\,dr\,d\theta\,z\,dz$.

We now sum across all of the elements to determine the total potential energy:

$$V = \int dV = \sigma g \left(\int_0^{2\pi} d\theta \right) \cdot \int_0^R \left(\int_0^{f(r)} z \, dz \right) r \, dr$$
$$= 2\sigma g \pi \int_0^R \frac{z^2}{2} \Big|_0^{f(r)} r \, dr$$
$$= \sigma g \pi \int_0^R (f(r))^2 r \, dr.$$

Calculating the kinetic energy T. If u represents the speed of an element of volume, then its kinetic energy is given by $dT = \frac{1}{2}u^2 dm$, where $dm \approx \sigma r \, dr \, d\theta \, dz$ is its mass. Since the angular speed ω is constant, the speed of an element at a distance r from the axis is given by $u = r\omega$. Thus the total kinetic energy of the system is

$$T = \int dT = \frac{1}{2}\sigma\omega^2 \left(\int_0^{2\pi} d\theta \right) \cdot \int_0^R \left(\int_0^{f(r)} dz \right) r^3 dr$$
$$= \sigma\pi\omega^2 \int_0^R f(r) r^3 \, dr.$$

Applying Hamilton's principle. Recall that Hamilton's principle aims to minimize the value of the integral $\int_{t_1}^{t_2} (T - V) dt$. Since we are in equilibrium, this integral will be minimized when the integrand $T - V$ is itself minimized. We have

$$T - V = \sigma\pi \int_0^R (f(r)\omega^2 r^3 - g(f(r))^2 r) \, dr,$$

which is of the form

$$\sigma\pi \int_0^R G(r, f, f') \, dr$$

with $G(r, f, f') = f(r)\omega^2 r^3 - g(f(r))^2 r$.

The minimization of I is subject to one constraint: the volume of the liquid must remain constant at $\text{Vol} = \pi R^2 h$. Since the surface of the liquid is a surface of revolution, this volume is given by

$$\text{Vol} = \int_0^{2\pi} d\theta \cdot \int_0^R \left(\int_0^{f(r)} dz \right) r \, dr = 2\pi \int_0^R r f(r) \, dr. \tag{14.41}$$

Theorem 14.18 allows us to resolve this problem under the volume constraint. We must replace G with the function $F(r, f, f') = \sigma\omega^2 f(r) r^3 - \sigma g (f(r))^2 r - 2\lambda r f(r)$. The Euler–Lagrange equation for F is

$$\frac{\partial F}{\partial f} - \frac{d}{dr}\left(\frac{\partial F}{\partial f'} \right) = 0.$$

Since the function F does not explicitly depend on f', in this particular case the equation may be simplified to $\frac{\partial F}{\partial f} = 0$, or

$$\sigma\omega^2 r^3 - 2\sigma g r f(r) - 2\lambda r = 0.$$

The function f is therefore

$$f(r) = \frac{\omega^2 r^2}{2g} - \frac{\lambda}{\sigma g}, \tag{14.42}$$

which describes a parabola. There are several interesting properties to note at this point. The form of the parabola depends only on the speed of the angular rotation and gravity, since the coefficient of r^2 is $\frac{\omega^2}{2g}$. It is somewhat surprising to note that the density σ of the liquid has absolutely no impact on the shape of the parabola. The term $\frac{\lambda}{\sigma g}$ represents a vertical translation of the parabola. Its specific value is determined by the volume of the liquid, which remains fixed.

It remains to calculate the value of λ using the constraint $\text{Vol} = \pi R^2 h$. The expressions for the volume of the liquid (14.41) and the profile f of the liquid (14.42) allow us to obtain

$$\begin{aligned}
\text{Vol} &= 2\pi \int_0^R \left(\frac{\omega^2 r^2}{2g} - \frac{\lambda}{\sigma g} \right) r \, dr \\
&= 2\pi \left[\frac{\omega^2 r^4}{8g} - \frac{\lambda r^2}{2\sigma g} \right]_0^R \\
&= \frac{\pi\omega^2 R^4}{4g} - \frac{\pi\lambda R^2}{\sigma g}.
\end{aligned}$$

Since the volume is constant ($\pi R^2 h$), this allows us to fix the constant λ as

$$\lambda = \frac{\sigma\omega^2 R^2}{4} - \sigma g h$$

and to give f its final form

$$f(r) = \frac{\omega^2 r^2}{2g} - \frac{\omega^2 R^2}{4g} + h.$$

We now have the equation defining the precise form of the paraboloid of revolution created by spinning the liquid at a constant speed. □

14.12 Exercises

The fundamental problem of calculus of variations

1. An airplane[5] must travel from point A to point B, both at zero altitude and separated from each other by a distance d. In this problem we assume that the surface of the Earth is actually a plane. An airplane costs more money to fly at a lower altitude than at a higher one. We wish to minimize the cost of a trajectory between the points A and B. The trajectory will be a curve through the vertical plane passing through the points A and B. The cost of traveling a distance ds at an altitude h is constant and given by $e^{-h/H}ds$.

(a) Choose a coordinate system that is well suited to this problem.

(b) Give an expression for the cost of the voyage between the points A and B, and express the problem of minimizing this cost as a variational problem.

(c) Derive the associated Euler–Lagrange or Beltrami equation, as appropriate.

The brachistochrone

2. What is the specific equation describing the cycloid on which a point mass will travel when falling between the points $(0, 0)$ and $(1, 2)$ in a minimum amount of time? How long will the particle take to travel this path? Use mathematical software to perform these calculations.

3. Calculate the area beneath an arch with a cycloidal profile. Is it related to the area of the circle that generated the cycloid?

4. Verify that the vector tangent to the cycloid $(a(\theta - \sin\theta), a(1 - \cos\theta))$ is vertical at $\theta = 0$.

5. Find out whether real half-pipes have a cycloidal profile.

6. **(a)** Let (x_1, y_1) and (x_2, y_2) be such that the brachistochrone between the two departs (x_1, y_1) vertically and arrives at (x_2, y_2) horizontally. Show that $\frac{y_2 - y_1}{x_2 - x_1} = \frac{2}{\pi}$.

(b) Show that if $\frac{y_2 - y_1}{x_2 - x_1} < \frac{2}{\pi}$, then the point mass traveling along a brachistochrone between the two points descends lower than y_2 before arriving at the point (x_2, y_2). Verify that such a solution still exists even for $y_1 = y_2$ (in the absence of friction). That is, the quickest path between two horizontal points descends below them.

7. **(a)** Calculate the time taken to descend from $(0, 0)$ to $P_\theta = (a(\theta - \sin\theta), a(1 - \cos\theta))$ by traveling along the straight line between the points. (Use equation (14.2) and replace y by the equation for the straight line.)

(b) Compare this with the time taken to travel along the brachistochrone between the two points, and show that the straight-line path always takes longer.

(c) Show that the time taken to travel along the straight line between the points tends to infinity as the line approaches being horizontal.

[5] This problem has been taken from course notes by Francis Clarke.

8. We are looking for the fastest way to travel between the point $(0,0)$ and a point on the vertical line $x = x_2$ to its right. We know that we must follow the path of a cycloid (14.19), but we do not know for which value of a.

(a) For a fixed a, show that the time taken to travel along the cycloid is $\sqrt{\frac{a}{g}}\theta$, where θ is determined implicitly by $a(\theta - \sin\theta) = x_2$.

(b) Show that the minimum occurs when $\theta = \pi$. In other words, show that the minimum occurs when the cycloid intersects the line $x = x_2$ horizontally.

An isochronous device

9. Here we explore another interesting property of the inverted catenary. In order to solve this problem you will have to draw inspiration from Huygens's isochronous device, as explored in Section 14.7.

(a) Show that the inverted catenary $y = -\cosh x + \sqrt{2}$ intersects the x axis at the points $x = \ln(\sqrt{2}-1)$ and $x = \ln(\sqrt{2}+1)$. Show that the slope is 1 at the point $x = \ln(\sqrt{2}-1)$ and -1 at the point $x = \ln(\sqrt{2}+1)$.

(b) Show that the curve between these two points has length 2.

(c) We construct a track consisting of a succession of such curves, connected one after the other as shown in Figure 14.18. Consider a bicycle with square wheels with side length 2. Show that as the bicycle travels along this track the center of its wheels will always remain at height $\sqrt{2}$. **Suggestion:** Consider a single square wheel rolling along the surface without slipping. At the point of departure, one of the corners of the wheel is situated at the junction between two connecting catenaries, such that it is tangent to both of them.

The fastest tunnel

10. We consider a circle $x^2 + y^2 = R^2$ with radius R and a smaller circle with radius $a < R$ rolling along the inside of the larger circle. At the point of departure the two circles are tangent at the point $P = (R, 0)$. Show that as the smaller circle rotates along the inside of the larger, the point P traces out a hypocycloid as described in (14.18) with $b = \frac{a}{R}$.

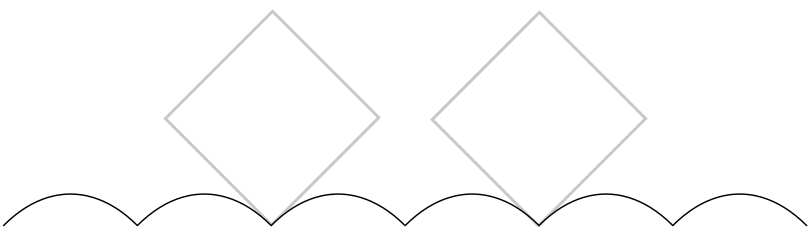


Fig. 14.18. The square wheels of a bicycle traveling along a path of inverted catenaries (see Exercise 9).

11. **(a)** In the case of $b = \frac{1}{2}$ verify that the movement of a particle traveling through the tunnel described by the hypocycloid of equation (14.18) is the same as the oscillations of a spring along a line (calculate the position of the particle as a function of time).
(b) Deduce that the period of the motion is independent of the height of the departure point.
(c) Determine the time taken for a point to travel between a point P and the antipodal point $-P$, traveling along a straight line through the center of the Earth and being acted upon only by the force of gravity. (The radius of the Earth is roughly 6365 km.)

12. Consider releasing a particle with zero initial velocity at height h in a hypocycloidal tunnel with parameter b. Show that for any value of b, the particle will oscillate in the tunnel with a period independent of h. That is, show that the motion of the particle through the tunnel is isochronous (see the discussion in Section 14.7). Determine the length of the period.

13. The exercise aims to calculate the travel time between New York and Los Angeles, assuming that we travel through a hypocycloidal tunnel between the cities. You might want to use the help of a mathematical software package to perform these calculations. The tunnel travels through the plane defined by the two cities and the center of the Earth. Assume that the radius of the Earth is given by $R = 6365$ km.
(a) New York is at roughly 41 degrees north latitude and 73 degrees west longitude. Los Angeles is situated approximately at 34 degrees north latitude and 118 degrees west longitude. Calculate the angle ϕ between the two vectors joining the center of the Earth to the two cities.
(b) Given a hypocycloidal as in (14.18) and an initial point $P_0 = (R, 0)$ corresponding to $\theta = 0$, calculate the first positive value θ_0 such that $P_{\theta_0} = (x(\theta_0), y(\theta_0))$ is on the circle with radius R. Calculate the angle ψ between the vectors $\overrightarrow{OP}_0$ and $\overrightarrow{OP}_{\theta_0}$.

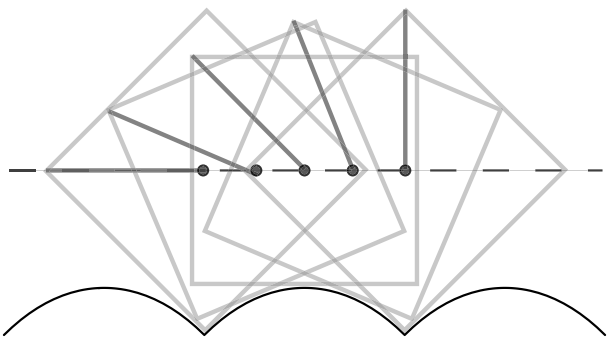


Fig. 14.19. A square wheel turning along a path of inverted catenaries (see Exercise 9). The positions of a spoke have been drawn.

(c) Setting $\phi = \psi$, calculate the parameter b of the hypocycloid corresponding to the tunnel between New York and Los Angeles.
(d) Calculate the time taken for a particle to travel along the hypocycloidal tunnel between New York and Los Angeles, under the effect of gravity only. (You may use the results of Exercise 12 to assist you in this).
(e) Calculate the maximum depth of the tunnel.
(f) Calculate the speed attained by the particle at the deepest point of the tunnel.

Hamilton's principle

14. **(a)** The potential energy stored in a compressed spring is proportional to the square of its deformation x from its position at equilibrium: $V(x) = \frac{1}{2}kx^2$, where k is a constant. This is called Hooke's law. We suppose that one end of a massless spring is attached to a rigid wall, and the other end is attached to a mass m. We fix the position x of m to be 0 when the spring is at equilibrium. Write the Lagrangian and the action integral describing the motion of this mass.
(b) Show that Hamilton's principle yields the classic equation for the motion of a mass attached to a spring: $x'' = -kx/m$, where x'' is the second derivative of the position of the mass.
(c) Assuming the particle is released without speed at the position $x = 1$ and time $t = 0$, show that its trajectory is described by the equation $x(t) = \cos(t\sqrt{k/m})$.

Soap bubbles

15. Consider the surface created by rotating the curve $z = f(x)$ around the x axis, for $x \in [a, b]$. Show that its area is given by

$$2\pi \int_a^b f\sqrt{1 + f'^2}dx.$$

16. **(a)** Show that the area of a surface given by the graph $z = f(x, y)$ above a region of the plane D is given by the double integral

$$I = \iint_D \sqrt{1 + f_x^2 + f_y^2}dx\,dy,$$

where $f_x = \frac{\partial f}{\partial x}$ and $f_y = \frac{\partial f}{\partial y}$.
(b) Suppose that the domain D is a rectangle $[a, b] \times [c, d]$. Consider a function f satisfying the boundary conditions

$$\begin{cases} f(a, y) = g_1(y), \\ f(b, y) = g_2(y), \\ f(x, c) = g_3(x), \\ f(x, d) = g_4(x), \end{cases}$$

where g_1, g_2, g_3, g_4 are functions that satisfy $g_1(c) = g_3(a)$, $g_1(d) = g_4(a)$, $g_2(c) = g_3(b)$, $g_2(d) = g_4(b)$. Show that such a function f that minimizes I satisfies the Euler–Lagrange equation given by

$$f_{xx}(1 + f_y^2) + f_{yy}(1 + f_x^2) - 2f_x f_y f_{xy} = 0. \tag{14.43}$$

Suggestion: You need to work through an analogue of the proof to Theorem 14.4. Suppose that the integral attains a minimum at f^* and consider a variation $F = f^* + \epsilon g$ where g is zero-valued along the boundary of D. Then I becomes a function of ϵ, and you need to show that its derivative at $\epsilon = 0$ is zero. To this end, transform the double integral into an iterated integral in order to apply integration by parts. One part of the function will need to be integrated with respect to x and then y, while another part requires proceeding in the opposite order. There is a fair amount of work required.

17. Show that the helicoid given by $z = \arctan \frac{y}{x}$ is a minimal surface. To do this you must show that the function $f(x, y) = \arctan \frac{y}{x}$ satisfies equation (14.43).

Three cities and a soapy film: the problem of minimal Steiner trees

18. **(a)** Let A, B, C be the three corners of a triangle and let P be its associated Fermat point, that is, the point $P = (x, y)$ chosen such that $|PA| + |PB| + |PC|$ is minimum. Prove that

$$\frac{\overrightarrow{PA}}{|PA|} + \frac{\overrightarrow{PB}}{|PB|} + \frac{\overrightarrow{PC}}{|PC|} = 0.$$

Hint: Take the partial derivatives with respect to x and y.

(b) Show that the only way that three unit vectors can have a zero sum is if they form an angle of $\frac{2\pi}{3}$ between them.

(c) Consider the construction shown in Figure 14.12. Show that the three inscribed lines must intersect at a single point and that this point is in the triangle if and only if the three internal angles of the triangle are less than $\frac{2\pi}{3}$.

(d) If the three angles of the triangle ABC are less than $\frac{2\pi}{3}$, show that there exists a unique point P inside the triangle such that the vectors $\overrightarrow{PA}$, $\overrightarrow{PB}$, and $\overrightarrow{PC}$ intersect at angles of $\frac{2\pi}{3}$.

Hint: The locus of points that subtend the segment AB with a given angle θ consists of the union of two arcs of a circle, as shown in Figure 14.20. The point P is therefore at the intersection of three circular arcs, each of which subtends one of the sides of the triangle ABC with an angle of $\frac{2\pi}{3}$.

(e) If the three angles of the triangle ABC are less than $\frac{2\pi}{3}$, show that the three lines constructing the Fermat point intersect at an angle of $\frac{\pi}{3}$. *Hint:* Let A' (resp. B', C') be the third corner of the equilateral triangle constructed on BC (resp. AC, AB). Show that the three vectors $\overrightarrow{AA'}$, $\overrightarrow{BB'}$, and $\overrightarrow{CC'}$ intersect each other at an angle of $\frac{2\pi}{3}$. This can be done by calculating the scalar product between each pair of vectors. Without loss of generality, suppose that $A = (0, 0)$, $B = (1, 0)$, and $C = (a, b)$.

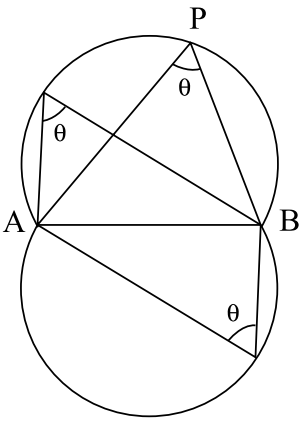


Fig. 14.20. The locus of points subtending the segment AB with angle θ (see Exercise 18).

(f) Deduce that the intersection points of these lines is a Fermat point only if it lies inside the triangle.
(g) Use the calculation in (e) to show that

$$|AA'| = |BB'| = |CC'|.$$

19. We consider the problem of finding the minimal Steiner tree for a set of four points situated at the corners of a square. The optimal solution is shown in Figure 14.21, in which all of the angles are 120 degrees. Showing that this network is the shortest possible is difficult. We will content ourselves with answering a subquestion.
(a) Show that the length of the network is smaller than the length of the two diagonals.

(b) Can you guess the minimal Steiner tree associated with the four corners of a rectangle?

Isoperimetric problems

20. Consider the graph of a function $y(x)$ that joins the points $(x_1, 0)$ and $(x_2, 0)$. We wish to maximize the area between the function and the x axis under the constraint that the perimeter of the region is L (see Example 14.17 discussed at the beginning of Section 14.10). Derive the Euler–Lagrange equation for the associated functional M of Theorem 14.18. Resolve the equation and show that the solution is an arc of a circle. What condition must be satisfied by L, x_1, and x_2?

21. The form of a suspension bridge. In contrast to a suspended cable, the form of the main cables in a suspension bridge are not catenary, but rather parabolic. The

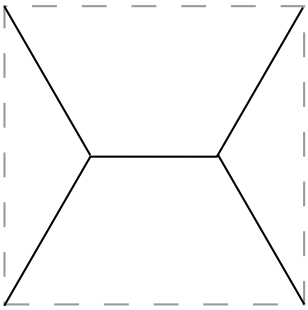


Fig. 14.21. The minimal Steiner tree for four points situated at the four corners of a square (see Exercise 19).

difference is that the weight of the cable is negligible compared to the weight of the attached bridge deck.

(a) Model the forces acting on the cable as in Example 14.20. Use the force diagram to deduce the differential equation that must be satisfied by the function defining the form of the curve. In this case, the weight P_x is proportional to dx and not to ds as in the case of the suspended cable.

(b) Show that the solution is a parabola.

References

[1] V. Arnold. *Mathematical Methods of Classical Mechanics*. Springer-Verlag, 1978.

[2] G.A. Bliss. *Lectures on the Calculus of Variations*. University of Chicago Press, 1946.

[3] J. Cox. The shape of the ideal column. *Mathematical Intelligencer*, 14:16–24, 1992.

[4] I. Ekeland. *The Best of All Possible Worlds*. University of Chicago Press, 2006.

[5] R.P. Feynman, R. Leighton, and M. Sands. *The Feynman Lectures on Physics*, volume II. Addison-Wesley, Reading, MA, 1964.

[6] B.K. Gibson. Liquid mirror telescopes. *Preprint UBC*.

[7] H.H. Goldstine. *A History of the Calculus of Variations from the 17th through the 19th Century*. Springer, New York, 1980.

[8] R. Weinstock. *Calculus of Variations*. Dover, New York, 1952.

15

Science Flashes

This chapter presents a variety of Science Flashes, small self-contained subjects that can each be covered in an hour or two. Most of these are geometric in nature, and many of these require little more than a familiarity with basic Euclidean geometry. Each section is independent. Several of the subjects may be treated as exercises: the lecturer can explain the problem in class, and the text can serve as an answer guide that is looked at only after the student has worked on the problem. Some of them are referred to as complementary material in the other chapters.

Notation. Throughout this chapter we will denote the length of a line segment AB by $|AB|$.

15.1 The Laws of Reflection and Refraction

The law of reflection describes the trajectory of a beam of light as it is reflected by a mirror. The law of refraction describes the trajectory of a beam of light as it passes from one uniform material to another (for example, from air into water). These two laws, seemingly quite different, can be united into one elegant principle.

The law of reflection. As a beam of light arrives at the surface of a mirror it is reflected such that the angle of incidence is equal to the angle of reflection (see Figure 15.1).

A simple principle allows us to reformulate the law of reflection: *light always travels the shortest path between two points A and B with one point on the mirror.*

We will show that this principle implies the law of reflection.

Theorem 15.1 *Let A and B be two points located on the same side of a mirror. Consider a beam of light going from point A to point B and touching the mirror in a point P. Then the shortest path is the one for which AP and PB make equal angles with the mirror as in the law of reflection.*

C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology*,
DOI: 10.1007/978-0-387-69216-6_15,

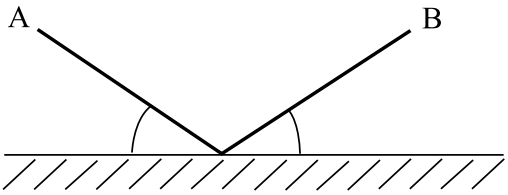


Fig. 15.1. The law of reflection.

PROOF: Let Q be a point of the mirror. Consider a path from A to B composed of segment AQ followed by segment QB as in Figure 15.2. The length of the path traveled by the beam is equal to $|AQ| + |BQ|$ (the length of AQ plus the length of QB). Let A' be the point symmetric to A with respect to the mirror. So AA' is perpendicular to the mirror and cuts the mirror in R such that $|AR| = |A'R|$. The two triangles ARQ and $A'RQ$ are congruent, since they have two equal sides $|AR| = |A'R|$ and RQ on both sides of an equal angle $\widehat{ARQ} = \widehat{A'RQ} = \frac{\pi}{2}$. It follows that $|AQ| = |A'Q|$. Then the length of the path traveled by the beam is equal to $|A'Q| + |QB|$. Compare

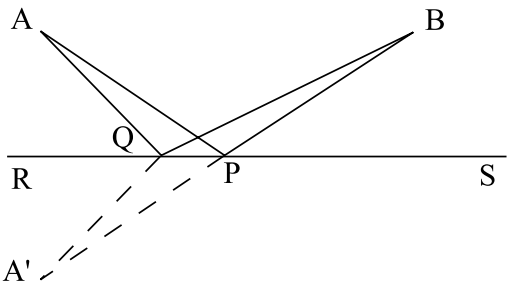


Fig. 15.2. The law of reflection and the shortest path.

this with the path AP and then PB, where $\widehat{APR} = \widehat{BPS}$. By taking $Q = P$ in the previous calculation we have that $|AP| = |A'P|$. Then the length of the path, given by $|AP| + |PB|$, is equal to $|A'P| + |PB|$. We have on one side $\widehat{APR} = \widehat{BPS}$ and on the other side $\widehat{APR} = \widehat{A'PR}$, since the triangles APR and $A'PR$ are congruent. This yields $\widehat{BPS} = \widehat{A'PR}$. We deduce that P lies on the segment $A'B$ by Lemma 15.2 below. Since P lies on the segment $A'B$, then $|A'P| + |BP| = |A'B|$. Since the line segment joining two points is the shortest path between the two points A' and B, we have for $Q \neq P$,

$$|A'P| + |PB| = |A'B| < |A'Q| + |QB| = |AQ| + |QB|.$$

□

Lemma 15.2 *We consider a line (D), a point P of (D), and two points A and B located on each side of (D) as in Figure 15.3. If $\widehat{APR} = \widehat{BPS}$, then A, P, and B are collinear.*

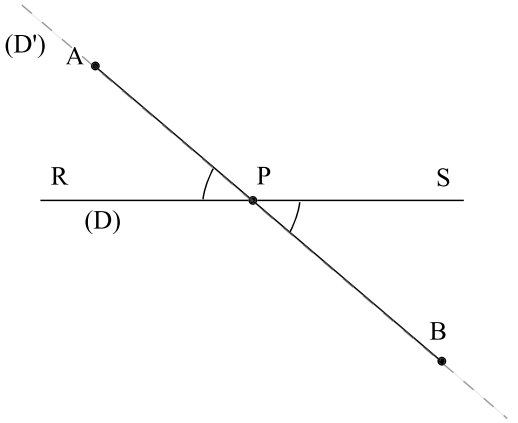


Fig. 15.3. If $\widehat{APR} = \widehat{BPS}$, then A, P, and B are aligned.

Proof. Consider Figure 15.3 and let us extend the line segment PA into a line (D'). The point P lies on (D'). Since two vertically opposite angles are equal, the angle between the lower part of (D') and PS is equal to $\widehat{APR}$. However, the segment PB also has this property. Hence PB is included in (D'). □

Remark. The geometric proof of Theorem 15.1 is very elegant. It uses the simple principle that the line segment between two points is the shortest path between them. We will see that the ideas introduced in this proof will be used in the proof of the remarkable properties of the parabola, ellipse, and hyperbola (Section 15.2).

The law of refraction. This second law allows us to calculate the deviation of a beam of light as it travels through a uniform material with speed v_1 and transitions into another uniform material where it travels with speed v_2. Let θ_1 be the angle of the beam of light through the first material, as measured from the perpendicular of the interface between the two materials. Similarly, let θ_2 be the angle of the beam of light through the second material, measured from the same perpendicular (see Figure 15.4). Then the law of refraction states that

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}.$$

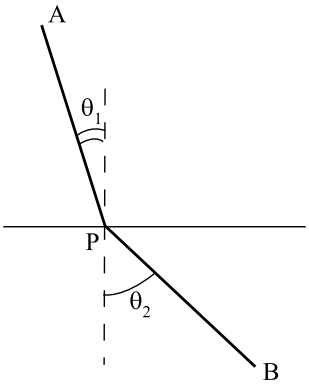


Fig. 15.4. The law of refraction.

It seems obvious that the previous principle, namely that light travels along the shortest path, does not accurately describe the law of refraction. As such, it does not unify the laws of reflection and refraction. However, when we were discussing the law of reflection, the speed of the beam did not change, since it was always traveling through a single uniform material. Thus, if the law of reflection seeks to minimize the length of the path between two points, this is entirely equivalent to minimizing the *time* taken to travel the path between the same two points. It is this principle that will unite the two seemingly distinct laws of reflection and refraction.

Principle: In the law of refraction, as in the law of reflection, light traveling between two points A and B follows the quickest possible path.

Theorem 15.3 *We consider two uniform materials separated by a plane. Let A and B be two points located on opposite sides of the separating plane. Let v_1 be the speed of light in the material containing A and v_2 the speed of light in the material containing B. The fastest path between A and B is the one that crosses the separating plane at the point P defined by the fact that the angles θ_1 and θ_2 between AP and PB and the normal to the separating plane are those given by the law of refraction, namely*

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}.$$

Proof: We will give the proof only for the planar problem (see Figure 15.5). The easiest proof uses differential calculus. Suppose that the beam of light transitions between media at the point Q with horizontal coordinate x (thus $|OQ| = x$) and let $l = |OR|$. Let $h_1 = |AO|$ and $h_2 = |BR|$. We calculate the travel time $T(x)$ between A and B. This time is equal to

$$T(x) = \frac{|AQ|}{v_1} + \frac{|QB|}{v_2} = \frac{\sqrt{x^2 + h_1^2}}{v_1} + \frac{\sqrt{(l-x)^2 + h_2^2}}{v_2}.$$

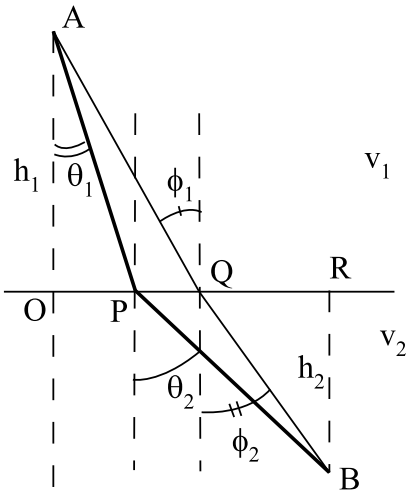


Fig. 15.5. The law of refraction and the quickest path.

To minimize this time, we are looking for a value of x such that $T'(x) = 0$. Since

$$T'(x) = \frac{x}{v_1\sqrt{x^2 + h_1^2}} - \frac{(l-x)}{v_2\sqrt{(l-x)^2 + h_2^2}},$$

then $T'(x_*) = 0$ for x_* satisfying

$$\frac{x_*}{v_1\sqrt{x_*^2 + h_1^2}} = \frac{(l-x_*)}{v_2\sqrt{(l-x_*)^2 + h_2^2}}.$$

The result follows by observing that

$$\frac{x_*}{\sqrt{x_*^2 + h_1^2}} = \sin\theta_1, \qquad \frac{(l-x_*)}{\sqrt{(l-x_*)^2 + h_2^2}} = \sin\theta_2.$$

We can easily verify that $T''(x_*) > 0$, and therefore that x_* is a minimum. In fact,

$$T''(x) = \frac{h_1^2}{v_1(x^2 + h_1^2)^{3/2}} + \frac{h_2^2}{v_2((l-x)^2 + h_2^2)^{3/2}}.$$

□

A beam of light always chooses the quickest path. We see right away the beauty of this principle: not only is it elegant in and of itself, but it allows us to consider new questions. For instance, we understand how to calculate the path traveled by light through heterogeneous media using differential and integral calculus.

The principle of optimization in physics. In fact, this is only one of many examples where the laws of physics seemingly obey a *principle of optimization*. All of Lagrangian mechanics is built upon a similar principle, as exploited by *variational calculus* (see Chapter 14). We give a few examples:

- A high-tension cable between two poles describes a curve. What is the formula of this curve? We can calculate its equation and see that it is a catenary, as described by a hyperbolic cosine. Recall that the hyperbolic cosine is defined as

$$\cosh x = \frac{e^x + e^{-x}}{2}.$$

  Why does it take this shape? Among all paths of the same length between the two poles, this is the one that minimizes the potential energy of the suspended cable. More details in Section 14.10 of Chapter 14.
- If we rotate a cylinder full of liquid at a constant angular velocity about its central axis, the surface of the liquid forms a paraboloid of revolution, or circular paraboloid. In this system we are not only considering potential energy but also kinetic energy. The surface of the liquid must be the one that minimizes the *Lagrangian* of the system, which is the difference between the potential and kinetic energies. This calculation is performed in Section 14.11 of Chapter 14.

We return to the law of refraction. If we know the angle θ_1 with the normal in the first material, we can calculate the angle θ_2 with the normal in the second material using

$$\sin\theta_2 = \frac{v_2 \sin\theta_1}{v_1}.$$

But does this equation always have a solution? If $v_2 > v_1$ and $\sin\theta_1 > \frac{v_1}{v_2}$, then $\frac{v_2 \sin\theta_1}{v_1} > 1$, which cannot be the sine of any angle. Thus, if the angle θ_1 is too large, meaning the beam arrives at too oblique an angle, then it will not actually enter the second material but will instead be reflected. How? Now we understand the power of the general principle stated above: to go from A to B the beam must follow the fastest path touching the separating surface between the two materials. Hence it must be reflected such that the angle of incidence is equal to the angle of reflection.

Fiber optics. Optical fibers are transparent cables within which light beams travel. Since the speed of light is slower in the cable than it is in air, the beams will be reflected if they arrive at the boundary with too great an angle with the normal (see Figure 15.6).

Fiber optics is often used in high-speed telecommunications networks because it allows the simultaneous transmission of many signals without any interference between them. Engineers face many challenges in designing and building fiber optic cables, and many of these can be the subject of a project (dispersion of waves, cables with refractive index varying with the distance to the axis of the fiber, signal separation when signals emerge, etc).

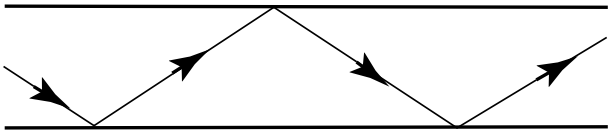


Fig. 15.6. The propagation of a beam of light in a fiber optic cable.

Short waves. Electromagnetic waves are roughly broken down into a variety of families including visible, ultraviolet, X rays, and radio waves. These families are defined based on the frequencies of the waves they encompass. For example, radio waves generally start at just a few hertz and go up to several hundred gigahertz.[1] In North America, commercial radio stations transmitting through *amplitude modulation* (AM) use frequencies around the 1 MHz[2] mark, while stations transmitting through *frequency modulation* (FM) use higher frequencies, around 100 MHz. Between these two spectra lies the family of waves known as *short waves*, from 3 to 30 MHz. Regardless of transmission power, the curvature of the Earth limits the reception radius of any antenna. Despite this, short waves (and other waves of lower frequency) are regularly transmitted much further than is possible by simple line of sight. This is because they are reflected by the higher layers of the ionosphere.

The atmosphere is a nonuniform medium. It is broken down into three major layers:

- the troposphere, from the Earth's surface to 15 km above it;
- the stratosphere, from 15 to 40 km; and
- the ionosphere, from 40 to 400 km.

In the higher levels of the ionosphere, ionized gases act as a mirror for short waves. The exact nature of these gases, and the reflections they produce, varies greatly depending on the time of day. Under favorable conditions it is possible for a signal to be reflected by the ionosphere and the Earth several times. The exact calculation of the trajectory taken by the signal must also take into account the layers below the ionosphere, since they refract the signal.

Localizing lightning strikes. In Section 1.3 of Chapter 1 it is seen that lightning strikes generate electromagnetic waves traveling through the atmosphere that are occasionally reflected by the ionosphere. When this happens, certain lightning strike detectors will detect the initial bolt of lightning, while others will detect its reflection.

[1] 1 gigahertz = 1 GHz = 10^9 Hz.
[2] 1 megahertz = 1 MHz = 10^6 Hz.

15.2 A Few Applications of Conics

15.2.1 A Remarkable Property of the Parabola

Legends say that Archimedes (287–212 BC) lit a Roman fleet of ships on fire as they were attacking Syracuse, his hometown on the island of Sicily. Supposedly, he did so by making use of the remarkable property of parabolas we will discuss below.

Most readers will certainly recall the basic equation of a parabola, $y = ax^2$, whose base lies at the origin and which is symmetric about the vertical axis. There exists an equivalent geometric formulation:

Definition 15.4 *A parabola is the locus of points in the plane that are at an equal distance to a point F (called the* focus *of the parabola) and a line (Δ), the* directrix *of the parabola (see Figure 15.7).*

Given a parabola with equation $y = ax^2$, it is relatively simple to identify both the focus and the directrix.

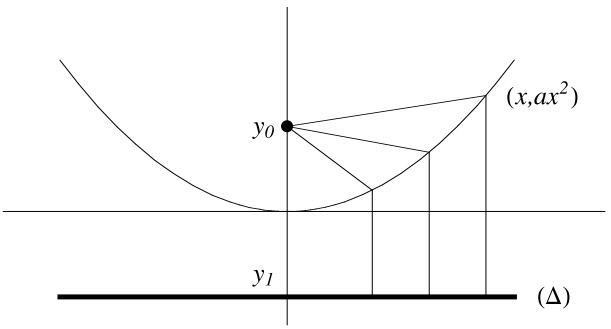


Fig. 15.7. The geometric definition of a parabola.

Proposition 15.5 *The focus of the parabola $y = ax^2$ is the point $(0, \frac{1}{4a})$, and the directrix is the line with equation $y = -\frac{1}{4a}$.*

Proof. By symmetry, the focus must be along the axis of symmetry of the parabola (the y axis in this case), and the directrix must be perpendicular to this axis. Thus

$$F = (0, y_0) \qquad \text{and} \qquad (\Delta) = \{(x, y_1) \mid x \in \mathbb{R}\}.$$

We can see already that $y_1 = -y_0$, since $(0, 0)$ is on the parabola. If a point belongs to the parabola it is of the form (x, ax^2), and its distance from both the focus and the directrix will be the same if

$$|(x, ax^2) - (0, y_0)| = |(x, ax^2) - (x, -y_0)|.$$

We square both sides to get rid of radicals,

$$|(x, ax^2) - (0, y_0)|^2 = |(x, ax^2) - (x, -y_0)|^2.$$

This yields $x^2 + (ax^2 - y_0)^2 = (x - x)^2 + (ax^2 + y_0)^2$, or equivalently

$$x^2 + a^2x^4 - 2ax^2y_0 + y_0^2 = a^2x^4 + 2ax^2y_0 + y_0^2,$$

which finally reduces to

$$x^2(1 - 4ay_0) = 0,$$

which must be satisfied for all x. Hence the coefficient of x^2 must be zero: $1 - 4ay_0 = 0$, which yields $y_0 = \frac{1}{4a}$. Thus, the focus is at $(0, \frac{1}{4a})$ and the directrix has the equation $y = -\frac{1}{4a}$. □

In order to understand the remarkable property about to be described we must first imagine that the interior of the parabola is a mirror. All beams of light reflecting off a point of the parabola will therefore satisfy the law of reflection: the angle of incidence of any such beam will be equal to the angle of reflection, both measured with respect to the line tangent to the parabola at that point. (See Section 15.1 for more on the law of reflection.) The following theorem describes the remarkable property of the parabola.

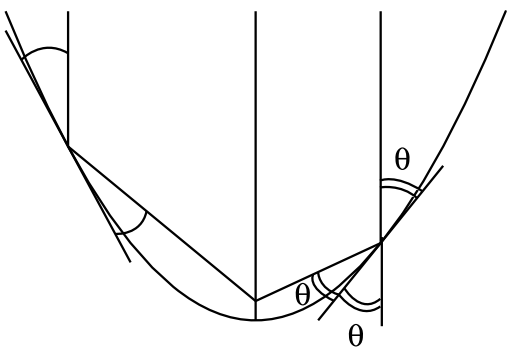


Fig. 15.8. A remarkable property of the parabola.

Theorem 15.6 The remarkable property of the parabola. *All beams parallel to the axis of the parabola and reflected on its surface will pass through the focus of the parabola.*

PROOF: Consider the parabola with the equation $y = f(x)$, where $f(x) = ax^2$. We will be considering the abstract function $f(x)$ for most of these calculations in order to allow us to reuse them in Theorem 15.7, which deals with the reciprocal of this theorem. Let (x_0, y_0) be a point on the parabola and let θ be the angle of incidence formed between the beam and the tangent to the parabola at the point (x_0, y_0). For reasons of symmetry we can limit ourselves to $x_0 \geq 0$. Looking at Figure 15.8 and using that vertically opposite angles are equal, we can see that the reflected beam will form an angle of 2θ with the vertical, thus an angle of $\frac{\pi}{2} - 2\theta$ with the horizontal. The equation of the reflected beam is therefore

$$y - y_0 = \tan\left(\frac{\pi}{2} - 2\theta\right)(x - x_0) \tag{15.1}$$

(this is where we make use of the fact that $x_0 \geq 0$, since we would have to add a negative sign in the case that $x_0 < 0$). We must calculate $\tan(\frac{\pi}{2} - 2\theta)$ as a function of x_0. The slope of the tangent to the parabola is given by $f'(x_0) = 2ax_0$. Since the angle between the tangent and the horizontal is $\frac{\pi}{2} - \theta$, we have that

$$\tan\left(\frac{\pi}{2} - \theta\right) = \cot\theta = f'(x_0) = 2ax_0.$$

Also

$$\tan\left(\frac{\pi}{2} - 2\theta\right) = \cot 2\theta.$$

Since $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\sin\theta\cos\theta$, we obtain that

$$\cot 2\theta = \frac{\cos^2\theta - \sin^2\theta}{2\sin\theta\cos\theta} = \frac{\frac{\cos^2\theta - \sin^2\theta}{\sin^2\theta}}{\frac{2\sin\theta\cos\theta}{\sin^2\theta}} = \frac{\cot^2\theta - 1}{2\cot\theta}.$$

This yields

$$\cot 2\theta = \frac{(f'(x_0))^2 - 1}{2f'(x_0)} = \frac{4a^2x_0^2 - 1}{4ax_0}.$$

The point of intersection between the reflected beam and the vertical axis of the parabola is found by substituting $x = 0$ into the equation (15.1) for the reflected beam and by observing that $y_0 = f(x_0)$. We obtain that

$$y = f(x_0) - x_0\frac{(f'(x_0))^2 - 1}{2f'(x_0)}.$$

We now use the fact that $f(x) = ax^2$. In doing so we obtain

$$y = \frac{1}{4a},$$

which is to say that the point of intersection $(0, y)$ of the reflected beam with the vertical axis is independent of the vertical incoming ray, and so of the point of reflection being

considered. Moreover, observe that the point of intersection of all reflected beams with the vertical axis, $(0, \frac{1}{4a})$, is precisely the focus of the parabola. □

The converse is also true:

Theorem 15.7 *The parabola is the only curve with the property that there exists a direction for which all incident beams parallel to this direction will be reflected by the curve through a single point.*

Discussion of the proof. This theorem is decidedly more advanced than the last. If we consider a curve with the equation $y = f(x)$ then we must resolve the differential equation we considered above,

$$f(x_0) - x_0 \frac{(f'(x_0))^2 - 1}{2f'(x_0)} = C,$$

where C is a constant. This is equivalent to the differential equation (we substitute $x_0 = x$ to have a more standard form)

$$2f(x)f'(x) - x(f'(x))^2 - 2Cf'(x) + x = 0.$$

We will not pursue the solution here. However, those readers familiar with the theory of differential equations will note that this is a nonlinear first-order equation. □

We will give a geometric proof of Theorem 15.6 using only the geometric definition of the parabola as introduced in Definition 15.4.

Geometric proof of Theorem 15.6. We reason with reference to Figure 15.9. We

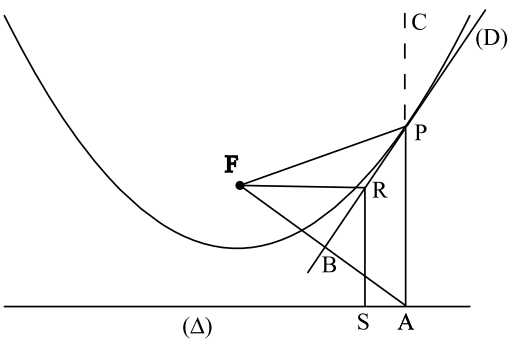


Fig. 15.9. The geometric proof of the remarkable property of the parabola.

consider a parabola with focus F and directrix (Δ). Let P be a point on the parabola and let A be its projection on the directrix (Δ). By the definition of the parabola

we know that $|PF| = |PA|$. Let B be the middle of the segment FA and let (D) be the line passing through P and B. Since the triangle FPA is isosceles, we know that $\widehat{FPB} = \widehat{APB}$. The theorem will be proved if we can show that the line (D) is tangent to the parabola at P. Indeed, consider the extension PC of PA, which is the incident beam. The angle that PC makes with (D) is equal to the angle $\widehat{APB}$ (vertically opposite angles), which is itself equal to the angle $\widehat{FPB}$. Thus, if the line (D) behaved as a mirror and if PC were the incident beam, then PF would be the reflected beam.

We must now prove that the line (D) defined above is tangent to the parabola at P. We will do this by showing that all of the points of (D), save P, lie below the parabola. Indeed, it is easy to convince oneself that any straight line through P other than the tangent line has some points lying above the parabola; see Figure 15.10.

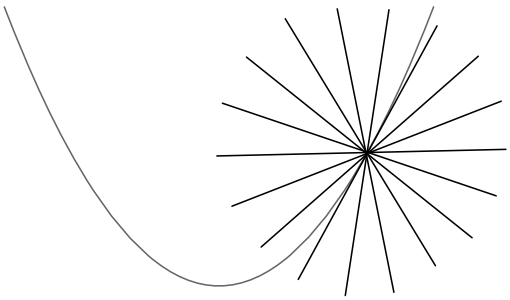


Fig. 15.10. The tangent line to the parabola at P is the only straight line through P that has no point above the parabola.

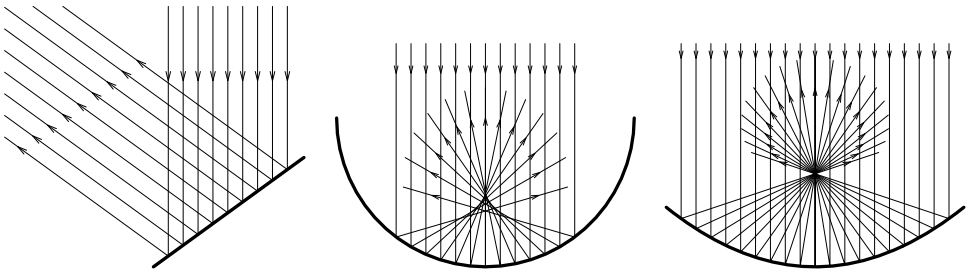
How do we prove that a point lies below the parabola? We come back to the geometric property defining a parabola, which can be rewritten as follows: let R be a point in the plane and let S be its orthogonal projection onto the directrix. Then we have

$$\begin{cases} |FR| < |SR| & \text{if } R \text{ is above the parabola,} \\ |FR| = |SR| & \text{if } R \text{ is on the parabola, and} \\ |FR| > |SR| & \text{if } R \text{ is below the parabola.} \end{cases} \tag{15.2}$$

Let R be a point of (D) distinct from P and let S be its projection on (Δ). The triangles FPR and PAR are congruent, since they have an equal angle between two equal sides. Thus $|FR| = |AR|$. Additionally, since AR is the hypotenuse of the right triangle RSA, then $|SR| < |AR|$. Thus, $|SR| < |FR|$, and by (15.2) we have that R is below the parabola. □

Is this property really all that remarkable? Theorem 15.7 affirms that it is, and that this property uniquely defines the parabola. How is this property used in practice? Consider Figure 15.11. A flat mirror reflects parallel beams of light as parallel beams

of light in another direction, a circular mirror reflects parallel beams into unfocused beams, while a parabola reflects all incoming beams parallel to its central axis through a unique focal point. Thus, it is no surprise that parabolas find many technological applications.



(a) flat mirror (b) circular mirror (c) parabolic mirror

Fig. 15.11. Comparing reflected beams with a flat mirror, a circular mirror, and a parabolic mirror.

Parabolic antennas. A parabolic antenna is usually oriented such that its central axis is pointed directly at the source of the signal (often a satellite) it is meant to receive. The physical receiver is then placed at the focal point of the antenna. Figure 15.12 shows a parabolic antenna at the entrance of the city of Höfn, Iceland. In Iceland, a country full of mountains and fjords, it is not always possible to aim an antenna directly at the desired satellite. Thus when passing some mountain gaps one observes pairs of parabolic antennas, each one aiming at a different valley floor below. One of the antennas is a receiver, relaying the received signal to the second antenna, which finally sends the signal to the antenna in the second valley floor.

Radar. Radar receivers also have a parabolic shape. The difference between these and standard satellite antennas is that the position of the axis is variable and the radar itself is the source of a signal that is emitted along its central axis. When the electromagnetic waves hit an object, they are reflected. A portion of these reflected waves will return to the transmitter (those that strike faces of the object that are perpendicular to the path of the signal). These beams will then strike the parabolic antenna and will be reflected to the receiver, situated at the focus. In order to cover many directions the radar is in constant rotation, with its axis remaining roughly horizontal.

Car headlights. The light bulb is located at the focus and emits light in all directions. All beams emitted behind the bulb are then reflected into beams parallel to the axis.

Telescopes. Once again we aim the telescope such that its axis is pointing toward the object or portion of the sky we wish to observe. The light is arriving from sufficiently far

Fig. 15.12. A parabolic antenna at the entrance to the city of Höfn, Iceland.

away that the beams are essentially parallel when they arrive at the receiver, where they are all reflected through the focus. Telescopes of this sort suffer from one big problem: the image is created at the focus of the mirror, which is itself above the mirror. But the observer (in this case the device capturing the image) should not be above the mirror, since it will obstruct and itself appear in the image. Thus a second mirror is used. There are two classical ways to proceed.

1. The first uses a flat mirror placed at an oblique angle, as shown in Figure 15.13. Such a telescope is called a *Newton telescope*.
2. The second type uses a convex (secondary) mirror situated above the large primary mirror. In this case the two mirrors are not necessarily parabolic, since it is the composition of the action of the two mirrors that focuses the image to a single point (see Figure 15.14). However, we may choose to construct the primary mirror as a parabolic mirror. In this case the secondary mirror is a convex hyperbolic mirror aligned such that the focus of the parabola is also a focus of the hyperbola. This particular choice for the secondary mirror is due to a remarkable property of hyperbolic mirrors that is discussed in Section 15.2.3. Such a telescope is called a *Schmidt–Cassegrain telescope*.
3. Recently there have appeared telescopes with liquid mirrors. Exercise 15.48 shows the plan of the telescope ALPACA to be installed on top of a Chilean mountain. For more on telescopes with liquid mirrors see Section 14.11 of Chapter 14.

Solar furnace: Solar furnaces are one method of using sunlight to produce electricity. Several of them have been constructed near the city of Odeillo, in the French Pyrenees. Odeillo is home to the PROMES laboratory of CNRS (Laboratoire PROcédés, Matériaux et Énergie Solaire du Conseil National de la Recherche Scientifique, or the

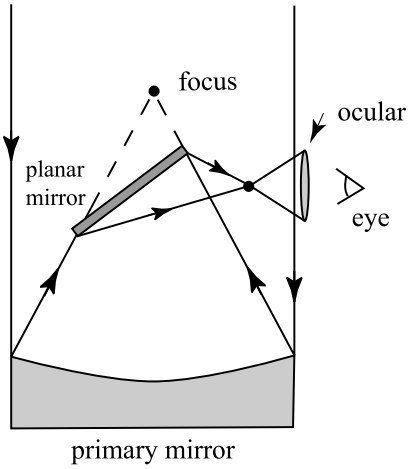


Fig. 15.13. Newton telescope.

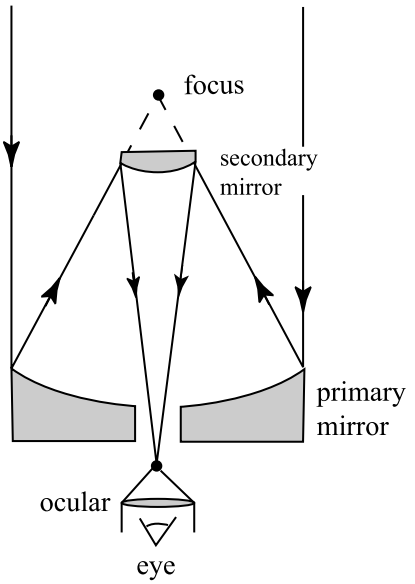


Fig. 15.14. Schmidt–Cassegrain telescope.

Processes, Materials, and Solar Energy Laboratory of the National Council on Scientific Research). The amount of sun received in the area is exceptional. The largest furnace

Fig. 15.15. The largest solar furnace at Odeillo. Several heliostats can be seen in the foreground. (Photo by Serge Chauvin.)

generates more than 1 megawatt[3] (see Figure 15.15). In comparison, there exist roughly 250 hydroelectric dams in France with power outputs between a few tens of kilowatts[4] to a few hundred megawatts. The largest hydroelectric dams in Quebec produce between 1000 and 2000 megawatts. Individual wind turbines can often produce around 600 kilowatts. The solar furnace shown in Figure 15.15 consists of a large parabolic mirror with a surface area of 1830 square meters. Its central axis is horizontal and the focus is situated 18 meters ahead of the mirror. Since it is not feasible to orient the entire mirror and furnace toward the sun, a set of 63 heliostats with a combined surface area of 2835 square meters is used instead (see Figure 15.16). A heliostat is simply a mirror driven by a clock mechanism that allows the mirror to reflect sunlight in a constant direction throughout the day. Heliostats are installed and programmed to ensure that they reflect the sun toward the parabola such that the beams are parallel to the central axis of the solar furnace at all times. This requires the solar furnace to be oriented to the north! The collected beams are then reflected toward the focus

[3] 1 megawatt $= 10^6$ watts.
[4] 1 kilowatt $= 10^3$ watts.

Fig. 15.16. Heliostats redirect the solar rays toward the primary parabolic mirror of Odeillo's solar furnace (photo by Serge Chauvin).

of the parabola, where they heat a container of hydrogen to very high temperature. This source of heat is transformed into mechanical power to run an electrical generator, the mechanism being called the "Stirling cycle." Research focuses on improving the net efficiency of the transformation of heat into electricity. Currently, such systems see roughly 18% efficiency.

A return to legend of Archimedes. Archimedes' use of parabolas (according to legend) was to construct large parabolic mirrors whose axes were pointed at the sun and whose foci were meant to be as close as possible to the ships of the Roman fleet. Modern technology would probably be capable of building mirrors of the scale and reflective quality necessary to ignite the sail of a distant ship. However, it is doubtful that the technology of the time was sufficiently advanced to build such defensive weapons, even using aligned polished metal shields. A group of engineers from the Massachusetts Institute of Technology, in Cambridge, recently tested the feasibility of such a device.[5] Using 127 one-square-foot mirrors (≈ 0.1 m^2) they succeeded, after a few attempts, to ignite a 10-foot-long (≈ 3 m) model of a boat situated roughly 100 feet (≈ 30 m)from the

[5] `http://web.mit.edu/2.009/www/experiments/deathray/10_ArchimedesResult.html`.

mirrors. The experiment was criticized because the engineers used modern materials that would not have been available in the time of Archimedes, and the target was positioned closer than reported by the legend. However, despite these criticisms the successful test indicates that the concept is not as absurd as it may at first seem.

Unlike the engineers at MIT, Archimedes could not simply buy hundreds of highly reflective mirrors from the local hardware store! However, could he have used hundreds of highly polished metal shields stacked side by side? Although doubtful, we are unable to exclude this possibility.

15.2.2 The Ellipse

Recall the geometric definition of an ellipse.

Definition 15.8 *An ellipse is the locus of points in the plane such that the sum of their distances from two points F_1 and F_2 (called the foci) is equal to some constant C, where $C > |F_1F_2|$.*

Ellipses have a remarkable property quite similar to that of parabolas.

Theorem 15.9 The remarkable property of the ellipse. *Any ray of light leaving one focus and reflected by the interior of the ellipse will arrive at the other focus.*

Proof. We will provide a geometric proof using only Definition 15.8, which may be rewritten as follows: if R is a point in the plane, then

$$\begin{cases} |F_1R| + |F_2R| < C & \text{if } R \text{ is inside the ellipse,} \\ |F_1R| + |F_2R| = C & \text{if } R \text{ is on the ellipse,} \\ |F_1R| + |F_2R| > C & \text{if } R \text{ is outside the ellipse.} \end{cases} \tag{15.3}$$

Imagine a beam originating at F_1 and consider the point P where it intersects the ellipse (see Figure 15.17). Let (D) be the line passing through P and making the same angle with both F_1P and F_2P. We must show that this line is tangent to the ellipse at P. Here again we will use the fact that any straight line through P other than the tangent line to the ellipse has points inside the ellipse (see Figure 15.18). So we must show that any point R along (D) except P satisfies $|F_1R| + |F_2R| > C$.

Let F be the point symmetric to F_1 with respect to (D). Since P and R are both on (D), we have that $|FP| = |F_1P|$ and $|FR| = |F_1R|$. Hence the triangles F_1PR and FPR are congruent, since they have three equal sides. Thus it follows that $\widehat{FPR} = \widehat{F_1PR}$. Since $\widehat{F_1PR} = \widehat{F_2PS}$ by definition of (D), we have that $\widehat{FPR} = \widehat{F_2PS}$, allowing us to conclude that F_2, F, and P are collinear by Lemma 15.2. It follows that $|FF_2| = |FP| + |PF_2|$ and

$$|F_1R| + |F_2R| = |FR| + |F_2R| > |FF_2|.$$

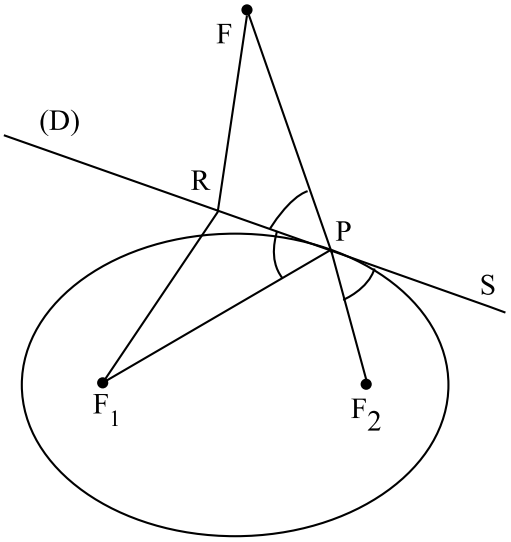


Fig. 15.17. A remarkable property of the ellipse.

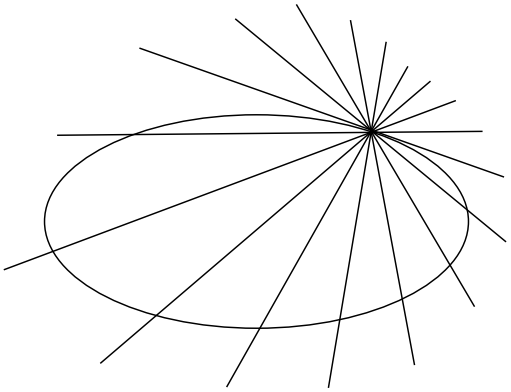


Fig. 15.18. The tangent line to the parabola at P is the only straight line through P that has no point inside the ellipse.

We also have

$$|FF_2| = |FP| + |PF_2| = |F_1P| + |PF_2| = C.$$

Hence $|F_1R| + |F_2R| > C$, allowing us to conclude that R is outside of the ellipse. □

Elliptical mirrors. Elliptical mirrors are studied in geometric optics and are currently used in a variety of applications. While parabolic mirrors are able to convert a point

source of light (for example, a light bulb) into a parallel beam of light (as is done in car headlights), an elliptical mirror reflects a pencil of rays originating from one point to a pencil of rays converging to another point. This property is used in certain types of film projectors where an elliptical mirror collects the light from the bulb and reflects it through the narrow aperture of the lens so that it passes through the film. Also, certain telescope designs employ elliptical secondary mirrors.

Elliptical arches. The described property of ellipses can also be observed with sound waves. For example, the arches in the Paris subway are roughly elliptical. Thus, if you are situated near the focus on one side of the tracks you can clearly hear a group of people situated near the focus on the other side of the tracks. In some cases, you can actually hear them more clearly than you would another person closer to you and on the same side of the tracks as you.

15.2.3 The Hyperbola

Recall the geometric definition of a hyperbola.

Definition 15.10 *A hyperbola is the locus of points in the plane such that the absolute value of the difference of their distances from two points F_1 and F_2 (called the foci) is equal to some constant C, where $C < |F_1F_2|$. In other words, P is on the hyperbola if and only if*

$$| \, |F_1P| - |F_2P| \, | = C.$$

A hyperbola has two branches. The branch attached to the focus F_1 is the set of points P such that $|F_2P| - |F_1P| = C$, while the branch attached to the focus F_2 is the set of points P such that $|F_1P| - |F_2P| = C$.

Hyperbolas have the following remarkable property:

Theorem 15.11 The remarkable property of the hyperbola. *Any beam aimed at the focus of one branch of a hyperbola and striking the exterior of this branch will be reflected toward the focus of the other branch (see Figure 15.19).*

Proof. We leave the proof to Exercise 4. It is quite similar to that of Theorem 15.9.□

Hyperbolic mirrors. Convex mirrors with a hyperbolic profile are studied in geometric optics and have various applications, one of which is their use in cameras. As discussed earlier, they are also used as the secondary mirror in Schmidt–Cassegrain-type telescopes (Figure 15.14). In such a telescope the first focus of the hyperbola is coincident with the focus of the parabolic primary mirror. The hyperbolic mirror serves to reflect the image through the second focal point of the hyperbola, which is situated below it.

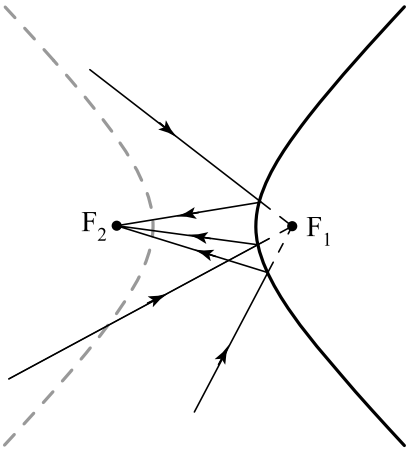


Fig. 15.19. A remarkable property of the hyperbola.

15.2.4 A Few Clever Tools for Drawing Conics

Given the general importance of conic sections, many ingenious methods for drawing them have been devised. The geometric definition of an ellipse allows it to be drawn quite easily by attaching a string of length C to the two foci F_1 and F_2 of the ellipse. We then draw the ellipse by ensuring that the string stays taut as we move the pencil (see Figure 15.20). This approach is not accurate unless the pencil is held perfectly perpendicular to the drawing surface. In Exercise 7 we will discuss a much more accurate approach. Exercise 8 presents a method for drawing a hyperbola similar to the string method for drawing an ellipse. Exercise 9 presents a method for drawing a parabola that makes use of a string and a carpenter's square.

15.3 Quadratic Surfaces in Architecture

Architects like creating audacious forms; just think of Gaudí's houses or the Montreal Olympic stadium. Other times it is engineers who, for structural reasons and optimization of strength, conceive of curved surfaces; consider cooling towers of nuclear reactors and hydroelectric dams, for example. Constructing the forms for pouring the concrete of these structures is a nontrivial problem, since the surfaces are not planar.

Certain mathematical surfaces, called *ruled surfaces*, have a remarkable mathematical property: they contain one or several families of lines such that any point on the surface will lie on at least one line in the family. A simple example of such a surface is a cone. This is our first example of a *quadratic surface* (also called *quadric*). However, not all quadratic surfaces are ruled surfaces. As examples, neither the sphere nor the ellipsoid contains even a single straight line.

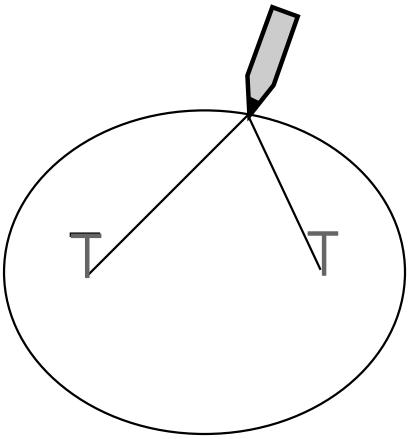


Fig. 15.20. Drawing an ellipse by attaching a cord to its foci.

The hyperboloid of one sheet (Figure 15.21) is another example of a ruled surface; in fact, it can be constructed by two distinct families of lines.

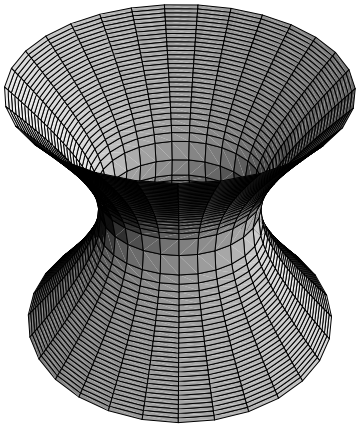
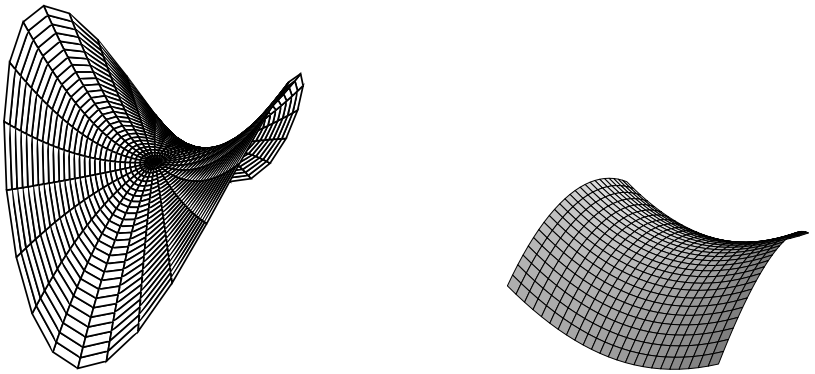


Fig. 15.21. A hyperboloid of revolution of one sheet.

Another quadratic surface often used in architecture is the *hyperbolic paraboloid*, or *saddle* (see Figure 15.22). Some roofs of buildings have been built with this form.

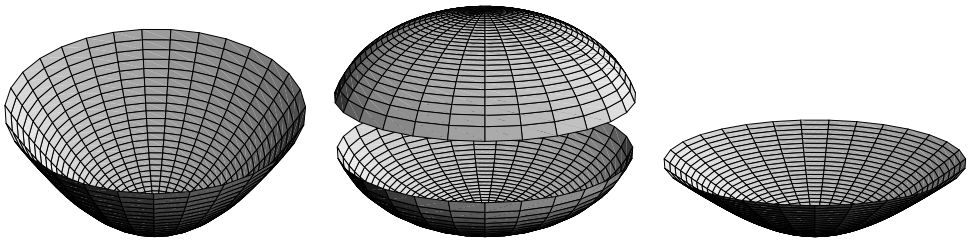
Earlier, when we were discussing parabolic mirrors, these were more precisely circular paraboloids (see Figure 15.23(a)). Elliptic mirrors are actually portions of ellipsoids of revolution (see Figure 15.23(b)), while hyperbolic mirrors are part of a hyperboloid



(a) (b)

Fig. 15.22. Two hyperbolic paraboloids, or saddle surfaces.

of revolution of two sheets (see Figure 15.23(c)). Thus we have identified three more quadratic surfaces with important technological applications.



(a) circular paraboloid (b) two portions of an ellipsoid (c) one sheet of a hyperboloid of two sheets

Fig. 15.23. Quadratic shapes often used as mirrors.

Here, we will be studying two quadratic ruled surfaces: the hyperboloid of one sheet and the hyperbolic paraboloid.

Definition 15.12 *A quadratic surface is a surface that may be described by the equation*

$$P(x, y, z) = 0,$$

where P is a degree-2 polynomial in the variables (x, y, z).

When studying quadratic surfaces one often encounters complicated polynomials P. So one often performs a change of coordinates that preserves both distances and angles (such a change of coordinates is called an isometry; see Chapter 2) in order to return the equation to a simpler canonical form in which we can read the geometry. It is the equivalent in three dimensions of what we do in two dimensions when we choose to consider the ellipse aligned to the axes with canonical equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

In this form the axes of symmetry of the ellipse are themselves the axes of the coordinate system.

The hyperboloid of one sheet. Under appropriately chosen orthonormal coordinates this surface has the canonical equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1. \tag{15.4}$$

If we intersect this surface with a plane containing the z axis, thus of the form $Ax+By = 0$, then the intersection describes a hyperbola in this plane. Alternatively, if we intersect this surface with a plane parallel to the xy plane, thus of the form $z = C$, then the intersection describes an ellipse in this plane.

Cooling towers of nuclear reactors often take on the form of a hyperboloid of one sheet of revolution: in this case we have $a = b$ in (15.4) (see Figure 15.21). We will discuss the advantages of this form after the following proposition.

Proposition 15.13 *We consider two circles $x^2 + y^2 = R^2$ situated in the planes $z = -z_0$ and $z = z_0$. Let $\phi_0 \in (-\pi, 0) \cup (0, \pi]$ be a fixed angle. Then the union of the lines (D_θ), where (D_θ) is the line joining the point $P(\theta) = (R\cos\theta, R\sin\theta, -z_0)$ on the first circle to the point $Q(\theta) = (R\cos(\theta + \phi_0), R\sin(\theta + \phi_0), z_0)$ on the second circle, is a hyperboloid of revolution of one sheet if $\phi_0 \neq \pi$ and is a cone if $\phi_0 = \pi$ (see Figure 15.24).*

Proof. The line (D_θ) passes through the point $P(\theta)$ in the direction $\overrightarrow{P(\theta)Q(\theta)}$. Thus, it is the set of points

$$\{(x(t,\theta), y(t,\theta), z(t,\theta)) | t \in \mathbb{R}\}$$

with

$$\begin{cases} x(t,\theta) = R\cos\theta + tR(\cos(\theta+\phi_0) - \cos\theta), \\ y(t,\theta) = R\sin\theta + tR(\sin(\theta+\phi_0) - \sin\theta), \\ z(t,\theta) = -z_0 + 2tz_0. \end{cases} \tag{15.5}$$

We must eliminate t and θ in order to find the equation of the surface. To do this we calculate $x^2(t,\theta) + y^2(t,\theta)$. We see that

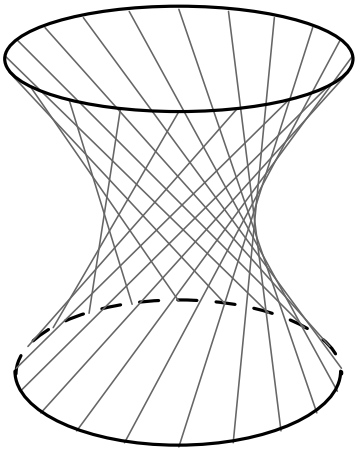


Fig. 15.24. The lines generating a hyperboloid of revolution of one sheet.

$$\begin{aligned} x^2(t,\theta) &= R^2[\cos^2\theta + t^2(\cos^2(\theta+\phi_0) - 2\cos(\theta+\phi_0)\cos\theta + \cos^2\theta) \\ &\quad +2t\cos\theta(\cos(\theta+\phi_0) - \cos\theta)] \end{aligned}$$

and

$$\begin{aligned} y^2(t,\theta) &= R^2[\sin^2\theta + t^2(\sin^2(\theta+\phi_0) - 2\sin(\theta+\phi_0)\sin\theta + \sin^2\theta) \\ &\quad +2t\sin\theta(\sin(\theta+\phi_0) - \sin\theta)], \end{aligned}$$

which yields

$$\begin{aligned} x^2(t,\theta) + y^2(t,\theta) &= R^2[1 + 2t^2(1 - (\cos\theta\cos(\theta+\phi_0) + \sin\theta\sin(\theta+\phi_0))) \\ &\quad -2t + 2t(\cos\theta\cos(\theta+\phi_0) + \sin\theta\sin(\theta+\phi_0))]. \end{aligned}$$

Observe that

$$\cos\theta\cos(\theta+\phi_0) + \sin\theta\sin(\theta+\phi_0) = \cos((\theta+\phi_0) - \theta) = \cos\phi_0,$$

yielding

$$\begin{aligned} x^2(t,\theta) + y^2(t,\theta) &= R^2[1 + 2t^2(1-\cos\phi_0) - 2t(1-\cos\phi_0)] \\ &= R^2[1 + 2(t^2-t)(1-\cos\phi_0)]. \end{aligned} \tag{15.6}$$

We have made progress: the parameter θ has been eliminated. In order to remove t we must now consider $z^2(t,\theta)$:

$$z^2(t,\theta) = z_0^2(1 + 4(t^2-t)),$$

from which it follows that

$$t^2 - t = \frac{z^2(t,\theta) - z_0^2}{4z_0^2}.$$

Substituting this into (15.6) and omitting the dependence on t and θ of x, y, z, we obtain

$$x^2 + y^2 = R^2 \left[1 + \frac{1}{2}(1 - \cos\phi_0)\frac{z^2 - z_0^2}{z_0^2}\right], \tag{15.7}$$

which is the equation of a hyperboloid of revolution of one sheet. In fact, to obtain the canonical form

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} - \frac{z^2}{c^2} = 1$$

it suffices to choose

$$\begin{cases} a = R\sqrt{\frac{1+\cos\phi_0}{2}}, \\ c = \frac{z_0\sqrt{1+\cos\phi_0}}{\sqrt{1-\cos\phi_0}}, \end{cases}$$

if $1 + \cos\phi_0 \neq 0$, equivalently $\cos\phi_0 \neq -1$ or again $\phi_0 \neq \pi$. For $\phi_0 = \pi$ we simplify to

$$x^2 + y^2 = \frac{R^2}{z_0^2} z^2,$$

which is the equation of a cone (see Exercise 10).

So we have shown that all lines (D_θ) lie on our quadratic surface (hyperboloid or cone). But does the quadratic surface contain other points? It is easy to show that this is not the case. Indeed, our surface is the union of circles located in the set of planes $z = z_1$, for $z_1 \in \mathbb{R}$ (in the case of the cone, one circle is reduced to a point when $z_1 = 0$). If we let $z = z_1$ in (15.5) we obtain $t = \frac{z_1+z_0}{2z_0}$. Replacing this value in $(x(t,\theta), y(t,\theta))$, we must show that the set of these points for $t = \frac{z_1+z_0}{2z_0}$ and $\theta \in [0, 2\pi]$ is a circle.

We will use the trigonometric formulas

$$\begin{aligned} \cos(a+b) &= \cos a \cos b - \sin a \sin b, \\ \sin(a+b) &= \sin a \cos b + \cos a \sin b. \end{aligned} \tag{15.8}$$

This allows us to write

$$\begin{cases} x(t,\theta) = R(1 + t(\cos\phi_0 - 1))\cos\theta - tR\sin\phi_0 \sin\theta, \\ y(t,\theta) = R(1 + t(\cos\phi_0 - 1))\sin\theta + tR\sin\phi_0 \cos\theta. \end{cases}$$

Let $\alpha = R(1 + t(\cos\phi_0 - 1))$ and $\beta = tR\sin\phi_0$. Let us write (α, β) in polar coordinates: $(\alpha, \beta) = (r\cos\psi_0, r\sin\psi_0)$. Then

$$\begin{cases} x(t,\theta) = r\cos\psi_0\cos\theta - r\sin\psi_0\sin\theta = r\cos(\theta + \psi_0), \\ y(t,\theta) = r\cos\psi_0\sin\theta + r\sin\psi_0\cos\theta = r\sin(\theta + \psi_0), \end{cases}$$

where the last equality again used (15.8). In this form it is clear that all points of the circle of radius r are attained when $\theta \in [0, 2\pi]$.

□

We have now seen that a hyperboloid of revolution of one sheet can be described as a family of straight lines. If you consider Figure 15.24 you can easily imagine a second family of such lines that is the mirror image of the first (see Exercise 12). Such a surface is said to be *doubly ruled*, since it may be constructed by either of two distinct families of straight lines. This is an advantage when such a form is realized in concrete. Not only can the pouring form be constructed using only straight pieces of wood, provided they are thin enough, but the concrete itself can be reinforced with two sets of straight pieces in two different directions. This greatly simplifies the construction of the pouring form and allows for an extremely solid structure.

The hyperbolic paraboloid. Under appropriately chosen orthonormal coordinates this surface has the canonical equation

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2}, \tag{15.9}$$

where $a, b > 0$ (see Figure 15.22). The intersection of this surface with a plane containing the z axis (a plane with the equation $Ax + By = 0$) is either a parabola or a horizontal line. On the other hand, the intersection of this surface with a plane parallel to the xy plane (a plane with the equation $z = C$) is a hyperbola in this plane if $C \neq 0$, and two straight lines if $C = 0$.

Proposition 15.14 *Let $B, C > 0$. Let (D_1) and (D_2) be the lines given by*

$$(D_1)\begin{cases} z = -Cx, \\ y = -B, \end{cases} \qquad (D_2)\begin{cases} z = Cx, \\ y = B. \end{cases}$$

We consider the line (Δ_{x_0}) joining the point $P(x_0) = (x_0, -B, -Cx_0)$ of (D_1) to the point $Q(x_0) = (x_0, B, Cx_0)$ of (D_2). Then the union of the family of lines (Δ_{x_0}) is a hyperbolic paraboloid (see Figure 15.25).

Proof. The line (Δ_{x_0}) passes through $P(x_0)$ with direction $\overrightarrow{P(x_0)Q(x_0)}$. Thus it is the set of points

$$(x(t, x_0), y(t, x_0), z(t, x_0)) = (x_0, -B + 2Bt, -Cx_0 + 2Ctx_0),$$

yielding

$$z = Cx_0(2t - 1) = \frac{C}{B}x_0 y = \frac{C}{B}xy.$$

If we substitute

$$\begin{cases} x = \frac{1}{\sqrt{2}}(X - Y), \\ y = \frac{1}{\sqrt{2}}(X + Y), \end{cases} \tag{15.10}$$

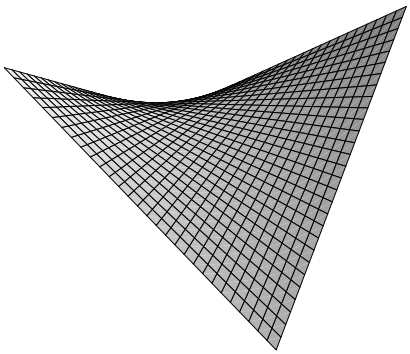


Fig. 15.25. Two families of straight lines on a hyperbolic paraboloid.

then the equation becomes

$$z = \frac{C}{B}xy = \frac{C}{2B}(X^2 - Y^2).$$

We immediately recognize the equation of a hyperbolic paraboloid. Remark: the change of variables of (15.10) is simply a rotation by $\frac{\pi}{4}$ of the coordinate system.

Here again we must show that any point of the hyperbolic paraboloid lies on one of the lines. Let (x, y, z) be a point on the hyperbolic paraboloid. It suffices to show that there exist x_0 and t such that $(x, y, z) = (x(t, x_0), y(t, x_0), z(t, x_0))$. Of course we choose $x_0 = x$. By letting $y = -B + 2Bt$ we get $t = \frac{y+B}{2B}$. Since z is on the hyperbolic paraboloid, we have $z = \frac{C}{B}xy$. This yields

$$z = \frac{C}{B}x(-B + 2Bt) = Cx(2t - 1) = z(t, x_0).$$

Hence $(x, y, z) = (x(t, x_0), y(t, x_0), z(t, x_0))$ for $x_0 = x$ and $t = \frac{y+B}{2B}$, which ensures that (x, y, z) is on the line (Δ_x). □

Proposition 15.14 suggests a method to construct a roof in the shape of a hyperbolic paraboloid. We place beams along (D_1) and (D_2) and we cover them with thinner beams or thin boards placed as the lines (Δ_{x_0}).

15.4 Optimal Cellular Antenna Placement in a Region

Cellular telephony is now a part of everyday life, with many companies offering service. In order to do this, each of these companies must first place antennas across the area they wish to serve in such a manner that (nearly) all points in the area may be served by a nearby antenna. At present, cellular services are quite reliable in and around large urban areas, but there are many remote regions that do not have access.

Suppose that a company wants to place antennas in a large territory so as to provide service to all points in the territory. In simpler terms, they wish to place the antennas in the territory in a manner such that every point will be no more than a distance r from the nearest antenna. The company considers several possible placement plans in order to determine which one requires the least number of antennas. For now we will consider only regular networks, and we will be comparing the following three schemes:

- placing antennas on a regular triangular network;
- placing antennas on a square network; and,
- placing antennas on a hexagonal network.

We will assume that the territory is sufficiently large and not too narrow so that we may safely ignore precisely what happens along its border.

Placing antennas on a regular triangular network. Consider covering a large city by placing antennas at the vertices of a regular triangular network. Two neighboring antennas are at a distance a, the side length of the equilateral triangles building up the network. In such a triangle the point that is the furthest away from the three corners is the center of the circle circumscribed about the triangle, which is the intersection point of the three perpendicular bisectors. Since the triangle is equilateral, this point is also the center of gravity situated at the intersection of the three medians. The length of the median is given by $h = a\cos\frac{\pi}{3} = \frac{\sqrt{3}}{2}a$. The second median crosses the first at the center of gravity of the triangle, which is situated two-thirds of the way along the median from a vertex. Thus, the center of gravity is at a distance of $\frac{2}{3}\frac{\sqrt{3}}{2}a = \frac{1}{\sqrt{3}}a$ from the vertices of the triangle. Because the antennas are at the vertices of the triangle and the center of gravity is the furthest point, each antenna must reach this point, thus requiring $r \geq \frac{1}{\sqrt{3}}a$. In order to minimize the number of antennas we take $r = \frac{1}{\sqrt{3}}a$. Hence we must take triangles with side lengths $a = \sqrt{3}r$. In conclusion, if the signal emitted by the antenna is usable up to a distance of r and the antennas are placed at the corners of a network of equilateral triangles, then we must take triangles with side lengths $a \leq \sqrt{3}r$ in order to ensure that all points in the territory will receive a usable signal.

Consider an $n \times n$ square territory for n much larger than r (see Figure 15.26). We will ignore exact behavior at the boundary. To traverse the square horizontally we need a line of $\frac{n}{\sqrt{3}r}$ points. Successive lines are situated a vertical distance h from one another. Since $h = \frac{\sqrt{3}a}{2} = \frac{3}{2}r$, we need $\frac{n}{h} = \frac{2n}{3r}$ lines to cover the entire square. Thus, we require

$$\frac{2}{3\sqrt{3}}\frac{n^2}{r^2} \approx 0.385\frac{n^2}{r^2} \tag{15.11}$$

points (or antennas) in total. This number is proportional to n^2, the area of the region to be covered.

In doing this calculation we neglected to discuss precisely how the points are aligned with respect to the boundary of the region. Should we put antennas along the boundary

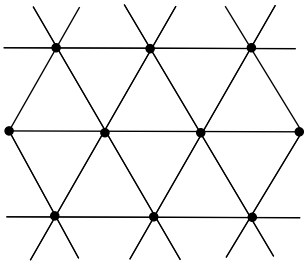


Fig. 15.26. A regular triangular network.

of the region or should we start the first row in its interior? At what lateral distance from the left boundary of the region should we place the first antenna? These questions are harder to answer than the simple calculation we performed above. However, we can easily convince ourselves that the difference in the number of antennas implied by the various possible placements near the boundaries is bounded above by Cn, for some positive constant C. If n is sufficiently large, then this difference quickly becomes negligible with respect to the bound given in equation (15.11), which is proportional to n^2. This remark is equally valid for the following discussion of regular square and hexagonal networks.

Placing antennas on a square network. Consider a square with side length a. In such a square the point that is the farthest away from the corners is the center of gravity situated at the intersection of the two diagonals. This point is at a distance of $r = \frac{1}{\sqrt{2}}a$ from the four corners. Thus we must use squares with side lengths of $a \leq \sqrt{2}r$.

Now consider an $n \times n$ square region for n much larger than r (see Figure 15.27). We will partition this region using a regular square network with side length a and place antennas at each of the nodes of the network. As discussed above, we may ignore

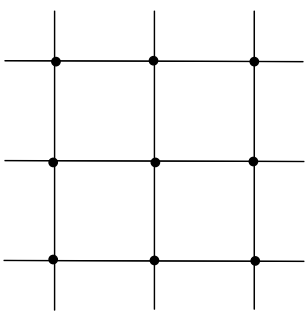


Fig. 15.27. Square network.

the details of positioning antennas near boundaries. We require a line of $\frac{n}{\sqrt{2}r}$ points to traverse the region horizontally and $\frac{n}{\sqrt{2}r}$ horizontal lines. Thus, we require

$$\frac{1}{2}\frac{n^2}{r^2} \approx 0.5\frac{n^2}{r^2}$$

antennas to cover the region.

Placing antennas on a hexagonal network. Now consider a regular hexagon with side length a. The point the farthest away from the vertices is the center of the hexagon situated at a distance a from each of the six vertices. Thus we must take hexagons with side length $a = r$.

To cover an $n \times n$ territory (see Figure 15.28), we will orient the hexagons such that

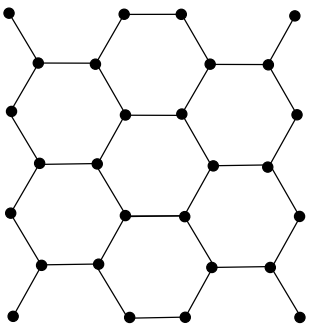


Fig. 15.28. Regular hexagonal network.

two of their edges are horizontal. We remark that along each horizontal line containing nodes of the network they are separated by distances $r, 2r, r, 2r, r, 2r, \ldots$. Thus the average distance between two successive points is $\frac{3}{2}r$. Hence, to traverse the region horizontally we require $\frac{2n}{3r}$ points. Each successive line is separated vertically by a distance h, where $h = \frac{\sqrt{3}}{2}r$. Thus we require $\frac{n}{h} = \frac{2n}{\sqrt{3}r}$ horizontal lines to cover the entire region. In total we require

$$\frac{4}{3\sqrt{3}}\frac{n^2}{r^2} \approx 0.770\frac{n^2}{r^2}$$

antennas, twice as many as are required using a triangular network.

If we compare the above three solutions, we see that the regular triangular partitioning is by far the most efficient, followed by the square partitioning and finally the hexagonal partitioning.

Just by visually inspecting the resulting networks we could have guessed that the triangular network would be exactly twice as efficient as the hexagonal network. In

fact, connecting the centers of the hexagons forms a regular triangular network (see Figure 15.29). The center of each triangle is situated at one of the nodes of the hexagonal

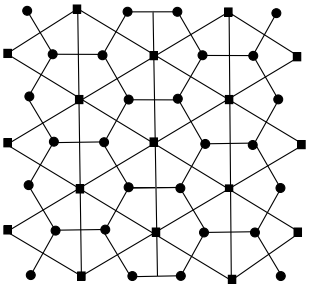


Fig. 15.29. Dual triangular and hexagonal networks.

network. This point is thus at exactly a distance r from the nearest triangle vertices. Along each horizontal line through the overlaid graphs we find two hexagon vertices for every triangle vertex.

15.5 Voronoi Diagrams

In this section we consider a problem that is in some sense the inverse to that in Section 15.4 (but you do not need to have read it). Suppose that we have a certain number of antennas distributed across a given region. We wish to divide this region into cells such that

- each cell contains exactly one antenna;
- each cell contains exactly the set of of points that are closer to the associated antenna than any other antenna (see Figure 15.30).

The set of cells obtained in this manner is called the Voronoi diagram of the antennas. In reality, antenna placement is subject to several constraints, both urban (zoning rules, availability of land, etc.) and geographic (antennas are more efficient if placed at peaks rather than in valleys). Drawing the Voronoi diagram for a network of antennas allows the planners to easily visualize poorly serviced areas and to plan new antenna placements.

A historical note. The Ukrainian mathematician Voronoi (1868–1908) defined the concept of Voronoi diagrams in arbitrary dimensions, but it was Dirichlet (1805–1859) who first studied them in detail in two and three dimensions. For this reason they are also called *Dirichlet tessellations*. Diagrams of this sort have actually been around since at least 1644, appearing in Descartes's notebooks.

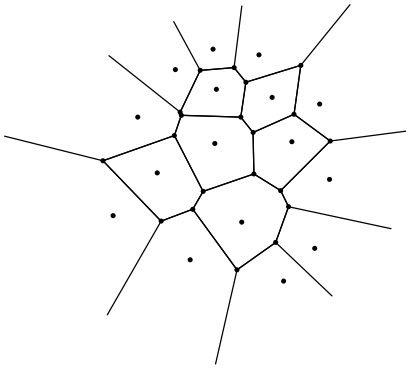


Fig. 15.30. A Voronoi diagram.

We can conceptually replace the antennas by post offices, hospitals, or even schools. In this last case it allows us to precisely determine the optimal school attendance areas such that each student will go to the closest school. As can be seen, Voronoi diagrams have numerous applications.

We describe the problem in mathematical terms.

Definition 15.15 *Let $S = \{P_1, \ldots, P_n\}$ be a set of distinct points in a region $\mathcal{D} \subset \mathbb{R}^2$. The points P_i are called* sites.

1. *For each site P_i the Voronoi cell of P_i, denoted by $V(P_i)$, is the set of points of $\mathcal{D}$ that are closer (or as close) to P_i than to any other site P_j:*

$$V(P_i) = \{Q \in \mathcal{D}, |P_iQ| \leq |P_jQ|, j \neq i\}.$$

2. *The Voronoi diagram of S, denoted by $V(S)$, is the decomposition of $\mathcal{D}$ into Voronoi cells.*

To decide how to approach the problem we will first consider the case $\mathcal{D} = \mathbb{R}^2$ and $S = \{P, Q\}$ with $P \neq Q$.

Proposition 15.16 *Let P and Q be two distinct points in the plane. The perpendicular bisector (or mediatrix) of the segment PQ is the locus of points at equal distance from P and Q. This locus is the straight line (D) that is normal to the segment PQ and that passes through its midpoint. All points R on one side of (D) satisfy $|PR| < |QR|$, while all those on the other side satisfy $|PR| > |QR|$. Thus, the Voronoi diagram of $S = \{P, Q\}$ is the partition of $\mathbb{R}^2$ into two closed half-planes bounded by (D) (see Figure 15.31).*

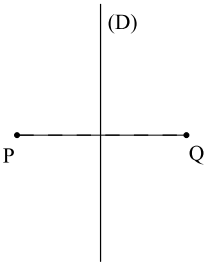


Fig. 15.31. Voronoi diagram of two points P and Q.

PROOF. The proof is left as an exercise to the reader. □

We now have the basic ingredients necessary to find the Voronoi cell $V(P_i)$ of a site P_i belonging to a collection of sites $S = \{P_1, \ldots, P_n\}$. We will limit ourselves to the case $\mathcal{D} = \mathbb{R}^2$, but the concept is similar in higher dimensions (see Figure 15.32).

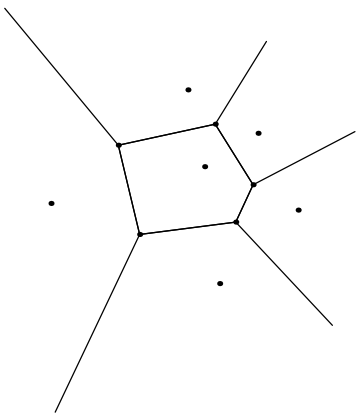


Fig. 15.32. A Voronoi cell.

Proposition 15.17 *Given a set of sites $S = \{P_1, \ldots, P_n\}$, for each pair of points (P_i, P_j) the perpendicular bisector of the segment P_iP_j divides the plane into two closed half-planes $\Pi_{i,j}$ and $\Pi_{j,i}$, the first containing P_i and the second containing P_j. The Voronoi cell $V(P_i)$ of the site P_i is the intersection of the half-planes $\Pi_{i,j}$ for $j \neq i$ (see Figure 15.32):*

$$V(P_i) = \bigcap_{j \neq i} \Pi_{i,j}.$$

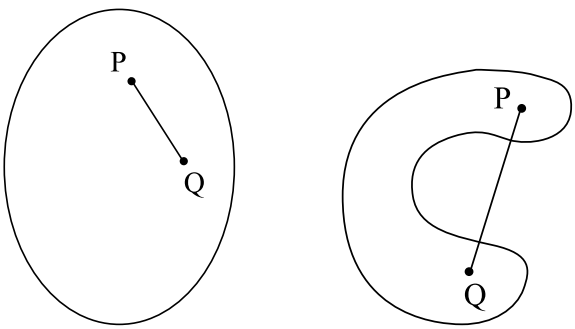
PROOF. The proof is simple. Let $\mathcal{R}_i = \bigcap_{j \neq i} \Pi_{i,j}$. We must show that $\mathcal{R}_i = V(P_i)$. Consider a point $R \in \mathcal{R}_i$. Then for all $j \neq i$ we have that $R \in \Pi_{i,j}$. Thus $|P_i R| \leq |P_j R|$ for all $j \neq i$. Hence $R \in V(P_i)$ by the definition of $V(P_i)$. So we have that $\mathcal{R}_i \subset V(P_i)$. Now suppose that $R \notin \mathcal{R}_i$: then there exists $j \neq i$ such that $R \notin \Pi_{i,j}$. Therefore $|P_i R| > |P_j R|$ and finally $R \notin V(P_i)$.

So we can conclude that $\mathcal{R}_i = V(P_i)$. □

We now consider the general form of Voronoi diagrams.

Definition 15.18 *A subset $\mathcal{D}$ of the plane is convex if for all points $P, Q \in \mathcal{D}$ the segment PQ lies within $\mathcal{D}$.*

Figure 15.33 gives an example of both a convex and a nonconvex set.



(a) A convex set (b) A nonconvex set

Fig. 15.33. Convex and nonconvex sets.

Proposition 15.19 *A Voronoi cell is a convex set. If the cell is finite (lies within some disk with finite radius r), then it is a polygon.*

PROOF. We present a rough idea of the proof, leaving the rest as an exercise. The entire proof centers on the following two facts: a half-plane is a convex set, and the intersection of convex sets is itself convex. □

Constructing Voronoi diagrams. It is not easy in practice to construct the Voronoi diagram of a set S of sites, especially when S is large. Research into algorithms for constructing these diagrams is ongoing and active in both combinatorial and computational geometry. However, there exists a large number of software packages and programming languages that allow for the efficient calculation of Voronoi diagrams. For example, Figures 15.30 and 15.32 were both created using a built-in function of Mathematica.

Voronoi diagrams are often displayed along with their "dual" *Delaunay triangulations*. An equally important problem in combinatorial geometry is to construct a partition of a set into triangles (called a *triangulation*), so that either two triangles have empty intersection or they share a common edge. Given a set S of sites and its Voronoi diagram, we can construct the Delaunay triangulation as follows: the vertices of the triangles are the sites S; we connect the sites P_i and P_j with the segment P_iP_j if the cells $V(P_i)$ and $V(P_j)$ share a common edge (see Figure 15.34).

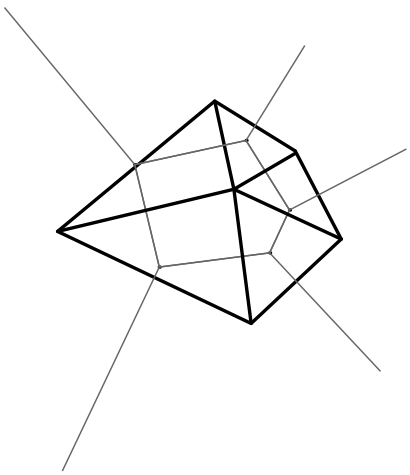


Fig. 15.34. The Delaunay triangulation (shown in thick black lines) associated with the Voronoi diagram of Figure 15.32 (shown in thin gray lines).

Given a set of sites, there exist many possible triangulations whose corners lie on the sites in S. However, the Delaunay triangulation is a triangulation with more equilateral (less flattened) triangles on average than others. Because of this property, Delaunay triangulations find use in many applied problems, in particular when meshes are required. (See also Exercise 24.)

The reciprocal problem. We have seen that given a set S of sites, we can calculate the associated Voronoi diagram that partitions the region into convex cells. More specifically, bounded cells are convex polygons, while nonbounded cells have a boundary consisting of a finite number of connected line segments and two half-rays. There is nothing stopping us from generalizing this problem to partitioning a surface rather than a planar region. The reciprocal problem, however, is harder: suppose that we have a partition of the plane (or a surface) into cells as described above. Under what conditions does there exist a set of sites S such that the provided diagram is the Voronoi diagram $V(S)$ of S? We can easily think of a modeling process that produces Voronoi diagrams. Suppose we were to light a small fire at each site, which was to spread outward in all

directions at a constant velocity. The points where the fire from two sites meet will describe the edges of the boundaries, while the points where the fires from three or more sites meet will be precisely the corners of the cells (see Chapter 4 for another problem using such a modeling technique, particularly Exercise 19 of that chapter). Another similar model is provided when sites are taken as points of a piece of blotting paper, and we put drops of ink on the sites that spread in all directions at constant velocity. The cell of a site is the set of points that have been reached first by the ink of that site. If we have some reason to think that the partition of the surface we are inspecting has been constructed by a process similar to those above, then it is likely that there will be an associated set of sites S. However, if we have no idea how the partition was created, then the problem must be approached in purely mathematical terms. We will discuss some simple cases in Exercise 26.

15.6 Computer Vision

In this section we consider only a small part of computer vision, which consists in reconstructing depth information starting from 2D images. We start with two photos taken by two observers situated at O_1 and O_2. In our model the images of the point P are P_1 and P_2 respectively. These points are situated at the intersection of the planes of projection and the lines (D_1) and (D_2) joining P to O_1 and O_2, respectively (see Figure 15.35). In Figure 15.35 the same plane of projection has been taken for each image, but this is not required. The plane of projection corresponds to the plane of the film or sensor of the camera.

The points O_i and P_i are known, so they uniquely define the line (D_i) as the line joining them. Since P_1 and P_2 are images of the same point, then (D_1) and (D_2) will intersect at a unique point P. This allows us to compute the location of P.

Let us do the details of the computation. We choose a system of axes such that O_1 and O_2 are located on the x axis and the origin lies exactly midway between the two. We choose the units such that $O_1 = (-1, 0, 0)$ and $O_2 = (1, 0, 0)$. We choose the y axis to be horizontal and scale it such that the planes of projection lie within the plane $y = 1$. The z axis is vertical and its scale can be chosen arbitrarily. Under this coordinate system the coordinates of the points P_i are $(x_i, 1, z_i)$. They are known because they can be measured directly from each of the photos.

Let (a, b, c) be the coordinates of P. These are the unknowns. To find them we will use the parametric equations of the lines (D_i). The line (D_1) passes through O_1 and its direction is given by the vector $\overrightarrow{O_1P_1} = (x_1 + 1, 1, z_1)$. Thus, (D_1) is the set of points

$$(D_1) = \{(-1, 0, 0) + t_1(x_1 + 1, 1, z_1) | t_1 \in \mathbb{R}\}.$$

Similarly we have that $\overrightarrow{O_2P_2} = (x_2 - 1, 1, z_2)$ and therefore

$$(D_2) = \{(1, 0, 0) + t_2(x_2 - 1, 1, z_2) | t_2 \in \mathbb{R}\}.$$

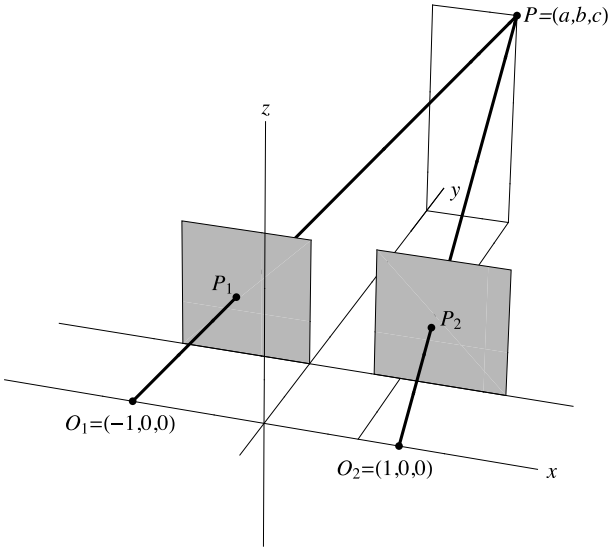


Fig. 15.35. Two photos taken from different points of view.

The point P is the intersection point of (D_1) and (D_2). To find it we look for t_1 and t_2 such that the point of (D_1) corresponding to t_1 coincides with the point of (D_2) corresponding to t_2:

$$\begin{cases} -1 + t_1(x_1 + 1) = 1 + t_2(x_2 - 1), \\ t_1 = t_2, \\ t_1 z_1 = t_2 z_2. \end{cases} \tag{15.12}$$

The second equation gives us $t_1 = t_2$. Replacing in the first equation yields

$$t_1 = \frac{2}{x_1 - x_2 + 2}. \tag{15.13}$$

Observe that $x_1 - x_2 + 2 > 0$; thus t_1 is positive. In fact, looking at Figure (15.35) we see that the distance between P_1 and P_2 is given by $x_2 - x_1$ and is smaller than the distance between O_1 and O_2, which is 2. Now consider the third equation of (15.12). Since $t_1 = t_2 \neq 0$, it tells us that $z_1 = z_2$: this is a necessary condition for the points P_1 and P_2 to be projections of the same point P. In fact, if we take two arbitrary points P_1 and P_2, the lines (D_1) and (D_2) will generally not intersect. The condition $z_1 = z_2$ ensures that the two lines are situated in the same plane $z = z_1 y$ and therefore that they will intersect if $x_1 - x_2 \neq 2$.

We now have located the point P:

$$\begin{aligned}(a,b,c) &= (-1,0,0)+\frac{2}{x_1-x_2+2}(x_1+1,1,z_1)\\ &= \left(\frac{x_1+x_2}{x_1-x_2+2},\frac{2}{x_1-x_2+2},\frac{2z_1}{x_1-x_2+2}\right).\end{aligned}$$

Remark. This is the mechanism behind our own depth perception. Our eyes observe the same scene from two points of view, and our brain uses geometry to "calculate" the depth of individual objects in the scene. Thus we must first understand the geometry behind depth perception before we can teach computers to do the same thing.

15.7 A Brief Look at Computer Architecture

Computers are built primarily with integrated circuits. The basic building block is the transistor, which may be roughly equated to an electrical switch. It is the precise layout and connection of millions of these transistors that allows a computer to do its work and in particular to compute operations.

We will consider only very simple electric circuits consisting entirely of switches. Each switch can take one of two positions, which we will associate with the numbers 0 and 1.

In this section we limit ourselves to showing how to construct circuits that can effectuate basic mathematical operations on the set $S = \{0, 1\}$. Programming languages are designed to allow compact and readable representations of complex calculations, which are in turn translated into long series of basic operations. Computers are designed to perform these basic operations, placing their results in appropriate places in memory. Early computers could perform only a single operation at a time, while modern computers typically perform many operations in parallel.

We will consider several basic operations performed by all modern computers and the electric circuits that realize them. Specifically, we will consider the Boolean operators NOT, AND, OR, and XOR (exclusive or), which operate on the set $S = \{0, 1\}$. The value 0 will be used to indicate an absence of electrical current, while 1 will mean that current is flowing.

The AND operator. The function AND : $S \times S \to S$ is given by the following table:

AND	0	1
0	0	0
1	0	1

(15.14)

Why do we call this operator "AND"? Suppose that A and B are two statements. We can assign each of them a truth value in S, 0 meaning that the statement is false and 1 meaning that it is true. Consider the logical statement A AND B. This statement is true only if A and B are both true. In the three other cases (A true and B false, A false and B true, A false and B false), the statement A AND B is false and therefore is

assigned the value 0. This is exactly the operation described in the above table. Notice that the AND operator is also equivalent to multiplication modulo 2, an operation of arithmetic modulo 2 that is used in several other chapters. A simple circuit modeling

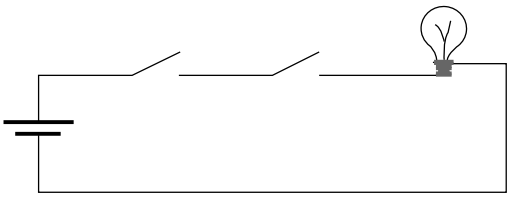


Fig. 15.36. A circuit realizing the AND operator.

this operation is shown in Figure 15.36. There are two switches in the circuit, each one corresponding to one of the two inputs. When the input is 1, the switch is closed and current flows through it. When the input is 0, the switch is open and no current may flow through it. It is easy to see that current will flow through the entire circuit if and only if both switches are closed. Current flowing through the entire circuit (and illuminating the bulb at the end) indicates an output of 1, while absence of current yields an output of 0.

Table (15.14) may be rewritten in the following form:

INPUT A	INPUT B	OUTPUT
0	0	0
0	1	0
1	0	0
1	1	1

(15.15)

The OR operator. This is the function OR : $S \times S \to S$ given in the following table:

OR	0	1
0	0	1
1	1	1

(15.16)

The statement A OR B is true when at least one of the statements A and B is true. Thus, the only time it is false is when the two statements A and B are both false. A simple circuit implementing this operation is shown in Figure 15.37. The rules of operation are the same as for the AND switch, but this time the two switches are in parallel. It is easy to see that current will flow through the circuit if either of the two switches is closed. Once again, current flowing through the circuit indicates a value of 1 or true, and vice versa. As with the AND operator we may rewrite table (15.16) as follows:

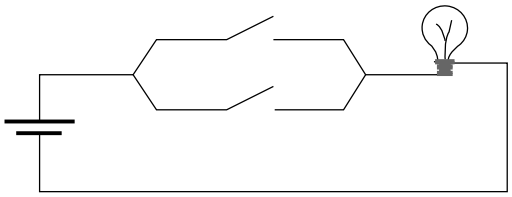


Fig. 15.37. A circuit realizing the OR operator.

INPUT A	INPUT B	OUTPUT
0	0	0
0	1	1
1	0	1
1	1	1

(15.17)

The XOR operator (sometimes written $\oplus$). The XOR operator is the function $\mathrm{XOR} : S \times S \to S$, given by the table

XOR	0	1
0	0	1
1	1	0

(15.18)

The statement A XOR B is true if and only if exactly one of the two statements A and B is true and the other is false, from which comes the name *exclusive or*. We remark that the truth table of the XOR operator is the same as that of addition modulo 2, which we have met in other chapters. A circuit implementing this operation is shown

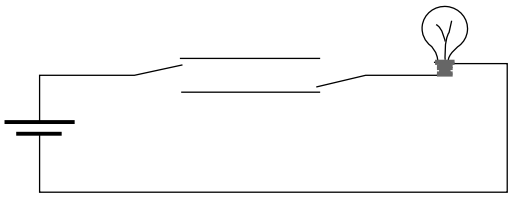


Fig. 15.38. A circuit realizing the XOR operator.

in Figure 15.38. This circuit is slightly more subtle than the others thus far. The left switch is in the upper position when the input is 1 (the switch is on), and in the lower position when the input is 0 (the switch is off). The right switch behaves in the opposite manner. Thus, we see that current will flow through the circuit when one of the switches is on and the other is off. We rewrite table (15.18) as follows:

INPUT A	INPUT B	OUTPUT
0	0	0
0	1	1
1	0	1
1	1	0

(15.19)

The NOT operator. The NOT operator is the function $\mathrm{NOT} : S \to S$ given by

$$\begin{cases} \mathrm{NOT}(0) = 1, \\ \mathrm{NOT}(1) = 0, \end{cases} \tag{15.20}$$

or equivalently

INPUT	OUTPUT
0	1
1	0

(15.21)

Consider Figure 15.39. There is exactly one switch that receives the input. The bulb

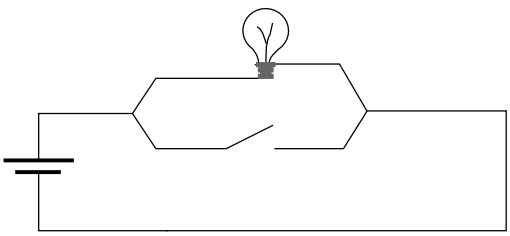


Fig. 15.39. A circuit realizing the NOT operator.

acts as a resistive load. The switch is situated on a parallel branch that has a lower resistance, than that of the bulb. When the input is 1 the switch is closed, the current will flow through the branch of less resistance, and the bulb will not be lit. However, when the switch is open (the input is 0), the current will flow through the only available branch and thus the bulb will be lit.

Some further thoughts. We pause to extract some deeper ideas from these simple examples.

1. When discussing the NOT operator we said that the bulb will not illuminate when the switch is closed. However, in real circuits a portion of the current will still flow through the upper branch and the bulb will in fact be dimly lit. Although we may consider our inputs and outputs as discrete values, current flow is effectively a continuous quantity. Thus, in real computers 0 and 1 values are distinguished through the use of a threshold. A current below the threshold value is interpreted as a 0, while one above it is considered as 1.

2. Each circuit considered so far has been self-contained, with inputs taking the form of switches, and outputs the form of light bulbs. It is easy to imagine that the switches acting as inputs may actually be controlled by some external process, for instance another circuit. Our input can then be the output of that circuit. Similarly, it is entirely possible that the output light bulbs act as inputs to yet other circuits. This is the case in modern computers, where the outputs could be used as inputs for further operations.

There exist other Boolean operators commonly used in computers: NAND and NOR. They are defined as

$$\begin{cases} A \text{ NAND } B \iff \text{NOT}(A \text{ AND } B), \\ A \text{ NOR } B \iff \text{NOT}(A \text{ OR } B). \end{cases} \tag{15.22}$$

Given their definition, we see that they may be implemented by combining a NOT circuit with an AND and OR circuit, respectively. However, they may be more efficiently realized by smaller circuits. As such, these two operators are often added to the list of basic Boolean operators. These two operators are called *universal*. Exercise 34 will explain why.

A first small step toward computers. Computers are built from transistors, which may be visualized as sophisticated switches. Analogously, we may consider them as "discriminators," working in only one direction, much as, for example, a door whose frame allows it to be opened in one direction only. A transistor can deliver an output without being affected by what happens afterward. For that, rather than interpreting the presence of a current as a 1, transistors use voltage differentials as input. When the voltage differential across its inputs is greater than a given threshold and has the proper sign, this creates a current that "opens" the door.

A word on very large scale integration systems (VLSI). Transistors can be used to create diverse logic families: TTL, ECL, NMOS, CMOS, etc. The beauty of these logic families is that transistors are assembled together to create "gates" that realize the AND, OR, XOR, and NOT operators (and often also the NAND and NOR operators). Each output can be used as the input of another circuit. This allows for the assembly of extremely complex circuits using many millions of gates and individual transistors. In most of these logic families the voltage differential represents the logical level and the current transports the charge that is required to attain these differentials. MOS transistors have historically been made in three layers: a layer of silicon, a layer of oxide (insulator), and a layer of metal (acting as the switch). Nowadays, the metal layer has been replaced by polycrystalline silicon and the insulating layer is extremely thin, being on the order of 12 Å (1 Å = 1 angstrom = 10^{-10} m). For comparison, a typical atomic bond has a length of approximately 2 Å. By far, CMOS is the most commonly used logic family. Along with NMOS/PMOS its efficiency resides in the fact that current flows only while the transistor state is in transition, in contrast to our simple circuits using light bulbs. Transition between logic states is effectuated by a transfer of charge,

carried by a current. Once the transfer of charge has been completed, current no longer flows. Thus, while in a steady state such transistors do not use energy. This allows for the construction of extremely large integrated circuits (more than a billion transistors) with reasonable energy consumption (< 150 W).

From a practical point of view, NAND and NOR gates are more important. This is because with CMOS technology, they are more easily and naturally constructed than other gates. Similarly, for practical reasons (NMOS transistors are better than PMOS transistors), NAND gates are preferred over NOR gates.

15.8 Regular Pentagonal Tiling of the Sphere

A few years ago, one of the authors (C.R.) was approached by Pierre Robert, called "Pierre the Juggler," woodworker and juggler, who constructs large balls for jugglers and acrobats to balance on. He had constructed a 50-cm-diameter wooden ball on which he wanted to paint five-pointed stars in a regularly tiled manner (see Figure 15.40). (In fact, in Quebec it is still common for woodworkers to work in imperial units; thus he had actually constructed a ball with a radius of 20 inches.)

Fig. 15.40. A circus ball painted with a regular tiling of five-pointed stars.

There exists a regular polyhedron whose 12 faces are regular pentagons, called the *dodecahedron* (see Figure 15.41). Since this polyhedron is regular, it may be inscribed in a sphere, meaning that all of its vertices lie along the surface of a sphere. Thus, the artist was in fact asking for a method of finding the vertices of the dodecahedron inscribed in the sphere he had constructed.

Drawing on a sphere. A woodworker who needs to draw on the surface of a sphere cannot do so using a ruler. However, a compass works quite well. So this will be our

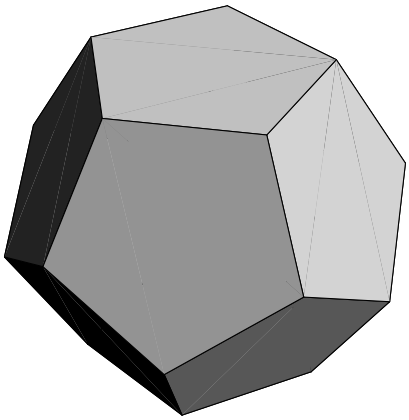


Fig. 15.41. A dodecahedron.

tool for finding the vertices of the inscribed dodecahedron. Once we have specified two points on the surface of the sphere, we can easily draw a great-circle arc between the two points by holding a string between the two points. Assuming that the friction between the string and the ball is negligible, the string will tend to follow a great-circle route. This method is sufficient if we do not require high precision. If we wish a more precise technique we must use a compass, calculating both its opening angle and the precise spot where to place its point (see Exercise 41).

Using a compass to draw on a sphere. If we place the point of a compass at a point N on the surface of a sphere and give it an opening of r', then we will draw a circle of radius $r \neq r'$ on the surface of the sphere (see Figure 15.45). The actual center P of the circle will lie in the interior of the sphere and is therefore not situated at N. However, all of the points along the circle just drawn will lie at a distance r' from N. We must pay close attention to this subtlety throughout our discussion. The actual relation between the radius of the circle r and the opening of the compass r' depends on the radius R of the sphere. It will be discussed later.

We will present a solution for drawing the vertices of the inscribed dodecahedron. Here are the symbols we will use in our discussion:

- R is the radius of the sphere;
- a is the length of an edge of the dodecahedron inscribed on the sphere;
- d is the length of a diagonal of a pentagonal face of the dodecahedron;
- r is the radius of the circle circumscribed about a pentagonal face;
- r' is the opening that must be given to a compass in order to draw a circle of radius r on a sphere of radius R.

Main ingredients of the solution. The first step is to calculate the length a of an edge and the length d of a diagonal of a pentagonal face of a dodecahedron, when the dodecahedron is inscribed in a sphere of radius R. As such it looks very difficult.

- Luckily we will be able to use a remarkable property of the dodecahedron: the diagonals of the pentagons are the edges of cubes inscribed on the dodecahedron. There are five such cubes (see Figure 15.42 and Exercise 44).

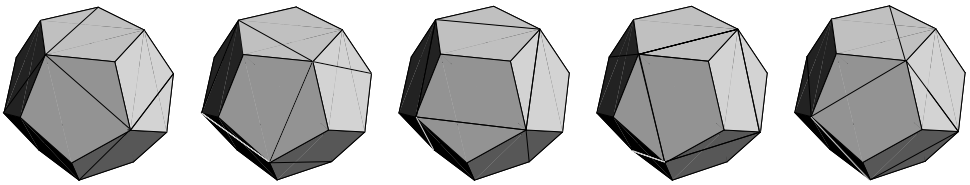


Fig. 15.42. The five cubes inscribed on a dodecahedron.

- Thus we have already reduced the problem to one that is slightly simpler. Each of these five cubes is itself inscribed in the sphere. Thus, we are looking for the edge length d of a cube inscribed in a sphere of radius R. We leave the actual calculation of this relationship to Exercise 39:

$$d = \frac{2}{\sqrt{3}} R.$$

- We must now find the relation between a and d. Since edges of the inscribed cube are diagonals of the inscribed pentagons, the problem is reduced to a planar one. Given a regular pentagon with side lengths a, find the length of its diagonal d (see Figure 15.43). The formula is evident after inspecting Figure 15.43 and noticing that the interior angles of the pentagon are $\frac{3\pi}{5}$. We leave this part to Exercise 36. The length is given by

$$d = 2a \cos \frac{\pi}{5}.$$

  Thus, we now know that

$$a = \frac{d}{2 \cos \frac{\pi}{5}} = \frac{R}{\sqrt{3} \cos \frac{\pi}{5}}.$$

Drawing the vertices of the dodecahedron. We have now seen the main ingredients necessary. We choose a random point P_1 on the sphere that will be one vertex of the dodecahedron. Each vertex is adjacent to three other vertices that are a distance a from P_1. Thus, we draw a circle C_1 centered at P_1 using a compass opened to a length of a. (P_1 is not in the plane of this circle!). We choose a random point P_2 along this

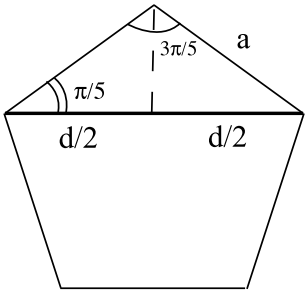


Fig. 15.43. A diagonal of a pentagon.

circle, which will be a second vertex of the dodecahedron. From this moment onward all of the vertices of the dodecahedron are uniquely determined. There are two other vertices P_3 and P_4 that lie along the circle C_1. Since these are situated a distance d from each other, we find them by finding the intersections of C_1 with the circle C_2 drawn by placing the point of the compass at P_2 and setting its opening to d. We continue this process by drawing a circle about the point P_2 using a compass opening of a, and then finding the two other vertices of the dodecahedron along this circle: they are located at distance d from P_1. We iterate this process for each of the other vertices (there are 20 vertices).

In our example we have that $R = 25$cm, yielding $a \approx 17.9$ cm and $d \approx 28.9$ cm.

The method given allows us to mark the vertices of the dodecahedron but not the centers of the pentagonal faces. In order to mark the center of each face we require one more ingredient. We proceed in two steps. We begin by finding the radius r of the circle circumscribed about a regular pentagon with side length a (see Figure 15.44). We leave

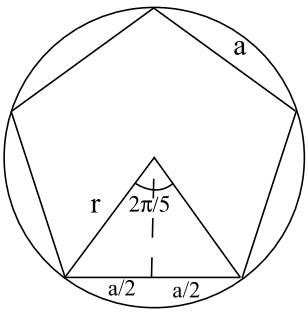


Fig. 15.44. Radius of the circle circumscribed about a regular pentagon.

it to Exercise 40 to show that this radius r is given by

$$r = \frac{a}{2\sin\frac{\pi}{5}}.$$

Thus we have that

$$r = \frac{R}{2\sqrt{3}\sin\frac{\pi}{5}\cos\frac{\pi}{5}} = \frac{R}{\sqrt{3}\sin\frac{2\pi}{5}}. \tag{15.23}$$

The missing ingredient at this step is the distance between the center (on the surface of the sphere) of the spherical pentagon and its vertices. This distance is the opening that must be given to the compass in order for it to draw the circle circumscribed about the pentagon when the point of the compass is placed at the center of the spherical pentagon. It can be determined as a special case of the following proposition.

Proposition 15.20 *We wish to draw a circle of radius r on a sphere of radius R. To do this we place the point of a compass at a point N, and give it an opening of*

$$r' = \sqrt{r^2 + \left(R - \sqrt{R^2 - r^2}\right)^2}. \tag{15.24}$$

Proof. We assume that the circle we wish to draw is located in a horizontal plane (see Figure 15.45). We must calculate the length $r' = |NA|$. We do this by applying the

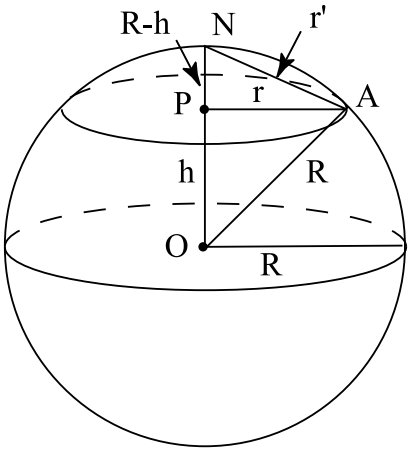


Fig. 15.45. To draw the circle centered at P with radius r we place the point of the compass at N and give the compass an opening of r'.

Pythagorean theorem to the two right triangles OPA and APN. This yields

$$h = \sqrt{R^2 - r^2},$$

and finally

$$r' = \sqrt{r^2 + (R-h)^2}.$$

□

In our case, the radius r of the circle circumscribed about a pentagon is given by equation (15.23), which yields

$$h = R\sqrt{1 - \frac{1}{3\sin^2 \frac{2\pi}{5}}}.$$

Thus

$$R - h = R\left(1 - \sqrt{1 - \frac{1}{3\sin^2 \frac{2\pi}{5}}}\right).$$

Using a calculator we obtain that $r' \approx 0.641R$. For $R = 25$ cm, this opening is $r' \approx 16.0$ cm.

An alternative method of drawing. Choose N on the surface of the sphere and draw the circle C obtained by placing the point of the compass at N and setting its opening to a length of r'. Choose a point A_1 on this circle that will be a vertex of the dodecahedron. Place the point of the compass at A_1 and set its opening to a length of a. Find the two points of intersection A_2 and A_3 between this circle and the circle C. Moving the compass first to A_2 and then to A_3 (while keeping the same opening a!) yields the two other vertices of the pentagonal face lying along the circle C.

We now look for the center of a second pentagonal face. Such a center is, for example, situated at a distance r' from each of the points A_1 and A_2. To find this we give the compass an opening of r' and draw the two circles centered at A_1 and A_2. These two circles will intersect at two points, one of them being the point N and the other being the center of the other pentagonal face containing the vertices A_1 and A_2. We repeat this process until we have found all of the vertices and all of the centers.

This problem contains one last piece of mathematics known in the ancient world. The formula for a makes use of the value $\cos \frac{\pi}{5}$, which we can easily calculate using a calculator. However, we will show that

Theorem 15.21

$$\cos \frac{\pi}{5} = \frac{1+\sqrt{5}}{4}.$$

Proof. The proof is simplified using Euler's formula and complex numbers:

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

We have that

$$e^{i\frac{\pi}{5}} = \cos\frac{\pi}{5} + i\sin\frac{\pi}{5}.$$

Moreover, using the properties of exponentials we have that

$$(e^{i\frac{\pi}{5}})^5 = e^{i\pi} = \cos\pi + i\sin\pi = -1. \tag{15.25}$$

On the other hand,

$$(e^{i\frac{\pi}{5}})^5 = \left(\cos\frac{\pi}{5} + i\sin\frac{\pi}{5}\right)^5.$$

Substituting $c = \cos\frac{\pi}{5}$ and $s = \sin\frac{\pi}{5}$, we obtain

$$(e^{i\frac{\pi}{5}})^5 = c^5 + 5ic^4s - 10c^3s^2 - 10ic^2s^3 + 5cs^4 + is^5. \tag{15.26}$$

Since the real and imaginary parts of equations (15.25) and (15.26) are independently equal, we obtain the following system of two equations:

$$\begin{aligned} c^5 - 10c^3s^2 + 5cs^4 &= -1, \\ 5c^4s - 10c^2s^3 + s^5 &= 0. \end{aligned} \tag{15.27}$$

The second equation of (15.27) can be factored as $s(5c^4 - 10c^2s^2 + s^4) = 0$, and since $s \neq 0$, we obtain

$$5c^4 - 10c^2s^2 + s^4 = 0. \tag{15.28}$$

Let $C = \cos\frac{2\pi}{5}$. We have the following trigonometric formula:

$$c^2 = \frac{1+C}{2}, \qquad s^2 = \frac{1-C}{2}. \tag{15.29}$$

Substituting into (15.28) yields

$$16C^2 + 8C - 4 = 4(4C^2 + 2C - 1) = 0.$$

This equation has both a positive and a negative root. Since $C = \cos\frac{2\pi}{5} > 0$, we have

$$C = \cos\frac{2\pi}{5} = \frac{-1+\sqrt{5}}{4}.$$

From this and (15.29) we can deduce

$$c^2 = \frac{3+\sqrt{5}}{8} \qquad \text{and} \qquad s^2 = \frac{5-\sqrt{5}}{8}. \tag{15.30}$$

The first equation of (15.27) can be rewritten as

$$c(c^4 - 10c^2s^2 + 5s^4) = -1,$$

from which it follows that

$$c = -\frac{1}{c^4 - 10c^2s^2 + 5s^4} = -\frac{1}{1-\sqrt{5}} = \frac{1+\sqrt{5}}{4}.$$

□

15.9 Laying Out a Highway

In civil engineering, when a highway is planned it is first drawn on a map. At some point, this layout of the highway must be marked out on the ground itself. To do this, the path of the highway is marked out using pickets, small wooden sticks typically brightly painted or with an attached brightly colored flag. Usually, the path of the highway will be closely approximated using straight-line segments and circular arcs.

Suppose that we wish to place pickets along the arc of a circle, each picket placed at a distance a from the next (for example, $a = 10$ m or $a = 30$ m). We assume that

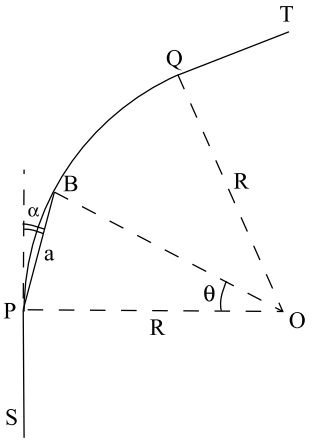


Fig. 15.46. Marking off a highway.

the segments SP and QT have already been marked out and that we need to mark an arc of a circle of radius R tangent to SP and QT. The plan of the engineers has been drawn such that such an arc exists! Then the center of the arc is the point O that is the intersection point of the line through P perpendicular to PS with the line through Q perpendicular to QT. If the plan is exact and has been accurately reproduced on the ground, then this point O is at distance R from P and Q. We now wish to place pickets along the arc of the circle centered at O with radius R and extreme points P and Q. The first point B to be marked will be at distance a from P. To mark it we need to determine the angle α between the line PS (the tangent of the circle at P) and the segment PB. Indeed, this allows us to mark the highway while staying on it.

Proposition 15.22 *(i)* $\alpha = \frac{\theta}{2}$*, where* θ *is the angle subtending the chord* PB*.*
(ii) $\alpha = \arcsin \frac{a}{2R}$.

Proof.

(i) Observe that OP is perpendicular to PS. Thus

$$\alpha = \frac{\pi}{2} - \widehat{OPB}.$$

Since the triangle OPB is isosceles, we have that

$$\widehat{OBP} = \widehat{OPB} = \frac{\pi}{2} - \alpha.$$

Moreover, the sum of the angles of the triangle is π. Thus

$$\widehat{OPB} + \widehat{OBP} + \widehat{POB} = 2\left(\frac{\pi}{2} - \alpha\right) + \theta = \pi - 2\alpha + \theta = \pi.$$

It follows that $2\alpha = \theta$, which proves (i).

(ii) Let X be the center of PB. Then OX is perpendicular to PB, since the triangle PBO is isosceles and

$$PX = \frac{a}{2} = R\sin\frac{\theta}{2}.$$

Since $\frac{\theta}{2} = \alpha$, then

$$a = 2R\sin\alpha,$$

proving (ii).

□

It suffices to place the picket B at a distance a from P along the straight line that forms an angle of $\alpha = \arcsin\frac{a}{2R}$ with the segment SP. This is a simple operation using standard surveying tools.

15.10 Exercises

The laws of reflection and refraction

1. We place two mirrors in the base of a box such that they form a right angle with each other. Show that any incoming vertical ray will be reflected parallel to itself (see Figure 15.47).

2. Exercise 18 of Chapter 1 discusses the operation of the sextant, a navigational instrument relying on the law of reflection. If you have not already done so, answer this question.

Conics

3. We already considered this problem in the plane. What we call a parabolic mirror is actually a circular paraboloid

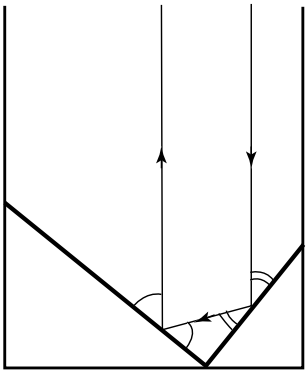


Fig. 15.47. Two perpendicular mirrors as in Exercise 1.

$$z = a(x^2 + y^2).$$

If all of the incoming rays arrive parallel to the axis of the mirror and are reflected according to the law of reflection, then show that all of the reflected rays pass through the same point, namely $(0, 0, \frac{1}{4a})$. To do this, use the planar result with the curve $z = ax^2$ and then make an argument for the general case by using the symmetry of the mirror for all rotations about its axis of revolution. The reflected ray will lie within the plane implied by the initial ray and the central axis of the paraboloid.

4. The remarkable property of the hyperbola. Consider a line L passing through a focal point of the hyperbola and a point P on the associated branch of the hyperbola. Let L' be the line symmetric to L about the tangent line to the hyperbola at P. Show that L' passes through the second focal point of the hyperbola (see Figure 15.19).

5. The telescope with liquid mirror from the ALPACA project. The plan of the telescope ALPACA to be installed on top of a Chilean mountain is given in Figure 15.48. Explain which conic shapes should be given to the three mirrors and how to place their respective foci. (More information on this telescope is given in Section 14.11 of Chapter 14.)

6. Rather than turning the large parabolic mirror, a solar furnace makes use of an array of smaller heliostats that reflect the sun's rays such that they strike the parabolic mirror parallel to its axis. For this exercise we assume that the heliostat consists of a flat mirror. At each point on the surface of this mirror a ray of light arrives that must be reflected parallel to the axis of the solar furnace.

(a) Show that the normal of the heliostat at a point P must bisect the angle between the incoming rays of sunlight striking P and the line originating from P that is parallel to the axis of the solar furnace.

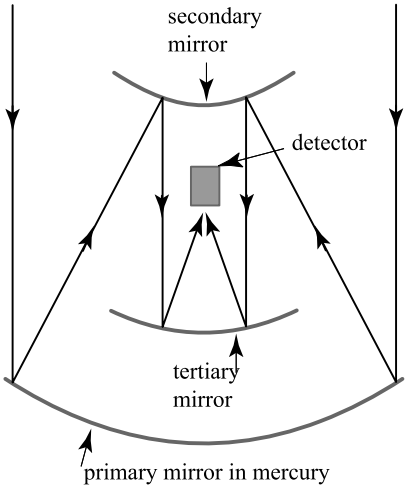


Fig. 15.48. The telescope from the ALPACA project (Exercise 5).

(b) In order to express directions we must first equip ourselves with a coordinate system. A direction is given by a unit vector. The tip of this vector lies along the unit sphere and may therefore be expressed in spherical coordinates as

$$(\cos\theta\cos\phi, \sin\theta\cos\phi, \sin\phi),$$

where $\theta \in [0, 2\pi]$ and $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Show that if

$$P_i = (\cos\theta_i\cos\phi_i, \sin\theta_i\cos\phi_i, \sin\phi_i), \qquad i = 1, 2,$$

then the direction of the bisector of the angle $\widehat{P_1OP_2}$ is given by the vector $\frac{\mathbf{v}}{|\mathbf{v}|}$, where $\mathbf{v} = \overrightarrow{OP_1} + \overrightarrow{OP_2}$.
Remark: The mirror on a heliostat is mounted to a gimbal, which is automatically adjusted according to the position of the sun during the course of the day. Using spherical coordinates shows that two rotations are sufficient for the mirror to be adjusted to any required orientation. (For more details on controlling motion about axes of rotation refer to Chapter 3.)

7. Here we discuss a tool used by carpenters and woodworkers for drawing ellipses. The tool consists of a square block within which there are two tracks in the shape of a plus sign. Each track houses a small block that is free to slide within it. The block labeled A slides vertically, and the block labeled B horizontally. From the centers A and B of each little block there is a small post perpendicular to the tool that attaches to an arm. The arm is rigid and moves in a plane parallel to the tool. Thus the distance between

the two posts is constant and equal to $d = |AB|$. The rigid arm has a total length of L. At the far end of the arm a pencil is attached. Refer to Figure 15.49 for a simple diagram of this device.

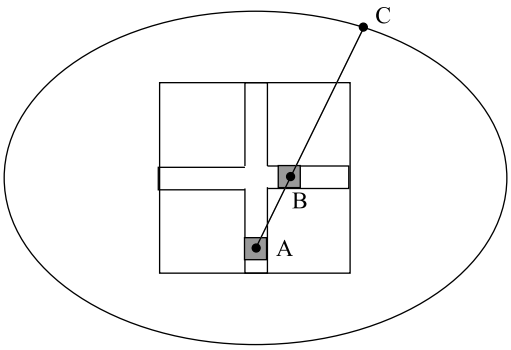


Fig. 15.49. A tool for drawing an ellipse (Exercise 7).

(a) Allowing the rigid arm to rotate about the vertical posts and letting the little blocks slide with the tracks, show that the pencil tip will draw an ellipse.
(b) How must d and L be chosen such that the drawn ellipse has semiaxes with lengths a and b?

8. A hyperbola is the set of points P in the plane whose absolute values of the differences between their distances from two points F_1 and F_2 are a constant r:

$$\big|\, |F_1P| - |F_2P| \,\big| = r. \tag{15.31}$$

We present a technique for drawing one branch of a hyperbola using only a straightedge, a pencil, and a piece of string. The straightedge is attached to and free to pivot around the first focal point F_1. At the far end of the straightedge A we attach a piece of string of length ℓ whose other end is attached to the second focal point F_2. The pencil is held tightly against the side of the straightedge such that the string remains taut, as shown in Figure 15.50.
(a) Show that the tip of the pencil will draw one branch of a hyperbola.
(b) What length ℓ must be chosen for the string if the straightedge is of length L, and we wish the drawn hyperbola to correspond to equation (15.31)?
(c) Describe how to draw the second branch of the hyperbola.

9. We describe a device for drawing a parabola. We affix a straightedge along a line (D). Along this we will slide a square. A string of length L is attached to the tip of the square at a height h above the straightedge, with the other end attached to a point O

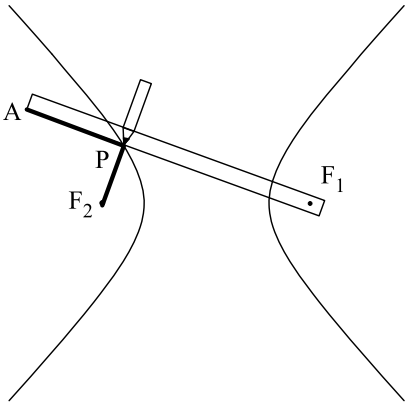


Fig. 15.50. Drawing a hyperbola with a straightedge (Exercise 8).

at a height h_1 above the straightedge. A pencil is held tightly against the vertical side of the square such that the string remains taut (see Figure 15.51). If the pencil is at P and the upper point of the straightedge is A, then, provided the string is taut, it follows that $|AP| + |OP| = L$. Let $h_2 = h - h_1$.

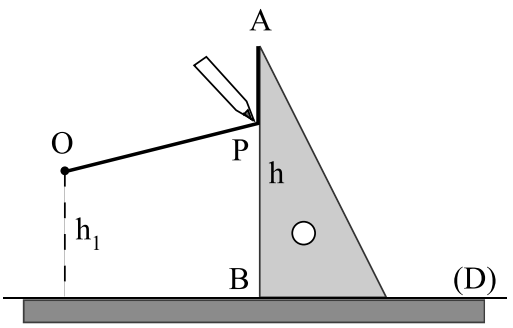


Fig. 15.51. Drawing a parabola (see Exercise 9).

(a) If $L > h_2$ show that the tip of the pencil will draw an arc of the parabola. (Hint: use a coordinate system centered at O and consider the coordinates (x, y) of the point P.)

(b) Show that the point O is the focal point of the parabola.

(c) Show that the arc of the parabola that will be drawn will be tangent to the straightedge (D) if $h_1 = \frac{L-h_2}{2}$. In this case, find the directrix of the parabola.

(d) Show that the bottom of the parabola is an extreme point of the drawn arc if and only if $\frac{L-h_2}{2} \leq h_1$.

Quadratic surfaces

10. Show that the equation

$$x^2 + y^2 = C^2 z^2$$

with $C > 0$ describes a cone with circular cross section.

11. Consider two ellipses $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ situated in the planes $z = -z_0$ and $z = z_0$. Let $\phi_0 \in (-\pi, 0) \cup (0, \pi]$ be a fixed angle. Let (D_θ) be the line between the point $P(\theta) = (a\cos\theta, b\sin\theta, -z_0)$ on the first ellipse and the point $Q(\theta) = (a\cos(\theta + \phi_0), b\sin(\theta + \phi_0), z_0)$ on the second ellipse. Show that the union of the lines (D_θ) is a hyperboloid of one sheet if $\phi_0 \neq \pi$ and a cone with elliptical cross section if $\phi_0 = \pi$. What is the surface if $\phi_0 = 0$?

12. **(a)** In Proposition 15.13 and Exercise 11 we constructed a hyperboloid of one sheet as the union of a family of straight lines. Show that there exists a second family of lines (D'_θ) whose union describes the same surface.
(b) Show that in the limiting case in which the family of lines describes a cone, these two families are actually one and the same.

13. Show that for any point on a hyperboloid of one sheet, the plane tangent to the hyperboloid at this point intersects the hyperboloid along two straight lines. (In particular, there exist points on the surface on each side of the tangent plane. This is a property of surfaces with negative Gaussian curvature.)

14. Show that for any point on a hyperbolic paraboloid, the plane tangent to the surface at this point intersects it along two straight lines.

15. In this problem we use cylindrical coordinates $(x, y, z) = (r\cos\theta, r\sin\theta, z)$. The helicoid is defined by the parametric equations

$$\begin{cases} x = r\cos\theta, \\ y = r\sin\theta, \\ z = C\theta, \end{cases}$$

where C is a constant. Attempt to visualize this surface (drawing it if you can!) and show that it is a ruled surface. (We can use this surface as a base for constructing a spiral staircase.)

Partitioning a region

16. We consider regular triangular partitions of a large region in which the triangles are all congruent, but are *not* equilateral. In a regular partition we have horizontal rows of triangles, which alternate, one up, one down. Show that the equilateral triangular network is the most efficient in terms of antenna count.

17. We consider the same regular networks presented in Section 15.4: regular equilateral triangular networks, regular square networks, and regular hexagonal networks. However, in this exercise, we change the optimization constraint. We wish to use the network whose total edge length (the sum of the lengths of all edges in the network) is minimal, under the constraint that each cell has an area of A. Show that the hexagonal network is the most efficient, followed by the square network and finally the triangular network.

(Motivation: Honeycombs are hexagonal in shape. For a long time it was conjectured that this was to minimize the amount of wax needed and that bees had evolved to choose this form for that reason. In fact, if the individual cells are sufficiently deep (such that the wax required to build the bottom is negligible compared to that used to build the sides) then this is the optimal layout. However, it is now known that the form of the bottom constructed by the bees is not optimal.)

18. We fill a large planar region with nonoverlapping disks of radius r. We use two methods: in the first method we place the centers of the disks on a square network (Figure 15.52 (a)) and in the second method we place them on a regular triangular network of equilateral triangles (Figure 15.52 (b)). Which method gives the denser filling? Suggestion:

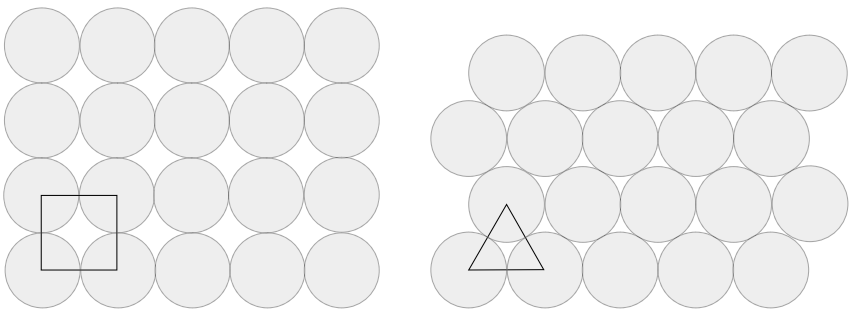


Fig. 15.52. The two methods for filling a planar region with disks (Exercise 18).

compute the proportion of each square covered by portions of disks in case (a) and the proportion of each triangle covered by portions of disks in case (b).

Voronoi diagrams

19. Generalize Proposition 15.17 to the case of an arbitrary region $\mathcal{D}$ in the plane.

20. We can also define Voronoi diagrams for a set of sites in $\mathbb{R}^3$. Propose a definition of such a diagram, and equivalents for Propositions 15.16 and 15.17.

21. Describe the Voronoi diagram for a set of three sites forming the corners of an equilateral triangle.

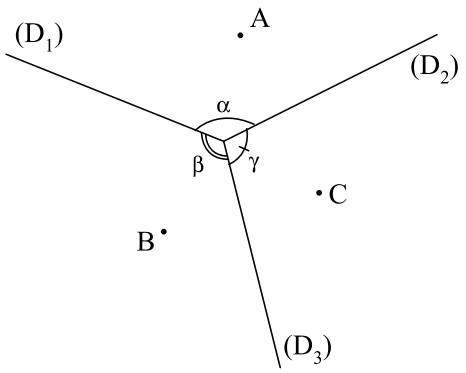
22. Give the conditions on the positions of a set of four points $S = \{P_1, P_2, P_3, P_4\}$ so that the Voronoi diagram of S contains a triangular cell.

23. Consider a convex polygon with n sides and a point P_1 in the interior of this polygon.
(a) Give an algorithm for adding n other points $P_2, \ldots, P_{n+1}$ such that the polygon will be the only closed cell of the Voronoi diagram of $S = \{P_1, \ldots, P_{n+1}\}$ (see Figure 15.32).
(b) Give an algorithm for adding the n half-lines needed to complete the Voronoi diagram.

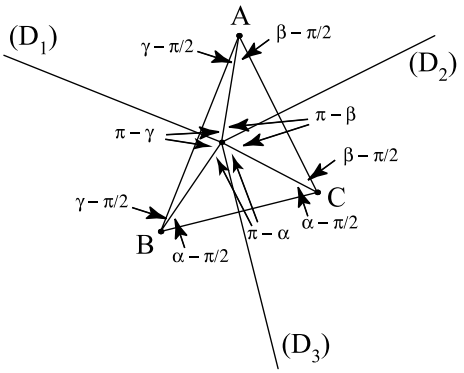
24. This exercise discusses the Delaunay triangulation, whose definition we recall here. Consider the Voronoi diagram of a set $S = \{P_1, \ldots, P_n\}$ of points. We connect points P_i and P_j if the cells $V(P_i)$ and $V(P_j)$ have an edge in common. The resulting set of lines forms the Delaunay triangulation of S.
(a) Verify that if each corner in the Voronoi diagram has at most three incoming edges, the described construction will create triangles.
(b) Verify that each corner P in the Voronoi diagram is the center of a circle circumscribed about a triangle in the Delaunay triangulation. Moreover, verify that the circumscribed circle passes through the three sites whose cells meet at P. (This question provides another way to show that the perpendicular bisectors of the three sides of a triangle meet at a point.)

25. Construct a set of sites S such that Figure 15.28 is its Voronoi diagram and construct the associated Delaunay triangulation.

26. Here we consider the inverse problem to finding a Voronoi diagram. Given a partitioning of the plane into cells, we wish to know whether there exists a set of sites S whose Voronoi diagram is given by the partitioning of the plane.
(a) We start by considering the case of three half-rays (D_1), (D_2), and (D_3), as in Figure 15.53(a). We are asking whether there exists a set of sites $S = \{A, B, C\}$ such that the half-rays form the Voronoi diagram of S. The discussion is different depending on whether the point of intersection O of (D_1), (D_2), and (D_3) lies within the triangle ABC. Show that a necessary condition for O to lie within the triangle ABC is that $\alpha, \beta, \gamma > \frac{\pi}{2}$ and show that if A, B, C exist, the angles of Figure 15.53(b) have the values given in the Figure.
(b) Show that if we choose A within the angle formed by (D_1) and (D_2), then there exist B and C such that (D_1), (D_2), and (D_3) form the Voronoi diagram of $S =$



(a) The half-rays (D_i) and the sites A, B, C

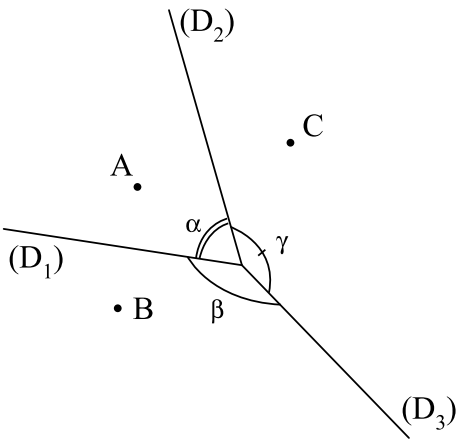


(b) The various angles of the diagram

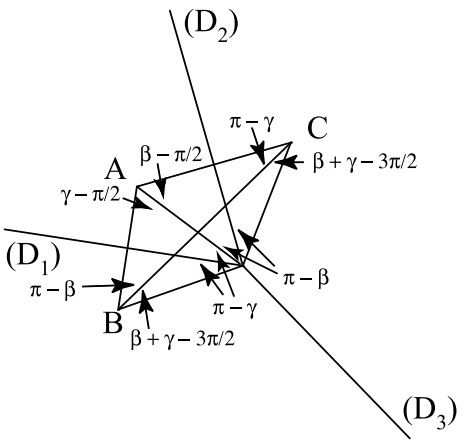
Fig. 15.53. The lines and angles of the Voronoi diagram of Exercise 26(a).

$\{A, B, C\}$ if and only if A lies along the half-line originating at O and making an angle of $\pi - \gamma$ with (D_1) and an angle of $\pi - \beta$ with (D_2).

(c) Now consider Figure 15.54(a) in the case that $\alpha < \frac{\pi}{2}$ and $\beta, \gamma > \frac{\pi}{2}$. Show that the various angles of the final diagram are those shown in Figure 15.54(b).



(a) The half-rays (D_i) and the sites A, B, C



(b) The various angles of the diagram

Fig. 15.54. The lines and angles of the Voronoi diagram of Exercise 26(c).

(d) Conclude that if we have a partitioning of the plane into cells as in Figure 15.55, then there does not always exist a set of sites $S = \{A, B, C, D\}$ such that the partitioning describes the Voronoi diagram of S.

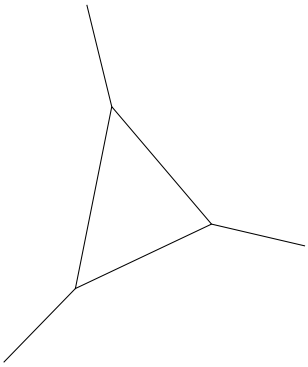


Fig. 15.55. A partitioning of the plane for Exercise 26(d).

(e) Can you describe what happens in the intermediate case of $\alpha = \frac{\pi}{2}$?

Computer vision

27. Consider Figure 15.35 with points $O_1 = (-1, 0, 0)$ and $O_2 = (1, 0, 0)$, and with the projections P_1 and P_2 of a point P both lying within the plane $y = 1$. The image of P on the ith photo is the intersection of the line O_iP with the projection plane $y = 1$.
(a) Show that the image of a vertical line is a vertical line on each of the projections.
(b) Describe the set of points in space that are hidden by P in the first projection. How will these points appear in the second projection?
(c) We consider an oblique line of the form $(a, b, c) + t(\alpha, \beta, \gamma)$, for $t \in \mathbb{R}$ where $\alpha, \beta, \gamma > 0$. Show that the image of the points on this line in the first projection is a line. Now consider only the image of the points (x, y, z) for the half-ray $y > 1$. Show that the image of the point at infinity on this half-ray depends only on (α, β, γ) and is independent of (a, b, c).

28. We have seen that if we take two photos from different points of view of the same point P, we can calculate the position of the point P. However, this is not possible if we have only one photo. A rather clever individual had the following idea for getting away with taking only one photo: he places a mirror in the scene such that points P in front of the mirror *and* their reflections P' both appear in the photo (see Figure 15.56). Assuming that the position and orientation of the mirror are known, explain how this information allows the observer to calculate the position of the point P.

A brief look at computer architecture

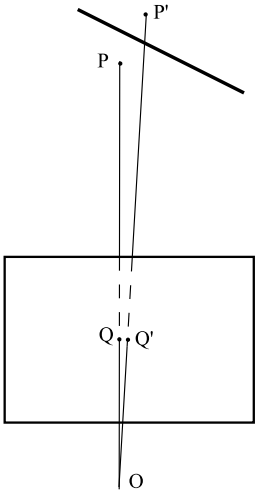


Fig. 15.56. A single photo using a mirror (for Exercise 28).

29. Design a simple electrical circuit that calculates

$$(A \text{ AND } B) \text{ OR } (C \text{ AND } D).$$

30. Design a simple electrical circuit that calculates

$$(A \text{ OR } B) \text{ AND } (C \text{ OR } D).$$

31. Design a simple electrical circuit that calculates

$$((A \text{ OR } B) \text{ AND } (C \text{ OR } D)) \text{ OR } (E \text{ AND } F).$$

32. **(a)** Show that we can define the OR and XOR operators using only the NOT and AND operators.
(b) Show that we can define the AND and XOR operators using only the NOT and OR operators.
(c) Show that we can define the AND and OR operators using only the NOT and XOR operators. (This question is more difficult than the first two.)

33. Construct the tables describing the NAND and NOR operators defined in (15.22).

34. The NAND and NOR operators are called the universal Boolean operators because just one of these operators can be used to construct all of the others. This exercise guides you through the first steps of this construction. Afterward, we may apply the constructions from Exercise 32.

(a) Show that we can define the NOT operation from the NAND operation alone.
(b) Show that we can define the NOT operation from the NOR operation alone.
(c) Show that we can construct the AND operation using only NAND operations.
(d) Show that we can construct the OR operation using only NAND operations.

35. A single fixture illuminates a stairwell. Two switches allow the light to be turned on or off, one at the bottom of the stairs and the other at the top. The electrician wired the switches using the circuit we constructed for one of the Boolean operators. Which one?

Regular pentagonal tiling of the sphere

36. **(a)** Show that each internal angle of a regular polygon with n sides is exactly $\frac{\pi(n-2)}{n}$.
(b) Deduce that the interior angles of a regular pentagon are $\frac{3\pi}{5}$ and that the length d of a diagonal of a pentagon with side lengths a (see Figure 15.43) is given by

$$d = 2a \cos \frac{\pi}{5}.$$

37. A tetrahedron is a regular polyhedron formed from four equilateral triangles (see Figure 15.57).

(a) Calculate the height of a regular tetrahedron with edge length a.

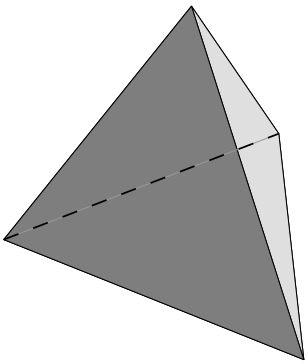


Fig. 15.57. A regular tetrahedron (see Exercise 37).

(b) What is the radius r of a circle circumscribed around an equilateral triangle with side length a?
(c) Consider the sphere of radius R circumscribed around a regular tetrahedron with edge length a. Calculate R as a function of a.
(d) Show that the distance from a vertex to the intersection points of the four altitudes of a regular tetrahedron is $\frac{3}{4}$ the length of the altitudes.

38. Show that an appropriate choice of diagonals of the faces of a cube forms a regular tetrahedron. How many different tetrahedra do we get?

39. **(a)** Show that the edge length d of a cube inscribed in a sphere with radius R is

$$d = \frac{2}{\sqrt{3}}R.$$

(b) With the help of a compass, explain how to draw the vertices of an inscribed cube on the surface of the circumscribing sphere.
(c) If we project (from the center of the sphere) the edges of the inscribed cube onto the surface of the sphere, we divide the surface of the sphere into six equal regions. The centers of these regions are the vertices of the regular octahedron (see Figure 15.58) inscribed in the sphere. Explain how to mark these vertices using a compass.

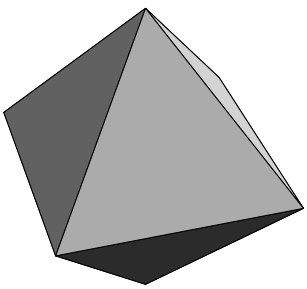


Fig. 15.58. An octahedron.

40. Show that the radius r of a circle circumscribed about a regular pentagon with side length a (see Figure 15.44) is given by

$$r = \frac{a}{2\sin\frac{\pi}{5}}.$$

41. **(a)** What is the opening R' that must be given to a compass in order to draw a great circle around a sphere with radius R?

(b) You are given two points P and Q on the surface of a sphere with radius R. Using only a compass, explain how to draw the great circle passing through P and Q. Under what condition on P and Q will this great circle be unique?

42. You have a sphere of diameter 30 cm on which you wish to reproduce a map of the Earth. You choose a random point that you label the North Pole.

(a) Using only a compass, explain how to draw the equator and find the South Pole.

(b) Explain how to draw the two tropics: these are the parallels of latitude at 23.5 degrees north and south of the equator.

(c) Explain how to draw the polar circles: these are the parallels of latitude at 66.5 degrees north and south of the equator.

(d) Explain how to draw any line of meridian, which you will then label the Greenwich meridian.

(e) Explain how to draw the meridian of longitude corresponding to 25 degrees west.

43. There exist five regular polyhedra: the tetrahedron, the cube, the octahedron, the dodecahedron and the icosahedron. The icosahedron is shown in Figure 15.59. It has 12 vertices and 20 faces, while the dodecahedron has 20 vertices and 12 faces.

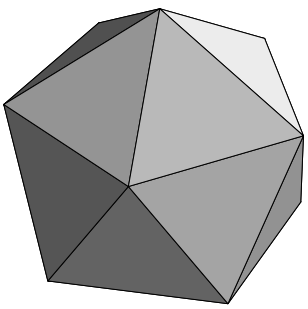


Fig. 15.59. An icosahedron.

(a) Show that the centers of the faces of a dodecahedron are the vertices of an icosahedron and vice versa. We say that these two polyhedra are *dual*.

(b) Describe a method for marking the vertices of an inscribed icosahedron on the surface of the sphere.

(c) Each vertex of an icosahedron is shared by five faces. Using only five different colors, there exists a method of coloring the faces of an icosahedron such that the five faces meeting at each vertex each have a different color. Can you propose such a coloring? Is it possible to color the faces such that each vertex, when seen from above, has its adjacent faces colored in the same order?

44. Explain why the diagonals of the pentagons of a dodecahedron form a cube. Hint: Consider the symmetries of the dodecahedron, for example the mediating plane of two such diagonals. It may actually be useful to construct a dodecahedron and draw all of the diagonals.

References

[1] V. Gutenmacher and N.B. Vasilyev. *Lines and Curves: A practical Geometry Handbook*. Birkhäuser, Boston, MA, 2004.

[2] C. Mead and L. Conway. *Introduction to VLSI Systems*. Addison-Wesley, Reading, MA, 1980.

Index