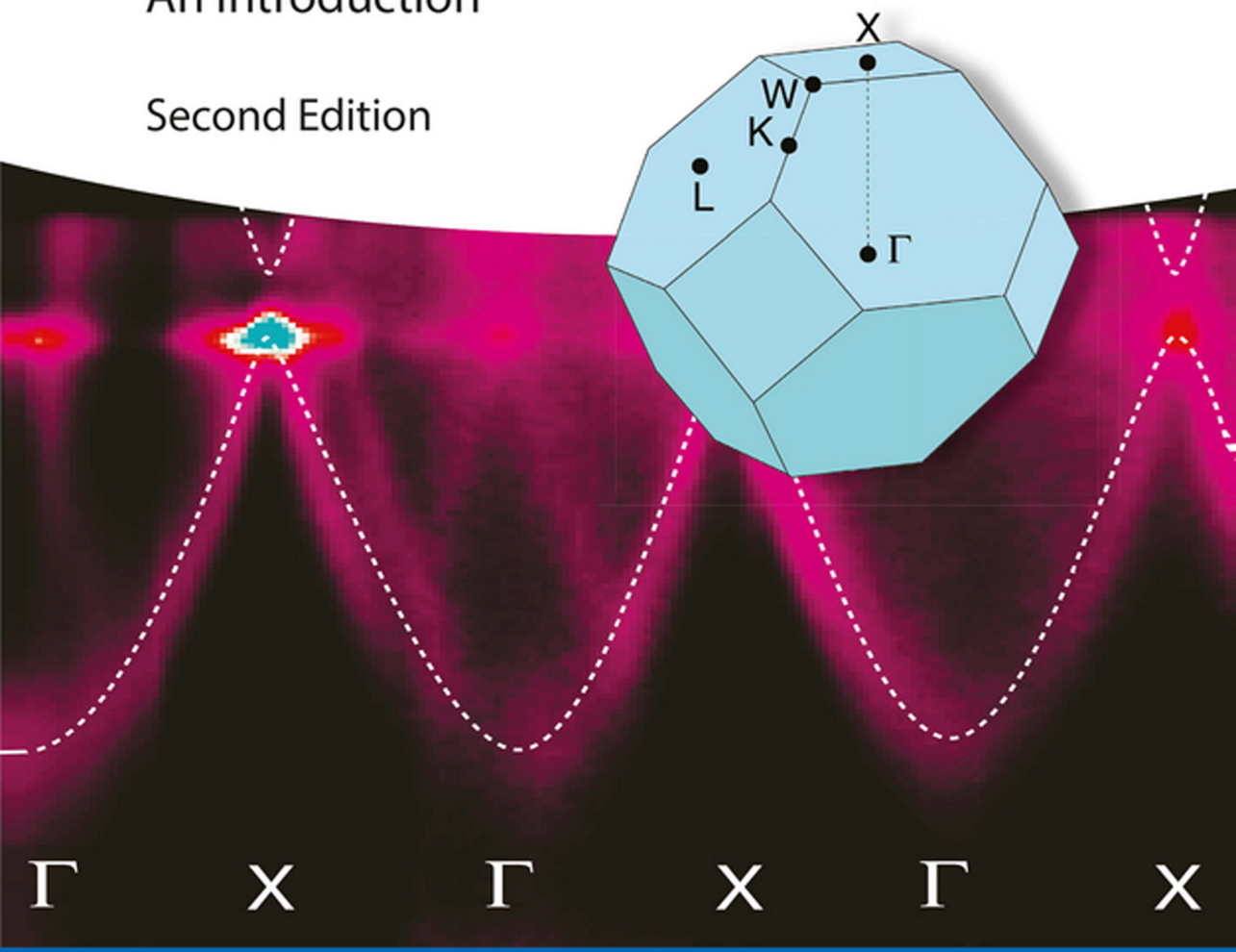
WILEY-VCH

Philip Hofmann

Solid State Physics

An Introduction

Second Edition



PHYSICS TEXTBOOK

Philip Hofmann

Solid State Physics

Related Titles

Callister, W.D., Rethwisch, D.G.

Materials Science and Engineering
An Introduction, Eighth Edition
8th Edition

2009

Print ISBN: 978-0-470-41997-7

Sze, S.M., Lee, M.

Semiconductor Devices
Physics and Technology, Third Edition
3rd Edition

2013

Print ISBN: 978-0-470-53794-7

Marder, M.P.

Condensed Matter Physics, Second Edition
2nd Edition

2011

Print ISBN: 978-0-470-61798-4

Kittel, C.

Introduction to Solid State Physics, 8th Edition
8th Edition

2005

Print ISBN: 978-0-471-41526-8

Kittel, C.

Quantum Theory of Solids, 2e Revised Edition
2nd Edition

1987

Print ISBN: 978-0-471-62412-7

Buckel, W., Kleiner, R.

Superconductivity
Fundamentals and Applications
2nd Edition

2004

Print ISBN: 978-3-527-40349-3

Mihály, L., Martin, M.

Solid State Physics
Problems and Solutions
2nd Edition

2009

Print ISBN: 978-3-527-40855-9

Würfel, P.

Physics of Solar Cells
From Basic Principles to Advanced Concepts
2nd Edition

2009

Print ISBN: 978-3-527-40857-3

Philip Hofmann

Solid State Physics

An Introduction

Second Edition

WILEY-VCH
Verlag GmbH & Co. KGaA

The Author

Dr. Philip Hofmann
Department of Physics and Astronomy
Aarhus University
Ny Munkegade 120
8000 Aarhus C
Denmark

Cover
Band structure of aluminum determined by angle–resolved photoemission. Data taken from Physical Review B 66, 245422 (2002), see also Fig. 6.12 in this book.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

Print ISBN: 978-3-527-41282-2
ePDF ISBN: 978-3-527-68203-4
ePub ISBN: 978-3-527-68206-5
Mobi ISBN: 978-3-527-68205-8

Typesetting Laserwords Private Limited, Chennai, India
Printing and Binding Betz-Druck GmbH, Darmstadt, Germany

Printed on acid-free paper

Contents

Preface of the First Edition

This book emerged from a course on solid state physics for third-year students of physics and nanoscience, but it should also be useful for students of related fields such as chemistry and engineering. The aim is to provide a bachelor-level survey over the whole field without going into too much detail. With this in mind, a lot of emphasis is put on a didactic presentation and little on stringent mathematical derivations or completeness. For a more in-depth treatment, the reader is referred to the many excellent advanced solid state physics books. A few are listed in the Appendix.

To follow this text, a basic university-level physics course is required as well as some working knowledge of chemistry, quantum mechanics, and statistical physics. A course in classical electrodynamics is of advantage but not strictly necessary.

Some remarks on *how to use this book*: Every chapter is accompanied by a set of "discussion" questions and problems. The intention of the questions is to give the student a tool for testing his/her understanding of the subject. Some of the questions can only be answered with knowledge of later chapters. These are marked by an asterisk. Some of the problems are more of a challenge in that they are more difficult mathematically or conceptually or both. These problems are also marked by an asterisk. Not all the information necessary for solving the problems is given here. For standard data, for example, the density of gold or the atomic weight of copper, the reader is referred to the excellent resources available on the World Wide Web.

Finally, I would like to thank the people who have helped me with many discussions and suggestions. In particular, I would like to mention my colleagues Arne Nylandsted Larsen, Ivan Steensgaard, Maria Fuglsang Jensen, Justin Wells, and many others involved in teaching the course in Aarhus.

Preface of the Second Edition

The second edition of this book is slightly enlarged in some subject areas and significantly improved throughout. The enlargement comprises subjects that turned out to be too essential to be missing, even in a basic introduction such as this one. One example is the tight-binding model for electronic states in solids, which is now added in its simplest form. Other enlargements reflect recent developments in the field that should at least be mentioned in the text and explained on a very basic level, such as graphene and topological insulators.

I decided to support the first edition by online material for subjects that were either crucial for the understanding of this text, but not familiar to all readers, or not central enough to be included in the book but still of interest. This turned out to be a good concept, and the new edition is therefore supported by an extended number of such notes; they are referred to in the text. The notes can be found on my homepage www.philiphofmann.net.

The didactical presentation has been improved, based on the experience of many people with the first edition. The most severe changes have been made in the chapter on magnetism but minor adjustments have been made throughout the book. In these changes, didactic presentation was given a higher priority than elegance or conformity to standard notation, for example, in the figures on Pauli paramagnetism or band ferromagnetism.

Every chapter now contains a "Further Reading" section in the end. Since these sections are supposed to be independent of each other, you will find that the same books are mentioned several times.

I thank the many students and instructors who participated in the last few years' Solid State Physics course at Aarhus University, as well as many colleagues for their criticism and suggestions. Special thanks go to NL architects for permitting me to use the flipper-bridge picture in Figure 11.3, to Justin Wells for suggesting the analogy to the topological insulators, to James Kermode for Figure 3.7, to Arne Nylandsted Larsen and Antonija Grubišić Čabo for advice on the sections on solar cells and magnetism, respectively.

Physical Constants and Energy Equivalents

Planck constant	h	$6.6260755 \times 10^{-34}$ J s
		$4.13566743 \times 10^{-15}$ eV s
Boltzmann constant	k_B	1.380658×10^{-23} J K^{-1}
		8.617385×10^{-5} eV K^{-1}
Proton charge	e	$1.60217733 \times 10^{-19}$ C
Bohr radius	a_0	5.29177×10^{-11} m
Bohr magneton	μ_B	$9.2740154 \times 10^{-24}$ J T^{-1}
Avogadro number	N_A	6.0221367×10^{23} particles/mol
Speed of light	c	2.99792458×10^{8} m s^{-1}
Rest mass of the electron	m_e	$9.1093897 \times 10^{-31}$ kg
Rest mass of the proton	m_p	$1.6726231 \times 10^{-27}$ kg
Rest mass of the neutron	m_n	$1.6749286 \times 10^{-27}$ kg
Atomic mass unit	amu	1.66054×10^{-27} kg
Permeability of vacuum	μ_0	$4\pi \times 10^{-7}$ V s A^{-1}m^{-1}
Permittivity of vacuum	ϵ_0	$8.854187817 \times 10^{-12}$ C^2 J^{-1} m^{-1}

1 eV = $1.6021773 \times 10^{-19}$ J
1 K = 8.617385×10^{-5} eV

1
Crystal Structures

Our general objective in this book is to understand the macroscopic properties of solids in a microscopic picture. In view of the many particles in solids, coming up with any microscopic description appears to be a daunting task. It is clearly impossible to solve the equations of motion (classical or quantum mechanical). Fortunately, it turns out that solids are often crystalline, with the atoms arranged on a regular lattice, and this symmetry permits us to solve microscopic models, despite the very many particles involved. This situation is somewhat similar to atomic physics where the key to a description is the spherical symmetry of the atom. We will often imagine a solid as one **single crystal,** a perfect lattice of atoms without any defects whatsoever, and it may seem that such perfect crystals are not particularly relevant for real materials. But this is not the case. Many solids are actually composed of small crystalline grains. These solids are called **polycrystalline,** in contrast to a macroscopic single crystal, but the number of atoms in a perfect crystalline environment is still very large compared to the number of atoms on the **grain boundary**. For instance, for a grain size on the order of 1000^3 atomic distances, only about 0.1% of the atoms are at the grain boundaries. There are, however, some solids that are not crystalline. These are called **amorphous**. The amorphous state is characterized by the absence of any long-range order. There may, however, be some short-range order between the atoms.

This chapter is divided into three parts. In the first part, we define some basic mathematical concepts needed to describe crystals. We keep things simple and mostly use two-dimensional examples to illustrate the ideas. In the second part, we discuss common crystal structures. At this point, we do not ask why the atoms bind together in the way that they do, as this is treated in the next chapter. Finally, we go into a somewhat more detailed discussion of X-ray diffraction, the experimental technique that can be used to determine the microscopic structure of crystals. X-ray diffraction is used not only in solid state physics but also for a wide range of problems in nanotechnology and structural biology.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

1.1 General Description of Crystal Structures

Our description of crystals starts with the mathematical definition of the **lattice**. A lattice is a set of regularly spaced points with positions defined as multiples of generating vectors. In two dimensions, a lattice can be defined as all the points that can be reached by the vectors $\mathbf{R}$, created from two vectors $\mathbf{a_1}$ and $\mathbf{a_2}$ as

$$\mathbf{R} = m\mathbf{a_1} + n\mathbf{a_2}, \tag{1.1}$$

where n and m are integers. In three dimensions, the definition is

$$\mathbf{R} = m\mathbf{a_1} + n\mathbf{a_2} + o\mathbf{a_3}. \tag{1.2}$$

Such a lattice of points is also called a **Bravais lattice**. The number of possible Bravais lattices that differ by symmetry is limited to 5 in two dimensions and to 14 in three dimensions. An example of a two-dimensional Bravais lattice is given in Figure 1.1. The lengths of the vectors $\mathbf{a_1}$ and $\mathbf{a_2}$ are often called the **lattice constants**.

Having defined the Bravais lattice, we move on to the definition of the **primitive unit cell**. This is any volume of space that, when translated through all the vectors of the Bravais lattice, fills space without overlap and without leaving voids. The primitive unit cell of a lattice contains only one lattice point. It is also possible to define **nonprimitive unit cells** that contain several lattice points. These fill space without leaving voids when translated through a subset of the Bravais lattice vectors. Possible choices of a unit cell for a two-dimensional rectangular Bravais lattice are given in Figure 1.2. From the figure, it is evident that a nonprimitive unit cell has to be translated by a multiple of one (or two) lattice vectors to fill space without voids and overlap. A special choice of the primitive unit cell is the **Wigner–Seitz cell** that is also shown in Figure 1.2. It is the region of space that is closer to one given lattice point than to any other.

The last definition we need in order to describe an actual crystal is that of a **basis**. The basis is what we "put" on the lattice points, that is, the building block for the real crystal. The basis can consist of one or several atoms. It can even consist of

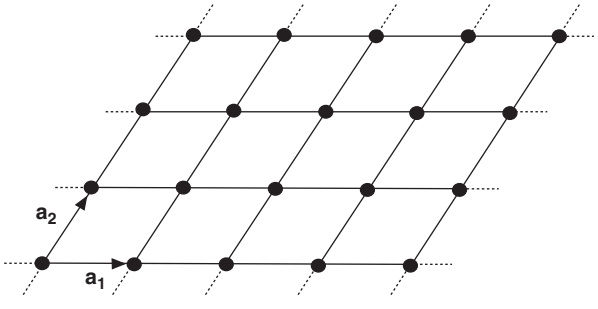


Figure 1.1 Example for a two-dimensional Bravais lattice.

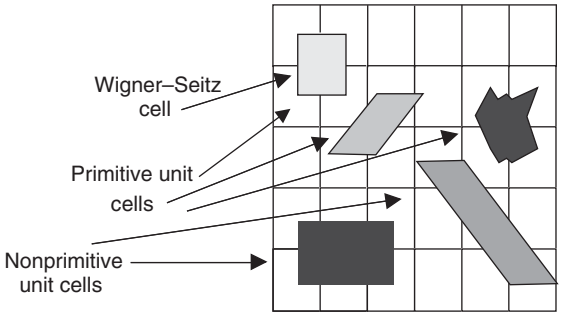


Figure 1.2 Illustration of unit cells (primitive and nonprimitive) and of the Wigner–Seitz cell for a rectangular two-dimensional lattice.

complex molecules as in the case of protein crystals. Different cases are illustrated in Figure 1.3.

Finally, we add a remark about symmetry. So far, we have discussed **translational symmetry**. But for a real crystal, there is also **point symmetry**. Compare the structures in the middle and the bottom of Figure 1.3. The former structure possesses a couple of symmetry elements that the latter does not have, for example, mirror lines, a rotational axis, and inversion symmetry. The knowledge of such symmetries can be very useful for the description of crystal properties.

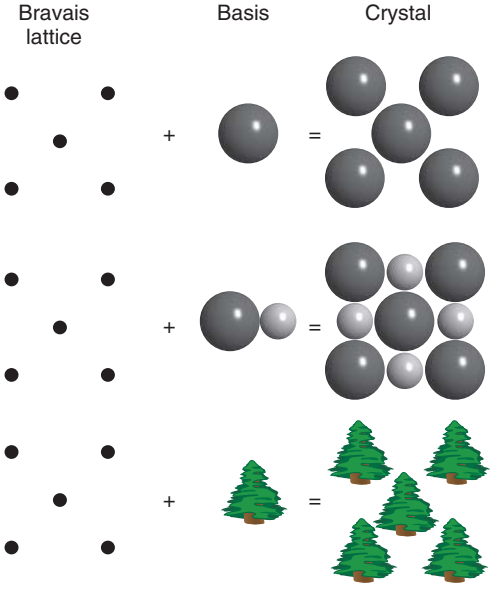


Figure 1.3 A two-dimensional Bravais lattice with different choices for the basis.

1.2
Some Important Crystal Structures

After this rather formal treatment, we look at a number of common crystal structures for different types of solids, such as metals, ionic solids, or covalently bonded solids. In the next chapter, we will take a closer look at the details of these bonding types.

1.2.1
Cubic Structures

We start with one of the simplest possible crystal structures, the **simple cubic structure** shown in Figure 1.4a. This structure is not very common among elemental solids, but it is an important starting point for many other structures. The reason why it is not common is its openness, that is, that there are many voids if we think of the ions as spheres touching each other. In metals, the most common elemental solids, directional bonding is not important and a close packing of the ions is usually favored. For covalent solids, directional bonding *is* important but six bonds on the same atom in an octahedral configuration are not common in elemental solids.

The packing density of the cubic structure is improved in the **body-centered cubic** (bcc) and **face-centered cubic** (fcc) structures that are also shown in Figure 1.4. In fact, the fcc structure has the highest possible packing density for spheres as we shall see later. These two structures are very common. Seventeen elements crystallize in the bcc structure and 24 elements in the fcc structure. Note that only for the simple cubic structure, the cube is identical with the Bravais lattice. For the bcc and fcc lattices, the cube is also a unit cell, but not the primitive one. Both structures are Bravais lattices with a basis containing one atom but the vectors spanning these Bravais lattices are not the edges of the cube.

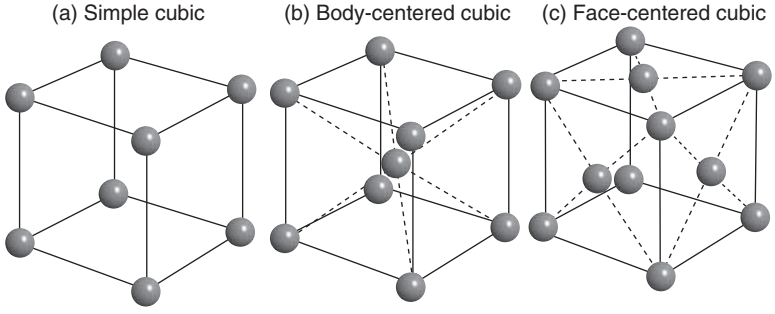


Figure 1.4 (a) Simple cubic structure; (b) body-centered cubic structure; and (c) face-centered cubic structure. Note that the spheres are depicted much smaller than in the situation of most dense packing and not all of the spheres on the faces of the cube are shown in (c).

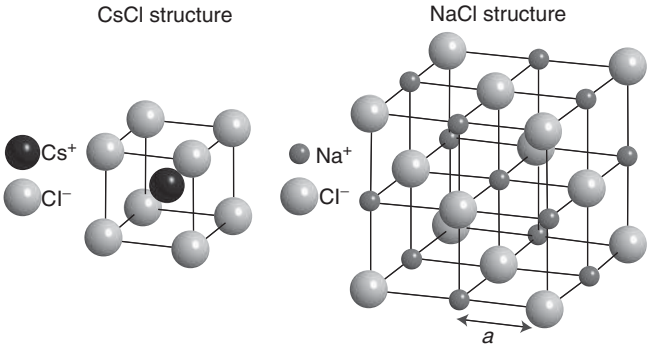


Figure 1.5 Structures of CsCl and NaCl. The spheres are depicted much smaller than in the situation of most dense packing, but the relative size of the different ions in each structure is correct.

Cubic structures with a more complex basis than a single atom are also important. Figure 1.5 shows the structures of the ionic crystals CsCl and NaCl that are both cubic with a basis containing two atoms. For CsCl, the structure can be thought of as two simple cubic structures stacked into each other. For NaCl, it consists of two fcc lattices stacked into each other. Which structure is preferred for such ionic crystals depends on the relative size of the ions.

1.2.2 Close-Packed Structures

Many metals prefer structural arrangements where the atoms are packed as closely as possible. In two dimensions, the closest possible packing of ions (i.e., spheres) is the hexagonal structure shown on the left-hand side of Figure 1.6. For building a three-dimensional close-packed structure, one adds a second layer as in the middle of Figure 1.6. For adding a third layer, there are then two possibilities. One can either put the ions in the "holes" just on top of the first layer ions, or one can put them into the other type of "holes." In this way, two different crystal structures can be built. The first has an ABABAB... stacking sequence, and the second has an ABCABCABC... stacking sequence. Both have exactly the

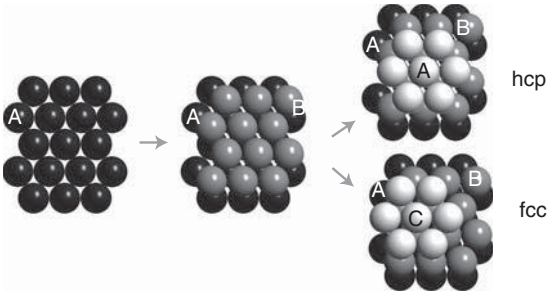


Figure 1.6 Close packing of spheres leading to the hcp and fcc structures.

same packing density, and the spheres fill 74% of the total volume. The former structure is called the **hexagonal close-packed structure** (hcp), and the latter turns out to be the fcc structure we already know. An alternative sketch of the hcp structure is shown in Figure 1.14b. The fcc and hcp structures are very common for elemental metals. Thirty-six elements crystallize as hcp and 24 elements as fcc. These structures also maximize the number of nearest neighbors for a given atom, the so-called **coordination number**. For both the fcc and the hcp lattice, the coordination number is 12.

An open question is why, if coordination is so important, not all metals crystallize in the fcc or hcp structure. A prediction of the actual structure for a given element is not possible with simple arguments. However, we can collect some factors that play a role. Not optimally packed structures, such as the bcc structure, have a lower coordination number, but they bring the second-nearest neighbors much closer to a given ion than in the close-packed structures. Another important consideration is that the bonding is not quite so simple, especially for **transition metals**. In these, bonding is not only achieved through the delocalized s and p valence electrons as in **simple metals**, but also by the more localized d electrons. Bonding through the latter has a much more directional character, so that not only the close packing of the ions is important.

The structures of many ionic solids can also be viewed as "close-packed" in some sense. One can arrive at these structures by treating the ions as hard spheres that have to be packed as closely to each other as possible.

1.2.3
Structures of Covalently Bonded Solids

In covalent structures, the atoms' valence electrons are not completely delocalized but shared between neighboring atoms and the bond length and direction are far more important than the packing density. Prominent examples are graphene, graphite, and diamond as displayed in Figure 1.7. Graphene is a single sheet of carbon atoms in a honeycomb lattice structure. It is a truly two-dimensional solid with a number of remarkable properties; so remarkable, in fact, that their

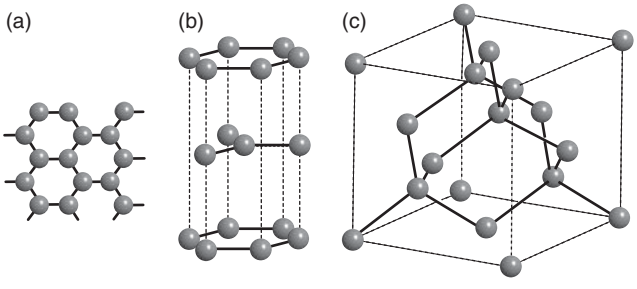


Figure 1.7 Structures for (a) graphene, (b) graphite, and (c) diamond. sp^2 and sp^3 bonds are displayed as solid lines.

discovery has lead to the 2010 Nobel prize in physics being awarded to A. Geim and K. Novoselov. The carbon atoms in graphene are connected by sp^2 hybrid bonds, enclosing an angle of 120°. The parent material of graphene is graphite, a stack of graphene sheets that are weakly bonded to each other. In fact, graphene can be isolated from graphite by peeling off flakes with a piece of scotch tape. In diamond, the carbon atoms form sp^3-type bonds and each atom has four nearest neighbors in a tetrahedral configuration. Interestingly, the diamond structure can also be described as an fcc Bravais lattice with a basis of two atoms.

The diamond structure is also found for Si and Ge. Many other isoelectronic materials (with the same total number of valence electrons), such as SiC, GaAs, and InP, also crystallize in a diamond-like structure but with each element on a different fcc sublattice.

1.3 Crystal Structure Determination

After having described different crystal structures, the question is of course how to determine these structures in the first place. By far, the most important technique for doing this is X-ray diffraction. In fact, the importance of this technique goes far beyond solid state physics, as it has become an essential tool for fields such as structural biology as well. There the idea is that, if you want to know the structure of a given protein, you can try to crystallize it and use the powerful methodology for structural determination by X-ray diffraction. We will also use X-ray diffraction as a motivation to extend our formal description of structures a bit.

1.3.1 X-Ray Diffraction

X-rays interact rather weakly with matter. A description of X-ray diffraction can therefore be restricted to single scattering, that is, incoming X-rays get scattered not more than once (most are not scattered at all). This is called the **kinematic approximation**; it greatly simplifies matters and is used throughout the treatment here. In addition to this, we will assume that the X-ray source and detector are very far away from the sample so that the incoming and outgoing waves can be treated as plane waves. X-ray diffraction of crystals was discovered and described by M. von Laue in 1912. Also in 1912, W. L. Bragg came up with an alternative description that is considerably simpler and serves as a starting point here.

1.3.1.1 Bragg Theory

Bragg treated the problem as the reflection of the incoming X-rays at flat crystal planes. These planes could, for example, be the close-packed planes making up the fcc and hcp crystals, or they could be alternating Cs and Cl planes making up the CsCl structure. At first glance, this has very little physical justification because the crystal planes are certainly not "flat" for X-rays that have a wavelength similar to the atomic spacing. Nevertheless, the description is highly successful, and we

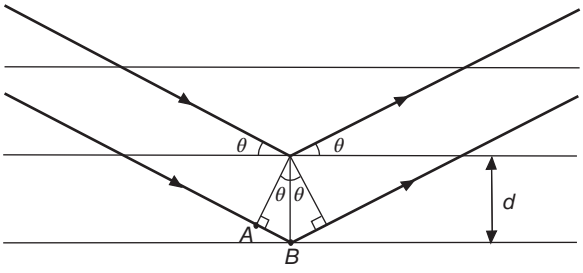


Figure 1.8 Construction for the derivation of the Bragg condition. The horizontal lines represent the crystal lattice planes that are separated by a distance d. The heavy lines represent the X-rays.

shall later see that it is actually a special case of the more complex Laue description of X-ray diffraction.

Figure 1.8 shows the geometrical considerations behind the Bragg description. A collimated beam of monochromatic X-rays hits the crystal. The intensity of diffracted X-rays is measured *in the specular direction*. The angle of incidence and emission is $90^\circ - \Theta$. The condition for constructive interference is that the path length difference between the X-rays reflected from one layer and the next layer is an integer multiple of the wavelength λ. In the figure, this means that $2AB = n\lambda$, where AB is the distance between points A and B and n is a natural number. On the other hand, we have $\sin\theta = AB/d$ such that we arrive at the **Bragg condition**

$$n\lambda = 2d\sin\theta. \tag{1.3}$$

It is obvious that if this condition is fulfilled for one layer and the layer below, it will also be fulfilled for any number of layers with identical spacing. In fact, the X-rays penetrate very deeply into the crystal so that thousands of layers contribute to the reflection. This results into very sharp maxima in the diffracted intensity, similar to the situation for an optical grating with many lines. The Bragg condition can obviously only be fulfilled for $\lambda < 2d$, putting an upper limit on the wavelength of the X-rays that can be used for crystal structure determination.

1.3.1.2 Lattice Planes and Miller Indices

The Bragg condition will work not only for a special kind of lattice plane in a crystal, such as the hexagonal planes in an hcp crystal, but for all possible parallel planes in a structure. We therefore come up with a more stringent definition of the term **lattice plane**. It can be defined as a plane containing at least three non-collinear points of a given Bravais lattice. If it contains three, it will actually contain infinitely many because of translational symmetry. Examples for lattice planes in a simple cubic structure are shown in Figure 1.9.

The lattice planes can be characterized by a set of three integers, the so-called **Miller indices**. We arrive at these in three steps:

1) We find the intercepts of the plane with the crystallographic axes in units of the lattice vectors, for example, $(1, \infty, \infty)$ for the leftmost plane in Figure 1.9.

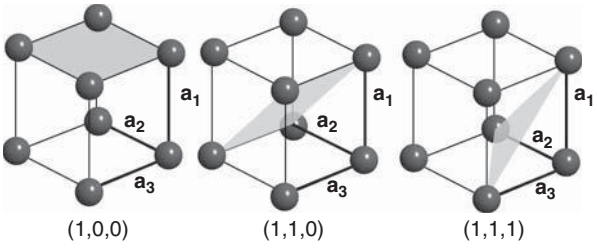


Figure 1.9 Three different lattice planes in the simple cubic structure characterized by their Miller indices.

2) We take the "reciprocal value" of these three numbers. For our example, this gives $(1, 0, 0)$.
3) By multiplying with some factor, we reduce the numbers to the smallest set of integers having the same ratio. This is not necessary in the example as we already have integer values.

Such a set of three integers can then be used to denote any given lattice plane. Later, we will encounter a different and more elegant definition of the Miller indices.

In practice, the X-ray diffraction peaks are so sharp that it is difficult to align and move the sample such that the incoming and reflected X-rays lie in one plane with the normal direction to a certain crystal plane. An elegant way to circumvent this problem is to use a powder of very small crystals instead of a large single crystal. This will not only ensure that some small crystals are orientated correctly to get constructive interference from a certain set of crystal planes, it will automatically give the interference pattern for all possible crystal planes.

1.3.1.3 General Diffraction Theory

The Bragg theory for X-ray diffraction is useful for extracting the distances between lattice planes in a crystal, but it has its limitations. Most importantly, it does not give any information on what the lattice actually consists of, that is, the basis. Also, the fact that the X-rays should be reflected by planes is physically somewhat obscure. We now discuss a more general description of X-ray diffraction that goes back to M. von Laue.

The physical process leading to X-ray scattering is that the electromagnetic field of the X-rays forces the electrons in the material to oscillate with the same frequency as that of the field. The oscillating electrons then emit new X-rays that give rise to an interference pattern. For the following discussion, however, it is merely important that something scatters the X-rays, not what it is.

It is highly beneficial to use the complex notation for describing the electromagnetic X-ray waves. For the electric field, a general plane wave can be written as

$$\boldsymbol{\mathcal{E}}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_0 e^{i\mathbf{k}\cdot\mathbf{r} - i\omega t}. \tag{1.4}$$

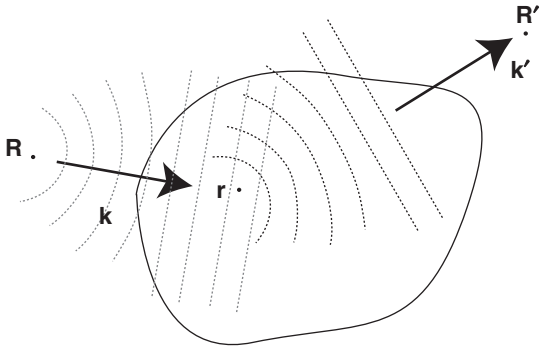


Figure 1.10 Illustration of X-ray scattering from a sample. The source and detector for the X-rays are placed at $\mathbf{R}$ and $\mathbf{R}'$, respectively. Both are very far away from the sample.

The wave vector $\mathbf{k}$ points in the direction of the wave propagation with a length of $2\pi/\lambda$, where λ is the wavelength. The convention is that the physical electric field is obtained as the real part of the complex field and the intensity of the wave is obtained as

$$I(\mathbf{r}) = |\mathcal{E}_0 e^{i\mathbf{k}\cdot\mathbf{r}-i\omega t}|^2 = |\mathcal{E}_0|^2. \tag{1.5}$$

Consider now the situation depicted in Figure 1.10. The source of the X-rays is far away from the sample at the position $\mathbf{R}$, so that the X-ray wave at the sample can be described as a plane wave. The electric field at a point $\mathbf{r}$ in the crystal at time t can be written as

$$\mathcal{E}(\mathbf{r}, t) = \mathcal{E}_0 e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})-i\omega t}. \tag{1.6}$$

Before we proceed, we can drop the absolute amplitude $\mathcal{E}_0$ from this expression because we are only concerned with relative phase changes. The field at point $\mathbf{r}$ is then

$$\mathcal{E}(\mathbf{r}, t) \propto e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} e^{-i\omega t}. \tag{1.7}$$

A small volume element dV located at $\mathbf{r}$ will give rise to scattered waves in all directions. The direction of interest is the direction towards the detector that shall be placed at the position $\mathbf{R}'$, in the direction of a second wave vector $\mathbf{k}'$. We assume that the amplitude of the wave scattered in this direction will be proportional to the incoming field from (1.7) and to a factor $\rho(\mathbf{r})$ describing the scattering probability and scattering phase. We already know that the scattering of X-rays proceeds via the electrons in the material, and for our purpose, we can view $\rho(\mathbf{r})$ as the electron concentration in the solid. For the field at the detector, we obtain

$$\mathcal{E}(\mathbf{R}', t) \propto \mathcal{E}(\mathbf{r}, t)\rho(\mathbf{r}) e^{i\mathbf{k}'\cdot(\mathbf{R}'-\mathbf{r})}. \tag{1.8}$$

Again, we have assumed that the detector is very far away from the sample such that the scattered wave at the detector can be written as a plane wave. Inserting (1.7) gives the field at the detector as

$$\mathcal{E}(\mathbf{R}', t) \propto e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})}\rho(\mathbf{r}) e^{i\mathbf{k}'\cdot(\mathbf{R}'-\mathbf{r})} e^{-i\omega t} = e^{i(\mathbf{k}'\cdot\mathbf{R}'-\mathbf{k}\cdot\mathbf{R})}\rho(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} e^{-i\omega t}. \tag{1.9}$$

We drop the first factor that does not contain $\mathbf{r}$ and will thus not play a role for the interference of X-rays emitted from different positions in the sample. The total wave field at the detector can finally be calculated by integrating over the entire volume of the crystal V. As the detector is far away from the sample, the wave vector $\mathbf{k}'$ is essentially the same for all points in the sample. The result is

$$\mathcal{E}(\mathbf{R}', t) \propto e^{-i\omega t} \int_V \rho(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} dV. \tag{1.10}$$

In most cases, it will only be possible to measure the intensity of the X-rays, not the field, and this intensity is

$$I(\mathbf{K}) \propto \left| e^{-i\omega t} \int_V \rho(\mathbf{r}) e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} dV \right|^2 = \left| \int_V \rho(\mathbf{r}) e^{-i\mathbf{K}\cdot\mathbf{r}} dV \right|^2, \tag{1.11}$$

where we have introduced the so-called scattering vector $\mathbf{K} = \mathbf{k}' - \mathbf{k}$, which is just the difference of outgoing and incoming wave vectors. Note that although the direction of the wave vector for the scattered waves $\mathbf{k}'$ is different from that of the incoming wave $\mathbf{k}$, the length is the same because we only consider elastic scattering.

Equation (1.11) is the final result. It relates the measured intensity to the electron concentration in the sample. Except for very light elements, most of the electrons are located close to the ion cores and the electron concentration that scatters the X-rays is essentially identical to the geometrical arrangement of the ion cores. Hence, (1.11) can be used for the desired structural determination. To this end, one could try to measure the intensity as a function of scattering vector $\mathbf{K}$ and to infer the structure from the result. This is a formidable task. It is greatly simplified if the specimen under investigation is a crystal with a periodic lattice. In the following, we introduce the mathematical tools that are needed to exploit the crystalline structure in the analysis. The most important one is the so-called reciprocal lattice.

1.3.1.4 The Reciprocal Lattice

The concept of the reciprocal lattice is fundamental to solid state physics because it permits us to exploit the crystal symmetry for the analysis of many problems. Here we will use it to describe X-ray diffraction from periodic structures and we will meet it again and again in the next chapters. Unfortunately, the meaning of the reciprocal lattice turns out to be hard to grasp. Here, we choose to start out with a formal definition and we provide some mathematical properties. We then discuss the meaning of the reciprocal lattice before we come back to X-ray diffraction. The full importance of the concept will become apparent throughout this book.

For a given Bravais lattice

$$\mathbf{R} = m\mathbf{a_1} + n\mathbf{a_2} + o\mathbf{a_3}, \tag{1.12}$$

we define the reciprocal lattice as the set of vectors $\mathbf{G}$ for which

$$\mathbf{R} \cdot \mathbf{G} = 2\pi l, \tag{1.13}$$

where l is an integer. Equivalently, we could require that

$$e^{i\mathbf{G}\cdot\mathbf{R}} = 1. \tag{1.14}$$

Note that this equation must hold for *any* choice of the lattice vector $\mathbf{R}$ and reciprocal lattice vector $\mathbf{G}$. We can write any $\mathbf{G}$ as the sum of three vectors

$$\mathbf{G} = m'\mathbf{b_1} + n'\mathbf{b_2} + o'\mathbf{b_3}, \tag{1.15}$$

where m', n' and o' are integers. The reciprocal lattice is again a Bravais lattice. The vectors $\mathbf{b_1}$, $\mathbf{b_2}$, and $\mathbf{b_3}$ spanning the reciprocal lattice can be constructed explicitly from the lattice vectors

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2\times\mathbf{a}_3}{\mathbf{a}_1\cdot(\mathbf{a}_2\times\mathbf{a}_3)}, \qquad \mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3\times\mathbf{a}_1}{\mathbf{a}_1\cdot(\mathbf{a}_2\times\mathbf{a}_3)}, \qquad \mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1\times\mathbf{a}_2}{\mathbf{a}_1\cdot(\mathbf{a}_2\times\mathbf{a}_3)}. \tag{1.16}$$

From this, one can derive the simple but useful property,[1)]

$$\mathbf{a}_i\cdot\mathbf{b}_j = 2\pi\delta_{ij}, \tag{1.17}$$

which can easily be verified. Equation (1.17) can then be used to verify that the reciprocal lattice vectors defined by (1.15) and (1.16) do indeed fulfill the fundamental property of (1.13) that defines the reciprocal lattice (see Problem 1.6).

Another way to view the vectors of the reciprocal lattice is as wave vectors that yield plane waves with the periodicity of the Bravais lattice because

$$e^{i\mathbf{G}\cdot\mathbf{r}} = e^{i\mathbf{G}\cdot\mathbf{r}}e^{i\mathbf{G}\cdot\mathbf{R}} = e^{i\mathbf{G}\cdot(\mathbf{r}+\mathbf{R})}. \tag{1.18}$$

Finally, one can define the **Miller indices** in a much simpler way using the reciprocal lattice: The Miller indices (i, j, k) define a plane that is perpendicular to the reciprocal lattice vector $i\mathbf{b_1} + j\mathbf{b_2} + k\mathbf{b_3}$ (see Problem 1.8).

1.3.1.5 The Meaning of the Reciprocal Lattice

We have now defined the reciprocal lattice in a proper way, and we will give some simple examples of its usefulness. The most important point of the reciprocal lattice is that it facilitates the description of functions that have the periodicity of the lattice. To see this, consider a one-dimensional lattice, a chain of points with a lattice constant a. We are interested in a function with the periodicity of the lattice, like the electron concentration along the chain $\rho(x) = \rho(x + a)$. We can write this as a Fourier series of the form

$$\rho(x) = C + \sum_{n=1}^{\infty}\left\{C_n\cos(x2\pi n/a) + S_n\sin(x2\pi n/a)\right\} \tag{1.19}$$

with real coefficients C_n and S_n. The sum starts at $n = 1$, that is, the constant part C has to be taken out of the sum. We can also write this in a more compact form

$$\rho(x) = \sum_{n=-\infty}^{\infty}\rho_n e^{ixn2\pi/a}, \tag{1.20}$$

1) δ_{ij} is Kronecker's delta, which is 1 for $i = j$ and zero otherwise.

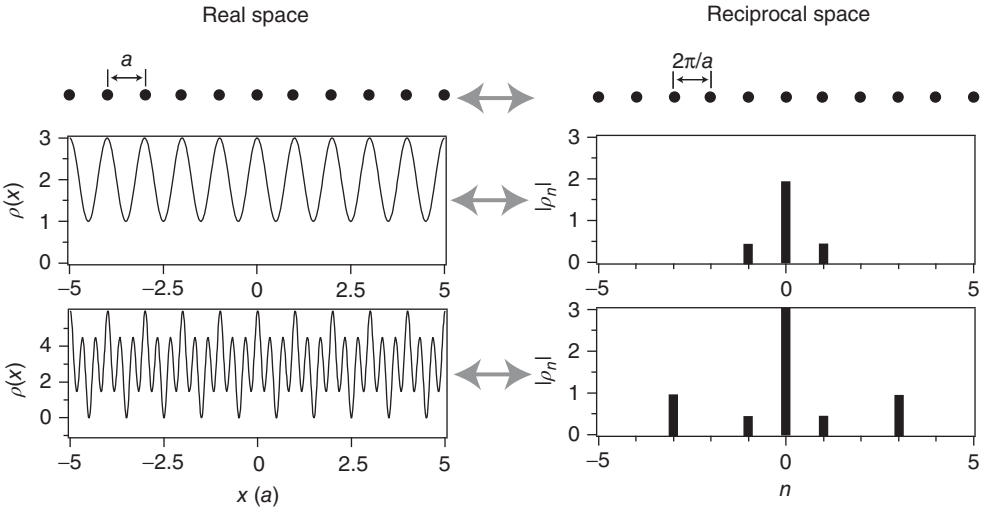


Figure 1.11 Top: Chain with a lattice constant a as well as its reciprocal lattice, a chain with a spacing of $2\pi/a$. Middle and bottom: Two lattice-periodic functions $\rho(x)$ in real space as well as their Fourier coefficients. The magnitude of the Fourier coefficients $|\rho_n|$ is plotted on the reciprocal lattice vectors they belong to.

using complex coefficients ρ_n. To ensure that $\rho(x)$ is still a real function, we have to require that

$$\rho_{-n}^{*} = \rho_n, \tag{1.21}$$

that is, that the coefficient ρ_{-n} must be the conjugate complex of the coefficient ρ_n. This description is more elegant than the one with the sine and cosine functions. How is it related to the reciprocal lattice? In one dimension, the reciprocal lattice of a chain of points with lattice constant a is also a chain of points with spacing $2\pi/a$ (see (1.17)). This means that we can write a general reciprocal lattice "vector" as

$$g = n\frac{2\pi}{a}, \tag{1.22}$$

where n is an integer. Exactly these reciprocal lattice "vectors" appear in (1.20). In fact, (1.20) is a sum of functions with a periodicity corresponding to the reciprocal lattice vectors, weighted by the coefficients ρ_n. Figure 1.11 illustrates these ideas by showing the lattice and reciprocal lattice for such a chain as well as two lattice-periodic functions, as real space functions and as Fourier coefficients on the reciprocal lattice points. The advantage of describing the functions by the coefficients ρ_n is immediately obvious: Instead of giving $\rho(x)$ for *every* point in a range of $0 \leqslant x < a$, the Fourier description consists only of three numbers for the upper function and five numbers for the lower function. Actually, it is only two and three numbers because of (1.21).

The same ideas also work in three dimensions. In fact, one can use a Fourier sum for lattice-periodic properties, which exactly corresponds to (1.20). For the

lattice-periodic electron concentration $\rho(\mathbf{r}) = \rho(\mathbf{r} + \mathbf{R})$, we get

$$\rho(\mathbf{r}) = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{1.23}$$

where $\mathbf{G}$ are the reciprocal lattice vectors.

With this we have seen that the reciprocal lattice is very useful for describing lattice-periodic functions. But this is not all: It can also simplify the treatment of waves in crystals in a very general sense. Such waves can be X-rays, elastic lattice distortions, or even electronic wave functions. We will come back to this point at a later stage.

1.3.1.6 X-Ray Diffraction from Periodic Structures

Turning back to the specific problem of X-ray diffraction, we can now exploit the fact that the electron concentration is lattice-periodic by inserting (1.23) in our expression (1.11) for the diffracted intensity. This gives

$$I(\mathbf{K}) \propto \left| \sum_{\mathbf{G}} \rho_{\mathbf{G}} \int_V e^{i(\mathbf{G}-\mathbf{K})\cdot\mathbf{r}} dV \right|^2. \tag{1.24}$$

Let us inspect the integrand. The exponential function represents a plane wave with a wave vector $\mathbf{G} - \mathbf{K}$. If the crystal is very big, the integration will average over the crests and troughs of this wave and the result of the integration will be very small (or zero for an infinitely large crystal). The only exception to this is the case where

$$\mathbf{K} = \mathbf{k}' - \mathbf{k} = \mathbf{G}, \tag{1.25}$$

that is, when the difference between incoming and scattered wave vector is equal to a reciprocal lattice vector. In this case, the exponential function in the integral is 1, and the value of the integral is equal to the volume of the crystal. Equation (1.25) is often called the **Laue condition**. It is central to the description of X-ray diffraction from crystals in that it describes the condition for the observation of constructive interference.

Looking back at (1.24), the observation of constructive interference for a chosen scattering geometry (or scattering vector $\mathbf{K}$) clearly corresponds to a particular reciprocal lattice vector $\mathbf{G}$. The intensity measured at the detector is proportional to the square of the Fourier coefficient of the electron concentration $|\rho_{\mathbf{G}}|^2$. We could therefore think of measuring the intensity of the diffraction spots appearing for all possible reciprocal lattice vectors, obtaining the Fourier coefficients of the electron concentration and reconstructing this concentration. This would give all the information needed and conclude the process of the structural determination. Unfortunately, this straightforward approach does not work because the Fourier coefficients are complex, not real numbers. Taking the square root of the intensity at the diffraction spot therefore gives the magnitude but not the phase of $\rho_{\mathbf{G}}$. The phase is lost in the measurement. This is known as the **phase problem** in X-ray diffraction. One has to work around it to solve the structure. One simple approach is to calculate the electron concentration for a structural model, obtain

the magnitude of the $\rho_{\mathbf{G}}$ values and thus also the expected diffracted intensity, and compare this to the experimental result. Based on the outcome, the model can be refined until the agreement is satisfactory.

More precisely, this can be done in the following way. We start with (1.11), the expression for the diffracted intensity that we had obtained before introducing the reciprocal lattice. But now we know that constructive interference is only observed in a geometry that corresponds to fulfilling the Laue condition and we can therefore write the intensity for a particular diffraction spot as

$$I(\mathbf{G}) \propto \left| \int_V \rho(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} dV \right|^2. \tag{1.26}$$

We also know that the crystal is made of many identical unit cells at the positions of the Bravais lattice $\mathbf{R}$. We can split the integral up as a sum of integrals over the individual unit cells

$$I(\mathbf{G}) \propto \left| \sum_{\mathbf{R}} \int_{V_{\text{cell}}} \rho(\mathbf{r}+\mathbf{R}) e^{-i\mathbf{G}\cdot(\mathbf{r}+\mathbf{R})} dV \right|^2 = \left| N \int_{V_{\text{cell}}} \rho(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} dV \right|^2, \tag{1.27}$$

where N is the number of unit cells in the crystal and we have used the lattice periodicity of $\rho(\mathbf{r})$ and (1.14) in the last step. We now assume that the electron concentration in the unit cell $\rho(\mathbf{r})$ is given by the sum of atomic electron concentrations $\rho_i(\mathbf{r})$ that can be calculated from the atomic wave functions. By doing so, we neglect the fact that some of the electrons form the bonds between the atoms and are not part of the spherical electron cloud around the atom any longer. If the atoms are not too light, however, the number of these valence electrons is small compared to the total number of electrons and the approximation is appropriate. We can then write

$$\rho(\mathbf{r}) = \sum_i \rho_i(\mathbf{r} - \mathbf{r}_i), \tag{1.28}$$

where we sum over the different atoms in the unit cell at positions $\mathbf{r}_i$. This permits us to rewrite the integral in (1.27) as a sum of integrals over the individual atoms in the unit cell

$$\int_{V_{cell}} \rho(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} dV = \sum_i e^{-i\mathbf{G}\cdot\mathbf{r}_i} \int_{V_{\text{atom}}} \rho_i(\mathbf{r}') e^{-i\mathbf{G}\cdot\mathbf{r}'} dV', \tag{1.29}$$

where $\mathbf{r}' = \mathbf{r} - \mathbf{r}_i$. The two exponential functions give rise to two types of interference. The first describes the interference between the X-rays scattered by the different atoms in the unit cell, and the second the interference between the X-rays scattered by the electrons within one atom. The last integral is called the **atomic form factor** and can be calculated from the atomic properties alone. We therefore see how the diffracted intensity for an assumed structure can be calculated from the atomic form factors and the arrangement of the atoms.

1.3.1.7 The Ewald Construction

In 1913, P. Ewald published an intuitive geometrical construction to visualize the Laue condition (1.25) and to determine the directions $\mathbf{k}'$ for which constructive

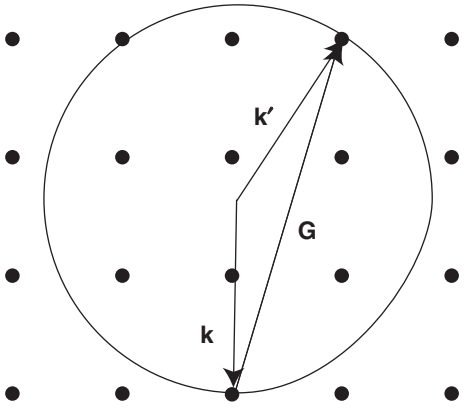


Figure 1.12 Ewald construction for finding the directions in which constructive interference can be observed. The dots represent the reciprocal lattice. The arrows labeled $\mathbf{k}$ and $\mathbf{k}'$ are the wave vectors of the incoming and scattered X-rays, respectively.

interference is to be expected. The construction is shown in Figure 1.12, which represents a cut through the reciprocal lattice; the black points are the reciprocal lattice points. The construction works as follows:

1) We draw the wave vector $\mathbf{k}$ of the incoming X-rays such that it ends in the origin of the reciprocal lattice (we may of course choose the point of origin freely).
2) We construct a circle of radius $|\mathbf{k}|$ around the starting point of $\mathbf{k}$.
3) Wherever the circle touches a reciprocal lattice point, the Laue condition $\mathbf{k}' - \mathbf{k} = \mathbf{G}$ is fulfilled.

For a three-dimensional crystal, this construction has to be carried out in different planes, of course. The figure clearly shows that (1.25) is a very stringent condition: It is not likely for the sphere to hit a second reciprocal lattice point, so that constructive interference is only expected for very few directions. As in the Bragg description, we see that the wavelength of the X-rays has to be short enough ($|\mathbf{k}|$ has to be long enough) for any constructive interference to occur.

Practical X-ray diffraction experiments are often carried out in such a way that many constructive interference maxima are observed despite the strong restrictions imposed by the Laue condition (1.25). This can, for example, be achieved by using a wide range of X-ray wavelengths, that is, non monochromatic radiation or by doing diffraction experiments not on one single crystal but on a powder of randomly oriented small crystals.

1.3.1.8 Relation Between Bragg and Laue Theory

We conclude our treatment of X-ray diffraction by showing that the Bragg description of X-ray diffraction is just a special case of the Laue description. We start by noting that the Laue condition (1.25) consists, in fact, of three separate conditions for the three components of the vectors. In the Bragg experiment, two of these

conditions are automatically fulfilled because of the specular geometry: The wave vector change parallel to the lattice planes is zero. So, the vector equation (1.25) reduces to the scalar equation

$$k'_{\perp} - k_{\perp} = 2k_{\perp} = 2\frac{2\pi}{\lambda}\sin\Theta = G_{\perp}, \tag{1.30}$$

where $G_{\perp}$ is a reciprocal lattice vector perpendicular to the lattice planes. We have seen in Section 1.3.1.4 that such a reciprocal lattice vector exists for any set of planes. The planes can be defined by their Miller indices (i, j, k) or by the reciprocal lattice vector $\mathbf{G}_{\perp} = i\mathbf{b_1} + j\mathbf{b_2} + k\mathbf{b_3}$ that is perpendicular to the planes (see Problem 1.8). The shortest possible $\mathbf{G}_{\perp}$ has a length of $2\pi/d$ with d being the distance between the planes, but any integer multiple of this will also work. If we thus insert $m2\pi/d$ for $G_{\perp}$ into (1.30), we obtain the usual form of the Bragg condition (1.3).

1.3.2
Other Methods for Structural Determination

While X-ray diffraction is arguably the most widespread and powerful method for structural determination, other techniques are used as well. Similar diffraction experiments can be carried out by making use of the wave character of neutrons or electrons. The former interact very weakly with matter because they are charge-neutral. They are also more difficult to produce than X-rays. However, the use of neutrons has two distinct advantages over X-rays: First, that their relative interaction strength with light atoms is stronger and second, that they carry a magnetic moment. They can therefore interact with the magnetic moments in the solid, that is, one can determine the magnetic order. Electrons, on the other hand, have the advantages that they are easy to produce and that one can use electron-optical imaging techniques, whereas making optical elements for X-rays is very difficult. On the other hand, their very strong interaction with matter causes a breakdown of the kinematic approximation, that is, multiple scattering events have to be taken into account. Because of the strong interaction with matter, low-energy electrons do not penetrate deeply into crystals either. Therefore, they are more appropriate for surface structure determination.

1.3.3
Inelastic Scattering

Our discussion has been confined to the case of elastic scattering. In real experiments, however, the X-rays or particles can also lose energy during the scattering events. This can be described formally by considering the diffraction from a structure that does not consist of ions at fixed positions but is time-dependent, that is, which fluctuates with the frequencies of the atomic vibrations. We cannot go into the details of inelastic scattering processes here, but it is important to emphasize that the inelastic scattering, especially of neutrons, can be used to measure the vibrational properties of a lattice.

1.4 Further Reading

The concepts of lattice-periodic solids, crystal structure, and X-ray diffraction are discussed in all standard texts on solid state physics, for example,

- Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Holt-Saunders.
- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.
- Kittel, C. (2005) *Introduction to Solid State Physics*, 8th edn, John Wiley & Sons, Inc.
- Rosenberg, H.M. (1988) *The Solid State*, 3rd edn, Oxford University Press.

For a more detailed discussion of X-ray diffraction, see, for example,

- Als-Nielsen, J. and McMorrow, D. (2011) *Elements of Modern X-Ray Physics*, 2nd edn, John Wiley & Sons, Ltd.

1.5 Discussion and Problems

Discussion

1) What mathematical concepts do you need to describe the structure of any crystal?
2) What are typical crystal structures for metals and why?
3) Why do covalent crystals typically have a much lower packing density than metal crystals?
4) How can the reciprocal lattice conveniently be used to describe lattice-periodic functions?
5) How can you determine the structure of crystals?
6) What is the difference between the Bragg and von Laue descriptions of X-ray diffraction?
7) How can you use the reciprocal lattice of a crystal to predict the pattern of diffracted X-rays?

Problems

1) *Fundamental concepts:* In the two-dimensional crystal in Figure 1.13, find (a) the Bravais lattice and a primitive unit cell, (b) a nonprimitive, rectangular unit cell, and (c) the basis.
2) *Real crystal structures:* Show that the packing of spheres in a simple cubic lattice fills 52% of available space.
3) *Real crystal structures*: Figure 1.14 shows the structure for a two-dimensional hexagonal packed layer of atoms, a hcp crystal, a two-dimensional sheet of carbon atoms arranged in a honeycomb lattice (graphene), and three-dimensional graphite. (a) Draw a choice of vectors spanning the Bravais

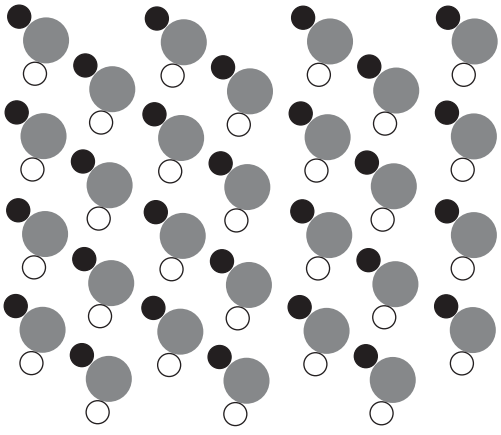


Figure 1.13 A two-dimensional crystal.

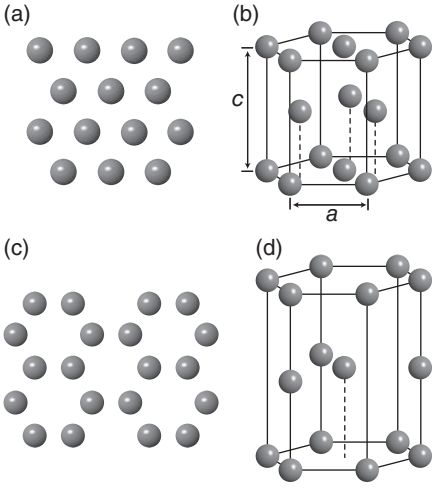


Figure 1.14 (a) Two-dimensional crystal structure of a hexagonal close-packed layer of atoms. (b) Crystal structure for a three-dimensional hcp crystal. (c) Two-dimensional crystal structure for graphene. (d) Three-dimensional crystal structure for graphite (strongly compressed along the *c* direction). The lines are a mere guide to the eye, not indicating bonds or the size of the unit cell.

lattice for the hexagonal layer of atoms and for graphene, and compare them to each other. (b) Show that the basis for the hexagonal layer contains one atom, while the bases for graphene and the three-dimensional hcp crystal contain two atoms. (c) (*) Choose the vectors for the Bravais lattice for graphite and show that the basis contains four atoms.

4) *Real crystal structures:* Consider the hcp lattice shown in Figure 1.14b. The Bravais lattice underlying the hcp structure is given by two vectors of length a in one plane, with an angle of 60° between them and a third vector of length c perpendicular to that plane. There are two atoms per unit cell. (a) Show that for the ideal packing of spheres, the ratio $c/a = (8/3)^{1/2}$. (b) (*) Construct the reciprocal lattice. Does the fact that there are two atoms per unit cell in the hcp crystal have any relevance? Hint: Use the result of Problem 1.7.
5) *X-ray diffraction:* (a) Determine the maximum wavelength for which constructive interference can be observed in the Bragg model for a simple cubic crystal with a lattice constant of 3.6 Å. (b) What is the energy of the X-rays in electron volts? (c) If you were to perform neutron diffraction, what would the energy of the neutrons have to be in order to obtain the same de Broglie wavelength? (d) You could argue that if you take X-rays with twice the wavelength, you would still get a Bragg peak because there would be constructive interference between the X-rays that are reflected from every other plane. Why is this argument not valid? (e) You could describe the same crystal by using a unit cell that is a bigger cube of twice the side length, containing eight atoms instead of one. The lattice constant would then be 7.2 Å. Discuss how this different description would affect the X-ray diffraction from the crystal.
6) *The reciprocal lattice:* Using the explicit definition of the reciprocal lattice (1.16), show first that (1.17) is fulfilled and then, using this relation, show that the reciprocal lattice defined by (1.16) does indeed fulfill the condition (1.13).
7) *The reciprocal lattice:* For a two-dimensional Bravais lattice

$$\mathbf{R} = m\mathbf{a_1} + n\mathbf{a_2}, \tag{1.31}$$

the reciprocal lattice is also two-dimensional:

$$\mathbf{G} = m'\mathbf{b_1} + n'\mathbf{b_2}. \tag{1.32}$$

Often, the most practical way to construct the reciprocal lattice is to use the relation

$$\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}, \tag{1.33}$$

which remains valid in the two-dimensional case. Find the reciprocal lattice for the three cases given in Figure 1.15.

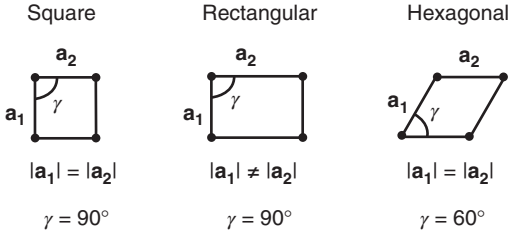


Figure 1.15 Two-dimensional Bravais lattices.

8) *Miller indices:* We have stated that the reciprocal lattice vector $m\mathbf{b_1} + n\mathbf{b_2} + o\mathbf{b_3}$ is perpendicular to the lattice plane given by the Miller indices (m, n, o). (a) Verify that this is correct for the lattice planes drawn in Figure 1.9. (b) (*) Show that this is true in general.

2
Bonding in Solids

After studying the structure of crystals, we now discuss the different mechanisms that lead to bonding between atoms such that they form these structures. We will encounter different scenarios such as ionic, covalent, or metallic bonding. It has to be kept in mind that these are just idealized limiting cases. Often mixed bonding types are found, for example, a combination of metallic and covalent bonding in the transition metals.

As in conventional chemistry, only a fraction of the electrons, the so-called **valence electrons**, participate in the bonding. These are the electrons in the outermost shell(s) of an atom. The electrons in the inner shells, or **core electrons**, are bound so tightly to the nucleus that their energies and wave functions are hardly influenced by the presence of other atoms in their neighborhood.

2.1
Attractive and Repulsive Forces

Two different forces must be present to establish bonding in a solid or in a molecule. An attractive force is necessary for any bonding. Different types of attractive forces are discussed in the following sections. A repulsive force, on the other hand, is required in order to keep the atoms from getting too close to each other. A simple expression for an **interatomic potential** can thus be written as

$$\phi(r) = \frac{A}{r^n} - \frac{B}{r^m}, \tag{2.1}$$

where r is the distance between the atoms and $n > m$, that is, the repulsive part has to prevail for short distances (sometimes, this is achieved by assuming an exponential repulsion potential). Such a potential and the resulting force are shown in Figure 2.1. The reason for the strong repulsion at short distances is the Pauli exclusion principle. For a strong overlap of the electron clouds of two atoms, the wave functions have to change in order to become orthogonal to each other, because the Pauli principle forbids having more than two electrons in the same quantum state. The orthogonalization costs much energy, hence the strong repulsion.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

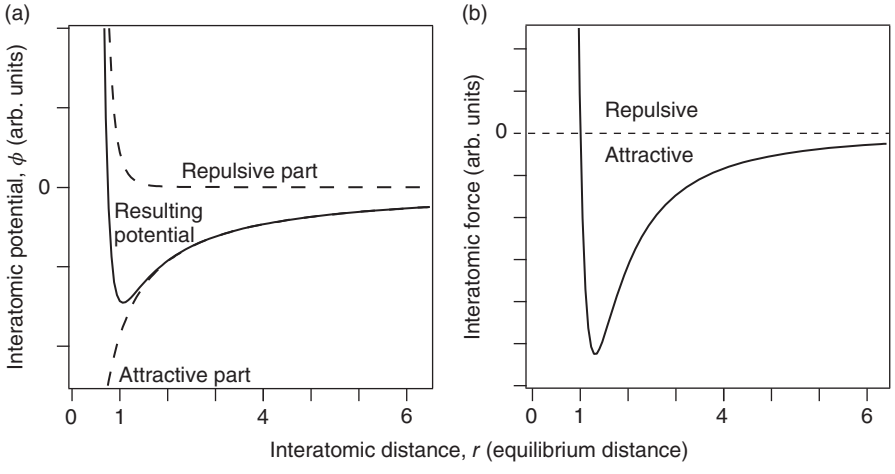


Figure 2.1 (a) Typical interatomic potential $\phi(\mathbf{r})$ for bonding in solids according to (2.1) with $n = 6$ and $m = 1$. (b) Resulting force, that is, $-\mathrm{grad}\phi(\mathbf{r})$.

2.2 Ionic Bonding

Ionic bonding involves the transfer of electrons from an electropositive atom to an electronegative atom. The bonding force is the Coulomb attraction between the two resulting ions. Turning the atoms into ions usually costs some energy. In the case of NaCl, the ionization energy of Na is 5.1 eV and the electron affinity of Cl is 3.6 eV. The net energy cost for creating a pair of ions is thus $5.1 - 3.6 = 1.5$ eV. The energy gain is given by the Coulomb interaction. For just one Na and one Cl ion separated by the distance found in the actual crystal structure of NaCl ($a = 0.28$ nm), this is $-e^2/4\pi\epsilon_0 a$, which amounts to 5.1 eV.

Knowing the crystal structure for NaCl, we can also calculate the electrostatic energy gain for forming an entire crystal. Consider one Na ion at the center of the NaCl cube in Figure 1.5. It has six Cl ions at a distance of $a = 0.28$ nm. They lead to an electrostatic energy gain of $-6e^2/4\pi\epsilon_0 a$. At a distance of $a\sqrt{2}$, there are 12 other Na ions that give rise to an energy increase of $+12e^2/4\pi\epsilon_0 a\sqrt{2}$. Then, one finds eight Cl ions that again decrease the energy. Eventually, this series converges and the total energy gain is

$$E_{\mathrm{Na}} = -1.748\frac{e^2}{4\pi\epsilon_0 a} = -M_d\frac{e^2}{4\pi\epsilon_0 a}. \tag{2.2}$$

M_d is called the **Madelung constant**. It is specific for a given structure (for the calculation of M_d, see Problem 2.3). For calculating the electrostatic energy gain per mole, we have to multiply (2.2) by Avogadro's number N_A. We also have to multiply it by a factor of 2 to account for the fact that we have both Na and Cl ions in the solid. But at the same time, we have to divide it by 2 in order to avoid a double counting of bonds when we evaluate the electrostatic energy gain. So in the end, the energy gain for 1 mol of NaCl is simply $-N_A 1.748e^2/4\pi\epsilon_0 a$. Note that M_d

is larger than 1 so that the energy gain for forming a solid is higher than that for an isolated dimer of ions. This is of course obvious since your salt shaker contains little crystals, not a molecular powder.

We can define the following contributions to the energy balance for forming the solid. The **cohesive energy** is the total energy difference between any solid and the isolated atoms it is made of. For an ionic crystal, the cohesive energy can be calculated in a simple way. First, we need to consider how much energy it costs to turn the atoms into ions using the **ionization energy** and **electron affinity** of the atoms. Then, the total electrostatic energy gain for the crystal needs to be calculated using the known crystal structure, as done above for NaCl. This energy gain is called the **lattice energy**. The cohesive energy is then simply the lattice energy minus the energy needed to turn the atoms into ions (see Problem 2.2).

It could appear as if we could calculate the cohesive energy for ionic solids from purely classical physics, but this is not correct. Note that we have used the *experimental* interatomic distance for the calculation of the lattice energy. The calculation of this distance would involve quantum mechanics because it contains the repulsive part of the potential. In fact, the presence of the repulsive potential also causes the actual potential minimum for a given interatomic distance a to be a bit shallower than expected from the pure Coulomb potential (by 10% or so). This can be seen in Figure 2.1 where the potential minimum lies above the Coulomb contribution to the potential at the equilibrium distance. In any event, ionic bonding is very strong. The cohesive energy per atom is on the order of several electron volts.

2.3 Covalent Bonding

Covalent bonding is based on the true sharing of electrons between different atoms. The simplest case is that of the hydrogen molecule that we will discuss quantitatively below. In solids, covalent bonding is often found for elements with a roughly half-filled outer shell. A prominent example is carbon that forms solids such as diamond, graphene, and graphite as well as complex molecules such as Buckminster Fullerene C_{60} or carbon nanotubes. The covalent bonds in diamond are constructed from a linear combination of the 2s orbital and three 2p orbitals. This results in four so-called sp^3 orbitals that stick out in a tetrahedral configuration from the carbon atoms. In graphene and graphite, the 2s orbital is combined with only two 2p orbitals, giving three sp^2 orbitals, all in one plane and separated by an angle of 120°, and one p orbital oriented perpendicular to this plane. This linear combination of orbitals already reveals an important characteristic for the covalent bonding: It is highly directional. In addition to this, it is also very stable and the cohesive energies for covalently bonded solids are typically several electron volts per atom.

An example for covalent bonding is the hydrogen molecule H_2 for which we will sketch a solution here. We go into some detail, as much of this will be useful

for the later discussion of magnetism in Chapter 8. However, understanding these details is not crucial at this point, and the reader could decide to jump to Section 2.4 instead and return here later.

As a starting point, take two hydrogen atoms with their nuclei at $\mathbf{R}_A$ and $\mathbf{R}_B$ and we call $|\mathbf{R}_B - \mathbf{R}_A| = R$. We do, of course, know the solution of the Schrödinger equation for each of the atoms. Let the ground-state wave functions be ϕ_A and ϕ_B, respectively. The Hamilton operator for the hydrogen molecule can be written as

$$H = -\frac{\hbar^2\nabla_1^2}{2m_e} - \frac{\hbar^2\nabla_2^2}{2m_e} + \frac{e^2}{4\pi\epsilon_0}\left\{\frac{1}{R} + \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} - \frac{1}{|\mathbf{R}_A - \mathbf{r}_1|} - \frac{1}{|\mathbf{R}_B - \mathbf{r}_2|} - \frac{1}{|\mathbf{R}_A - \mathbf{r}_2|} - \frac{1}{|\mathbf{R}_B - \mathbf{r}_1|}\right\}, \tag{2.3}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the coordinates of the electrons belonging to the A and the B nucleus, respectively. The first two terms refer to the kinetic energy of the two electrons. The operators ∇_1^2 and ∇_2^2 act only on the coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively. The electrostatic term contains the repulsion between the two nuclei and the repulsion between the two electrons, as well as the attraction between each electron and each nucleus.

The solution of this problem is not simple. It would be greatly simplified by removing the electrostatic interaction between the two electrons because then the Hamiltonian could be written as the sum of two parts, one for each electron (the fixed nuclei would merely contribute with an energy offset). If the last two terms in (2.3) are also removed, the problem could be solved by a product of the two wave functions that are solutions to the two individual atomic Hamiltonians. The two-particle wave function would look like $\phi(\mathbf{r}_1, \mathbf{r}_2) = \phi_A(\mathbf{r}_1)\phi_B(\mathbf{r}_2)$.

Actually, this is not quite right because such a wave function is not in accordance with the Pauli principle. Since the electrons are fermions, the total wave function must be antisymmetric with respect to particle exchange and the simple product wave function does not fulfill this requirement. The total wave function consists of a spatial part and a spin part and, therefore, there are two possibilities for forming an antisymmetric wave function. We can either choose a symmetric spatial part and an antisymmetric spin part or vice versa. This is achieved by constructing the spatial wave function of the form

$$\Psi_{\uparrow\downarrow}(\mathbf{r}_1, \mathbf{r}_2) \propto \phi_A(\mathbf{r}_1)\phi_B(\mathbf{r}_2) + \phi_A(\mathbf{r}_2)\phi_B(\mathbf{r}_1) \tag{2.4}$$

$$\Psi_{\uparrow\uparrow}(\mathbf{r}_1, \mathbf{r}_2) \propto \phi_A(\mathbf{r}_1)\phi_B(\mathbf{r}_2) - \phi_A(\mathbf{r}_2)\phi_B(\mathbf{r}_1), \tag{2.5}$$

The plus sign in (2.4) returns a symmetric spatial wave function that we can combine with an antisymmetric spin wave function with the total spin equal to zero (the so-called **singlet state**); the minus in (2.5) results in an antisymmetric spatial wave function for a symmetric spin wave function with the total spin equal to 1 (the so-called **triplet state**).

The antisymmetric wave function (2.5) vanishes if $\mathbf{r}_1 = \mathbf{r}_2$, that is, the two electrons cannot be at the same place simultaneously. This leads to a depletion of the electron density between the nuclei and hence to an antibonding state. For the

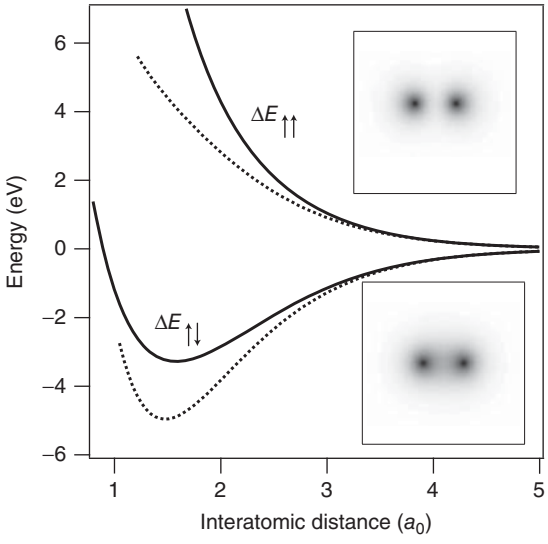


Figure 2.2 The energy changes $\Delta E_{\uparrow\uparrow}$ and $\Delta E_{\uparrow\downarrow}$ for the formation of the hydrogen molecule. The dashed lines represent the approximation for long distances. The two insets show gray scale images of the corresponding electron probability density.

symmetric case, on the other hand, the electrons have opposite spins and can be at the same place, which leads to a charge accumulation between the nuclei and hence to a bonding state (see Figure 2.2).

An approximate way to calculate the eigenvalues of (2.3) was suggested by W. Heitler and F. London in 1927. The idea is to use the known single-particle 1s wave functions for atomic hydrogen for ϕ_A and ϕ_B to form a two-electron wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2)$, which is given by either (2.4) or (2.5). These wave functions might not be entirely correct because the atomic wave functions will certainly be modified by the presence of the other atom. However, even if they are only approximately correct, we can obtain the molecular energy levels as

$$E = \frac{\int \Psi^*(\mathbf{r}_1, \mathbf{r}_2) H \Psi(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2}{\int \Psi^*(\mathbf{r}_1, \mathbf{r}_2) \Psi(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2}. \tag{2.6}$$

According to the variational principle in quantum mechanics, the resulting energy will be higher than the correct ground-state energy but it will approach it for a good choice of the trial wave functions.

The calculation is quite lengthy and shall not be given here.[1)] The resulting ground-state energies for the singlet and triplet states can be written as

$$E_{\text{singlet}} = 2E_0 + \Delta E_{\uparrow\downarrow}, \tag{2.7}$$

$$E_{\text{triplet}} = 2E_0 + \Delta E_{\uparrow\uparrow}. \tag{2.8}$$

1) For the full calculation, see online note on www.philiphofmann.net.

E_0 is the ground-state energy for one hydrogen atom that appears here twice because we start with two atoms. The energies $\Delta E_{\uparrow\uparrow}$ and $\Delta E_{\uparrow\downarrow}$ are also shown in Figure 2.2. $\Delta E_{\uparrow\uparrow}$ is always larger than zero and does not lead to any chemical bonding. $\Delta E_{\uparrow\downarrow}$, on the other hand, shows a minimum below zero at approximately 1.5 times the Bohr radius a_0. This is the bonding state.

For long distances between the nuclei, (2.7) and (2.8) can be rewritten to give

$$E = 2E_0 + C \pm X, \tag{2.9}$$

where the +(−) sign is applied for the singlet (triplet) state. Now the energy change upon bonding has two parts, one that does depend on the relative spin orientations of the electrons ($\pm X$) and one that does not (C). The energy difference between the two states is then given by $2X$, where X is called the **exchange energy**. In the case of the hydrogen molecule, the exchange energy is always negative. Equation (2.9) is a remarkable result because it means that the energy of the system depends on the relative orientation of the spins, even though these spins did not actually enter the Schrödinger equation.

We will encounter similar concepts in the chapter about magnetism where the underlying principle for magnetic ordering is very similar to what we see here: The total energy of a system of electrons depends on their relative spin directions through the exchange energy and, therefore, a particular ordered spin configuration is favored. For two electrons, the "magnetic" character is purely given by the sign of X. For a negative X, the coupling with two opposite spins is favorable (the "antiferromagnetic" case), whereas a positive X would lead to a situation where two parallel spins give the lowest energy (the "ferromagnetic" case).

2.4 Metallic Bonding

In metals, the valence electrons are removed from the ion cores, but in contrast to ionic solids, there are no electronegative ions to bind them. Therefore, they are free to migrate between the ion cores. These delocalized valence electrons are involved in the conduction of electricity and are therefore often called **conduction electrons**. One can expect metals to form from elements for which the energy cost of removing outer electrons is not too big. Nevertheless, this removal always costs some energy that has to be more than compensated by the bonding. Explaining the energy gain from the bonding in an intuitive picture is difficult, but we can at least try to make it plausible. The ultimate reason must be some sort of energy lowering.

One energy contribution that is lowered is the kinetic energy of the conduction electrons. Consider the kinetic energy contribution in a Hamiltonian, $T = -\hbar^2\nabla^2/2m_e$. A matrix element $\langle\Psi|T|\Psi\rangle$ measures the kinetic energy of a particle $T\Psi$ is proportional to the second spatial derivative of the wave function, that is, the curvature. For an electron that is localized to an atom, the curvature of the wave function is much higher than for a nearly free electron in a metal and this is where the energy gain comes from.

The other contribution to the electron energy is the potential energy. One should think that the average electrostatic potential of any single electron in a solid is almost zero because there are (almost) as many other electrons as there are ions with the same amount of charge. But this turns out to be wrong. In fact, the electrons see an attractive potential. The reason is again partly due to the Pauli principle that, loosely speaking, does not allow two electrons with the same spin direction to be at the same place (see (2.5)) and, therefore, the electrons go "out of each other's way." In addition to this, there is also a direct Coulomb interaction between the electrons, which makes them avoid each other. We will discuss this in more detail when dealing with magnetism.

We can also understand why metals prefer close-packed structures. First of all, the metallic bonding does not have any directional preference. Second, close-packed structures secure the highest possible overlap between the valence orbitals of the atoms, maximizing the delocalization of the electrons and thereby the kinetic energy gain. The structures also maximize the number of nearest neighbors for any given atom, again giving rise to strongly delocalized states.

Typically, metallic bonding is not as strong as covalent or ionic bonding but it amounts to a few electron volts per atom. Stronger bonding is found in transition metals, that is, metals with both s and p conduction electrons and a partially filled d shell. The explanation for this is that we have a mixed bonding. The s and p electrons turn into delocalized metallic conduction electrons, whereas the d electrons create much more localized, covalent-type bonds.

2.5 Hydrogen Bonding

Hydrogen atoms have only one electron and can form one covalent bond. If the bond is formed with a very electronegative atom (like F or O), the electron is mostly located close to that atom and the hydrogen nucleus represents an isolated positive (partial) charge. This can lead to a considerable charge density because of the small size, and it can therefore attract negative (partial) charges in other molecules to form an electrostatic bond. This type of bonding is called hydrogen bonding. It is usually quite weak but in some cases, the cohesive energy can be up to several hundred meV per atom. It is responsible for the intermolecular attraction in water ice and for the bonding of the double helix in DNA.

2.6 van der Waals Bonding

The term van der Waals bonding refers to a weak and purely quantum mechanical effect. The electron cloud around an atom or a molecule has no static charge distribution but one governed by quantum mechanical fluctuations. A simple atom with a closed shell can thus be viewed as a fluctuating dipole. The field

of this dipole can polarize other atoms nearby, and the interaction of the two neighboring dipoles reduces the total energy, that is, it can lead to bonding. This type of interaction is present in every solid but it is much weaker than ionic, covalent, or metallic bonding. Typical binding energies per atom are in the meV range and, therefore, van der Waals bonding is only observable for solids that do not show other bonding behavior, for example, noble gases. Pure van der Waals crystals can only exist at very low temperatures.

2.7 Further Reading

Several of the bonding types discussed here are identical to those relevant for the formation of molecules (with the exception of metallic bonding). They are therefore discussed in great depth in the literature for chemistry and molecular physics. A good overview on bonding in solids is given in

- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.
- Kittel, C. (2005) *Introduction to Solid State Physics*, 8th edn, John Wiley & Sons, Inc.

2.8 Discussion and Problems

Discussion

1) Why is a typical interatomic potential, such as in Figure 2.1, so asymmetric?
2) Which elements are likely to form crystals through ionic bonding?
3) What kind of forces are important for ionic bonding?
4) How does the lattice energy in an ionic crystal depend on the interatomic distance?
5) Explain the difference between cohesive energy and lattice energy.
6) Which elements are likely to form metals?
7) Where does the energy gain in metallic bonding come from?
8) What is the difference between a simple metal and a transition metal (definition and typical physical properties)?
9) Why is van der Waals bonding much weaker than most other bonding types?

Problems

1) *Metallic bonding:* The most important contribution to the stability gained by metallic bonding is the lowering of kinetic energy. To see this, consider an electron in a one-dimensional box. The potential shall be zero and infinite inside and outside the box, respectively. Consider first a box with a length corresponding to the size of an atom, say, twice the Bohr radius, and calculate

Figure 2.3 One-dimensional chain of ions.

the lowest energy eigenvalue. Give the result in electron volts. Clearly, this energy is only kinetic energy. By how much is the kinetic energy lowered when you increase the size of the box by a factor of 10, so that it is roughly the size of the interatomic spacing in a crystal?

2) *Ionic bonding:* Calculate the potential energy for an ion in a sodium chloride crystal (the interatomic distance a is 2.81 Å) in units of electron volts and joules. Neglect the influence of the repulsive potential. From this, calculate the lattice energy of sodium chloride and compare the result to the experimental value of 776 kJ mol^{-1}. Also, calculate the cohesive energy in the same units.
3) *Ionic bonding:* The Madelung constant for a three-dimensional crystal of NaCl was presented in Section 2.2. (a) Derive the Madelung constant analytically for a one-dimensional chain of NaCl, as shown in Figure 2.3. (b) (*) Calculate the Madelung constant numerically for a one-dimensional, two-dimensional, and three-dimensional NaCl lattice and plot the result as a function of the number of neighbor "shells" included in the computation. Compare the result for the one-dimensional case to the analytical value from (a).
4) *van der Waals force:* Show that the bonding energy due to the van der Waals force between two atoms depends on their distance r as r^{-6}. Hint: The van der Waals force is caused by the mutual interaction of fluctuating dipoles. Suppose that one atom forms a spontaneous dipole moment at some time. This can be modeled as two point charges, separated by a distance d. This electric dipole gives rise to an electric field $\boldsymbol{\mathcal{E}}(\mathbf{r})$ and the other atom is polarized in this field, such that a dipole moment $\mathbf{p}$ is induced in this second atom. $\mathbf{p}$ is proportional to the field, that is, $\mathbf{p} = \alpha\boldsymbol{\mathcal{E}}(\mathbf{r})$ (see (9.4)). The potential energy of an electric dipole in an electric field is $U = -\boldsymbol{\mathcal{E}} \cdot \mathbf{p} = -\boldsymbol{\mathcal{E}}(\mathbf{r}) \cdot \alpha\boldsymbol{\mathcal{E}}(\mathbf{r})$. Therefore, all you have to show is that the electric field caused by the dipole in the first atom decays as r^{-3} for $r \gg d$.

3
Mechanical Properties

In this chapter, we discuss the macroscopic behavior of a solid that is subject to mechanical stress, and we will try to explain it in a microscopic picture. Our discussion is restricted to isotropic solids, that is, solids for which the direction of the applied mechanical stress with respect to the crystal lattice is not important.

We start out with some fundamental definitions that are all illustrated in Figure 3.1. The applied **stress** on a solid σ is defined as the force F per area A perpendicular to the direction of the applied force (Figure 3.1a). Depending on the force direction, one can distinguish between tensile and compressive stress. The stress has the same dimension as a pressure, that is, Nm^{-2} or Pa. The solid responds to the stress by a deformation called **strain** ϵ. In the case of the tensile stress applied in Figure 3.1a, the response is a length extension Δl_1 in the direction of the force. The strain is defined as the relative length extension $\epsilon = \Delta l_1 / l_1$. It is therefore dimensionless, but in technical texts sometimes the unit meter per meter is found. The strain $\Delta l_1 / l_1$ in Figure 3.1a is frequently accompanied by length changes in the two other directions Δl_2 and Δl_3. In most cases, the solid contracts in these directions. We shall discuss this in more detail below.

A **shear stress** τ is defined in a similar way as the applied force F per area A, but now the force is applied tangentially to the area (Figure 3.1b). Again, the material deforms as a consequence of the shear stress. The deformation is described by the angle α shown in the figure. The last situation illustrated in Figure 3.1c is the exposure of the solid to hydrostatic pressure from all sides. This leads to a reduction of the volume. There are other mechanical deformations such as torsion, and these lead to similar definitions, but we do not discuss them here.

If we consider only the relation between stress and strain, the typical response of a solid is illustrated in Figure 3.2. It shows the resulting stress as a function of applied strain. This type of plot can seem rather odd at first. If you think of the strain as a consequence of the applied stress, you might be tempted to draw the curve with swapped axes. For the interpretation of the curve as it is displayed, one should adopt another point of view: The solid's length is increased and the stress that is "pulling back" is measured for every extension, very much like the force upon the extension of a mechanical spring.

Different regions in the curve can be distinguished. For a very small strain, typically much smaller than 1%, the deformation is **elastic**, that is, the solid goes back

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

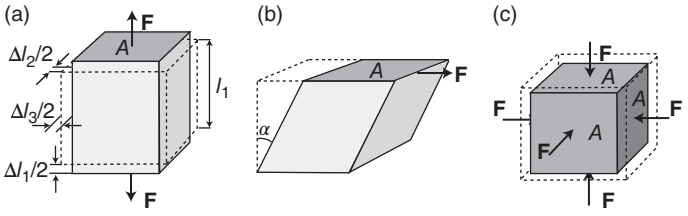


Figure 3.1 (a) Illustration of stress $\sigma = F/A$ and strain $\epsilon = \Delta l_1/l_1$. Δl_2 and Δl_3 are the length changes in the direction perpendicular to **F**. (b) Illustration of shear stress $\tau = F/A$ and shear angle α; and (c) of hydrostatic pressure. Note that the dark shaded areas represent the area A and that the force is perpendicular and parallel to A in (a), (c), and (b), respectively.

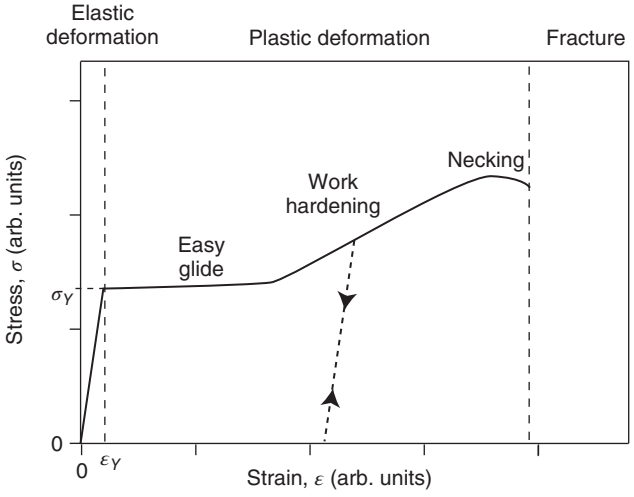


Figure 3.2 Typical stress of a solid as a function of applied strain. ϵ_Y and σ_Y denote the yield strain and yield stress, respectively. The dashed line with the arrows illustrates the behavior when the stress is released and reapplied after increasing the strain into the work hardening region.

to its initial shape once the stress is released. In this region, the stress is also a linear function of the strain and this will allow the definition of various elastic constants in the next section. Beyond a certain **yield strain** ϵ_Y or **yield stress** σ_Y, **plastic** deformation sets in. This means that the deformation is permanent; once the stress is released, the solid does not return to its original shape. It only contracts slightly. The curve's shape in the region of plastic deformation will be discussed in a later section. Eventually, the strain becomes so high that the material fractures. This, naturally, defines the end of the stress/strain curve.

While the region of elastic deformation is usually quite small, the amount of possible plastic deformation can vary widely. Some materials, such as glass or cast iron, will fracture immediately at the end point of the elastic limit. Such materials

are called **brittle**. Materials that do show plastic deformation before they fracture are called **ductile**. Most metals are ductile.

3.1 Elastic Deformation

The elastic regime of deformation is small, but it is of high technical importance because most applications require the deformation of materials to remain elastic. Apart from exploring the limits of elastic deformation, an interesting question is how strongly the material resists such a deformation. This is described by the macroscopic elastic constants that we shall introduce now. We will also see that these constants can be connected to the picture of interatomic bonding that we have encountered earlier.

3.1.1 Macroscopic Picture

3.1.1.1 Elastic Constants

The linear behavior for the small deformations in the elastic regime leads to a few definitions of macroscopic elastic constants. The relation between stress and strain is given by **Young's modulus** Y:

$$Y = \frac{\sigma}{\epsilon} = \frac{F}{A}\frac{l}{\Delta l}. \tag{3.1}$$

Young's modulus has therefore the same unit as the stress, that is, Pascal. It is the slope of the initial stress/strain curve in Figure 3.2. The values of Young's modulus are very high, typically in the gigapascal region.

The possibility to define Young's modulus is equivalent to the validity of Hooke's law that is commonly used to describe a "spring-like" force response. Suppose you extend a spring by some small amount. It is going to respond by a force that is proportional to the extension. This is equivalent to

$$\sigma = Y\epsilon. \tag{3.2}$$

Multiplication by A gives

$$F = \frac{YA}{l}\Delta l, \tag{3.3}$$

so that the usual spring constant is YA/l. The advantage of using Y instead of the spring constant is that it depends only on the material, not on the geometry.

The shearing of a solid can also be described by an elastic constant. The **modulus of rigidity** G is defined by

$$G = \frac{\tau}{\alpha}. \tag{3.4}$$

Finally, the exposure of the solid to hydrostatic pressure leads to the definition of the **bulk modulus** K via

$$K = -p\frac{V}{\Delta V}, \tag{3.5}$$

where $p = F/A$ is the pressure and V the volume. The minus sign is introduced in order to obtain a positive bulk modulus for a decrease in volume. Note that both G and K have the unit Pascal, just like Y.

3.1.1.2 Poisson's Ratio

When mechanical stress is applied to a solid, the strain in the direction of the stress is not the only consequence. In addition to this, the solid's dimensions may change in the directions perpendicular to the stress, as illustrated in Figure 3.1a. This change is described by Poisson's ratio ν, which is defined as

$$\frac{\Delta l_2}{l_2} = \frac{\Delta l_3}{l_3} = -\nu\frac{\Delta l_1}{l_1} = -\nu\epsilon. \tag{3.6}$$

As we discuss only isotropic solids, the fractional changes $\Delta l_2/l_2$ and $\Delta l_3/l_3$ are the same. The minus sign in the definition assures that ν is positive in the "normal" situation where the solid contracts sideways upon tensile stress. However, there are some exotic materials with a negative Poisson's ratio, for example, special types of molecular foam that expand sideways upon being exposed to tensile stress and which contract when subject to compressive stress.

Poisson's ratio cannot take all possible values. It is limited to a range between -1 and $+0.5$. The lower limit is not so relevant as materials with a negative Poisson's ratio are rather rare. The upper limit is caused by the fact that a solid cannot decrease its volume when we pull on one side and that it cannot increase its volume when we press it from one side. To calculate the upper limit of ν, consider the volume of the solid after the application of stress

$$(l_1 + \Delta l_1)(l_2 + \Delta l_2)(l_3 + \Delta l_3). \tag{3.7}$$

For small changes, we can neglect higher order terms in the Δl's and this becomes

$$l_1 l_2 l_3 + \Delta l_1 l_2 l_3 + l_1 \Delta l_2 l_3 + l_1 l_2 \Delta l_3. \tag{3.8}$$

So the change in volume is

$$\begin{aligned}\Delta l_1 l_2 l_3 + l_1 \Delta l_2 l_3 + l_1 l_2 \Delta l_3 &= \Delta l_1 l_2 l_3 + l_1\left(-\nu\frac{\Delta l_1}{l_1} l_2\right) l_3 + l_1 l_2\left(-\nu\frac{\Delta l_1}{l_1} l_3\right) \\ &= (1-2\nu)\Delta l_1 l_2 l_3. \end{aligned} \tag{3.9}$$

For a positive Δl_1, this must not be negative, which can only be achieved for values of ν smaller than or equal to 0.5.

Typical values of Poisson's ratio range between 0.2 and 0.4 for most materials. Rubber has a ν very close to 0.5, that is, it is an almost ideal noncompressible solid. Cork has $\nu \approx 0$, which is advantageous when you try to put a wine cork back into the bottle.

3.1.1.3 Relation between Elastic Constants

In a broader mathematical context, all the macroscopic constants can be derived from a few fundamental elastic properties. Not surprisingly, they are therefore related to each other. For example,[1)]

$$G = \frac{Y}{2(1+\nu)}. \tag{3.10}$$

A similar relation exists between the bulk modulus and Young's modulus (3.17), and Problem 3.1 enables us to understand the origin of this relation. The close connection between different elastic properties has an advantage for our task to explain the macroscopic behavior in a microscopic picture. We will, in most cases, restrict ourselves to explaining one type of mechanical property and we are allowed to do so without great loss of generality. In any case, we see from (3.10) and (3.17) that for a given material with ν in the "normal" range, the elastic constants Y, G, and K have the same order of magnitude.

3.1.2 Microscopic Picture

The elastic deformation of a solid can be explained in terms of changing interatomic distances. According to Figure 2.1, the equilibrium distance between two atoms corresponds to the minimum in the interatomic potential ϕ and the force at this distance, which is just the negative spatial derivative of the potential, is zero. Upon the application of a compressive stress, the distance between the atoms is decreased. This results in a force that presses the atoms away from each other. For tensile stress, it is the other way round. Once the stress is released, the atoms return to their equilibrium distance.

This explains why the behavior is elastic, but why is it linear? A linear force for distance changes close to the equilibrium can readily be seen in Figure 2.1b. More formally, we can expand the potential for distances close to the equilibrium $x = a$ as a Taylor series:

$$\phi(x) = \phi(a) + \frac{\phi'(a)}{1!}(x-a) + \frac{\phi''(a)}{2!}(x-a)^2 + \frac{\phi'''}{3!}(x-a)^3 + \dots \tag{3.11}$$

The first term is simply an offset of the absolute energy scale and therefore irrelevant here. The second term is zero because the derivative of ϕ vanishes at the equilibrium distance. The third term is responsible for the elastic behavior. It states that the potential close to the equilibrium is proportional to the square of the distance change, that is, the force depends linearly on the distance change. Moreover, it is the curvature of the potential that gives rise to the interatomic force constant. The fourth and higher order terms are usually neglected.

For distances $(x - a)$ that are sufficiently large, terminating the Taylor series after the third term may become imprecise, and one would expect to see nonlinear elastic deformation. In this case, the elastic constants would depend on the actual

1) For a derivation of this equation, see online note on www.philiphofmann.net.

change of the atomic separation, for example, Young's modulus would depend on the applied stress. It turns out, however, that this is very rarely observed. For most solids, the plastic deformation sets in for a strain of less than 1%, and this is before higher order terms in (3.11) become important.

The interatomic force constant that is calculated from the Taylor series (3.11) also allows for harmonic vibrations of the atoms. Indeed, we will later see that it is possible to relate the vibrational properties of a solid to its elastic properties (see Section 4.1.5.2).

We conclude our treatment of the elastic regime by looking at typical values of Young's modulus in Figure 3.3. As stated earlier, Y is very high for most materials, on the order of many gigapascal. It is also apparent how different bonding types lead to different values of Young's modulus. Metals and alloys are all in the range between 15 and 300 GPa. As a tendency, transition metals have a higher Y than simple metals, consistent with a stronger bonding due to localized d electrons. W and Mo have particularly strong bonding, something that leads to a high Y and high melting points, as we shall see later. Solids with covalent bonding span a much wider range. The sp^2 and sp^3 bonds in graphite and diamond lead to particular strength, but note that graphite also appears at the lower end of the range. This is because we have neglected the possibility of anisotropy in solids. Graphite is strongly bonded parallel to the sp^2-linked planes but very weakly perpendicular to these planes. Not surprisingly, graphene, the single layer of graphite, has the same Young's modulus as graphite in the two-dimensional plane. Polymers show low values of Y. The reason is that a reversible length extension in a polymer does not have to be achieved by extending interatomic bonds. It is sufficient to change the angles of the many bonds in a polymer, that is, to "unfold" it. Composite materials and fibers cover a wide range of Y, from carbon nanotubes that have an extremely high Y, just like graphene (they can be viewed as rolled-up graphene sheets), to wood perpendicular to the grains that has $Y < 1$ GPa.

Note that Young's modulus does not reflect the same physical quantity as the cohesive energy discussed in Chapter 2. The cohesive energy measures how deep the potential minimum in Figure 2.1 is. Young's modulus, on the other hand, corresponds to the curvature of the potential around the minimum. Obviously, these two are related as the properties in Figure 3.3 correspond well to what we have discussed in connection with the cohesive energies. The covalent bonds in diamond, for example, give rise to a high cohesive energy, and at the same time, they strongly resist small changes in the bonding distance.

3.2 Plastic Deformation

Now we address the plastic deformation part of the stress/strain curve. We will be able to establish a link to microscopic models, but a detailed understanding of all phenomena cannot be accomplished on the basis of the perfect crystal. It will be necessary to introduce different types of imperfections, such as point defects and

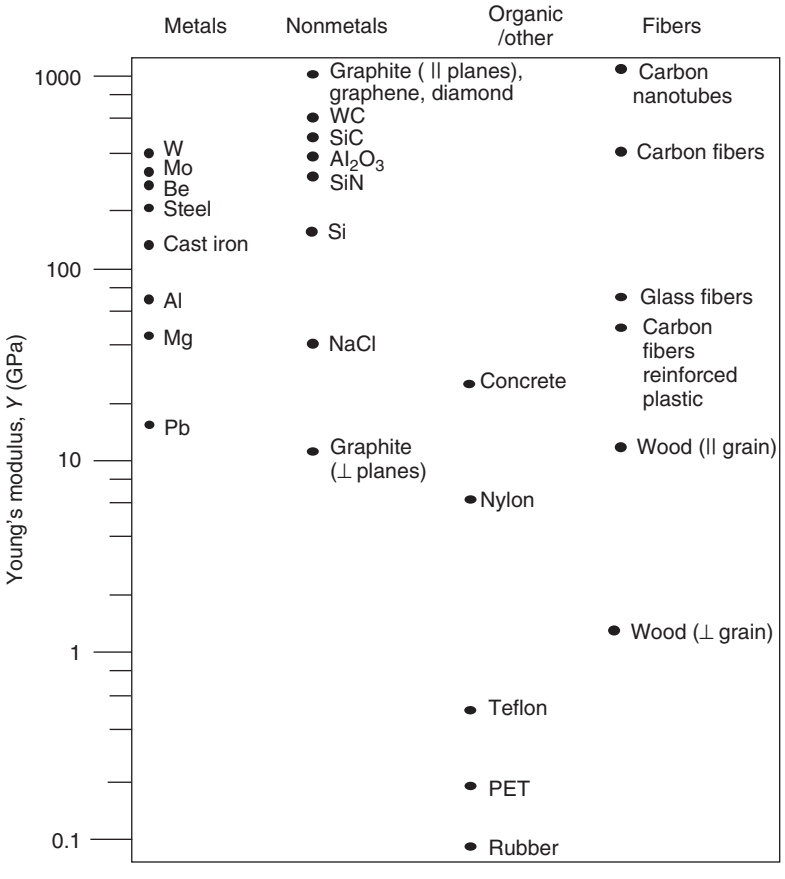


Figure 3.3 Young's modulus for different materials. The values are merely a guide, as strong variations are possible.

dislocations. Such imperfections will also be important in our later treatment of electrical resistance.

3.2.1 Estimate of the Yield Stress

The most important practical question for many applications of materials is where plastic deformation starts to set in. Can we estimate the yield stress or yield strain? On the atomic scale, this is particularly simple for a shear deformation of a crystal.

Figure 3.4a shows two atom rows in a hexagonal crystal plane that in Figure 3.4b is subject to shear stress. This leads to a deformation with a shear angle α in which the atomic rows are pulled away from their lowest energy position. For small shear angles, we can relate α to the interlayer distance a and the displacement x:

$$\alpha = \tan^{-1}\left(\frac{x}{a}\right) \approx \frac{x}{a}. \tag{3.12}$$

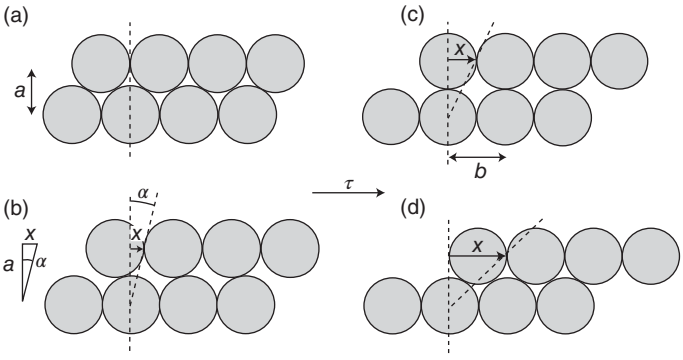


Figure 3.4 Estimate of the yield stress for shearing a solid. (a) Atoms in equilibrium position. (b) Distortion for a small shear stress. (c) Meta-stable equilibrium. (d) New stable equilibrium for the sheared solid.

This expression of α is inserted in the definition of the modulus of rigidity (3.4):

$$\tau = G\alpha \approx \frac{Gx}{a}, \tag{3.13}$$

which directly establishes a link between macroscopic quantities and the microscopic displacements.

For small values of α, the atoms will go back into their equilibrium positions as soon as the shear stress is released; the deformation is elastic. But if the shear stress is increased more and more, the rows of atoms will eventually start to glide over each other. This leads to an unstable equilibrium when x is equal to half the interatomic distance b, as in Figure 3.4c. Stable equilibrium is reached as $x = b$ (Figure 3.4d). In fact, the microscopic positions of the atoms with respect to each other in the crystal do now exactly correspond to the starting point before applying any shear stress, but the crystal has undergone plastic deformation. Consequently, the shear stress must be a periodic function of x with a period of b. We assume a simple sine dependence and write

$$\tau = C \sin\left(\frac{2\pi x}{b}\right), \tag{3.14}$$

where C is the highest value of the shear stress that has to be overcome in order to have the planes glide on top of each other. In other words, when the applied $\tau \geq C$, the solid can be plastically deformed and thus C is equal to the shear stress at the yield point τ_Y. Equation (3.14) can be approximated for small x by replacing the sine with its argument and combined with (3.13) to give

$$C\frac{2\pi x}{b} = \frac{Gx}{a}, \tag{3.15}$$

that is,

$$C = \tau_Y = \frac{Gb}{2\pi a}. \tag{3.16}$$

We can now estimate the order of magnitude for τ_Y. We set $a \approx b$ and $2\pi \approx 10$ and find $\tau_Y \approx 0.1G$. Also, since the modulus of rigidity G has the same order of magnitude as Young's modulus Y (see (3.10)), we are able to state $\tau_Y \approx 0.1G \approx 0.1Y$ and obtain a direct estimate from data such as given in Figure 3.3. It turns out that the mechanism of gliding crystal planes is also responsible for the yield upon tensile stress, so that we can simultaneously estimate that $\sigma_Y \approx \tau_Y \approx 0.1Y$.

But these estimates, however crude, cannot be reconciled with the experimental results. In fact, the measured yield stress is not found to be merely a factor of 10 lower than Young's modulus but several orders of magnitude. For aluminum, for example, Young's modulus is ≈ 70 GPa but its yield stress is only ≈ 30 MPa. What is the reason for this disagreement? It turns out that there is nothing wrong with the calculation for the simple model here; the problem is that we have assumed a perfect defect-free crystal. Even a qualitative understanding of the stress/strain curve in the plastic regime requires the introduction of defects.

3.2.2 Point Defects and Dislocations

Defects, or crystal imperfections, in solids are a wide research subject of their own, and we have to take them into account to some degree, even though we are mostly concerned with perfect crystals. Defects are more than small annoying perturbations of the perfect crystal. In the present context, they are essential for explaining the mechanical properties, and later we will see that they are also indispensable for phenomena like electrical resistance or for the electronic properties of semiconductors. We broadly distinguish between very localized point defects and extended defects such as grain boundaries or dislocations.

A crystal can have different types of point defects and we mention only a few here. There can be an atom missing in the otherwise perfect crystal structure. Such a defect is called a **vacancy**. Atoms of a different kind can be present, either on the original lattice sites instead of the "correct" atoms or in between lattice sites. These defects are called **substitutional** and **interstitial**, respectively. The former play an important role for changing the conductivity of semiconductors; the latter are often used to design alloys with improved mechanical properties (see below).

Dislocations are line-type defects and therefore much more extended. These lines can extend through the whole crystal, or they can have the shape of a loop. For the mechanical properties of a solid, the **edge dislocation** shown in Figure 3.5 is of particular importance. This type of dislocation is caused by one extra sheet of atoms in the crystal. It can move within the **slip plane** as we will explain next.

3.2.3 The Role of Defects in Plastic Deformation

The presence of edge dislocations can explain why the yield stress of a real crystal is much smaller than predicted by (3.16). This is illustrated in Figure 3.6a. As the solid with a dislocation is exposed to shear stress, a plastic yielding can be

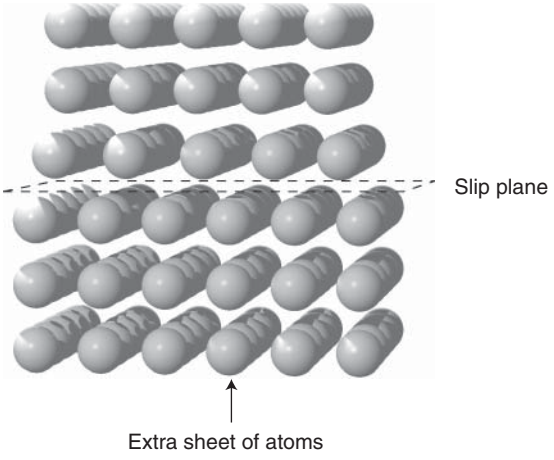


Figure 3.5 An edge dislocation formed by an extra sheet of atoms. The dislocation can move in the slip plane.

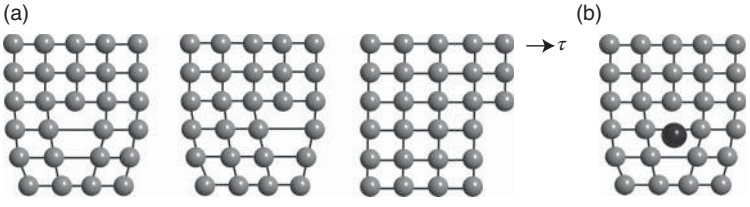


Figure 3.6 (a) Shearing of a solid in the presence of an edge dislocation. The dislocation moves through the solid by breaking only one row of bonds at a time. (b) A point defect can pin a dislocation such that it cannot move.

achieved by moving the dislocation through the crystal. It is immediately evident why this is much easier than the process shown in Figure 3.4: When a dislocation is present, the plastic deformation can proceed by breaking one row of bonds at a time instead of all bonds between two planes of atoms simultaneously. Since edge dislocations are always present in real materials, the observed yield stress is the stress at which dislocations start to move. It is thus far lower than the yield stress from (3.16).

The yield stress of materials can therefore be increased by hindering the movement of dislocations. Frequently, this is achieved by impurities that can "pin" a dislocation as shown in Figure 3.6b. In fact, impurities often gather in the extra space available in dislocations and simultaneously hinder their movement. Impurities are therefore frequently introduced into real materials. Examples are carbon, turning iron into steel, or beryllium that can stop dislocation movement in copper.

The presence of dislocations and defects now allows us to understand the details of the plastic deformation in the stress/strain curve of Figure 3.2 up to the point

of fracture. Once the yield stress is overcome, dislocation-assisted glide sets in. This is the so-called **easy-glide region**. The stress increases only very little for a big strain increase, that is, the curve is rather flat.

The next part of the curve is called the **work hardening region**. Here, the stress/strain curve is considerably steeper. The meaning of the term work hardening becomes obvious when we consider what happens as the stress is released (see Figure 3.2): The material will contract very little as the stress goes to zero. Upon a new application of stress, the material will deform *elastically* until the original stress/strain curve is reached again. But this point is reached at a higher stress than for the original material. This means that the yield stress is higher and hence the term work hardening. The microscopic picture behind the work hardening is that the number of dislocations increases for higher strain values, for reasons not discussed here. At some point, the dislocations hinder each other's movement, and the slope of the stress/strain curve increases. Work hardening can be a useful technique to increase the strength of materials by pre-straining them.

Not only the presence of defects but also the solid's temperature has a major significance for its mechanical properties. At elevated temperature, the role of the entropy in a crystal becomes more important and defects are generated in order to minimize the Gibbs free energy. In addition to this, activation barriers, such as the one needed to move a dislocation across a point defect, can be overcome more easily.

3.3 Fracture

At the end of the stress/strain curve, the material will fracture. Immediately before fracturing, the stress might even decrease. This is due to a phenomenon called **necking**, in which the material narrows somewhere between the points at which the stress is applied. The narrower cross section means that the local stress is even higher than elsewhere. This causes a self-amplification that leads to fracture.

So far, we have only discussed ductile materials that have a stress/strain curve similar to the one shown in Figure 3.2. What happens in brittle materials that do not show any plastic deformation at all, but fracture at the end of the elastic part of the curve? This so-called **brittle fracture** follows a different physical mechanism. It is associated with the presence of fine cracks in the material, perpendicular to the direction of the applied stress. At the end of a crack, the local stress is higher than the average stress in the material. More specifically, it is increased by a factor $\approx 2\sqrt{l/r}$, where l is the length of the crack and r the radius at the end. For a ductile material, such local stress can be relieved by a plastic deformation that work hardens the material in this area. Then, the crack is stopped and cannot proceed further. If such a plastic deformation cannot happen, the crack will propagate through the whole material. Indeed, as the stress is increasing for deeper cracks,

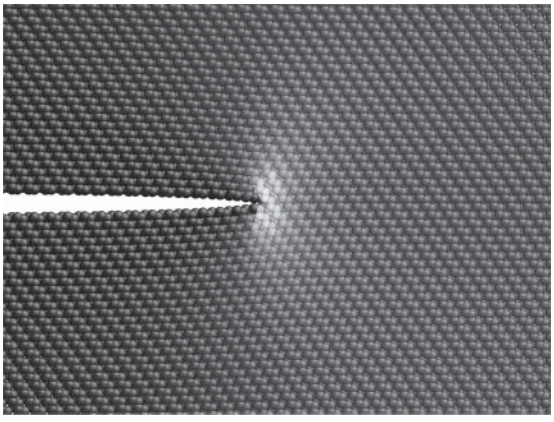


Figure 3.7 Calculated local stress field for a crack along the (1,1,1) plane in silicon. The stress per atom is encoded as grayscale. Bright corresponds to high stress. Image courtesy of James Kermode (www.jrkermode.co.uk).

this is a self-amplifying process that can lead to the spontaneous breaking of very big structures (e.g., entire ocean-going ships).

It is even possible to calculate the local stress field on an atomic scale, and the result of such a simulation is shown in Figure 3.7 for a crack in silicon along the crystal plane with the Miller indices (1,1,1). The image encodes the local stress in the grayscale of the atoms, and it is easy to see what drives the propagation of the crack: The presence of the crack completely relaxes the stress in the crystal on the left-hand side, above, and below the crack. However, it also leads to a strongly increased stress near the tip of the crack, causing fracture there and thus further propagation of the crack.

Again, temperature is important for the behavior of materials. At elevated temperatures, the propagation of dislocations is easier and materials that behave brittle at low temperature can be ductile as the temperature is raised. A prominent example is glass, which is usually brittle, but at high temperature, it is so ductile that its shape can be changed by blowing.

3.4 Further Reading

A discussion of the mechanical properties of solids is found in the general solid-state physics books.

- Kittel, C. (2005) *Introduction to Solid State Physics*, 8th edn, John Wiley & Sons, Inc.
- Myers, H.P. (1990) *Introductory Solid State Physics*, 2nd edn, Taylor & Francis Ltd.
- Rosenberg, H.M. (1988) *The Solid State*, 3rd edn, Oxford University Press.
- Turton, R.J. (2000) *The Physics of Solids*, Oxford University Press.

More detailed information can be found in

- Callister, W.D. Jr. and Rethwisch, D.G. (2009) *Materials Science and Engineering: An Introduction*, 8th edn, John Wiley & Sons, Inc.

3.5 Discussion and Problems

Discussion

1) What typically happens when a crystal is exposed to a small stress?
2) How can an elastic deformation of a crystal be described microscopically, and why would you expect Hooke's law to hold for a small strain?
3) How do the stress/strain curves look for a typical ductile and brittle material?
4) (*)Young's modulus can be estimated from the microscopic force constants between atoms (or the other way round). Why and how?
5) The yield stress of a solid estimated from a simple calculation is often much higher than the observed yield stress. Explain why.
6) Explain the phenomenon of work hardening.

Problems

1) *Elastic constants:* (a) Consider a cube of isotropic material and show that the bulk modulus (3.5) is related to Young's modulus (3.1) and Poisson's ratio (3.6) by

$$K = \frac{Y}{3(1-2\nu)}. \tag{3.17}$$

   (b) What happens when ν reaches its upper limit of 0.5?
2) *Elastic constants:* We have stated that the numerical value of Poisson's ratio is always between +0.5 and −1, but we have proven only the upper limit. Use (3.10) to argue why −1 is the lower boundary for the Poisson ratio.
3) *Elastic constants:* The metals with the highest values of Young's modulus in Figure 3.3 are also those with the highest cohesive energies and melting temperatures (see Figure 4.16a). Are these two aspects of the same thing?

4
Thermal Properties of the Lattice

In this chapter, we discuss some thermal properties of solids such as their heat capacity, thermal conduction, thermal expansion, and melting. For now, we only consider the contribution of the lattice, that is, the effects caused by the motion of the atoms around their equilibrium position. For some thermal effects, the motion of the free electrons in metals can be very significant (e.g., thermal conduction), but we neglect this for now and come back to it in the next two chapters.

4.1
Lattice Vibrations

The atoms in a crystal can vibrate around their equilibrium position. The restoring force can be derived from the interatomic potential, as expressed in the Taylor series (3.11). In most cases, it is sufficient to assume a linear restoring force, considering only the first three terms in the series. This leads to a description of the lattice vibrations as harmonic oscillators and is therefore called the **harmonic approximation**.

4.1.1
A Simple Harmonic Oscillator

When inspecting the interatomic potential in Figure 2.1 and the Taylor series for the potential (3.11), one might be tempted to describe the vibrations of a solid with N atoms simply as $3N$ independent harmonic oscillators. The factor of 3 comes from the three different directions the atoms can oscillate in. While this is clearly much too simple (the oscillators are all coupled to each other), it is surprising how far one gets with this picture.

If the force constant γ is equal to $\phi''(a)$ in (3.11) and x is the displacement from the equilibrium position (for convenience, we set the origin of the coordinate system such that a in (3.11) is zero), the equation of motion is

$$M\frac{d^2x}{dt^2} = -\gamma x, \tag{4.1}$$

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

where M is the mass of the vibrating atom. This leads to a harmonic motion with the frequency[1)]

$$\omega = \sqrt{\frac{\gamma}{M}}. \tag{4.2}$$

The total energy for a one-dimensional harmonic oscillator is the sum of kinetic and potential energies:

$$E = \frac{1}{2}Mv^2 + \frac{1}{2}\gamma x^2. \tag{4.3}$$

Assuming classical motion, we can estimate the amplitude of the vibration by using the **equipartition theorem** of statistical mechanics. This theorem states that every generalized momentum or position coordinate, which appears squared in the Hamilton function (or in the total energy in our case), contributes with a mean energy of $k_B T/2$ to the system. Here, we have two squared coordinates and, therefore, the mean energy of the oscillator in contact with a heat bath is $k_B T$. At the highest displacement, the oscillator has only potential energy and so

$$\frac{1}{2}\gamma x_{\max}^2 = k_B T, \tag{4.4}$$

and thus

$$x_{\max} = \left(\frac{2k_B T}{\gamma}\right)^{1/2}. \tag{4.5}$$

This is usually a small percentage of the lattice spacing.

4.1.2 An Infinite Chain of Atoms

A more realistic model for crystal vibrations is a chain of atoms. This is still very simple but now the oscillators are coupled to each other. If we restrict ourselves to longitudinal vibrations in this chain, the model is only one-dimensional. However, we can learn a lot about vibrations in three-dimensional solids already from this. We start with infinite chains of one and two different atoms per unit cell before passing on to chains of finite length.

4.1.2.1 One Atom Per Unit Cell

Consider a one-dimensional atomic lattice with one atom per unit cell and a lattice constant a. The atoms can move out of their equilibrium position along the direction of the chain, as shown in Figure 4.1a. The atoms at the lattice sites shall be connected to their neighbors with springs of a force constant γ. The equation of motion for atom n is

$$M\frac{d^2 u_n}{dt^2} = -\gamma(u_n - u_{n-1}) + \gamma(u_{n+1} - u_n), \tag{4.6}$$

1) For simplicity, we will often speak merely of "frequency" when "angular frequency" would be the correct term.

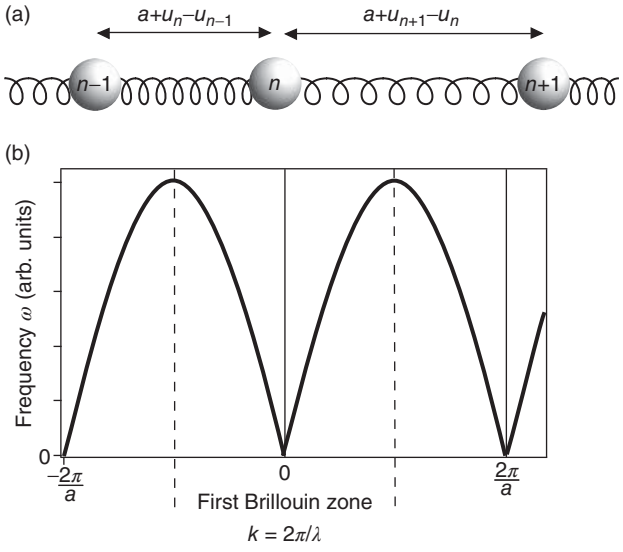


Figure 4.1 (a) One-dimensional chain with one atom per unit cell. (b) Allowed vibrational frequencies ω as a function of the wave vector k.

where u_n is the displacement of the nth atom in the chain (see Figure 4.1a), or

$$M\frac{d^2u_n}{dt^2} = -\gamma[2u_n - u_{n-1} - u_{n+1}]. \tag{4.7}$$

This can be solved by a kind of wave that is only defined on the lattice sites:

$$u_n(t) = ue^{i(kan-\omega t)}, \tag{4.8}$$

where $k = 2\pi/\lambda$ is the one-dimensional wave vector with the wavelength λ and u is the amplitude of the oscillation. Substituting this into the equation of motion gives

$$-M\omega^2 e^{i(kan-\omega t)} = -\gamma\left[2 - e^{-ika} - e^{ika}\right]e^{i(kan-\omega t)} = -2\gamma(1-\cos ka)e^{i(kan-\omega t)}, \tag{4.9}$$

and this has a solution if we choose the $\omega = \omega(k)$ such that

$$\omega(k) = \sqrt{\frac{2\gamma(1-\cos ka)}{M}} = 2\sqrt{\frac{\gamma}{M}}\left|\sin\frac{ka}{2}\right|. \tag{4.10}$$

The resulting $\omega(k)$ is plotted in Figure 4.1b. The solutions given by (4.8) now solve the equation of motion if such an $\omega(k)$ is chosen for a given wave vector k. They describe waves propagating along the chain. What is special about these waves is that they are only defined on the actual lattice sites.

Relations of the type of (4.10), which connect a frequency or energy to a wave vector, are called **dispersion relations**. We will encounter them many more times, for example, in connection with electronic states. A particular solution with $\omega(k)$ is called a **normal mode** of the chain. Note that such a vibration is not localized to one particular atom in the chain. All the atoms move, and they do so with the

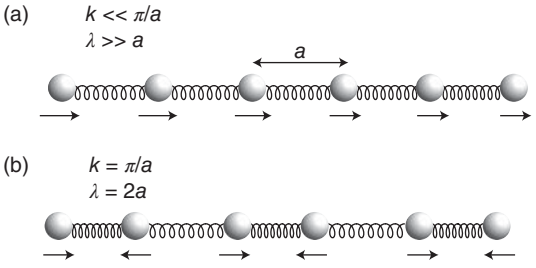


Figure 4.2 Motion of the atoms in the chain for (a) $k \ll \pi/a$ and (b) $k = \pi/a$.

same frequency ω. How do the atoms actually move? We study two important cases, the situation for a very small $k \ll \pi/a$ and that for $k = \pi/a$. All we need for doing this is the relation between the wave vector and the wavelength $k = 2\pi/\lambda$. For a small k, the wavelength of the mode must be much longer than the lattice constant. Therefore, atoms that are close to each other must very nearly move in phase. If we pick a certain instant in time, and the leftmost atom of the chain in Figure 4.2a happens to move to the right, the atoms in its vicinity perform essentially the same motion. In fact, atoms moving in the opposite direction are only found many lattice constants ($\pi/(ak)$) away from the atom under consideration. For modes with a very small k, the particular atomic structure is thus not important, it would be sufficient to view the chain as a macroscopic elastic medium. A small k also allows us to replace the sine in (4.10) by its argument to obtain the linear relation

$$\omega(k) = \sqrt{\frac{\gamma}{M}}ak = vk, \tag{4.11}$$

where v has the dimension of a velocity. If we now apply the usual definitions of the **phase velocity** (ω/k) and **group velocity** ($\partial\omega/\partial k$) for a wave, we see that (4.11) describes a situation in which phase velocity and group velocity are the same (v).[2)] The propagation speed of the waves does, therefore, not depend on their frequency. It turns out that this long wavelength limit corresponds to the propagation of sound waves with v being the **speed of sound**. In fact, the situation is very similar to long-wavelength sound propagation in air, but since the atoms in a solid are much closer packed than in air, the speed of sound is considerably higher.

The limit of short wavelengths is also very instructive. The shortest possible wavelength must be two lattice spacings, such that $\lambda = 2a$ and $k = \pi/a$. If we again assume that the leftmost atom in the chain of Figure 4.2b moves to the right, then the atom two lattice spacings away must perform the same motion. The atom on the neighboring lattice site, on the other hand, is half a wavelength away and must therefore move in the opposite direction. For $k = \pm\pi/a$, the group velocity of the wave is zero since the dispersion curve in Figure 4.1b is flat, meaning that the solutions of (4.7) are standing waves, consistent with the motion in Figure 4.2b (Problem 4.1 asks you to show this formally).

2) For a detailed discussion of the phase velocity and group velocity, see online note on www.philiphofmann.net.

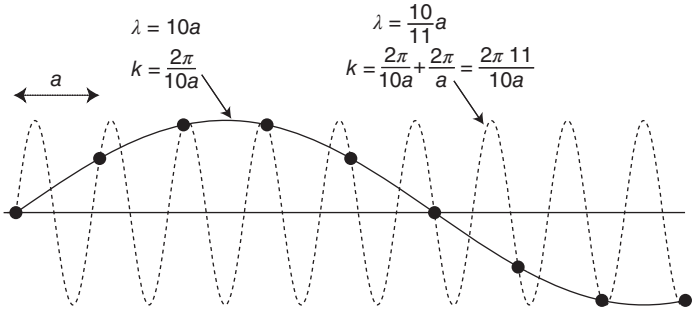


Figure 4.3 Instantaneous position of atoms in a chain for two different wavelengths: $\lambda = 10a$ and $\lambda = (10/11)a$. Note that the wave is transverse for illustrative purposes. Otherwise, we have only considered longitudinal waves in one-dimensional chains.

According to (4.10), $\omega(k)$ is periodic in k with a period of $2\pi/a$. This periodicity corresponds precisely to one reciprocal lattice "vector" in our one-dimensional crystal. If this is so, we could suspect that the actual motion of the atoms is also unaffected if we add a reciprocal lattice vector to k. Amazingly, this is really the case: It is illustrated in Figure 4.3, which shows the instantaneous displacement of the atoms for two waves with wave vectors that differ by a reciprocal lattice vector. The displacement of the atoms is the same even though the two waves differ in wavelength. We see that this is so because the wave (4.8) is only defined on the lattice sites. Another way of viewing this is to note again that the shortest possible wavelength in a lattice of spacing a is $\lambda = 2a$ or $k = \pi/a$, which corresponds to the situation where neighboring atoms move exactly out of phase. Any wave that has an even shorter wavelength can be equivalently described by one with a longer wavelength.

4.1.2.2 The First Brillouin Zone

The most remarkable result of the last section is probably that we have easily managed to describe the motion of all the atoms in an *infinite* chain just by making use of the chain's periodicity. It turned out that dispersion relation $\omega(k)$ and even the motion of the atoms themselves is unaffected if we change the wave vector by multiples of the reciprocal lattice vector $2\pi/a$. Therefore, it is sufficient to know the solution of the equation of motion only in an interval of length $2\pi/a$. One could even argue that an interval of π/a is sufficient. This is due to the left/right symmetry of the chain. It does not matter if the wave travels to the left or to the right, that is, if k is positive or negative.

This does not only show the usefulness of the reciprocal lattice for describing waves in crystals, it also motivates another definition that appears rather formal right now but turns out to be very useful. We call the region between $k = -\pi/a$ and $k = \pi/a$ the **first Brillouin zone** of the lattice. The first Brillouin zone and similar constructions are often said to be placed in **reciprocal space** or **k-space**. The first Brillouin zone is indicated in Figure 4.1b. For a definition of the first Brillouin zone in three dimensions, see Section 4.1.5.

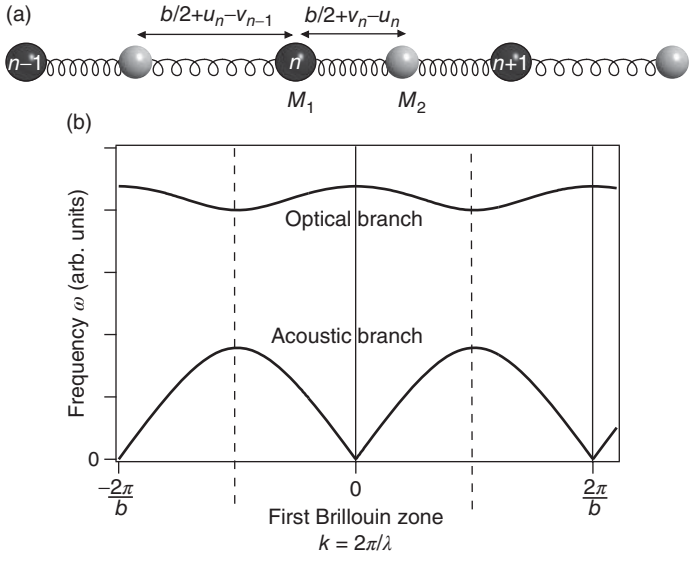


Figure 4.4 (a) One-dimensional chain with two atoms per unit cell. (b) Allowed vibrational frequencies ω as a function of the wave vector k.

4.1.2.3 Two Atoms per Unit Cell

We also discuss the vibrations of a chain with two atoms per unit cell as shown in Figure 4.4a. The calculation is very similar to the case of one atom per unit cell. Now we call the lattice constant b and the length of the reciprocal lattice vector is $2\pi/b$. One of the two atoms in the unit cell is placed at the origin and the other at $b/2$. We write down the forces on each atom in a similar manner as above and obtain two equations of motion, one for each type of atom.

$$M_1 \frac{d^2u_n}{dt^2} = -\gamma[2u_n - v_{n-1} - v_n], \qquad M_2 \frac{d^2v_n}{dt^2} = -\gamma[2v_n - u_n - u_{n+1}], \quad (4.12)$$

where u_n and v_n are the displacements of the first and second atom in the nth unit cell, respectively. This can again be solved by wave-type functions of the form

$$u_n(t) = ue^{i(kbn-\omega t)}, \qquad v_n(t) = ve^{i(kbn-\omega t)}. \quad (4.13)$$

When this is inserted into the equations of motion, we obtain a homogeneous linear system of equations for the amplitudes u and v:

$$-\omega^2 M_1 u = \gamma v(1 + e^{-ikb}) - 2\gamma u, \qquad -\omega^2 M_2 v = \gamma u(e^{ikb} + 1) - 2\gamma v. \quad (4.14)$$

A solution exists if the determinant of the coefficient matrix vanishes, that is,

$$\begin{vmatrix} 2\gamma - \omega^2 M_1 & -\gamma(e^{-ikb} + 1) \\ -\gamma(1 + e^{ikb}) & 2\gamma - \omega^2 M_2 \end{vmatrix} = 0. \quad (4.15)$$

This happens when

$$\omega^2 = \gamma\left(\frac{1}{M_1} + \frac{1}{M_2}\right) \pm \gamma\left[\left(\frac{1}{M_1} + \frac{1}{M_2}\right)^2 - \frac{4}{M_1 M_2}\sin^2\frac{kb}{2}\right]^{1/2}, \quad (4.16)$$

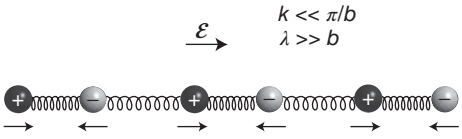


Figure 4.5 Motion of the atoms for $k \approx 0$ in the optical branch. $\mathcal{E}$ represents an external electric field.

and these solutions are shown in Figure 4.4b. Again, the solutions have the periodicity of the reciprocal lattice $2\pi/b$, that is, it is sufficient to know them within the first Brillouin zone. What is new is that we have two branches of solutions. The solution that goes to zero for small k is called the **acoustic branch**. As before, it corresponds to the propagation of sound waves through the crystal.

The solution that has a finite ω at $k = 0$ is called the **optical branch**. It is called like this because of the possibility to couple these vibrations to the oscillating electric field of an electromagnetic wave. To see this, consider the motion of the atoms in the optical branch for $k = 0$. For this wave vector, the two atoms in the unit cell vibrate exactly out of phase, as shown in Figure 4.5 (see Problem 4.2). The phase difference between the vibration of a given atom and the corresponding atom in the neighboring unit cell is zero, and so the wavelength of the mode must be infinite, consistent with $k = 0$. Figure 4.5 illustrates the particular situation of an ionic crystal in which the two different ions in the unit cell carry opposite charges. An electromagnetic field (as indicated by the $\boldsymbol{\mathcal{E}}$-vector) can couple to this motion. For the direction of $\boldsymbol{\mathcal{E}}$ in the figure, the ions will move as indicated by the arrows.

Let us assume that the electric field has a time dependence $\boldsymbol{\mathcal{E}}(t) = \boldsymbol{\mathcal{E}}_0 \exp(i\omega t)$, like the field of an electromagnetic wave. If ω is small, the field will slowly change and the ions will follow from side to side. If ω matches the frequency of the optical branch at $k = 0$, the electromagnetic wave can excite this vibrational mode very efficiently. This situation is discussed in detail in Chapter 9. As we shall see below, typical vibrational frequencies in the optical branch are of the order 10^{13} s^{-1}. Therefore, the exciting radiation must be in the infrared spectral range and the corresponding wavelength is very long compared to the unit cell length b (on the order of 10 μm). This implies that the field moves all ions in phase over a very long distance, and so it is only the $k \approx 0$ mode that can be excited by electromagnetic radiation.

4.1.3 A Finite Chain of Atoms

For describing the properties of real solids, the models discussed so far have a fundamental problem because the chains are infinite. This will, for example, lead to infinite heat capacities. What we really want is a finite but long chain of atoms. This can be done by limiting the length and introducing boundary conditions. For example, one can hold the atoms at the ends fixed. This leads to standing waves in the chain because we have fixed the nodes. Although this approach would not

be wrong, it would be more convenient to start with traveling wave solutions if we are to describe phenomena such as heat transport by lattice vibrations later.

The most convenient boundary conditions solving the problem of what to do at the ends of the chain have been introduced in 1912 by M. Born and T. von Kármán. For a chain with N atoms, these conditions state that

$$u_{N+n}(t) = u_n(t). \tag{4.17}$$

This can be visualized as a finite chain of atoms in which the end is tied to the beginning. Therefore, the conditions are also called **cyclic boundary conditions** or **periodic boundary conditions**. In three dimensions, this simple visualization does not work, but we can think of a crystal of finite size that has identical crystals with identical motions attached to its sides. In this way, we end up with an infinite crystal lattice again but since it is made from finite crystals, we can use it to describe the properties of one of these finite crystals and simultaneously get rid of the crystals' surfaces.

The dispersion relations $\omega(k)$ are not affected by the chain being finite but the periodic boundary conditions restrict the possible k values for the waves in the crystal. First of all, it is clear that the longest possible wavelength for a chain of N atoms with a spacing of a is Na. More precisely, we have to require that

$$e^{ikan} = e^{ika(N+n)}, \tag{4.18}$$

so that

$$e^{ikNa} = 1, \tag{4.19}$$

and for this to be fulfilled, the possible values of k must be

$$k = \frac{2\pi}{aN}m, \tag{4.20}$$

where m is an integer. We have seen that the vibrations are unaffected by adding multiples of $2\pi/a$ (i.e., reciprocal lattice vectors) to k and, therefore, we only get N possible different values for k and hence no more than N different vibrational frequencies ω per dispersion branch. The possible values of k can be chosen to lie in the first Brillouin zone, that is, $-\pi/a \leq k < \pi/a$ (or $0 \leq k < 2\pi/a$ if only positive values of k are desired). The allowed k values and corresponding frequencies $\omega(k)$ for a finite chain are illustrated in Figure 4.6. Note that for a macroscopic solid, the number of atoms in any direction is very large. Therefore, the distances between the allowed k points are very small and the discrete vibrational frequencies closely resemble the continuum of states for the infinite chain (see Problem 4.3).

If we take N free atoms that can only move in one dimension, each atom has one degree of freedom. The total number of degrees of freedom is conserved when we put the atoms into a chain linked with springs, since we also get N different normal modes, one for each allowed k vector in the first Brillouin zone. We can make the same argument for a chain with two atoms per unit cell. If we have N unit cells, we again get N different k values. But now there are two vibrational modes for each k: the acoustic and the optical mode. We thus obtain $2N$ normal modes, and

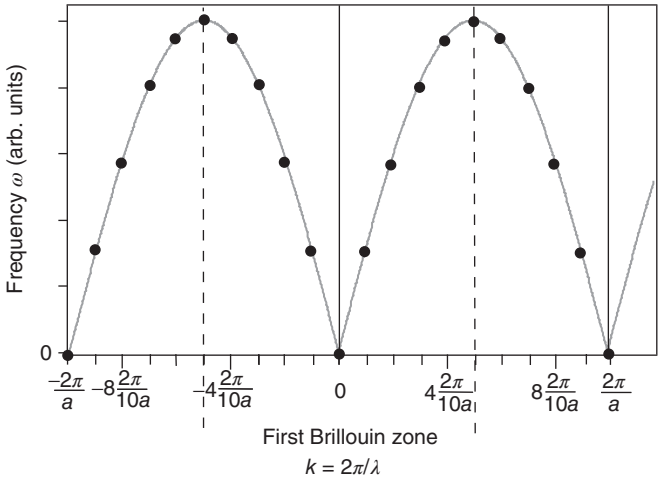


Figure 4.6 Vibrational spectrum for a finite chain of atoms with a length of 10 unit cells and a unit cell length of a. The light gray line represents the vibrational spectrum for an infinite chain of lattice constant a. The black markers represent the vibrational frequencies that are actually allowed.

again the number of degrees of freedom is conserved. Indeed, it is very useful to view the normal modes of a chain as the fundamental possible vibrations, each with a frequency $\omega(k)$. When doing this, we only have to emphasize again that the normal modes involve the vibration of all the atoms in the chain, all with the same ω.

4.1.4 Quantized Vibrations, Phonons

So far, we have neglected the quantized character of the lattice vibrations but taking this into account turns out to be essential for the correct description of many properties, for example, the heat capacity.

For one harmonic oscillator, as described by (4.1), the quantization is very simple. The frequency remains $\omega = (\gamma/M)^{1/2}$ and the quantized energy levels are given by

$$E_l = \left(l + \frac{1}{2}\right)\hbar\omega, \tag{4.21}$$

with $l = 0, 1, 2, \ldots$. These energy levels are displayed in Figure 4.7a.

For a chain with N unit cells and one atom per unit cell, the quantization can performed in exactly the same way if we start from the N normal modes with frequencies $\omega(k)$ and $k = 2\pi m/aN$. For this system, the quantized energy levels become

$$E_l(k) = \left(l + \frac{1}{2}\right)\hbar\omega(k). \tag{4.22}$$

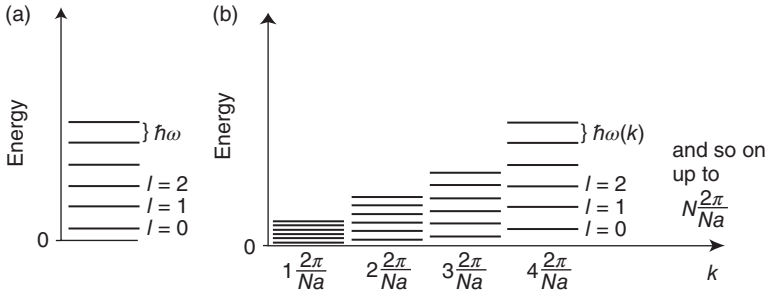


Figure 4.7 (a) Energy level diagram for one harmonic oscillator. (b) Energy level diagram for a chain of atoms with one atom per unit cell and a length of N unit cells.

The first few of these energy levels are shown in Figure 4.7b. If we have more than one atom per unit cell, the equation can be generalized by adding an index to ω that marks the branch of the dispersion (acoustic or optical).

The notation in (4.22) lends itself to an alternative interpretation of k. So far, we have viewed k as the one-dimensional wave vector. But here it becomes apparent that k also takes the role of a quantum number, just as l. k takes only discrete values and can be used to "label" different normal modes. The combination of k and l describes one vibrational excitation of the chain, the normal mode k that is excited to the level l. The interpretation of k as a quantum number has a lot to do with the symmetry of the system. In atoms, we have spherical symmetry that gives rise to the quantum numbers l and m. In solids, we have translational symmetry and the appropriate quantum number is k.

In the quantum mechanical picture of the chain, normal modes can thus be excited in discrete energy quanta of $\hbar$ in front of $\omega(k)$. These excitation are called **phonons** in analogy to photons, the quantized excitations of the electromagnetic field,[3] and relations such as (4.10) and (4.16) are often called **phonon dispersion relations**. Depending on the type of experiment, photons can have wave character as well as particle character and the same is true for phonons. So far, our description mostly emphasized the wave character but if we want to describe properties like thermal conductivity, we need to think of phonons as "particles" that are generated at the hot end of some sample and conducted to the cold end. As in the case of photons, the wave and particle character can be reconciled if we describe the "particle" as a superposition of waves, which is localized in a certain volume of space.[4] An additional similarity between phonons and photons is that both are bosons and therefore not subject to the Pauli exclusion principle. The quantization of the excitation energies and the Bose–Einstein statistics for phonons will become important when we evaluate the heat capacity of solids.

3) The similarity is quite far-reaching. Even mathematically (in quantum field theory), photons can be viewed as excitations of quantum mechanical harmonic oscillators.

4) This is also discussed in the online note on phase velocity and group velocity on www.philiphofmann.net.

The concept of phonons also permits an alternative view on the excitation of optical vibrations by light (see Figures 4.4 and 4.5). Such an excitation involves the creation of a phonon and the annihilation of a photon. For this to be allowed, the phonon and photon must have the same energy and wave vector. The energy of the photon $h\nu$ must be in the infrared regime to match $\hbar\omega(k)$. The wave vector of the photon has to be $k = \omega(k)/c$, where c is the speed of light. As $\omega(k)$ is quite small and c is very high, the photon's k is extremely small. This implies that only phonons with $k \approx 0$ can be excited by light (convince yourself of this by inserting approximate numbers!).

4.1.5 Three-Dimensional Solids

Our discussion of atomic chains already contains most of the important physics for the vibrational properties of three-dimensional crystals. Here we briefly generalize the discussion to three dimensions, not least because we wish to establish a link to measured quantities of real solids.

4.1.5.1 Generalization to Three Dimensions

The concepts discussed for atomic chains are easily generalized to three dimensions. We do not go into much detail because little additional physical insight is gained and the equations get quite messy because one has to keep track of many indices. The wave-type ansatz (4.13) also solves the equation of motion for three-dimensional solids. However, the notation has to be more complex in order to accommodate not only more atoms per unit cell but also more directions of motion. For instance, for a three-dimensional solid with two atoms per unit cell (i.e., a basis containing two atoms), the matrix corresponding to (4.15) will be a 6×6 matrix, giving six vibrational frequencies ω for every value of k. There will be three acoustic branches, one with longitudinal polarization as in one dimension and two with transverse polarization. Similarly, there will also be three optical branches. If the crystal has a basis containing only one atom, there will only be three acoustic branches.

In three dimensions, the one-dimensional k turns into a true wave vector $\mathbf{k}$ with three components, but it does of course retain its additional interpretation as a quantum number. For a simple cubic crystal with a lattice spacing a and N atoms in every direction, the generalization of the periodic boundary conditions (4.20) gives

$$\mathbf{k} = (k_x, k_y, k_z) = \frac{2\pi}{aN}(n_x, n_y, n_z) = \left(\frac{n_x 2\pi}{L}, \frac{n_y 2\pi}{L}\frac{n_z 2\pi}{L}\right), \tag{4.23}$$

with n_x, n_y, n_z being integers and L the macroscopic side length of the crystal (the restriction to a macroscopic cube makes life easier without any loss of generality).

As in one dimension, it is sufficient to describe the vibrational states only within the **first Brillouin zone**. In three dimensions, the first Brillouin zone is defined

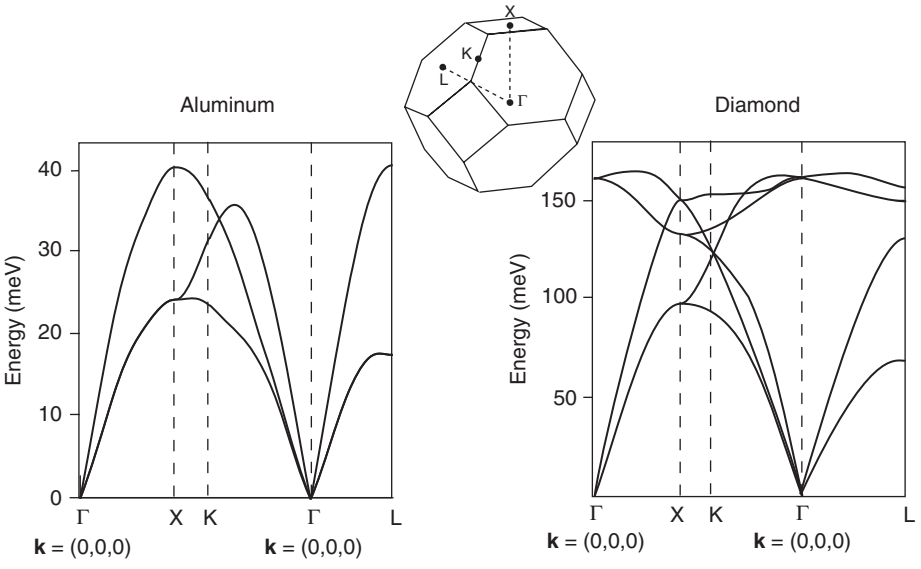


Figure 4.8 Phonon dispersion in Al and diamond along several directions in reciprocal space. The inset shows the first Brillouin zone, which has the same shape for both materials. Reproduced from Grabowski, Hikkel and Neugebauer (2007), Mounet and Marzari (2005).

as all points that are closer to a given reciprocal lattice point (the origin) than to any other. We recognize that this definition corresponds exactly to that of the Wigner–Seitz cell in real space: The first Brillouin zone is the Wigner–Seitz cell of the reciprocal lattice. The geometrical shape of the first Brillouin zone can be quite complicated. For a face-centered cubic (fcc) crystal, it is the truncated octahedron that is shown in the inset of Figure 4.8. In the Brillouin zone and on the Brillouin zone boundary, points of high symmetry are abbreviated by certain letters. The letter Γ always stands for the center of the Brillouin zone, that is, for $\mathbf{k} = (0, 0, 0)$, and the zone center is often referred to as the Γ point.

Figure 4.8 shows the phonon dispersion curves for aluminum and diamond. In both cases, we can clearly identify the acoustic phonon branch with a linear dispersion near the Γ point. For diamond, there are also optical phonons, that is, phonons with a finite energy at Γ but these are not found for Al. The reason for this is simple: Both materials have an fcc Bravais lattice but Al can be described with only one atom as basis, whereas two atoms are needed for diamond.

4.1.5.2 Estimate of the Vibrational Frequencies from the Elastic Constants

We have previously used the idea of a harmonic potential between the atoms in order to explain the elastic deformation of solids and the linear relation used to define Young's modulus. We can now relate Young's modulus to the force constant γ that appears in the description of atomic vibrations. Consider a simple cubic crystal with a lattice constant a and imagine that we cut a rod from this crystal with a side length of just one lattice constant, as shown in Figure 4.9. When we

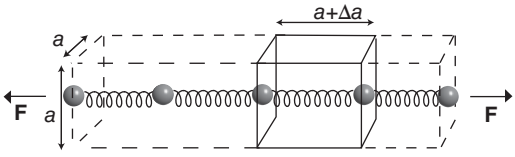


Figure 4.9 Obtaining the interatomic force constant from Young's modulus for a simple cubic solid.

pull on this rod, the stress is

$$\sigma = \frac{F}{a^2}. \tag{4.24}$$

For simplicity, we consider only one unit cell, as shown by the solid lines in Figure 4.9. Upon applying the stress, the unit cell expands by Δa and since only one atomic spring is expanded, we can write

$$F = \gamma \Delta a, \tag{4.25}$$

so that

$$\sigma = \frac{\gamma \Delta a}{a^2}. \tag{4.26}$$

The strain for one unit cell is simply $\epsilon = \Delta a/a$ and

$$Y = \frac{\sigma}{\epsilon} = \frac{\gamma \Delta a}{a^2} \frac{a}{\Delta a} = \frac{\gamma}{a}. \tag{4.27}$$

From Young's modulus, we can therefore estimate the interatomic force constant and the vibrational frequencies corresponding to this force constant. The result can be compared to the experimental values determined by other techniques. Table 4.1 shows such a comparison. We estimate the atomic force constant for diamond and lead from (4.27), using Young's modulus and the nearest neighbor distance. Then, we assume that the highest vibrational frequency is $2\sqrt{\gamma/M}$, as in a one-dimensional chain. While this estimate is admittedly rather crude, ignoring the true three-dimensional nature of the problem and the existence of optical phonons, it gives the right order of magnitude. We also see that diamond, which has light atoms and strong bonds, has much higher vibrational frequencies than lead, which has heavy atoms and weak bonds.

Table 4.1 Comparison between vibrational frequencies estimated from Young's modulus (ω_{calc}) and the experimental result (ω_{measr}) for diamond and lead.

	Diamond	**Pb**
Mass	12 u	207 u
Nearest neighbor distance	1.55 Å	3.50 Å
Young's modulus	950 GPa	15 GPa
ω_{calc}	9×10^{13} Hz	4×10^{12} Hz
ω_{measr}	2×10^{14} Hz	1×10^{13} Hz

4.2 Heat Capacity of the Lattice

Historically, understanding the heat capacity of solids was one of the biggest early successes of quantum theory. In the beginning of the last century, the situation was extremely puzzling. Classical statistical mechanics explained the heat capacity of insulators at room temperature fairly well, but it failed for lower temperatures, and it totally failed for metals. Metals were expected to have a much higher heat capacity than insulators because of the many free electrons, but it turned out that a metal's heat capacity at room temperature is similar to that of an insulator, as if the electrons were not there. We shall see later why this is so, and we focus on the lattice now.

We ignore the difference between heat capacities at constant volume and at constant pressure. For solids, this difference is usually quite small but not entirely negligible. Experimentalists usually like to measure the heat capacity at constant pressure, for example, at ambient pressure. Theorists, on the other hand, prefer calculations at constant volume because otherwise all the quantum mechanical eigenvalues have to be recalculated for every volume.

4.2.1 Classical Theory and Experimental Results

Classically, we can use two different approaches to calculate the heat capacity of a solid. The worrying thing is that they do not give the same answer. Classical thermodynamics does not tell us anything about the value of the heat capacity at finite temperature, but it can be shown that it should vanish for zero temperature. The argument leading to this is based on very few general principles, notably on the requirement of a finite entropy at zero temperature.

Classical statistical mechanics, on the other hand, gives us the possibility to calculate a numerical value for the heat capacity of a solid via the equipartition theorem: For a one-dimensional harmonic oscillator in contact to a heat bath, the mean energy is $\langle E\rangle = k_B T$. For a three-dimensional oscillator, it must be $\langle E\rangle = 3k_B T$. This means that the heat capacity for a "solid" containing one atom is $\partial\langle E\rangle/\partial T = 3k_B$ and for 1 mol of atoms it is $3k_B N_A = 3R = 24.9\ \mathrm{JK}^{-1}$ *independent of the temperature and the material*.[5] This result is called the **rule of Dulong–Petit**. Table 4.2 shows the heat capacities for a number of solids. The agreement with the Dulong–Petit value is rather good at room temperature but less good at the boiling point of nitrogen where the heat capacity is generally smaller.

So, there already is a conflict between different types of classical theories. The Dulong–Petit law predicts a temperature-independent heat capacity, whereas the heat capacity has to vanish at zero temperature according to classical thermodynamics. A vanishing heat capacity at zero temperature is also supported by

5) While we have seen that the atoms do not behave as independent oscillators, we also know that the number of independent normal modes is still three times the number of atoms.

Table 4.2 Molar heat capacity of different solids at the boiling point of nitrogen and at room temperature compared to the Dulong–Petit value of 24.9 JK^{-1}.

Material	77 (K)	273 (K)
Cu	12.5	24.3
Al	9.1	23.8
Au	19.1	25.2
Pb	23.6	26.7
Fe	8.1	24.8
Diamond	0.1	5.2

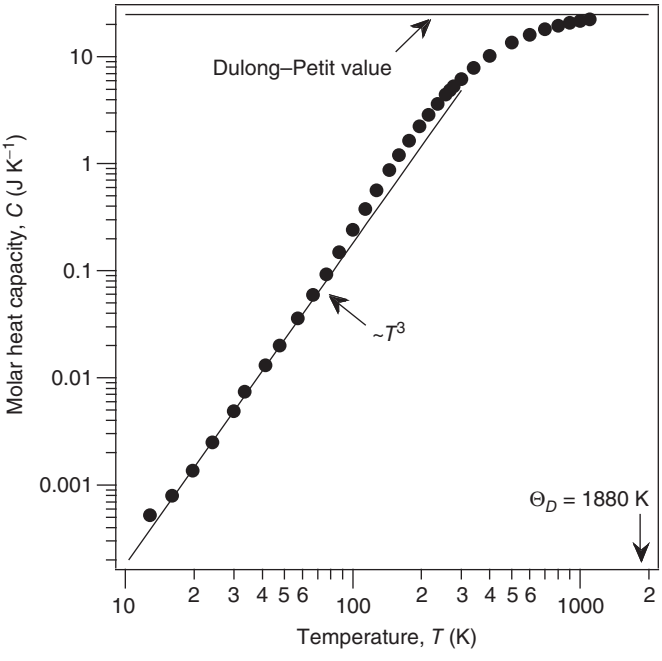


Figure 4.10 Temperature-dependent heat capacity of diamond. Data from Desnoyers and Morrison (1958), Victor (1903).

experimental results. Already Table 4.2 points in this direction. Another example is shown in Figure 4.10, which shows the heat capacity of diamond as a function of temperature. At high temperatures, the heat capacity approaches the Dulong–Petit value but at lower temperatures it drops to zero. The figure is plotted using a double logarithmic scale. The low-temperature limit is a line in this plot, suggesting a power law behavior. From the slope, we can directly read that $C \propto T^3$. A microscopic theory of the heat capacity should be able to reproduce this behavior.

4.2.2 Einstein Model

The breakthrough to understanding the temperature-dependent heat capacity of solids was made by A. Einstein. His idea was to approach the problem using quantum mechanics to describe the oscillators in the solid. The calculation starts out by assuming that the solid's vibrations are represented by independent harmonic oscillators that all have the same frequency, the Einstein frequency ω_E, so that their energy levels are

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega_E. \tag{4.28}$$

We are interested in the mean energy for $3N_A$ of these oscillators per mole of atoms, which are in contact to a heat bath. This is given by $3N_A$ times the mean energy for one oscillator:

$$\langle E\rangle = 3N_A\left(\langle n\rangle + \frac{1}{2}\right)\hbar\omega_E. \tag{4.29}$$

The mean quantum (or phonon) number $\langle n\rangle$ can be found using the Bose–Einstein distribution since lattice vibrations are of bosonic character, that is, there is no limit on the number of quanta n per state.

$$\langle n\rangle = \frac{1}{e^{\hbar\omega_E/k_BT} - 1}. \tag{4.30}$$

The resulting mean energy for $3N_A$ oscillators is therefore

$$\langle E\rangle = 3N_A\left(\frac{1}{e^{\hbar\omega_E/k_BT} - 1} + \frac{1}{2}\right)\hbar\omega_E. \tag{4.31}$$

The heat capacity is found by differentiation:

$$C = \frac{\partial\langle E\rangle}{\partial T} = 3R\left(\frac{\hbar\omega_E}{k_BT}\right)^2 \frac{e^{\hbar\omega_E/k_BT}}{(e^{\hbar\omega_E/k_BT} - 1)^2}. \tag{4.32}$$

The result of this calculation is shown in Figure 4.11 together with a curve representing typical experimental results similar to those in Figure 4.10. In the high-temperature limit, the Einstein model correctly reproduces the Dulong–Petit value. "High temperature" means that the temperature must be at least as high as the **Einstein temperature** which is the temperature corresponding to the vibrational frequency of the oscillators, that is, $\Theta_E = \hbar\omega_E/k_B$. The heat capacity also drops to zero at lower temperatures, in agreement with the experimental data. The only problem is that it drops too quickly to zero. At low temperatures it shows an exponential behavior, whereas the experiment shows a power law behavior $C \propto T^3$. On the linear temperature scale, this does not show up too clearly but on the double log scale, the problem is evident.

Can we understand this behavior in simple terms? In the high-temperature limit, the thermal energy is much greater than the spacing between the energy levels $\hbar\omega_E$ and the quantized nature of the problem becomes insignificant. This is

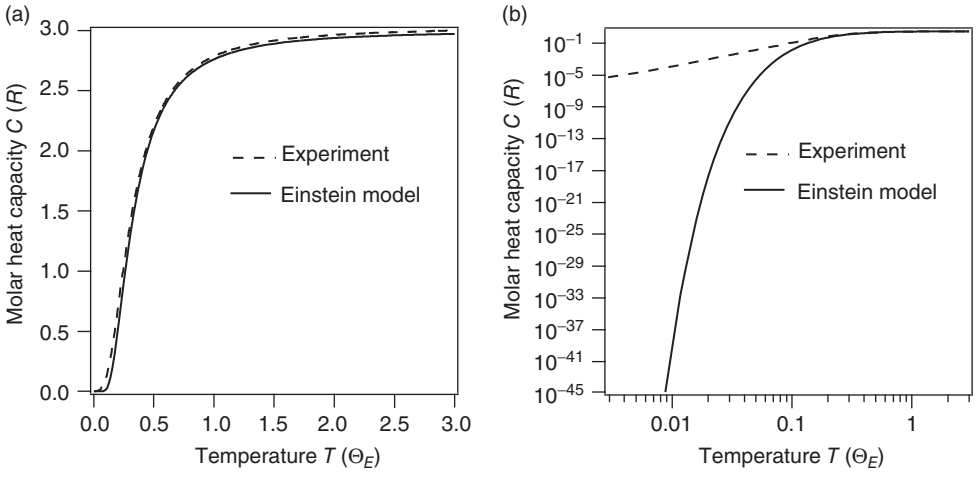


Figure 4.11 Temperature-dependent heat capacity in the Einstein model compared to a typical experimental result for an insulator (a) linear scale (b) log–log scale.

why we recover the Dulong–Petit value for the specific heat. The only condition is that the temperature must be higher than the Einstein temperature. This result will clearly still hold for a more complicated model with many different vibrational frequencies, as long as the temperature is higher than the temperature corresponding to the highest vibrational energy. At sufficiently low temperatures, almost all the oscillators are in their ground state. If the temperature is raised just a little, by much less than $\hbar\omega_E$, nothing will change and the heat capacity is essentially zero. In fact, the probability of occupation of the first excited level of the oscillators follows an exponential behavior:

$$p_1 \propto e^{-\hbar\omega_E/k_B T}, \tag{4.33}$$

which is the low-temperature limit of (4.30). Therefore, the exponential decrease of the heat capacity when cooling an Einstein solid originates from the fact that the oscillators are "frozen" into their ground state.

4.2.3 Debye Model

We have seen that the key problem in the Einstein model is the low-temperature heat capacity. It falls off too quickly because all the Einstein oscillators get "frozen out" below Θ_E when there is not enough thermal energy available to supply the $\hbar\omega_E$ required to excite them out of their ground state. The problem cannot be solved by choosing a lower ω_E because this shifts the transition to the Dulong–Petit regime to lower temperatures, too.

P. Debye noticed that the problem can be cured by using a more realistic model for the lattice vibrations. We have already seen that every solid has an acoustic phonon branch and this gives rise to quantum mechanical oscillators with

very small energy level spacings near $k = 0$: The wave vector for the first oscillator shown in Figure 4.7b is $k = 2\pi/aN$. Since N is very large for a macroscopic solid, k is very small and so is $\omega(k)$ (see Problem 4.3). These oscillators can thus be excited even at very low temperatures, and we avoid the Einstein model's problem of "freezing out" all the vibrations.

The excitations with the lowest energies near $k = 0$ are the sound waves for which the dispersion of the acoustic branch (4.10) is approximated by the linear dispersion (4.11). The basic assumption in the Debye model is now that this dispersion holds for all values of k. This is clearly inaccurate for the excitations at higher k, as we have seen in Figure 4.1b. It is totally incorrect for a unit cell containing more than one atom because it ignores the existence of the optical branches (see Figure 4.4). However, these modes are not excited at low temperatures anyway. At high temperatures, they may be excited but this does not matter so much. From the Einstein model, we have seen that the high-temperature heat capacity approaches the classical value, independent of the actual oscillator frequencies. For low temperatures, the Debye assumption is appropriate and it leads to good results as we shall see now.

We need to calculate the mean thermal energy for a set of oscillators with frequencies given by (4.11). For one oscillator with frequency ω, we know that the result is

$$\langle E\rangle = \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1} \tag{4.34}$$

(see (4.31)). The zero point energy $\hbar\omega/2$ is neglected right away as it does not contribute to the heat capacity. For a three-dimensional solid, it is now tempting to write the mean energy for all oscillators as

$$\langle E\rangle = 3\int_0^{\omega_D} \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1} d\omega, \tag{4.35}$$

where ω_D is the highest phonon frequency in the material. The factor of 3 stems from the fact that there are three possibilities for the wave polarization for a given ω: two different transverse and one longitudinal polarization. Here, we assume that all three waves follow the dispersion $\omega = v|\mathbf{k}|$.

But there are two problems: The first is that there may be more oscillators in, say, the frequency interval $\omega_1 + d\omega$ than in $\omega_2 + d\omega$. This must be included by a weighting factor $g(\omega)$ in the integral. This factor is called the **density of states**, for obvious reasons. The second problem is that we have to establish the upper limit ω_D for the integration; in other words, we have

$$\langle E\rangle = 3\int_0^{\omega_D} \frac{g(\omega)\hbar\omega}{e^{\hbar\omega/k_B T} - 1} d\omega \tag{4.36}$$

and we are looking for $g(\omega)$ and ω_D.

We start by calculating the density of states $g(\omega)$ for a three-dimensional solid. It is not sufficient to do this in one dimension because $g(\omega)$ depends on the dimensionality of the problem, and we want to explain the experimental data for real solids. The density of states $g(\omega)$ is the number of states in a small frequency interval $d\omega$ around ω. Therefore, the strategy to calculate it is to first figure out the total

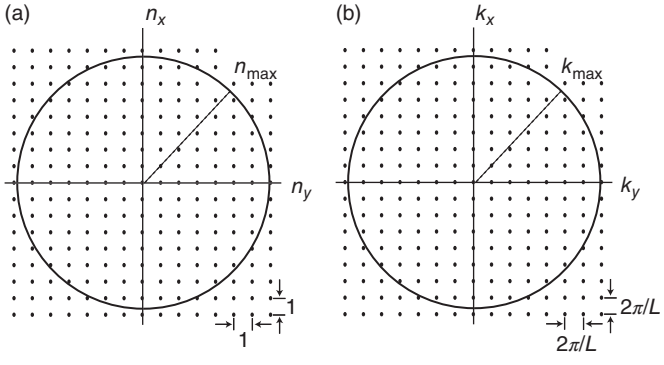


Figure 4.12 (a) Points of integers (n_x, n_y, n_z) that represent the allowed vibrational states and (b) the wave vectors (k_x, k_y, k_z) corresponding to (n_x, n_y, n_z). The sketch is a two-dimensional cut for $n_z = 0$ or $k_z = 0$ and the circle represents a cut through a sphere for a certain highest value of n_{max} or k_{max}, such that a total of N states is enclosed.

number of states below a certain frequency ω. We call this $N(\omega)$ and obtain $g(\omega)$ by differentiating $N(\omega)$ with respect to ω.

We consider a cube of solid with a macroscopic side length L and use the periodic boundary conditions (4.23). We can think of each vibrational state as being defined by a triple of ks, (k_x, k_y, k_z) or, equivalently, by a triple of ns, (n_x, n_y, n_z).[6] How many possible states N do we get for a given highest k_{max} or n_{max}? If n_{max} is large, this comes down to a simple geometrical problem as illustrated in Figure 4.12: We want to know how many states lie within a sphere of radius n_{max} or, alternatively, a sphere with radius k_{max}. This is given by the volume of the sphere :

$$N = \frac{4}{3}\pi n_{max}^3, \tag{4.37}$$

or expressed in terms of k_{max}:

$$N = \frac{4}{3}\pi\left(\frac{Lk_{max}}{2\pi}\right)^3. \tag{4.38}$$

To calculate the density of states, we need to express N as a function of frequency ω and this can be done by using the dispersion $\omega(k) = vk$:

$$N(\omega) = \frac{4}{3}\pi\left(\frac{L\omega}{2\pi v}\right)^3 = \frac{V}{6\pi^2 v^3}\omega^3, \tag{4.39}$$

where $V = L^3$ is the volume of the crystal. From this, we can get the density of states by differentiation:

$$g(\omega) = \frac{dN}{d\omega} = \frac{\omega^2 V}{2\pi^2 v^3}. \tag{4.40}$$

6) Actually, every such triple would characterize three different vibrational states, but we have already taken care of this by the factor of 3 in (4.36).

Now we have to address the question of the upper integral limit ω_D for (4.36). Whatever the nature of the excitations, the limit of the integral must be chosen such that we recover the correct number of normal modes, so for N atoms in the solid, we must have

$$3N = 3\int_0^{\omega_D} g(\omega)d\omega. \tag{4.41}$$

Using (4.40) and performing the integration results in

$$\omega_D^3 = 6\pi^2 \frac{N}{V} v^3. \tag{4.42}$$

ω_D is called the **Debye frequency** and the corresponding temperature $\Theta_D = \hbar\omega_D/k_B$ is called the **Debye temperature**. With this (4.36) becomes

$$\langle E\rangle = 3\int_0^{\omega_D} \frac{\omega^2 V}{2\pi^2 v^3} \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1} d\omega = \frac{3V\hbar}{2\pi^2 v^3} \int_0^{\omega_D} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1} d\omega, \tag{4.43}$$

and with the substitution $x = \hbar\omega/k_B T$ and $x_D = \hbar\omega_D/k_B T$ we obtain

$$\langle E\rangle = \frac{3Vk_B^4 T^4}{2\pi^2 v^3 \hbar^3} \int_0^{x_D} \frac{x^3}{e^x - 1} dx = 9Nk_B T \left(\frac{T}{\Theta_D}\right)^3 \int_0^{x_D} \frac{x^3}{e^x - 1} dx. \tag{4.44}$$

From this, the heat capacity of the solid can be determined by differentiation with respect to temperature. Instead of writing down an expression for the heat capacity, we focus on the low- and high-temperature limits.

For high temperatures, x in (4.44) is small and the exponential function in the integral can be approximated by $1 + x$. The integral is then merely over x^2 and the resulting energy is $\langle E\rangle = 3Nk_B T$. For 1 mol of atoms, this is equal to $3RT$, that is, it leads to the Dulong–Petit result. This was of course expected from the Einstein model: The Dulong–Petit result is always reached at sufficiently high temperatures.

The less obvious limit is that for low temperatures. Here, x is large and we can make the approximation to carry out the integration to infinity instead of x_D. Then, the integral has a value of $\pi^4/15$ and after differentiation, we find the heat capacity to be

$$C = \frac{12\pi^4}{5} Nk_B \left(\frac{T}{\Theta_D}\right)^3. \tag{4.45}$$

This is the Debye T^3 law that fits the experimental data far better than the exponential behavior of the Einstein model (see Figure 4.10). The reason that the Debye model works so well at low temperatures has been mentioned in the preceding text: It provides a good description of the vibrational modes at low energies and long wavelengths while it is inaccurate for higher energies. But at low temperatures, only the low-energy modes are excited anyway, and therefore it works well.

The Debye temperature Θ_D plays a similar role as the Einstein temperature in the Einstein model: It sets a temperature scale. Only for temperatures sufficiently above the Debye temperature, the Dulong–Petit law is obeyed. Moreover, the Debye frequency has been defined as the highest vibrational frequency of the

Table 4.3 Debye temperatures and frequencies for selected materials.

Material	Θ_D(K)	ω_D(Hz)
Pb	105	1.37×10^{13}
Cu	343	4.49×10^{13}
Si	645	8.45×10^{13}
Diamond	1860	2.44×10^{14}

material, assuming a linear dispersion, and it is very often used as a measure of the maximum frequency or energy associated with vibrations in a given material. The Debye temperatures and frequencies for a number of selected materials are given in Table 4.3. The Debye temperatures follow a tendency that is consistent with intuition: Heavy atoms and weak (i.e., metallic) bonds give rise to lower vibrational frequencies than light atoms with strong (i.e., covalent) bonds. The standard examples are again lead and diamond, and we find that ω_D's order of magnitude is consistent with the highest measured vibrational frequencies in Table 4.1.

4.3 Thermal Conductivity

In this section, we address the transport of heat through a crystal by lattice vibrations (phonons), that is, the thermal conductivity κ_p. Daily experience tells us that metals are usually much better thermal conductors than insulators. Therefore, the contributions of the free electrons to the thermal conductivity could be thought to be much more important than the lattice contribution. This, however, is not always the case. A classic example is the insulator diamond that has one of the highest thermal conductivities of all materials at room temperature. It may not be close to daily experience, but it would not be good to make teaspoons out of diamond. We will discuss the metals' free electron contribution to the thermal conductivity κ_e later. Fortunately, the two contributions just add. The total thermal conductivity is the sum of the lattice and the electronic thermal conductivity κ_p and κ_e:

$$\kappa = \kappa_p + \kappa_e. \tag{4.46}$$

Before developing a model for κ_p, we have to state more precisely what we mean by thermal conductivity. Suppose we have a rod with a cross-sectional area A. One end of the rod should be in a heat bath with temperature T, while the other end is constantly heated with a power $\partial Q/\partial t$. Once dynamic equilibrium is established after some time, we measure the temperature difference ΔT between two points in the middle of the rod, which are separated by Δx. The thermal conductivity κ is then defined as

$$\kappa = \frac{1}{A}\frac{\partial Q}{\partial t}\frac{\Delta x}{\Delta T}. \tag{4.47}$$

Note the similarity of this expression with the usual definition of the electrical conductivity. Here, $\Delta T/\Delta x$ plays the role of the electric field and $(1/A)(\partial Q/\partial t)$ the role of the current density.

It is not straightforward to describe thermal conduction using the wave picture of phonons we have discussed so far. If we consider a particular vibrational normal mode $\omega(k)$ for a fixed k, this may be a traveling wave, but the wave amplitude is the same over the entire crystal. What we need, however, are vibrational excitations that can be generated at the "hot" end of the solid and which then propagate to the "cold" end, that is, we need to describe the phonons as particles. This can be achieved by the superposition of normal modes to generate wave packets that then travel through the crystal with a certain group velocity.[7)]

In this sense, lattice vibrations can be viewed as a type of particles that travel through the solid. Often, such wave packets are also called "phonons." Indeed, it turns out that the thermal conductivity of the solid can be described by assuming that these phonons are a gas of particles bouncing through the solid. For describing the thermal conductivity due to the gas of phonons, we adopt a result from kinetic gas theory:

$$\kappa_p = \frac{1}{3} c \lambda_p v_p, \tag{4.48}$$

where c is the heat capacity of the solid per unit volume, λ_p is the mean free path of the phonons, and v_p is the phonon speed. In order to evaluate κ_p, we can take v_p to be the speed of sound v, and for the heat capacity, we can take the results of the previous section. The only quantity that is unknown is the mean free path of the phonons. We discuss it in the following.

As the phonons propagate through the crystal, they can be scattered by imperfections of the lattice, such as point defects, dislocations, and the like. It is possible to grow crystals of such high perfection that these scattering effects become unimportant. Then, the scattering of the phonons at the sample boundaries can be observed. Scattering from crystal imperfection is the dominant mechanism at low temperatures.

At high temperatures, another scattering process becomes important: The number of phonons increases and phonons can be scattered from other phonons. This causes the mean free path λ_p, and thereby also κ_p, to decrease. At low temperatures, on the other hand, the heat capacity in (4.48) decreases, causing κ_p to decrease as well. This means that there must be an intermediate temperature at which κ_p reaches a maximum. This can typically be found at about 10% of the Debye temperature. An example is the thermal conductivity of silicon given in Figure 4.13. Note the very strong temperature variation of κ_p.

Some numerical values for the thermal conductivity of solids at room temperature are given in Table 4.4. Evidently, not only metals show high thermal conductivity but also some insulators, especially diamond. Diamond has a nearly perfect crystal structure such that defect scattering of phonons is unimportant.

7) The formation of wave packets is illustrate in the online note on phase velocity and group velocity on www.philiphofmann.net. Problem 6.12 is also concerned with this subject for electron waves.

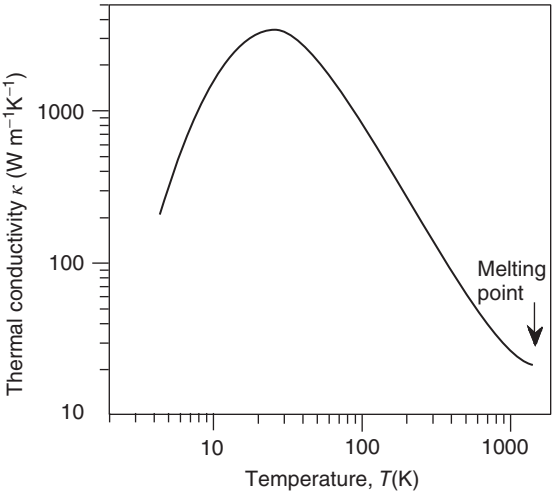


Figure 4.13 Temperature-dependent thermal conductivity of Si. Reproduced from Glassbrenner and Slack (1964).

Table 4.4 Thermal conductivity κ for some metals and insulators at room temperature.

Material	κ(W m^{-1}K^{-1})
Copper	386
Aluminum	237
Steel	50
Diamond	2300
Quartz	10
Glass	0.8
Polystyrene	0.03

Its bonding and structure are similar to Si, which also shows a very high thermal conductivity but at a lower temperature (see Figure 4.13). In diamond, the maximum in $\kappa(T)$ is shifted to a higher temperature with respect to Si because of the much higher Debye temperature. Because of this, phonon–phonon scattering only becomes important at higher temperatures, explaining diamond's good thermal conductivity at room temperature.

Phonon–phonon scattering deserves a comment: In the case of purely harmonic vibrations and waves, this cannot happen. Consider, for example, water waves or low-intensity light waves (i.e., anything but lasers). There, the principle of superposition holds. The field amplitude at a given point of space is just the superposition of the field amplitudes from different waves propagating through space. The waves "propagate through each other." This principle also holds for phonons since these are lattice waves for the harmonic solid. Phonon–phonon scattering can only happen in the **anharmonic case**, that is, if the amplitude of the oscillations

becomes so large that the fourth and higher order terms in (3.11) become important. This may seem problematic because our whole treatment so far is based on the assumption that the vibrations are harmonic. In fact, the whole concept of a phonon only makes sense for a harmonic solid. If the anharmonic effects are not too strong, however, we can save the phonon picture to some extent. We can still think of phonons propagating through the crystal just that they now have a certain finite lifetime after which they decay into other phonons.

4.4 Thermal Expansion

We have seen that anharmonic effects are responsible for limiting the thermal conductivity of a crystal. In this section, we will encounter another result of **anharmonic vibrations**, which is well known from daily experience: the thermal expansion of solids. We restrict our treatment to the linear expansion of isotropic solids. The **coefficient of thermal expansion** α can be defined as

$$\frac{\Delta l}{l} = \alpha \Delta T, \tag{4.49}$$

where $\Delta l/l$ is the fractional length change of the solid and ΔT is the temperature change, usually with respect to room temperature. α is always quite small (see Table 4.5). It is also temperature-dependent and can be shown to vanish at zero temperature, something that limits the validity of (4.49) to small temperature intervals. At room temperature, most materials have an α on the order of 10^{-5} K^{-1}. A remarkable exception is Invar, a nickel–iron alloy that has one of the lowest coefficients of thermal expansion of all metallic compounds. Its discovery by C.-E. Guillaume around 1900 was a major technical breakthrough because it permitted the construction of highly stable measurement instruments, clocks, and the like.

We can understand the thermal expansion of solids by an inspection of the Taylor series for the interatomic potential (3.11). The first anharmonic term is the cubic term, which turns the potential asymmetric and therefore results in a change of the equilibrium distance for different temperatures. This is easily seen in

Table 4.5 Coefficient of thermal expansion α at room temperature.

Material	$\alpha(10^{-5}\mathrm{K}^{-1})$
Pb	2.9
Al	2.4
Cu	1.7
Steel	1.1
Glass	0.9
Invar	0.09

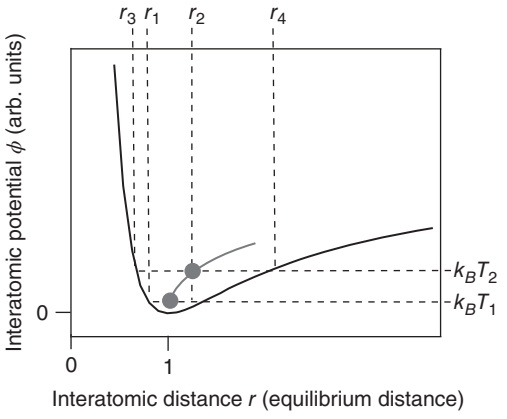


Figure 4.14 Classical picture for the thermal expansion of a solid. The interatomic potential Φ is shown as a function of interatomic distance. The gray line marks the temperature-dependent mean interatomic distance.

a classical picture. Figure 4.14 shows a scaled-up version of the interatomic potential in Figure 2.1. We choose the energy scale such that the potential minimum is at zero. According to the equipartition theorem, the mean energy of the oscillator at a temperature T is $k_B T$. For a low temperature (T_1), the oscillation thus takes place between the positions r_1 and r_2. As the potential is roughly symmetric around its minimum, the average interatomic spacing is equal to the equilibrium distance. For a higher temperature (T_2), the oscillation takes place between r_3 and r_4. The potential is not symmetric anymore, and the interatomic distance expands slightly on average. In effect, it follows the gray line for higher temperatures.

This picture is not changed qualitatively in a quantum mechanical treatment. The energy levels are then discrete. For a purely harmonic oscillator, they are equidistant and the mean interatomic distance is the same for all levels. For a nonharmonic oscillator, the energy level separation is not constant and the mean interatomic distance depends on the energy level.

4.5 Allotropic Phase Transitions and Melting

In our description of crystal structures, we have merely argued that the optimal structure should be that with the strongest possible binding, that is, with the lowest total energy. This is only true at zero temperature. For higher temperatures, entropy effects have to be taken into account. Here, we discuss two types of structural changes at higher temperatures caused by this: allotropic phase transitions in which a crystal structure is transformed to another crystal structure and melting.

We start with a thermodynamic picture. In most cases, we are interested in the structure at a given temperature, pressure, and particle number. This means that

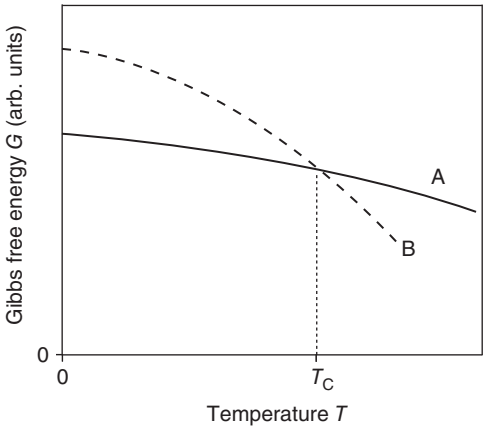


Figure 4.15 Gibbs free energy for two competing phases, A and B. At the temperature T_C a phase transition occurs.

we have to minimize the Gibbs free energy:

$$G = U + PV - TS. \tag{4.50}$$

The situation for two competing phases is illustrated in Figure 4.15. For low temperatures, the A phase has the lower Gibbs free energy, but at high temperatures, the phase B has. There must be a phase transition between the two structures at a transition temperature T_C. This is the idea behind the so-called **allotropic phase transitions** from one crystal structure to another. As an example, iron crystallizes in a body-centered cubic (bcc) structure at low temperatures but transforms into a fcc structure at 1185 K and again into a bcc structure at 1667 K.

Sometimes, structures with a high Gibbs free energy exist even though alternative structures with a lower G are possible under the same conditions. An example is diamond, which has a higher Gibbs free energy than graphite. Still, diamond exists under normal conditions because there is a high activation barrier for the transformation to graphite. Diamond is thus a **metastable structure**.

The **melting** of a crystal can be described in the same picture, only that the B phase is taken to be the liquid. If we want to predict the melting temperature of a solid, we therefore have to consider the energy and entropy of both the liquid and the solid phase as a function of temperature, a formidable task.

There have also been attempts for a more simplistic prediction of the melting temperature T_m, which neglects entropy effects. First of all, one might suspect that T_m is related to the cohesive energy of a solid. There is indeed a strong correlation, as shown in Figure 4.16a. We can easily understand the trends in the figure: The noble gas crystals are merely bonded by the van der Waals interaction. This results in a low cohesive energy and melting temperature. On the other extreme, we have the refractory transition metals such W or Mo for which covalent bonding is important, as well as covalent materials such as Si. Simple metals like the alkali metals are found in the middle of the range.

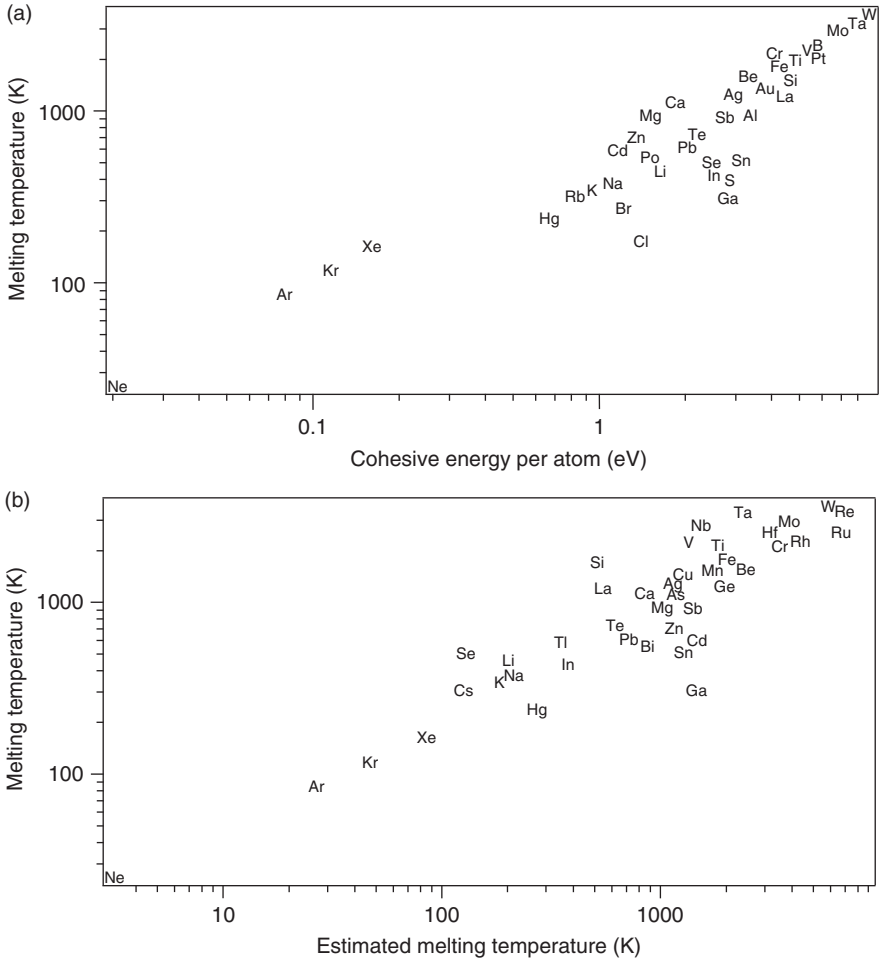


Figure 4.16 (a) Melting temperature as a function of cohesive energy for the elements. (b) Melting temperature as a function of the estimated melting temperature from the Lindemann criterion (4.52).

For the prediction of the melting temperature of a solid, the relation to the cohesive energy is not very useful because the cohesive energy has to be known in the first place. An alternative idea was developed by F. Lindemann in 1910. He suggested that melting would occur when the amplitude of the interatomic vibration $x_{\max}$ becomes too large, that is, when it reaches a certain fraction of the interatomic spacing. Using (4.5), we try this idea, guessing that the solid melts once $x_{\max}$ reaches 5% of the interatomic distance a, that is,

$$T_m = \frac{(0.05a)^2\gamma}{2k_B} = \frac{(0.05a)^2\omega^2 M}{2k_B}. \tag{4.51}$$

For ω, we can use the Debye frequency ω_D. If we now write the above expression in terms of the Debye temperature, we get

$$T_m = \frac{(0.05a)^2 \Theta_D^2 k_B M}{2\hbar^2}. \tag{4.52}$$

The result of this is shown in Figure 4.16b. With the given choice of 5% of the interatomic distance as a melting criterion, the simple model reproduces the trend correctly even though it neglects entropy effects as well as the influence of the detailed atomic structure.

References

Desnoyers, J.E. and Morrison, J.A. (1958) *Philos. Mag.*, **3**, 42.
Glassbrenner, C.J. and Slack, G.E. (1964) *Phys. Rev.*, **134**, A1058.
Grabowski, B., Hickel, T., and Neugebauer, J. (2007) *Phys. Rev. B*, **76**, 024309.
Mounet, N. and Marzari, N. (2005) *Phys. Rev. B*, **71**, 205214.
Victor, A.C. (1961) *J. Chem. Phys.*, **36**, 1903.

4.6 Further Reading

The subject of this chapter is central to solid state physics and discussed in detail in the standard literature, such as

- Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Holt–Saunders.
- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.
- Kittel, C. (2005) *Introduction to Solid State Physics*, 8th edn, John Wiley & Sons, Inc.
- Myers, H.P. (1990) *Introductory Solid State Physics*, 2nd edn, Taylor & Francis Ltd.
- Omar, M.A. (1993) *Elementary Solid State Physics*, Addison–Wesley.

These books all include a more detailed discussion of the phonon dispersion for three-dimensional solids.

For a more detailed discussion of the heat capacity, see, for example,

- Mandl, F. (1988) *Statistical Physics*, 2nd edn, John Wiley & Sons, Ltd.

4.7 Discussion and Problems

Discussion

1) What does the phonon dispersion, that is, the vibrational frequency as a function of wave vector $\omega(k)$, look like for an infinite chain of atoms with one atom per unit cell?

2) What does it look like for two atoms per unit cell? Why does one speak of optical and acoustic branches?
3) The phonon dispersion for a one-dimensional chain with one and two atoms per unit cell is given by (4.10) and (4.16), respectively. What about the amplitude of the vibrations?
4) Why is the movement of the atoms in the chain the same when multiples of $2\pi/a$ are added to, or subtracted from, the wave vector?
5) Explain the meaning and use of periodic boundary conditions to describe the properties of finite solids.
6) What predictions does the so-called Dulong–Petit law make about the heat capacity of a solid and its temperature dependence (and why)? How does it compare to the experiment?
7) At room temperature, do metals have a noticeably higher heat capacity than insulators because of the mobile electrons?
8) Explain the Einstein model for the heat capacity of a lattice. How do its predictions compare to the experiment?
9) Explain the Debye model for the heat capacity of a lattice. In which respect does it work better than the Einstein model and why?
10) What is the definition of the Debye temperature (or frequency) and what are the typical values?
11) Which one has a higher Debye temperature, lead or diamond, and why?
12) The thermal conductivity of an insulator has a maximum at about 10% of the Debye temperature. Why does it decrease for lower temperatures? Why does it decrease for higher temperatures?
13) Explain why a solid undergoes thermal expansion in a microscopic model.
14) Thermal expansion is a so-called anharmonic process. Why is it called so?

Problems

1) *One-dimensional chain with one atom per unit cell:* We have determined the phonon dispersion relation for an infinite chain of atoms with lattice spacing a and one atom per unit cell (mass M). The result is (4.10), and (4.11) for a small k, describing sound waves. (a) Show that light waves in vacuum have the same dispersion relation as (4.11) when replacing the speed of sound with the speed of light. (b) For k close to 0 ($k \ll \pi/a$) and for k at the Brillouin zone boundary ($k = \pi/a$), we have argued that the atoms move as in Figure 4.2. Show this formally using (4.8).
2) *One-dimensional chain with two atoms per unit cell:* For two atoms per unit cell of length b, we get two branches in the dispersion, the acoustic and the optical branch. The solutions are given by (4.16). (a) Plot these solutions inside the first Brillouin zone for $M_2 = 0.2\ M_1$, $M_2 = 0.9\ M_1$, and $M_2 = M_1$. (b) Consider the case where $M_1 \neq M_2$. For longitudinal vibrations, sketch how the atoms would move for k close to 0 and for k at the Brillouin zone boundary ($k = \pi/b$). (c) Which movement corresponds to which solution of (4.16)? (d) Explain what happens in the case of $M_2 = M_1$ discussed in (a).

3) *Periodic boundary conditions:* Periodic boundary conditions lead to a restriction of the possible k values. Consider a linear chain of copper atoms. The length of the chain should be 1 cm, the lattice spacing should be 0.36 nm, and the force constant should be 50 Nm^{-1}. Calculate the smallest possible finite wave vector k. What is the vibrational angular frequency for this wave vector? What is the corresponding energy in electron volts and temperature in Kelvin?
4) *The Debye model:* The phonon dispersion for a one-dimensional chain of atoms is given by (4.10) and shown in Figure 4.1b. What would the dispersion look like in the Debye model?
5) *Atomic force constants and Debye temperature:* (a) Estimate the value of the force constant γ and the angular frequency ω for the vibrations of the atoms in copper. Use that Young's modulus $Y = 130$ GPa and that the cubic lattice constant is 0.36 nm (ignore that copper actually has an fcc structure instead of a simple cubic structure). (b) Estimate the angular frequency for the vibrations that correspond to the Debye temperature of 343 K and compare it to the result in (a). (c) Estimate the amplitude of the vibrations as a fraction of the lattice spacing at room temperature.
6) *(*) Heat capacity of graphene:* For a three-dimensional (3D) solid, we have found the low-temperature heat capacity to be proportional to T^3. How does it depend on the temperature for the two-dimensional (2D) graphene? Hint: In order to give a correct answer to this question, you have to know a curious fact about the phonon dispersion in graphene. We have seen that the acoustic phonon branches in both one dimension and 3D have a dispersion with $\omega(k) \propto k$. For graphene, this is not so: Graphene has three acoustic branches and one of them has a dispersion for which $\omega \propto k^2$. For very low temperatures, this is the important branch and you should base your calculation on this dispersion only. The reason for this unusual behavior is that graphene may be 2D, but it exists in a 3D world. It therefore has two "normal" in-plane acoustic phonon branches, one longitudinal and one transverse. In addition to these, it has a phonon branch that corresponds to a "flexing" motion out of the plane and this is the one with the unusual dispersion.
7) *(*) Thermal expansion:* Explain why the coefficient of thermal expansion for a solid vanishes at $T = 0$.
8) *Thermal expansion:* In a Bragg reflection experiment using copper, a sharp peak is observed at an angle of 25.23° at 300 K. At 500 K, the same peak is observed at an angle of 25.14°. Use this information to calculate the coefficient of linear expansion for copper.

5
Electronic Properties of Metals: Classical Approach

Here and in Chapters 6, 7, and 9, we are concerned with the electrical properties of metals, semiconductors, and insulators. It would be natural to start out with a definition of the difference between these types of materials, but this is actually not so easy and we need to postpone it. Consider a couple of simple possibilities: One could argue that metals are good conductors of heat and electricity, whereas semiconductors and insulators are not. In the case of heat conduction, we have already seen that diamond, which is an insulator, conducts even better than most metals. Electrical conductivity is not of much help either: Some semiconductors such as silicon conduct electricity reasonably well. Yet another possibility is to define metals by the fact that they look "shiny" or "metallic." But this applies to some semiconductors, too, and again silicon can be taken as an example. It turns out that a proper definition has to wait until we treat electronic states in a quantum mechanical model in the next chapter. Here we start out with a classical description of metals.

5.1
Basic Assumptions of the Drude Model

In 1900, only 3 years after the discovery of the electron by J. J. Thomson, P. Drude suggested a simple model to explain many of the observed properties of metals. He did this by combining the existence of electrons as charge carriers with the highly successful kinetic gas theory. We will later see that the Drude model has many shortcomings, but it is still of fundamental importance for the concepts associated with electrical conductivity. The model is based on the following assumptions:

- The electrons in a solid behave like a classical ideal gas. They do not interact with each other at all: There is no Coulomb interaction and, as opposed to a classical gas model, they do not collide with each other either. This is known as the **independent electron approximation**. We will later see that this approximation is quite a good one: The electrons do indeed not interact much with each other.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

- The positive charge is located on immobile ion cores. The electrons can collide with the ion cores. These collisions instantaneously change their velocity. However, in between collisions, the electrons do not interact with the ions either. This is known as the **free electron approximation**. We will see that this approximation is not very good. Indeed, the whole picture of the electrons colliding with the ions is problematic. In a perfect crystalline solid at low temperatures, the electrons do not collide with the ions at all, as we shall see later.
- The electrons reach thermal equilibrium with the lattice through the collisions with the ions. According to the equipartition theorem, their mean kinetic energy is

$$\frac{1}{2}m_e v_t^2 = \frac{3}{2}k_B T. \tag{5.1}$$

  At room temperature, this results in an average speed of $v_t \approx 10^5$ ms^{-1}.
- In between collisions, the electrons move freely. The mean length of this free movement is called the **mean free path** λ. Knowing the typical packing density of the ions, we can estimate that $\lambda \approx 1$ nm. Given the average speed v_t, the mean free path also corresponds to a mean time between collisions given by $\tau = \lambda / v_t$. τ is called the **relaxation time** and plays a fundamental role in the theory. With $\lambda = 1$ nm and v_t at room temperature, we estimate that $\tau \approx 1 \times 10^{-14}$ s.

For the description of almost all properties within the Drude model, it is essential to know the density of the gas formed by the free electrons. This is known as the **conduction electron density** n, that is, the number of conduction electrons per volume. n is calculated by assuming that every atom contributes Z_V conduction electrons, that is, electrons from its outermost shell, to the metallic bonding. The core electrons remain bound to the metal ions. For the alkali metals, Z_V is 1, for the alkaline earth metals, it is 2, and so on. There are ρ_m / M atoms per cubic meter, where ρ_m is the density of the solid in kg m^{-3} and M is the atomic mass in

Table 5.1 Number of conduction electrons per atom Z_V, calculated conduction electron density n, and measured Hall coefficient R_H of selected metals.

Metal	Z_V	$n(10^{28}$ m$^{-3})$	Measured R_H divided by $-1/ne$
Li	1	4.7	0.8
Na	1	2.7	1.2
K	1	1.3	1.1
Rb	1	1.2	1.0
Cs	1	0.9	0.9
Cu	1	8.5	1.5
Ag	1	5.9	1.3
Be	2	24.7	−0.2
Mg	2	8.6	−0.4
Al	3	18.1	−0.3
Bi	5	14.1	≈40 000

kilograms per atom. Consequently, the conduction electron density n is $Z_V \rho_m / M$. Values of n for selected metals are given in Table 5.1.

5.2 Results from the Drude Model

We now show how several properties of metals can be explained by the Drude model. In Section 5.3, we will discuss the limitations of the model and the most significant disagreements with the experimental results.

5.2.1 DC Electrical Conductivity

To explain the DC conductivity of metals, consider the behavior of an electron when an electric field $\mathcal{E}$ is applied. The equation of motion is[1)]

$$m_e \frac{d\mathbf{v}}{dt} = -e\boldsymbol{\mathcal{E}}, \tag{5.2}$$

with the solution

$$\mathbf{v}(t) = \frac{-e\boldsymbol{\mathcal{E}}t}{m_e}, \tag{5.3}$$

that is, an accelerated drift motion in the direction opposite to the field. If we assume that the drift motion is destroyed in a collision with the ions and that on average the time for a collision-free drift is τ, the average drift velocity is[2)]

$$\overline{\mathbf{v}} = \frac{-e\boldsymbol{\mathcal{E}}\tau}{m_e}. \tag{5.4}$$

We can estimate the order of magnitude for $|\overline{\mathbf{v}}|$: For an electric field of $\mathcal{E} \approx 10\ \mathrm{V\,m^{-1}}$, we get a drift velocity of $|\overline{\mathbf{v}}| = 10^{-2}\ \mathrm{m\,s^{-1}}$. This is *very* slow compared to the thermal movement of the electrons. The result justifies our simple approach because the drift motion induced by the electric field will not have a significant effect on the relaxation time.

Having the drift velocity, we can calculate the conductivity. Consider an area A perpendicular to the electric field. The number of electrons passing through the area per unit time is

$$n|\overline{\mathbf{v}}|A. \tag{5.5}$$

The amount of charge passing through the area is therefore

$$-en|\overline{\mathbf{v}}|A. \tag{5.6}$$

1) Note that the charge of the electron is $-e$ throughout this book.
2) It is not obvious that (5.4) is the correct result for the average drift velocity. One could be tempted to think that it is too high by a factor of 2. See Further Reading.

With this we can calculate the current density

$$\mathbf{j} = -en\bar{\mathbf{v}}, \tag{5.7}$$

and with (5.4) we get

$$\mathbf{j} = \frac{ne^2\tau}{m_e}\boldsymbol{\mathcal{E}} = \sigma\boldsymbol{\mathcal{E}} = \frac{\boldsymbol{\mathcal{E}}}{\rho}, \tag{5.8}$$

that is, the current density is in the direction of the electric field and proportional to the field strength. This result is the familiar **Ohm's law**, and the constant of proportionality σ is called the **conductivity** of the material. Its inverse ρ is called the **resistivity**.

Consider now the explicit expressions for the conductivity and resistivity, which we have obtained. The conductivity is

$$\sigma = \frac{ne^2\tau}{m_e}, \tag{5.9}$$

and the resistivity

$$\rho = \frac{m_e}{ne^2\tau}. \tag{5.10}$$

Note that the elementary charge appears squared in these equations. The reason for this is that it is needed both to couple to the electric field, dragging the electrons along, and in the definition of the current. The fact that it is squared means that we would get the same result for charge carriers with a positive charge $+e$ instead of $-e$. We will come back to this when discussing semiconductors where positively charged carriers do in fact appear.

Another useful definition is the **mobility** of the electrons μ. It is given by

$$\mu = \frac{e\tau}{m_e}, \tag{5.11}$$

and the conductivity and resistivity can, of course, also be defined using this mobility:

$$\sigma = n\mu e, \quad \rho = \frac{1}{n\mu e}. \tag{5.12}$$

Why do we need this definition? The concept of mobility can be useful for solids in which the electron concentration can be changed by some external parameter without changing the scattering mechanism inside the solid, that is, without changing the relaxation time. The mobility also has a simple physical meaning: It is the ratio of drift velocity to applied electric field, as can be seen when dividing (5.8) by $-ne$.

The Drude model thus explains Ohm's law qualitatively. We can also perform a quantitative comparison of the predicted and measured conductivities. Figure 5.1 shows this for some selected metals at two different temperatures. The calculations have been made assuming a mean free path of $\lambda = 1$ nm for all elements. For $T = 273$ K, the calculation (solid line) reproduces the right order of magnitude and lies in the middle of the scattered experimental data points. Some elements lie

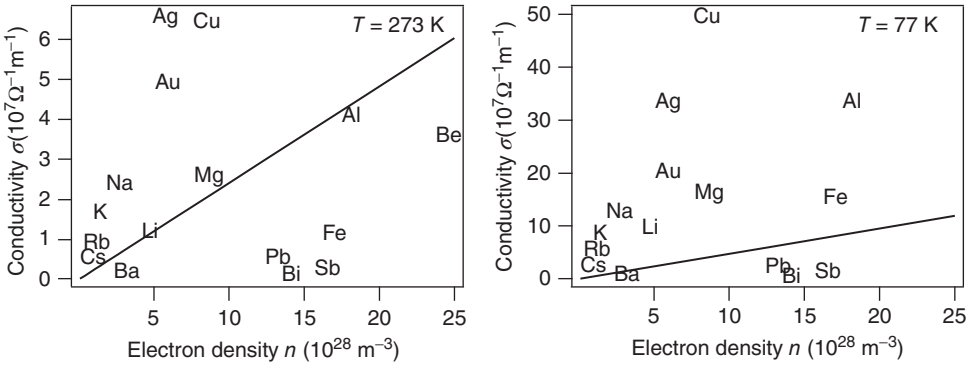


Figure 5.1 Measured and calculated electrical conductivities of metals as a function of conduction electron density for two different temperatures. The measured data are marked by the elements' names; the calculations are the solid lines.

far away from the calculation, notably the noble metals and the group V semimetals Sb and Bi. One could be tempted to conclude that the Drude model does not reproduce the details, but the general trend is correct anyway. For lower temperatures, the situation becomes more problematic. At 77 K, the calculated conductivity increases because ν_t gets smaller, but the measured conductivity increases much more. At even lower temperature, the comparison becomes increasingly unfavorable.

5.2.2 Hall Effect

Another result of the Drude model is that it can explain the Hall effect. This effect was discovered by E. Hall in 1879 when he investigated the influence of a magnetic field on the current in a conductor. It is illustrated in Figure 5.2a. Hall found that

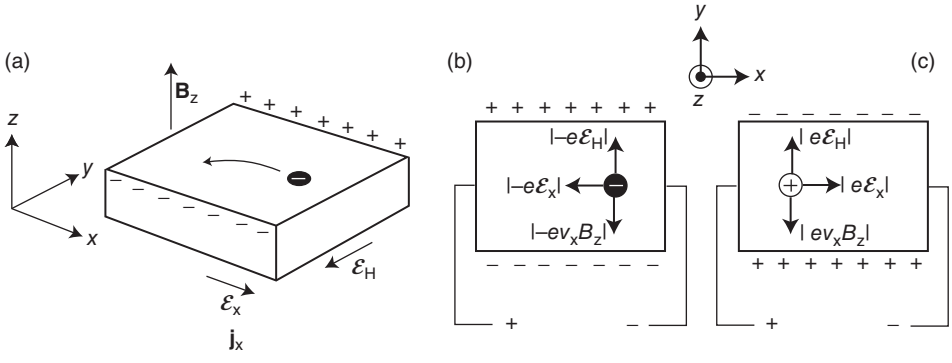


Figure 5.2 (a) Illustration of the Hall effect. (b) Equilibrium between Lorentz force and force caused by the Hall field for electrons passing through the sample (charge $-e$). (c) The same for positively charged carriers passing through the sample (charge $+e$).

an electric field $\mathcal{E}_H$ is built up, which is perpendicular to both the magnetic field and the current density. The magnitude of this Hall field is proportional to both current density j_x and magnetic field B_z:

$$\mathcal{E}_H = R_H j_x B_z, \tag{5.13}$$

where R_H is called the **Hall coefficient**. This is explained quite easily in the steady state (see Figure 5.2b). For the electrons to pass through the sample, the Hall field $\mathcal{E}_H$ must exactly compensate the Lorentz force in the opposite direction; thus,

$$|-e\mathcal{E}_H| = |-eB_z v_x|. \tag{5.14}$$

Using this and the definitions of the current density (5.7) and R_H, we obtain

$$R_H = \frac{\mathcal{E}_H}{j_x B_z} = \frac{\mathcal{E}_H}{-env_x B_z} = \frac{v_x B_z}{-env_x B_z} = \frac{-1}{ne}. \tag{5.15}$$

Therefore, measuring the Hall coefficient provides direct experimental access to the conduction electron density.

We can compare the measured Hall coefficients to those calculated from the electron density for different elements. The result is shown in Table 5.1 where the measured R_H has been divided by $-1/ne$ for easier comparison. For the alkali metals, the result is close to the expected value of 1, and for the noble metals, the agreement is also acceptable. It is very bad for Bi. The very high value means that for some reason the true conduction electron density must be much smaller than the calculated value. In a sense, the agreement is even worse for Be, Mg, and Al because not only does the magnitude not quite fit, the measured R_H is even positive, not negative. In this context, it is important to note again that the sign of the charge carriers is irrelevant for the conductivity of the sample, but it shows up in the Hall effect. It appears therefore that the current in Be, Mg, and Al is carried by positive charges: Imagine that we had positive charge carriers with a density p (see Figure 5.2c). Then, it is easy to show that

$$R_H = \frac{1}{pe}, \tag{5.16}$$

that is, we get a positive R_H. The notion of positive carriers does not make sense in the Drude model, but we will see that the quantum model of the electronic states is able to give an intuitive picture of positive carriers. This will be particularly useful for treating semiconductors, but we will also come back to the positive Hall coefficient in metals in Section 7.2.2.

5.2.3 Optical Reflectivity of Metals

The Drude model can also explain why metals reflect light and therefore appear shiny. Before we discuss this, we briefly state some fundamental relations from optics. We shall need these equations again when we discuss the optical properties of insulators in Chapter 9. Some of the concepts used here are explained in more detail in the beginning of Chapter 9.

Light can be described as a transverse, plane electromagnetic wave. We can write the electric field for a wave propagating in the z direction as

$$\boldsymbol{\mathcal{E}}(z,t) = \boldsymbol{\mathcal{E}}_0 e^{i(kz-\omega t)}, \tag{5.17}$$

with the amplitude $\boldsymbol{\mathcal{E}}_0$ in the $x-y$ plane and the wave vector

$$k = \frac{2\pi N}{\lambda_0}, \tag{5.18}$$

where λ_0 is the wavelength in vacuum, and

$$N = n + i\kappa \tag{5.19}$$

is the complex **index of refraction**. n, the real part of N (not to be confused with the electron density), describes the change of the wavelength in matter and thereby the refraction at an interface, and the imaginary part κ accounts for the damping inside the material.[3] In general, N depends on the frequency ω, a phenomenon called **dispersion**. A familiar consequence of this is the separation of light into different colors when refracted by a glass prism.

An alternative way to describe the optical properties of materials is to use the complex **dielectric function** ϵ instead of the refractive index N. You may be familiar with the static dielectric constant ϵ that appears in the description of capacitors. The dielectric function is the same quantity, but it accounts for a frequency dependence, that is, in general ϵ depends on ω. ϵ is related to N via

$$N = \sqrt{\epsilon} = \sqrt{\epsilon_r + i\epsilon_i}, \tag{5.20}$$

where ϵ_r and ϵ_i are the real and imaginary parts of ϵ, respectively. With this, (5.17) can be written as

$$\boldsymbol{\mathcal{E}}(z,t) = \boldsymbol{\mathcal{E}}_0 e^{i((2\pi N/\lambda_0)z-\omega t)} = \boldsymbol{\mathcal{E}}_0 e^{i((\omega\sqrt{\epsilon}/c)z-\omega t)}. \tag{5.21}$$

In the last step, we have used that $\lambda_0\omega/2\pi = c$, with c being the speed of light in vacuum.

Having these basic equations, we can proceed to explain the reflectivity of metals. Consider an electron in the electromagnetic AC field given by the optical light wave. If the angular frequency ω of the light is very small, we basically retain the DC behavior. If, on the other hand, ω is so high that $2\pi/\omega$ is much shorter than the relaxation time τ, then the electron is wiggled many times back and forth by the field before a scattering process occurs. We can then ignore the collisions with the ions altogether and treat the electrons as completely free. As τ is on the order of 10^{-14} s, this condition is fulfilled reasonably well for optical frequencies.

We treat one single electron in the electric field of an electromagnetic wave. The polarization should be such that the $\boldsymbol{\mathcal{E}}$ field lies in the x direction, and the time-dependent magnitude of field is $\mathcal{E}_0 e^{-i\omega t}$. The electron will move according to the equation of motion

$$m_e \frac{d^2x(t)}{dt^2} = -e\mathcal{E}_0 e^{-i\omega t}. \tag{5.22}$$

3) This can be seen by inserting $N = n + i\kappa$ into (5.18) and (5.17) where it leads to a damping factor $\exp(-2\pi\kappa z/\lambda_0)$, that is, a damping for increasing z.

A good ansatz for the solution of (5.22) appears to be

$$x(t) = Ae^{-i\omega t}, \tag{5.23}$$

where A is a (complex) amplitude. By inserting (5.23) back into (5.22), we see that it is indeed a solution if the amplitude is chosen to be

$$A = \frac{e\mathcal{E}_0}{m_e\omega^2}. \tag{5.24}$$

The electron is now periodically displaced from its position, and this leads to a changing dipole moment $-ex(t)$. For a solid with a conduction electron density of n, the macroscopic polarization $P(t)$ resulting from these dipole moments is[4)]

$$P(t) = -nex(t) = -neAe^{-i\omega t} = -\frac{ne^2\mathcal{E}_0 e^{-i\omega t}}{m_e\omega^2}. \tag{5.25}$$

On the other hand, we know the general relation between the electric field $\mathcal{E}$ and the dielectric displacement field D, which is

$$D = \epsilon\epsilon_0\mathcal{E} = \epsilon_0\mathcal{E} + P, \tag{5.26}$$

such that

$$\epsilon = 1 + \frac{P(t)}{\epsilon_0\mathcal{E}_0 e^{-i\omega t}}. \tag{5.27}$$

Using this and our result for the polarization (5.25), we obtain an expression for the dielectric function

$$\epsilon = 1 - \frac{ne^2}{\epsilon_0 m_e\omega^2} = 1 - \frac{\omega_P^2}{\omega^2}, \tag{5.28}$$

with the so-called **plasma frequency** ω_P given by

$$\omega_P^2 = \frac{ne^2}{m_e\epsilon_0}. \tag{5.29}$$

This expression for the dielectric function is our final result. Why does this explain that metals reflect visible light? To see this, consider (5.21) and (5.28). We have to distinguish between two cases: For $\omega < \omega_P$, ϵ is a real and negative number. Therefore, $\sqrt{\epsilon}$ is purely imaginary and (5.21) represents an exponentially damped penetration of the wave into the solid. Equivalently, we see that for a negative ϵ, the complex index of refraction (5.19) contains only the imaginary component $i\kappa$. The damping cannot be due to inelastic losses because our (5.22) does not take such processes into account. Since the light is not transmitted through the solid either, and energy is conserved, it must be reflected (for a more formal treatment, see Problem 9.4). For $\omega > \omega_P$, on the other hand, ϵ is real and positive and (5.21) represents a light wave that propagates into the metal. The bottom line is that metals are reflecting low-frequency light, but they become transparent for high-frequency light. The transition happens at the plasma frequency. The low-frequency behavior is not surprising because it should essentially be the same as

4) For a definition of the polarization, see (9.2) in Chapter 9.

Table 5.2 Observed values of the plasma energy $\hbar\omega_P$ together with the values calculated from the Drude model.

Metal	Measured $\hbar\omega_P$ (eV)	Calculated $\hbar\omega_P$ (eV)
Li	6.2	8.3
K	3.7	4.3
Mg	10.6	10.9
Al	15.3	15.8

in the electrostatic case for which it is assumed that metals are free of electric fields.

The plasma frequency can be calculated solely from the conduction electron density of the metal. Instead of the plasma frequency ω_p, one commonly uses the plasma energy $\hbar\omega_p$. Calculated and measured values are given in Table 5.2. We see that the agreement between experiment and prediction is fairly good and that the plasma frequency for metals lies in the far ultraviolet region, that is, metals are reflecting visible light but transmitting ultraviolet radiation.

5.2.4 The Wiedemann–Franz Law

In Drude's time, one of the most convincing pieces of evidence for his theory appeared to be that it yielded a quantitatively correct description of the Wiedemann–Franz law. In 1853, G. H. Wiedemann and R. Franz found that the ratio of thermal to electrical conductivity is constant for all metals at a given temperature. Later, it was found by L. Lorenz that this constant is proportional to the temperature; thus,

$$\frac{\kappa}{\sigma} = LT, \tag{5.30}$$

where L is the so-called **Lorenz number**.

In the Drude model, the ratio of thermal and electrical conductivity is readily calculated. The thermal conductivity is that of a classical gas and can be described by an equation similar to (4.48), only using corresponding properties (heat capacity, speed, and mean free path) for the electron gas. The electrical conductivity is given by (5.9). The result is

$$\frac{\kappa}{\sigma} = \frac{3}{2}\frac{k_B^2}{e^2}T = LT, \tag{5.31}$$

which is just the Wiedemann–Franz law (see Problem 5.5). L, as calculated here, is roughly a factor of 2 smaller than the value obtained by experiments (or by a proper quantum mechanical calculation, see (6.19)). Drude, however, had made a mistake of a factor of 2 in his calculation such that L came out almost correct. Therefore, his theory was in impressive quantitative agreement with the experimental data. It should be pointed out that Drude's mistake was rather subtle (it had

to do with the scattering probabilities of the electrons) and therefore not readily discovered by other researchers.

5.3 Shortcomings of the Drude Model

Despite its great success, the Drude model has a number of serious shortcomings. We discuss several of them here to motivate the quantum treatment of metals in the next chapter. Even before starting our work on the Drude model, several assumptions could have raised suspicion. Take, for example, the nature of the scattering. There is no justification for a missing electrostatic interaction between the electrons and the lattice, and it is also not clear why the electrons collide only with lattice ions and not among themselves. In addition to this, the de Broglie wavelength for electrons with a thermal energy is on the order of nanometers. The criterion for treating the electrons as classical particles, however, is that their de Broglie wavelength is much smaller than the typical dimensions of the structures they are moving in. This is clearly not fulfilled.

As for a comparison to experimental data, we have already seen that the predicted conductivity at low temperatures is not high enough. When assuming a fixed mean free path, the Drude model does give a higher conductivity at low temperatures because of the increased relaxation time (see Figure 5.1) but the measured conductivity increases much more. It turns out that the assumption of a fixed mean free path, given by the atomic spacing, is completely wrong. In fact, at low temperatures, the mean free path of electrons in very pure and perfect crystals can become macroscopic, micrometers, or even millimeters. Apparently, the electrons manage to sneak past all other electrons and all ions as well. This appears quite mysterious, but we will be able to explain it in the next chapter. Another problem is that the Drude model cannot explain the conductivity of alloys. Alloying a small amount of impurities into an otherwise pure metals can drastically reduce the conductivity. This happens even if the impurity atoms are quite similar to the host and would be expected to give rise to a similar electron concentration (e.g., Au in Cu).

The historically most important issue associated with the classical treatment of electrons in a metal is that these electrons should give a considerable contribution to the heat capacity, but this is not observed. In the previous chapter, we have seen that the experimentally determined heat capacity of most solids, including metals, agrees with the Dulong–Petit value at room temperature (see Table 4.2). We could also understand the Dulong–Petit rule as the high-temperature limiting case for the heat capacity of the solid's lattice. In this classical picture, the presence of free electrons should lead to an increased heat capacity for metals: For 1 mol of classical metal, the heat capacity of the lattice would still be given by the Dulong–Petit rule as $3N_A k_B = 3R$ but the electrons would be expected to contribute to the total heat capacity as well. Each electron has three translational degrees of freedom, each contributing with $k_B/2$ to the heat capacity. If the metal has one

conduction electron per atom, these electrons would contribute to the molar heat capacity with $3R/2$ and the total heat capacity would thus be $9R/2$. This is significantly higher than the Dulong–Petit value that is actually observed. For metals with more conduction electrons per atom, the agreement with the experimental result would be even poorer. The fact that the Dulong–Petit value is observed for many metals therefore suggests that the electrons do not contribute to the heat capacity, even though they are free to move and they do contribute to the conduction of electricity. This puzzle can only be resolved by a quantum mechanical treatment of metals.

5.4 Further Reading

The Drude model is treated in many standard texts on solid state physics. A particularly good and in-depth description, including the issue of the factor of 2 in (5.4), is found in

- Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Holt-Saunders.

For a more detailed discussion of a metal's reflectivity, see

- Fox, M. (2010) *Optical Properties of Solids*, 2nd edn, Oxford University Press.

5.5 Discussion and Problems

Discussion

1) Describe the basic assumption of the Drude model for metals. Explain the relaxation time and the mean free path of the electrons.
2) How fast do the electrons move in the Drude model, and how does their speed depend on the temperature?
3) Describe the electrical conduction in the Drude model. Where does the electrical resistance come from?
4) How does the measured voltage drop along a metal wire as a function of current through the wire compare to the prediction of the Drude model qualitatively? How is the quantitative agreement?
5) (*) List cases in which Ohm's law is not valid.
6) When an electric field is applied, how does the additional speed of the electrons compare to their thermal speed (at room temperature)?
7) What is the Hall effect, and what can be measured by it?
8) Explain qualitatively why the sign of the Hall coefficient depends on the charge of the particles carrying the current (positive or negative).
9) Why do metals not transmit light? Is this so for all light frequencies?

10) What is the Wiedemann–Franz law?
11) Which properties of metals are not described adequately by the Drude model?

Problems

1) *Classical versus quantum description of metals:* Calculate the classical mean kinetic energy for the electrons in Na at room temperature. From this, determine their de Broglie wavelength λ. For a classical description to be valid, we have to require that λ is much smaller than the mean separation d of the particles. Show that this is not the case.
2) *Ohm's law:* Explain how expression (5.8) is related to the more familiar form of Ohm's law $I = U/R$.
3) *Optical reflectivity of metals:* (a) Suggest a way of measuring the plasma energy in a metal. (b) Suppose that you want to develop a metallic thin-film coating for windows such that they would transmit visible light but not infrared radiation. What properties would you require the coating material to have?
4) *Optical reflectivity of metals:* Estimate how deeply visible light penetrates into aluminum. This penetration depth is defined as the depth at which the intensity of the incoming light wave has dropped to $1/e$ of the original intensity.
5) *Wiedemann–Franz law:* Show that the Wiedemann–Franz coefficient L in the Drude model is indeed given by (5.31), that is,

$$L = \frac{\kappa}{\sigma T} = \frac{3}{2}\frac{k_B^2}{e^2}. \tag{5.32}$$

6) *Resistivity:* We have seen that the Drude model gives the correct order of magnitude for the resistivity of many metals near room temperature, but what about the temperature dependence of the resistivity? Experimentally, it is found that this temperature dependence is linear near room temperature, that is,

$$\rho(T) = \rho_0(1 + \alpha(T - T_0)), \tag{5.33}$$

where $\rho(T)$ is the temperature-dependent resistivity, ρ_0 the resistivity at room temperature T_0, and α the so-called thermal resistance coefficient. Show that this experimental finding is in qualitative disagreement with the Drude model, that is, that the Drude model does not give rise to a linear temperature dependence.
7) *Phonons in metals:* The Drude model can be used to estimate the force constants, vibrational frequencies, and related properties in metals. We use a crude model to describe the vibration of a single ion in a monovalent metal: The ion is assumed to be a positive point charge in a spherical unit cell that is filled with electrons of the appropriate density n. All the rest of the crystal is ignored. (a) Consider a small displacement of the ion from the center of the unit cell. Show that the restoring force is proportional to the magnitude of

the displacement. Hint: Use Gauss' law to calculate the force. (b) What is the vibrational frequency for a single ion of sodium? (c) Perform a crude estimate of the speed of sound in sodium, and compare your result to the experimental value of 3200 m s^{-1}.

6
Electronic Properties of Solids: Quantum Mechanical Approach

In the previous chapter, we have discussed the Drude model in which the electrons in a metal are treated as classical, free, and independent particles. We have seen the success and the limitations of this approach. Now we take the quantum mechanical nature of the problem into account, and we will see how this fixes many of the shortcomings of the classical description. We will also see that, going beyond the assumption of free electrons, it is possible to explain not only metals but also nonmetallic solids. Indeed, we will come up with a more formal definition of what a metal is in the first place. We will, however, retain the approximation that the electrons move independently from each other. This works surprisingly well for many solids, and we will try to understand why.

Finding the quantum mechanical eigenstates for the electrons in the solid is a formidable problem: We would have to construct a wave function that depends on the coordinates of all the electrons and also of all the ions, which make up the positive part of the potential. This is clearly hopeless! The first approximation we make is to ignore the motion of the ions by "freezing" them into their equilibrium position. We know that there are thermal vibrations and that this approximation appears poorly justified. However, it turns out to work rather well. The reason is the mass difference between the electrons and the ions. Suppose that the ions are in some given position and the electrons are in their ground state. When the ions move out of position, their motion is so slow that the fast electrons will be able to readjust their distribution such that they stay in a modified ground state, but still in the ground state. When the ions move back, the electrons adjust themselves to the old ground state. Therefore, the electronic and ionic motions can be effectively separated. This is called the **Born–Oppenheimer approximation**, and it is also often used for treating molecules.

The other fundamental simplification we make is that we do not consider the correlated motion of the electrons. We merely calculate the electronic states for one electron that is moving in an effective potential $U(\mathbf{r})$, given by all the ions and all the other electrons. The stationary Schrödinger equation for the one-electron states then becomes

$$-\frac{\hbar^2\nabla^2}{2m_e}\psi(\mathbf{r}) + U(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}). \tag{6.1}$$

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

Admittedly, this **one-electron approximation** creates the problem of finding the right $U(\mathbf{r})$! Amazingly, we will find that for many metals, $U(\mathbf{r})$ is quite small, that is, the electrons behave as if they were nearly free.

One great help for finding the electronic energy levels is the symmetry of the lattice. No matter how complicated $U(\mathbf{r})$ is, at least we know that it must be lattice-periodic, that is,

$$U(\mathbf{r}) = U(\mathbf{r} + \mathbf{R}), \tag{6.2}$$

where $\mathbf{R}$ can be any vector of the Bravais lattice.

Finally, when we have found the eigenstates in the one-electron picture, we fill them with all the electrons according to the Pauli principle. This gives the correct occupation of the states but only for zero temperature. At higher temperatures, the statistical occupation of the states is given by the Fermi–Dirac distribution.

Before we start with a detailed quantum mechanical description of solids along these lines, we consider very simple models for the electronic states in solids. We motivate the idea of electronic energy bands, and we give an informal but intuitive picture of the difference between metals, semiconductors, and insulators.

6.1 The Idea of Energy Bands

Let us consider the solid as a type of giant molecule and ask about the possible energy levels in such a molecule. This approach is intuitive and would give the correct results, but it is not very practical for treating crystals, in particular because their high degree of symmetry is not exploited. For simplicity, we build the molecule from Na atoms that have only one valence electron. Figure 6.1 illustrates what happens as we assemble an ever bigger cluster of Na atoms. For two atoms, the situation is similar to that of the hydrogen molecule in Figure 2.2: As the atoms approach each other, bonding and antibonding molecular orbitals are formed[1)]. Each Na atom has one 3s electron and the two electrons are accommodated in the bonding orbital. They have opposite spins in order to fulfill the Pauli principle's requirements (Figure 6.1a). Figure 6.1b illustrates the position of the energy levels as a function of the interatomic separation. The energy scale is set to zero for a very large separation of the atoms. For the separation a, the bonding level reaches the lowest energy. Since only the bonding level is occupied by two electrons, the energy of the antibonding state is irrelevant and the energy gain is maximized for the separation a.

What happens if we take more than two atoms? The interaction of two atomic states leads to the formation of two levels that are delocalized over the entire

1) There is a conceptual difference between our treatment of the hydrogen molecule and the treatment of a Na cluster here. In the hydrogen molecule, we have calculated the energy levels for a genuine two-electron wave function. In this chapter, we stick to the one-electron approximation, that is, we always calculate the electronic states of one electron in an effective potential of the ions and the other electrons. Then, we fill these one-electron states according to the Pauli principle.

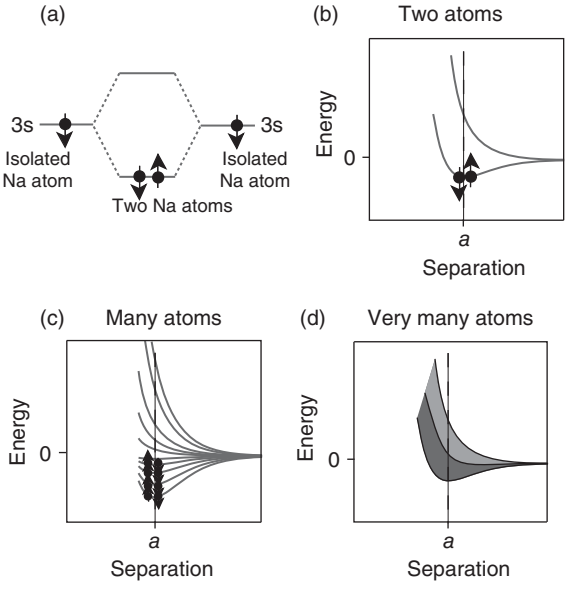


Figure 6.1 The formation of energy bands in solids. (a) Bonding and antibonding energy levels and their occupation for a molecule constructed from two Na atoms. The black dots and arrows symbolize the electrons with their spin. (b) The molecule's energy levels as a function of interatomic separation. (c) The energy levels for a cluster of many Na atoms as a function of their separation. (d) For very many atoms, there is a quasi-continuum between the lowest and highest energy levels. This energy band is half-filled with electrons (dark area) and half-empty (bright area).

molecule. The situation is similar for N atoms. The N atomic energy levels split up into N nondegenerate molecular levels.[2] $N/2$ of these levels are then occupied by two electrons each. This is shown in Figure 6.1c. For a very large N, the same principles apply. There is a quasi-continuum of states between the lowest and the highest level, which is half-filled (Figure 6.1d). This quasi-continuum is called an **energy band**.

Now we can qualitatively see why Na should show metallic behavior. The energy band of the valence electrons is exactly half-filled. When an electric field is applied to a sample of Na, the electrons experience a force opposite to the field direction. In order to move in that direction, they have to increase their kinetic energy by a bit, that is, they have to go into a state with a slightly higher energy. For the electrons in the highest occupied states, this is easily possible because there are plenty of unoccupied states available at slightly higher energies.

To see a quite different behavior, consider the formation of energy bands in Si, which is shown in Figure 6.2. The valence electrons involved in the bonding are the two 3s and the two 3p electrons. As the Si atoms approach each other, the

2) Some of these levels might be degenerate because of symmetry, but this is not important for this qualitative discussion.

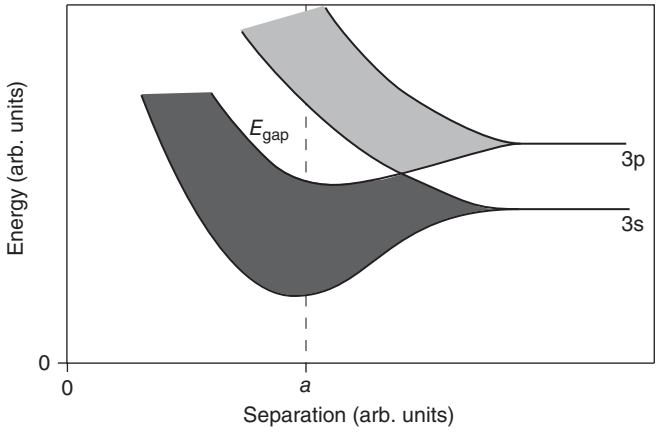


Figure 6.2 Band formation in Si. The lower band corresponds to the sp^3 states and is completely filled.

orbitals hybridize and form two bands of states at equilibrium distance a, each containing four states per Si atom, that is, a total of eight states per atom, derived from two and six atomic s and p states, respectively. The lower band consists of the sp^3 orbitals, and these are fully occupied by the four valence electrons of each Si atom. The upper band is completely unoccupied, and between the two bands, there is an energy region without any states, a so-called **band gap**. This explains the insulating behavior of Si: When a voltage is applied, the electrons in the filled sp^3 band cannot increase their kinetic energy by a small amount because there are no vacant states with energies slightly above the band.

This picture allows us to group materials in two classes: metals and nonmetals. Nonmetals can be further divided into semiconductors and insulators (see Chapter 7). Unfortunately, the simple model lacks any predictive power. If we take carbon that has the same number of valence electrons as Si, we would come to the same picture for the bonding. This is correct for diamond that is sp^3 bonded and has a band gap. But for graphite, the situation is quite different. There are also bands formed in graphite but there is no band gap.

6.2 Free Electron Model

6.2.1 The Quantum Mechanical Eigenstates

The free electron model is the quantum mechanical analogue to the Drude model. Its objective is to obtain a simple descriptions of metals, assuming that the electrons are free in the sense that they are not interacting with the ions or with each other (the model is therefore also called the **free electron gas**). Treating free

electrons in a quantum model comes down to the standard problem of a free particle in a box. The task is to solve (6.1) with $U(\mathbf{r}) = 0$, assuming certain boundary conditions. What boundary conditions should we choose? The situation is very similar to that of a finite chain of atoms described in Section 4.1.3. The simplest boundary conditions are that the wave function has to vanish at the boundaries. This corresponds to holding the atoms at the end of a finite chain fixed. In both cases, it leads to standing waves. As in the case of lattice vibrations, it would be inconvenient to use such boundary conditions here because, ultimately, we will be interested in traveling solutions to account for electrical and thermal conductivity. Therefore, the periodic boundary conditions (4.17) are a better choice (for a more in-depth discussion of the different boundary conditions, see Problem 11.1). We consider the three-dimensional case right away rather than discussing the problem in one dimension first, because a number of properties depend on the dimensionality. For simplicity, let us assume that we have a cubic box with a macroscopic side length L and volume $V = L^3$. Then, the periodic boundary conditions are

$$\psi(\mathbf{r}) = \psi(x, y, z) = \psi(x + L, y, z) = \psi(x, y + L, z) = \psi(x, y, z + L). \tag{6.3}$$

The solutions to the stationary Schrödinger equation are plane waves, normalized such that the integrated probability to find an electron in the box is 1:

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{6.4}$$

This is very similar to genuinely free electrons, but there are restrictions on the allowed values of $\mathbf{k}$ imposed by the periodic boundary conditions. These are the same as for crystal vibrations (see (4.23)), that is,

$$\mathbf{k} = (k_x, k_y, k_z) = \left(\frac{n_x 2\pi}{L}, \frac{n_y 2\pi}{L}, \frac{n_z 2\pi}{L} \right), \tag{6.5}$$

where n_x, n_y, and n_z are integers. The energy levels are

$$E(\mathbf{k}) = \frac{\hbar^2 k^2}{2m_e} = \frac{\hbar^2}{2m_e}(k_x^2 + k_y^2 + k_z^2). \tag{6.6}$$

These are shown as a function of n_x or k_x in Figure 6.3 for $n_y = n_z = 0$. In the figure, it appears as if the level separation increases for higher energies or higher values of n_x, but this is misleading and due to the fact that we have held n_y and n_z constant at zero. Actually, the separation between the energy levels is of the order

$$\frac{\hbar^2}{2m_e} \left(\frac{2\pi}{L} \right)^2, \tag{6.7}$$

(for all energies), which is very small because L is a macroscopic distance. Therefore, this model already gives rise to a quasi-continuum (or a band) of energy levels, as qualitatively described in the previous section.

These calculated energy levels are one-electron levels. We can put in the electrons according to the Pauli principle. We start by filling two electrons into the

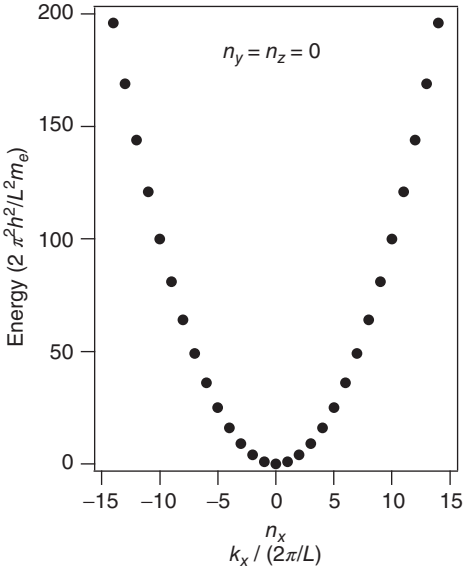


Figure 6.3 Electronic states in the free electron model. The increasing energy separation between the points at higher energies is an artifact caused by holding $n_y = n_z = 0$.

lowest energy state with $\mathbf{k} = (0, 0, 0)$ and $E(\mathbf{k}) = 0$. Then, we proceed by occupying levels at higher $\mathbf{k}$, for example, $\mathbf{k} = (0, 0, 2\pi/L)$ until we have used up all the electrons.

We want to know the highest occupied energy that we get when filling up the states. If the number of electrons to be distributed is very large, we can use the same geometric construction as for the vibrational states (see Figure 4.12). Suppose that we have N electrons in our enclosed volume, such that the conduction electron density is $n = N/V = N/L^3$. These can be accommodated on the $N/2$ states with the lowest energy since we can have two electrons per state. We have to fill all states that lie within a sphere of radius $n_{\max}$ or, alternatively, a sphere with radius $k_{\max}$ such that

$$\frac{N}{2} = \frac{4}{3}\pi n_{\max}^3, \tag{6.8}$$

which gives

$$n_{\max} = \left(\frac{3N}{8\pi}\right)^{1/3}. \tag{6.9}$$

From this, we can calculate the energy of the highest occupied electron states to

$$E_{\max} = \frac{\hbar^2 k_{\max}^2}{2m_e} = \frac{\hbar^2}{2m_e}\left(\frac{2\pi}{L}\right)^2 n_{\max}^2. \tag{6.10}$$

This energy has a special name: It is called the **Fermi energy** E_F. For most metals, it is a few electron volts, that is, in the range of typical chemical binding energies. Similarly, $k_{\max}$ is called the **Fermi wave vector** k_F. $n_{\max}$ is not used very much

because it depends on the size of the system. A useful expression that follows from (6.10) is the relation between Fermi energy E_F and conduction electron density n:

$$E_F = \frac{\hbar^2}{2m_e}\left(3\pi^2 n\right)^{2/3}. \tag{6.11}$$

The Fermi energy is the highest kinetic energy of the electrons in the solid. The corresponding **Fermi velocity** v_F can be calculated from $v_F^2 = 2E_F/m_e$. The result is on the order of 10^6 ms^{-1}. This is very high, especially if we keep in mind that everything has so far been calculated for a temperature of 0 K.

Finally, we can calculate the **density of states** $g(E)$ in the free electron model, that is, the energy-dependent number of available states per energy interval dE. We will need this for the correct description of the situation at finite temperature and for many other things. From (6.9) and (6.10), the highest occupied energy for N electrons can be written as

$$E(N) = \frac{\hbar^2}{2m_e}\left(\frac{3\pi^2 N}{V}\right)^{2/3}, \tag{6.12}$$

from which we get the total number of states $N(E)$ for a given highest energy E and with this

$$g(E) = \frac{dN}{dE} = \frac{V}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2} E^{1/2}. \tag{6.13}$$

This density of states for the free electron model is shown in Figure 6.4a.

So far, we have only considered the situation at zero temperature. At any finite temperature, electrons will be thermally excited from their ground state. For fermions, the occupation probability of the states is given by the Fermi–Dirac distribution $f(E, T)$

$$f(E, T) = \frac{1}{e^{(E-\mu)/k_B T} + 1}, \tag{6.14}$$

where μ is the **chemical potential**. For metals, we can set $\mu = E_F$ and do not distinguish between the Fermi energy and the chemical potential at all. At zero temperature, this is exactly true, and at finite temperature, it is a very good approximation. The Fermi–Dirac distribution is shown in Figure 6.4b. At zero temperature, it is represented by the dashed line that has a value of 1 for energies smaller than μ and 0 for energies higher than μ, meaning that all states below the chemical potential are occupied and all others are empty. This is consistent with our above discussion of how the states are filled. At finite temperature, the Fermi–Dirac distribution develops a "soft zone" around μ in which the occupation probability is no longer 1 or 0 but something in between. The soft zone is symmetric around μ (or E_F), and it has a width of about $4k_B T$. At room temperature, $k_B T \approx 25$ meV, such that the soft zone is around 100 meV wide.

It is instructive to compare the electrons' mean kinetic energy in the quantum model to the result of the Drude model. In the quantum model, it must be some fraction of the Fermi energy E_F (see Problem 6.1), whereas it is given by (5.1) in the Drude model. The most important difference is not that the kinetic energy

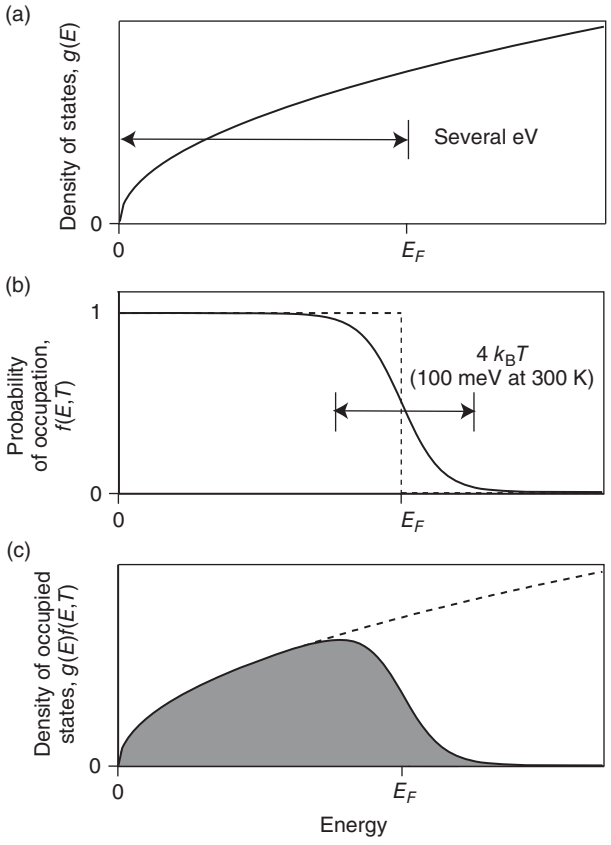


Figure 6.4 (a) Density of states for a free electron gas $g(E)$. (b) Fermi–Dirac distribution function $f(E, T)$ at $T = 0$ (dashed line) and at a finite temperature (solid line). (c) Density of occupied states $g(E)f(E, T)$. Note that the relative width of Fermi distribution's "soft zone" ($\approx 4k_BT$) is exaggerated in the sketch for temperatures around room temperature.

in the quantum model is fairly high but that it is so (almost) *independent of the temperature*.

Finally, the **density of occupied electron states** at a given energy and temperature can be found by multiplying the density of states $g(E)$ with the Fermi–Dirac distribution $f(E, T)$, see Figure 6.4c. This definition gives us a way of calculating the chemical potential at any temperature because the total number of electrons N must be given by

$$N = \int_0^\infty g(E)f(E, T)dE. \tag{6.15}$$

As pointed out earlier, however, the chemical potential in a metal depends only very weakly on the temperature (see Problem 7.3).

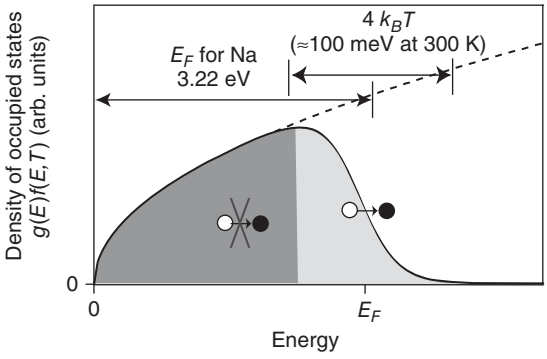


Figure 6.5 Most of the electrons in a metal (roughly those in the dark gray zone) cannot change their energy by a small amount because the reachable states are already occupied by other electrons. As in Figure 6.4, the width of the Fermi–Dirac distribution's "soft zone" is not drawn to scale.

It is important to notice how different the energy scales are. The Fermi energy is several electron volts, while the soft zone of the Fermi–Dirac distribution is only 100 meV wide at room temperature. This means that the relative number of electrons in the soft zone is very small indeed. This turns out to be the key to understanding many properties of metals, for example, their heat capacity.

6.2.2 Electronic Heat Capacity

The fact that the Dulong–Petit rule is not only valid for insulators but also for many metals (see Table 4.2) suggests that the contribution of the free electrons to the heat capacity of a metal is very small. The Drude model does not explain this, but now we can understand why. When the temperature of the solid is raised, only a very small fraction of the electrons can be thermally excited. This is illustrated by Figure 6.5. Assume that the temperature of the solid is raised from zero to some finite temperature T. Classical particles would increase their kinetic energy by $3k_BT/2$. Here, this is impossible for most of the electrons because they are trapped: There are already other electrons occupying the states at slightly higher energies. A contribution to the heat capacity is, in fact, only possible for the electrons near the Fermi energy.

Let us try a "quick-and-dirty" estimate of the electronic heat capacity. The number of electrons in the soft zone is of the order $k_BTg(E_F)$. If we say that the mean thermal energy of these electrons is $3k_BT/2$, the total mean thermal energy is

$$\langle E\rangle = \frac{3}{2}k_BTg(E_F)k_BT \tag{6.16}$$

plus some offset, which does not depend on the temperature. This gives

$$C = \frac{\partial\langle E\rangle}{\partial T} = 3k_B^2Tg(E_F). \tag{6.17}$$

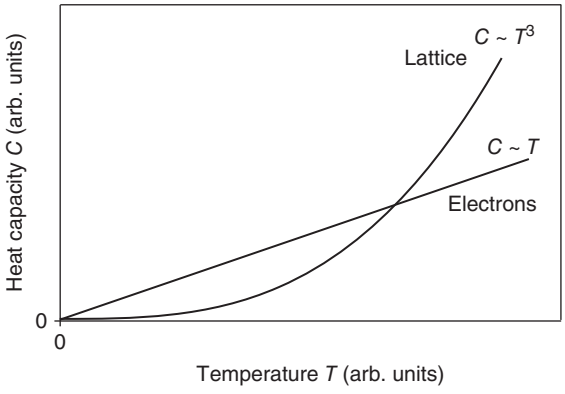


Figure 6.6 Sketch of the electronic and lattice contributions to the heat capacity. At sufficiently low temperatures, the electronic contribution dominates.

This is quite close to the correct result, which is

$$C = \frac{\pi^2}{3} k_B^2 T g(E_F) \tag{6.18}$$

(see Problem 6.3). This expression has a number of interesting implications. The heat capacity is proportional to the density of states at the Fermi energy $g(E_F)$. This is easy to understand because only the electrons close to the Fermi energy can participate in thermal excitations. Since these electrons constitute only a small fraction of all electrons, the free electrons in a metal do not usually lead to a strong deviation from the Dulong–Petit behavior at high temperatures. On the other hand, (6.18) is linear in T, whereas the (low temperature) heat capacity of the lattice (4.45) is proportional to T^3. This means that at low temperatures, the lattice contribution vanishes faster than the electronic contribution and the latter can, in fact, be measured. This is illustrated in Figure 6.6 (see Problem 6.3 to calculate the cross-over temperature).

6.2.3
The Wiedemann–Franz Law

The free electron model correctly reproduces the Wiedemann–Franz law and also gives the correct Lorenz number L. This can be seen by inserting the appropriate expressions into (5.30). The thermal conductivity can be taken to have the same form as (4.48) with appropriate modifications for the velocity, which should be the Fermi velocity, and the heat capacity. For the electrical conductivity, we take the expression from the Drude model. Both conductivities can be written such that they contain the relaxation time τ. We do not know anything about τ, but fortunately, it cancels out in the final expression. Working out the details of this is left to the reader (see Problem 6.4). The final result is

$$\frac{\kappa}{\sigma} = \frac{\pi^2}{3} \frac{k_B^2}{e^2} T = LT. \tag{6.19}$$

This gives $L = 2.45 \times 10^{-8}\ \mathrm{W\Omega\,K^{-2}}$, which agrees very well with the experimental data for many metals.

6.2.4 Screening

An important feature of metals is their ability to screen out electric fields. In fact, for the purpose of classical electrostatic field theory, it is commonly assumed that the metals are internally field-free. In the Drude model, we have seen that this is also a good approximation for AC electric fields with frequencies below the plasma frequency. On the atomic scale, this is not quite so simple, but a metal is still very effective in screening out external fields. Before we describe this quantitatively, we develop a simple picture of the screening effect. Consider first the Coulomb potential due to a positive point charge q in vacuum:

$$\phi_0(r) = \frac{1}{4\pi\epsilon_0}\frac{q}{r}, \tag{6.20}$$

where r is the distance from the charge. As we will see in Chapter 9, this result does not change very much when we put the point charge into an insulator. We only have to substitute ϵ_0 with $\epsilon\epsilon_0$, where ϵ is the dielectric constant of the insulator. The total potential is thus reduced by a constant factor of ϵ.

However, if we put the positive point charge into a metal, it will attract the surrounding electrons. In sharp contrast to the situation in an insulator, these electrons are free to move toward the point charge. This leads to a negative electron cloud around the point charge, strongly reducing the total potential at larger distances. This is the effect of metallic screening that we will now describe quantitatively. We consider a positive point charge in a metal with the potential given by (6.20). This potential is spherically symmetric, and it leads to an electron cloud around the impurity with a potential $\phi_s(r)$ that also has spherical symmetry. Superposition dictates that the total potential is $\phi(r) = \phi_0(r) + \phi_s(r)$.

If we assume that $e\phi(r)$ is small compared to the Fermi energy E_F, that $\phi(r)$ is slowly varying in space, and that the temperature is $T = 0$ K, the screening can be described by the picture in Figure 6.7. Far away from the impurity, where $\phi(r)$ is essentially zero, the free electron metal states are filled up to the Fermi energy. Close to the impurity, however, the electrons "feel" the additional electrostatic energy from $e\phi(r)$, such that the energies of all the states are lowered. This corresponds to a shift of the density of states to lower energies (for a positive point charge). This, in turn, leads to a situation where electrons from the rest of the metal can move to the available lower energy states close to the impurity and occupy them (light gray area on the figure) until equilibrium is reached. In the vicinity of r, the accumulated charge density is thus given by $-e$ times the size of the light gray area, resulting in

$$\rho(r) = -e^2(1/V)g(E_F)\phi(r). \tag{6.21}$$

One might think that this flow of charge would have to lead to a charge reduction in the rest of the material. This is also correct, but the volume of the solid is

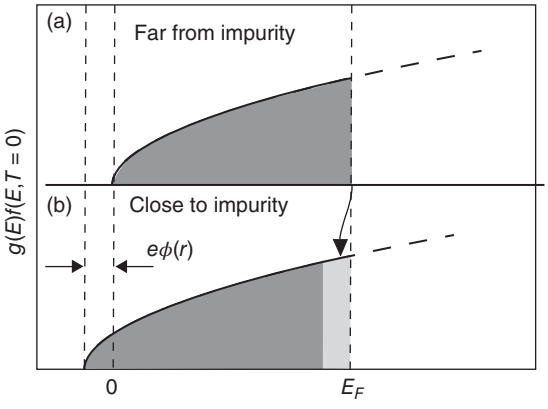


Figure 6.7 Screening of a positively charged impurity in a metal. (a) The occupied density of states of a free electron metal at $T = 0$ K. (b) A positively charged point impurity locally shifts the energy of the electronic states. This allows electrons from the rest of the metal to flow into the newly available states below E_F (light gray area).

assumed to be very large compared to the area around the impurity such that this reduction can be neglected.

While we now know the charge density created by the total potential, we still do not know the potential. We can find it using the Poisson equation $\nabla^2\phi(\mathbf{r}) = -\rho(\mathbf{r})/\epsilon_0$ and (6.21). Since we have spherical symmetry, everything only depends on the distance r and there is no angular dependence. The easiest way to solve the problem is to write the Laplace operator ∇^2 in spherical coordinates and use that $\phi(\mathbf{r})$ depends only on r, not on the direction. We then obtain

$$\nabla^2\phi(\mathbf{r}) = \frac{\partial^2\phi(r)}{\partial r^2} + \frac{2}{r}\frac{\partial\phi(r)}{\partial r} = \frac{e^2}{V\epsilon_0}g(E_F)\phi(r). \tag{6.22}$$

We now need to find a solution to this differential equation, and it is easy to show (by inserting into (6.22)) that such a solution is

$$\phi(r) = c\frac{1}{r}e^{-r/r_{\mathrm{TF}}}, \tag{6.23}$$

where c is a constant and r_{TF} is the so-called **Thomas–Fermi screening length** that is given by

$$r_{\mathrm{TF}} = \sqrt{\frac{V\epsilon_0}{e^2 g(E_F)}}. \tag{6.24}$$

The constant c in (6.23) can be fixed by requiring that the bare Coulomb potential for the positive point charge $\phi_0(r)$ in (6.20) is recovered for a situation in which $g(E_F)$ goes toward zero, that is, when the metal becomes similar to vacuum. For a small $g(E)$, r_{TF} would be very large and the exponential function in (6.23) would approach unity. Therefore, we have to choose c such that

$$\phi(r) = \frac{1}{4\pi\epsilon_0}\frac{q}{r}e^{-r/r_{\mathrm{TF}}}, \tag{6.25}$$

which is the final result.

In most metals r_{TF} is very small, on the order of 1 Å, and the exponential part in (6.25) causes the screened potential to decay on the same length scale, much faster than the bare Coulomb potential (see Figure 6.8). This confirms the expected result that an electrostatic potential is screened out very rapidly in a metal. In fact, the effective screening can be used as an argument to explain why the electrons are free within a metal in the first place. With such an effective screening, it is not possible to localize them close to the potential of an ion because the screened ionic potential is too weak.

6.3
The General Form of the Electronic States

The free electron model appears to describe certain phenomena quite well, but it still has some obvious shortcomings. In particular, it appears to be a fair description for metals, but what about nonmetallic compounds such as diamond or silicon? As we have seen in Figure 6.2, their characteristic is that the sp^3 band is completely filled and that there are no states immediately above the top of the band, in which electrons can be excited. This is not captured by the free electron model that gives a continuum of states from the lowest energy to infinity. One could argue that even for a metal, this might not be correct: As we have seen in Figure 6.1d, the 3s band for Na is half-filled but it also has a finite width, that is, there are no states at all possible energies. This is hardly relevant for most physical phenomena such as conduction or heat capacity in which the excitation energies of

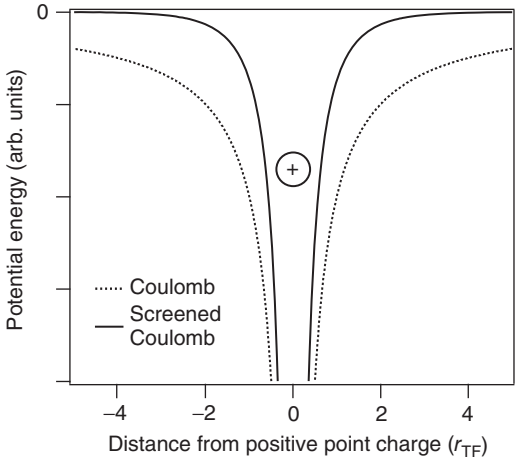


Figure 6.8 Potential due to a positive point charge in a metal compared to the Coulomb potential in free space.

the electrons are very small compared to the band width. But even in such cases, the free electron model poses some problems. Take, for example, Al, which is a simple metal (not a transition metal) and should be described quite well by the free electron model. Yet, not even the sign of the Hall coefficient in Table 5.1 is correct and going from the classical to the quantum free electron description does not cure this problem.

Finally, the quantum mechanical free electron model is still not able to resolve some of the most basic questions of electron motion in solids. One example is that the mean free path of the electrons can reach macroscopic distances at low temperatures, and we do not understand how the electrons can move through the lattice of ions without scattering. The free electron model does not help here since it simply ignores the presence of the ions, but the question remains valid.

In order to make some progress, we have to describe the motion of the electrons in a nonvanishing lattice-periodic potential (6.2) and solve (6.1). F. Bloch showed that the general wave function solving this problem has the simple form

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}), \tag{6.26}$$

where $u_{\mathbf{k}}(\mathbf{r})$ is a function with the periodicity of the Bravais lattice:

$$u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r} + \mathbf{R}). \tag{6.27}$$

Be careful not to confuse $u_{\mathbf{k}}(\mathbf{r})$ with the lattice-periodic potential $U(\mathbf{r})$! The index $\mathbf{k}$ refers to the fact that the function $u_{\mathbf{k}}(\mathbf{r})$ can be changing, depending on the wave vector $\mathbf{k}$. One often refers to (6.26) as **Bloch's theorem** and calls $\psi_{\mathbf{k}}(\mathbf{r})$ a **Bloch wave function**. Another way of stating Bloch's theorem is to use the lattice periodicity of $u_{\mathbf{k}}(\mathbf{r})$ and require that

$$\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{R}} \psi_{\mathbf{k}}(\mathbf{r}). \tag{6.28}$$

Before we prove Bloch's theorem, we mention one of its most important consequences. The solution (6.26) is very similar to the free electron solution; it is a plane wave modulated by a lattice-periodic function. This is an amazing fact because it means that the electronic states are spread out over the whole crystal, even if we turn on the lattice-periodic potential! In fact, the probability density for finding an electron may vary within one unit cell, but it is identical for the corresponding positions within every unit cell in the solid (see Problem 6.6). This means that the electrons travel through the crystal without bouncing into the lattice ions at all, immediately explaining the possibility of a very long mean free path, much longer than the distance between the ions. In fact, if the electrons in a metal are not scattered by the ions, we could expect the resistivity of a perfectly periodic metal crystal to be zero. We will later see that this would indeed be the case, and we will discuss which mechanisms cause a finite resistivity.

We now prove Bloch's theorem (6.26). Like in the free electron model, we use a cubic crystal of side length L and periodic boundary conditions. The allowed values of $\mathbf{k}$ are then given by (6.5). Every solution of the Schrödinger equation (6.1) consistent with these boundary conditions can be written as a sum of plane waves:

$$\psi(\mathbf{r}) = \sum_{\mathbf{k}} c_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{6.29}$$

where $\mathbf{k}$ are the values consistent with the boundary conditions, and the coefficients $c_{\mathbf{k}}$ are assumed to take care of the wave function's normalization. The lattice-periodic potential can also be written as a Fourier series, using the reciprocal lattice vectors $\mathbf{G}$:

$$U(\mathbf{r}) = \sum_{\mathbf{G}} U_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}. \tag{6.30}$$

Since we want the potential to be a real quantity, we must further require that

$$U_{-\mathbf{G}} = U_{\mathbf{G}}^{*}. \tag{6.31}$$

The two expansions can now be inserted into the Schrödinger equation (6.1). The kinetic energy term then becomes

$$-\frac{\hbar^2\nabla^2}{2m_e}\psi(\mathbf{r}) = \sum_{\mathbf{k}} \frac{\hbar^2 k^2}{2m_e} c_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{6.32}$$

The potential energy term becomes

$$\begin{aligned} U(\mathbf{r})\psi(\mathbf{r}) = \left(\sum_{\mathbf{G}} U_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}\right)\left(\sum_{\mathbf{k}} c_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}\right) &= \sum_{\mathbf{kG}} U_{\mathbf{G}} c_{\mathbf{k}} e^{i(\mathbf{G}+\mathbf{k})\cdot\mathbf{r}} \\ &= \sum_{\mathbf{k'G}} U_{\mathbf{G}} c_{\mathbf{k'}-\mathbf{G}} e^{i\mathbf{k'}\cdot\mathbf{r}}. \end{aligned} \tag{6.33}$$

In the last step, we have changed the summation index from $\mathbf{k}$ to $\mathbf{k}' = \mathbf{G} + \mathbf{k}$ in order to obtain the same plane wave form as in the expression for the kinetic energy. We are allowed to "shift" the indices by reciprocal lattice vectors as we wish since the sum in question extends over all wave vectors consistent with the boundary conditions, and the reciprocal lattice vectors are clearly a subset of these. If we now rename the index $\mathbf{k}'$ in the potential energy expression back to $\mathbf{k}$, we can write the whole Schrödinger equation in the new form:

$$\sum_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} \left\{ \left(\frac{\hbar^2 k^2}{2m_e} - E \right) c_{\mathbf{k}} + \sum_{\mathbf{G}} U_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} \right\} = 0. \tag{6.34}$$

Since the plane waves with different $\mathbf{k}$ are orthogonal, every coefficient in the equation has to vanish in order for the sum to vanish. So, the Schrödinger equation is reduced to a set of equations:

$$\left(\frac{\hbar^2 k^2}{2m_e} - E \right) c_{\mathbf{k}} + \sum_{\mathbf{G}} U_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} = 0. \tag{6.35}$$

Since the summation in (6.34) runs over all $\mathbf{k}$ consistent with the periodic boundary conditions, we could choose $\mathbf{k}$ in (6.35) to lie in the first Brillouin zone. Technically, the sum over the reciprocal lattice in (6.35) is infinite. In practice, however, the potential can often be described by very few nonzero Fourier coefficients $U_{\mathbf{G}}$, so that the sum is rather short. Equation (6.35) then gives a relation between $c_{\mathbf{k}}$ and the values $c_{\mathbf{k}-\mathbf{G}}$ for which $U_{\mathbf{G}} \neq 0$. It is clear that there will be similar equations

for each of these $c_{\mathbf{k}-\mathbf{G}}$ coefficients. If, for example, $U_{\mathbf{G}'} \neq 0$, we will also have to consider the equation

$$\left(\frac{\hbar^2(|\mathbf{k}-\mathbf{G}'|^2}{2m_e} - E\right)c_{\mathbf{k}-\mathbf{G}'} + \sum_{\mathbf{G}} U_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}'-\mathbf{G}} = 0. \tag{6.36}$$

The task is then to find the set of coefficients $c_{\mathbf{k}}, c_{\mathbf{k}+\mathbf{G}}, c_{\mathbf{k}-\mathbf{G}} \ldots$ that solves all these equations simultaneously. This will be illustrated in the next section.

The problem of solving the Schrödinger equation is now reduced to solving a set of equations such as (6.35), (6.36)... for every $\mathbf{k}$ in the first Brillouin zone. For a given $\mathbf{k}$, these only contain the coefficients $c_{\mathbf{k}}, c_{\mathbf{k}+\mathbf{G}}, c_{\mathbf{k}-\mathbf{G}} \ldots$, determining only these coefficients. This means that for a certain $\mathbf{k}$, the wave function (6.29) also only contains nonvanishing terms with these coefficients and, therefore, it can be written as

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{i(\mathbf{k}-\mathbf{G})\cdot\mathbf{r}}. \tag{6.37}$$

This is equivalent to

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}}\left(\sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}}\right). \tag{6.38}$$

We now realize that the term in brackets is a Fourier series over the reciprocal lattice vectors and, therefore, a lattice-periodic function, hence we have proven Bloch's theorem.

From this proof, we immediately obtain another important property of the Bloch functions. If we take (6.37) and shift $\mathbf{k}$ by an arbitrary reciprocal lattice vector $\mathbf{G}'$, we get

$$\psi_{\mathbf{k}+\mathbf{G}'}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}+\mathbf{G}'} e^{i(\mathbf{k}-\mathbf{G}+\mathbf{G}')\cdot\mathbf{r}} = \sum_{\mathbf{G}''} c_{\mathbf{k}-\mathbf{G}''} e^{i(\mathbf{k}-\mathbf{G}'')\cdot\mathbf{r}}, \tag{6.39}$$

with $\mathbf{G}'' = \mathbf{G} - \mathbf{G}'$. We still sum over all reciprocal lattice vectors, so this is exactly the same as the $\psi_{\mathbf{k}}(\mathbf{r})$ we started with. Therefore,

$$\psi_{\mathbf{k}+\mathbf{G}'} = \psi_{\mathbf{k}}(\mathbf{r}), \tag{6.40}$$

and when we insert this in the Schrödinger equation, we get

$$E(\mathbf{k}+\mathbf{G}') = E(\mathbf{k}). \tag{6.41}$$

The periodicity of the potential in real space translates into a periodicity of the solutions in reciprocal space, something that we had also seen for the lattice vibrations.

6.4 Nearly Free Electron Model

Our proof of Bloch's theorem has also given us a rewritten form of the Schrödinger equation (6.35). Remarkably, all we need to do in order to determine the electronic wave functions and their energies for any three-dimensional solid is to find the

correct coefficients $c_{\mathbf{k}}$, assuming that we already know the potential. The crucial difficulty is of course that we do not know the potential for real solids.

However, it is most instructive to solve the (6.35) for a one-dimensional solid with a lattice constant a, assuming a simple potential. The reciprocal lattice for this solid is spanned by "vectors" of length $g = 2\pi/a$ and the potential can be written as a Fourier series

$$U(x) = \sum_n U_n e^{ingx}, \tag{6.42}$$

where the sum runs over all integers. We shall use a very simple potential: U_0 can be set to zero because a constant potential offset does not change anything apart from a rigid shift of the energy eigenvalues. The only coefficients we are going to keep are $U_1 = U_{-1}$, and we call them simply U.

We start out by using a very small U. In practice, this means that we are treating free electrons but with the symmetry of the lattice. For a given k, we can write down many equations of the type (6.35). We require that only c_k, c_{k-g}, and c_{k+g} are different from zero, giving us a set of three equations. At present, there is no justification for this and we explore the consequences of including more coefficients and equations further down. We get

$$\begin{aligned}
\left(\frac{\hbar^2(k-g)^2}{2m_e} - E\right)c_{k-g} + Uc_k &= 0,\\
\left(\frac{\hbar^2k^2}{2m_e} - E\right)c_k + Uc_{k-g} + Uc_{k+g} &= 0,\\
\left(\frac{\hbar^2(k+g)^2}{2m_e} - E\right)c_{k+g} + Uc_k &= 0.
\end{aligned} \tag{6.43}$$

This is a linear system of equations that has three solutions for every value of k. The solutions are shown in Figure 6.9a. We find three parabolas that are identical to the free electron result in Figure 6.3. The parabolas are centered on the reciprocal lattice points 0, g, and $-g$. This periodicity is expected from (6.41). The only obvious problem is that there are no parabolas centered on higher order reciprocal lattice vectors, such as $2g$ and $-2g$, and this is actually caused by the fact that we have only used three coefficients and three equations in (6.43). We can extend (6.43) to five equations with five coefficients by also considering c_{k-2g} and c_{k+2g}. The result of this calculation is shown in Figure 6.9b. It is essentially the same as in Figure 6.9a, only that we now also have parabolas centered on $2g$ and $-2g$ (but still none at the higher reciprocal lattice points). We see that neglecting higher coefficients has two consequences in the present case: We do not get the correct result outside the first Brillouin zone and we do not get the correct result in the first Brillouin zone at high energies, because we lack the parabolas from the neighboring Brillouin zones, which reach back into the first Brillouin zone. In any case, the result is very much like the free electron result, but it also fulfills the symmetry requirement (6.41) imposed by the lattice, at least to some extent. If we only concentrate on the first Brillouin zone, the periodicity (6.41) has the effect that the parabolas appear to be **back-folded** at the Brillouin zone boundary, such that the

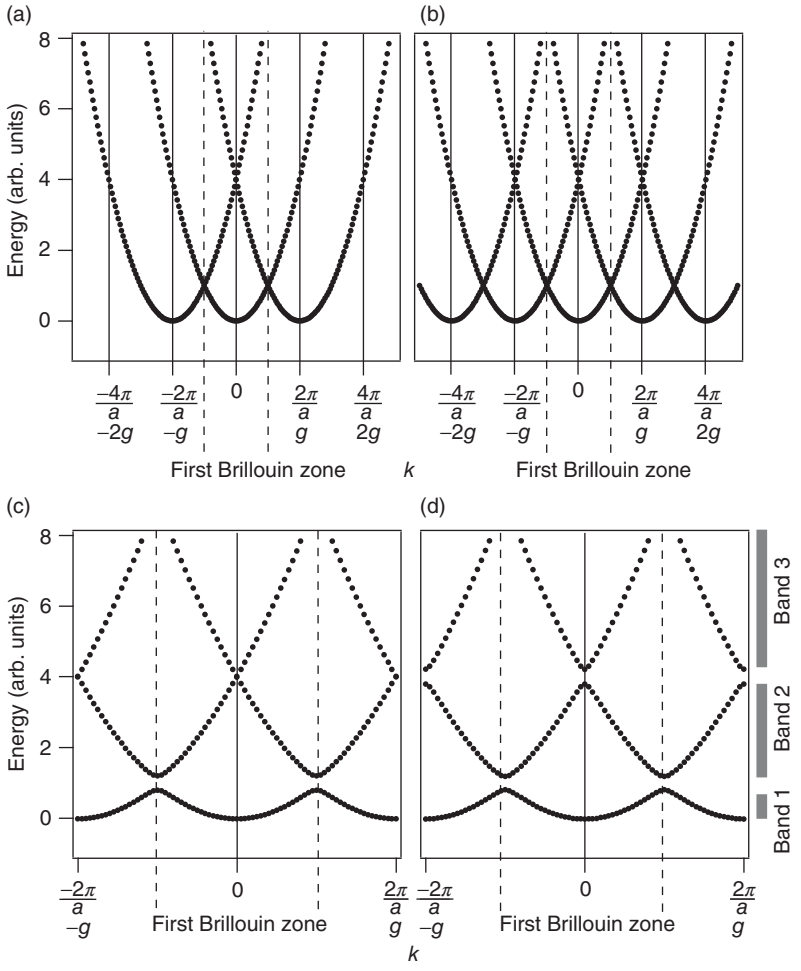


Figure 6.9 Electronic states in the nearly free electron model for a one-dimensional chain with unit cell length a. (a) Solutions for three equations (6.43), using a nearly vanishing $U = U_1 = U_{-1}$. (b) Solutions of five equations similar to (6.43), using a nearly vanishing $U = U_1 = U_{-1}$. (c) Same as (b) but for a larger value of U. (d) Same as (b) but for larger values for both $U_1 = U_{-1}$ and $U_2 = U_{-2}$. The gray bars symbolize the ranges where a quasi-continuum of energies is available (bands). In between these, there are band gaps.

second lowest band in the first Brillouin zone is essentially the same band as the first, but originating in the neighboring zone.

What happens when we now turn on the lattice potential U? This is shown in Figure 6.9c, again for the case of five equations of the type (6.35). The main consequence of the finite U is the opening of gaps between the parabolas at the Brillouin zone boundary; the other states are largely unaffected and appear still very much

free electron like. However, this is a major change from the free electron model because it means that the solid no longer has a continuum of states from the lowest energy to infinity. In fact, we obtain a **band gap**, a range of energies for which there are no states at all.

The effect of turning on higher order contributions of the potential is shown in Figure 6.9d. Here, not only $U_1 = U_{-1}$ but also $U_2 = U_{-2}$ are chosen to have a finite value. The main effect of a finite $U_2 = U_{-2}$ is an additional gap opening, now at the crossing points of the parabolas at higher energies at the center of the Brillouin zone ($k = 0$).

Quite generally, the solution to a system of equations like (6.43) gives us n energy eigenvalues for every value of k, or n relations of the type $E_n(k)$. n runs over all positive integers and k over the values consistent with the periodic boundary conditions. The relations $E_n(k)$ are the **dispersion relations** for the electronic states and usually called the **electronic band structure** of the solid. In this context, n plays the role of a band index. In Figure 6.9d, for example, we see the two lowest bands and a part of the third band.

$\mathbf{k}$ can still be interpreted as the wave vector of the Bloch wave, but we can also view it as a quantum number of the electronic states, in complete analogy with the role of $\mathbf{k}$ in the case of the lattice vibrations. Compare the situation to atomic physics. There n, the main quantum number, specifies the shell containing the electrons. The other quantum numbers l and m are the parameters of the spherical harmonics functions, which describe the angular part. In this sense, l and m are quantum numbers that are related to the spherical symmetry of the atom. In the solid, the symmetry is given by the periodic lattice and $\mathbf{k}$ can be viewed as the quantum number related to this symmetry.

We have now two different interpretations of $\mathbf{k}$. It can be viewed as the wave vector of the Bloch wave or as a quantum number describing the state containing the electron. It is also very tempting also to interpret $\hbar\mathbf{k}$ as the momentum of the electron, as in the case of free electrons. But this is *wrong*. This can be seen quite easily. We apply the momentum operator $-i\hbar\nabla$ on the Bloch wave (6.26) to obtain

$$-i\hbar\nabla\psi_{\mathbf{k}}(\mathbf{r}) = \hbar\mathbf{k}\psi_{\mathbf{k}}(\mathbf{r}) - e^{i\mathbf{k}\cdot\mathbf{r}}i\hbar\nabla u_{\mathbf{k}}(\mathbf{r}). \tag{6.44}$$

We see that $\hbar\mathbf{k}$ is only an eigenvalue to the momentum operator when $u_{\mathbf{k}}(\mathbf{r})$ is constant, that is, when the Bloch wave is a free electron wave. In fact, we already know that $\hbar\mathbf{k}$ cannot be the momentum of a Bloch wave because the states do not change if we add or subtract a reciprocal lattice vector (see (6.40) and (6.41)), something that we have also seen in the discussion of phonons. However, $\hbar\mathbf{k}$ is still a useful quantity because conservation rules for $\mathbf{k}$ apply for scattering processes in solids. Instead of simply "momentum", $\hbar\mathbf{k}$ it is called the **crystal momentum**. In contrast to regular momentum that is completely conserved, crystal momentum can only be conserved within a reciprocal lattice vector. As an example, consider a process in which an electron with energy E and wave vector $\mathbf{k}$ is scattered by absorbing a phonon with $\hbar\omega$ and $\mathbf{q}$. The scattered electron has an energy $E + \hbar\omega$ and a wave vector $\mathbf{k} + \mathbf{q} + \mathbf{G}$. Therefore, we have a conservation of crystal momentum (or wave vector sum) that is very similar to momentum conservation. We will

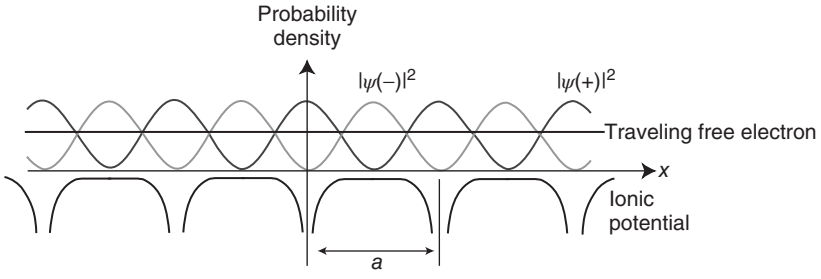


Figure 6.10 Qualitative explanation for the gap openings at the Brillouin zone boundary. Shown are the probability densities for two possible standing electron waves with k corresponding to the zone boundary π/a. These are either accumulated or depleted in the vicinity of the ion cores compared to a traveling free electron wave that has a constant probability density.

understand the meaning of **k** somewhat better when we discuss the transport of electricity via Bloch states.

The occurrence of energy gaps at the Brillouin zone boundary can also be made plausible by a very simple argument. Consider a free electron traveling perpendicular to a set of lattice planes separated by a distance a. If x is the direction perpendicular to the planes, the electron has the wave function $\psi(x) \propto e^{ikx}$ in this direction, that is, it behaves like a plane wave with a wavelength $\lambda = 2\pi/k$. Such a wave fulfills the Bragg condition (1.3) for a value of $k = n\pi/a$. This means that the lattice will reflect the wave back to some degree. Since the solid is very big, the amplitude of the back-reflected wave will eventually be the same as for the forward-moving wave, so that the total wave function has the form $\psi(x) \propto e^{ikx} + Ae^{-ikx}$ with $|A| = 1$. The left/right symmetry of the crystal assumed to be present here also requires A to be real and so there are two possible results:

$$\psi(+) \propto e^{i(\pi/a)x} + e^{-i(\pi/a)x} = 2\cos\left(\frac{\pi}{a}x\right), \tag{6.45}$$

$$\psi(-) \propto e^{i(\pi/a)x} - e^{-i(\pi/a)x} = 2i\sin\left(\frac{\pi}{a}x\right). \tag{6.46}$$

Both represent standing electrons waves. Their probability densities $|\psi(+)|^2$ and $|\psi(-)|^2$ are shifted with respect to the positive ion potential, as shown in Figure 6.10. $\psi(+)$ shows an accumulation of probability near the ion cores, whereas $\psi(-)$ shows a depletion. Therefore, $\psi(+)$ has a lower energy than $\psi(-)$ even though both have the same wave vector $k = n\pi/a$. $\psi(+)$ and $\psi(-)$ thus correspond to the solution just below and above the energy gap at the Brillouin zone boundary, respectively. Note that these probability densities are quite different from the case of a free electron wave where $|\psi|^2$ it is constant.

When discussing lattice vibrations, we have stated that the group velocity of the lattice waves is given by $d\omega/dk$, where ω is the frequency of the wave and k the wave vector. This expression for the group velocity is of very general character in

the theory of waves, and it can be shown that it also holds for Bloch waves. There it is convenient to write it as

$$v_g = \frac{d\omega(k)}{dk} = \frac{1}{\hbar}\frac{dE(k)}{dk}. \tag{6.47}$$

In other words, the group velocity is given by the slope of the bands. If we now consider Figure 6.9c and d, we see that the group velocity of the bands with a finite periodic potential is zero at the Brillouin zone boundaries. This means that we have standing waves there, perfectly consistent with the argument we have just made. Note that a group velocity of zero has to be seen in a quantum mechanical sense. If we could measure the velocity v_g of an electron wave packet at the zone boundary, the expectation value would be zero, so we could not say if it moves to the right or the left side. However, this does not mean that the electron does not move. The expectation value for the kinetic energy is not zero.

6.5 Tight-binding Model

We have started this chapter by discussing a qualitative model for the electronic structure of solids in which atomic energy levels from very many atoms were combined to give a continuous band of states. For a quantitative description, however, we have abandoned this picture and treated the electrons first as entirely free and then as nearly free. This did indeed lead to a quasi-continuous distribution of energy levels with gaps in between them. We now return to the description that starts with atomic states by constructing a Bloch wave function through a **linear combination of atomic orbitals**. This method is known as the **tight-binding approach**. The nearly free electron approach from the last section is a more natural starting point to describe metals, while the tight-binding approach is the obvious starting point for covalently bonded crystals or for the more localized electrons in metals, such as the d electrons in transition metals. Eventually, both are mere approximations and refining them will lead to the same result from both ends. But discussing the tight-binding approach here gives us some deeper insight into the meaning of the band structure of solids.

We sketch the tight-binding approximation in its simplest form. We start with the Hamiltonian for the atoms making up the solid (considering only one kind of atom for simplicity). It is given by

$$H_{\mathrm{at}} = -\frac{\hbar^2\nabla^2}{2m_e} + V_{\mathrm{at}}(\mathbf{r}), \tag{6.48}$$

where V_{at} is the atomic one-electron potential. Atoms have different energy levels E_n and corresponding wave functions. When we put the atoms together to form a solid, we expect that each energy level turns into a band in the solid. We could, for example, think of the Na atoms from the beginning of the chapter and consider the band that is derived from the 3s state with the energy $E_{3\mathrm{s}}$ and the wave function $\phi_{3\mathrm{s}}(\mathbf{r})$.

If we have an atom on every point **R** of the Bravais lattice, the Hamiltonian for the solid can be written as

$$H_{\text{sol}} = -\frac{\hbar^2\nabla^2}{2m_e} + \sum_{\mathbf{R}} V_{\text{at}}(\mathbf{r}-\mathbf{R}) = -\frac{\hbar^2\nabla^2}{2m_e} + V_{\text{at}}(\mathbf{r}) + \sum_{\mathbf{R}\neq 0} V_{\text{at}}(\mathbf{r}-\mathbf{R}). \quad (6.49)$$

The first term is the kinetic energy of the single electron we consider; the second is the sum of the atomic potentials of all the atoms in the solid. The potential in this Hamiltonian has the periodicity of the lattice, as it must. The right-hand side of the equation shows that we can split this potential up in any way we like, for example, as the potential of the atom at the origin $V_{\text{at}}(\mathbf{r})$ plus the potential of the rest of the solid. This can also be written as

$$H_{\text{sol}} = -\frac{\hbar^2\nabla^2}{2m_e} + V_{\text{at}}(\mathbf{r}) + \nu(\mathbf{r}) = H_{\text{at}} + \nu(\mathbf{r}), \quad (6.50)$$

where

$$\nu(\mathbf{r}) = \sum_{\mathbf{R}\neq 0} V_{\text{at}}(\mathbf{r}-\mathbf{R}). \quad (6.51)$$

This can be viewed as the Hamiltonian for an atom at the origin plus some correction potential from all the other atoms. Consider the situation in which the atoms are quite far from each other. In this case, we can try to use the atomic wave functions $\phi_n(\mathbf{r})$ belonging to the atomic energy levels E_n to calculate the energy eigenvalues of the solid. We obtain

$$\int \phi_n^*(\mathbf{r}) H_{\text{sol}} \phi_n(\mathbf{r}) d\mathbf{r} = E_n + \int \phi_n^*(\mathbf{r})\nu(\mathbf{r})\phi_n(\mathbf{r})d\mathbf{r} = E_n - \beta, \quad (6.52)$$

where $-\beta$ is a small shift of the atomic energy level due to the presence of the other atoms' potentials. If the atoms are sufficiently far away from each other, $\beta = 0$ because the wave function $\phi_n(\mathbf{r})$ will have dropped to zero before the potential $\nu(\mathbf{r})$ from the neighboring atoms at $\mathbf{R} \neq 0$ becomes appreciably larger than zero. It is easy to see that the atomic wave function centered on any other site **R** will also solve the Schrödinger equation for the Hamiltonian (6.49). We merely have to rewrite the Hamiltonian such that it is centered on the atom at **R** plus the potential from all the other atoms. So, the result of this treatment is that, for a solid of N atoms, we obtain N degenerate solutions for every energy eigenvalue of the atomic Hamiltonian. This is of course what one would expect if the atoms are placed so far from each other that they do not interact. The "band structure" of this result would be consisting of a "band" at the energy E_n with no dispersion at all.

We now discuss the more interesting situation where there is some interaction between the neighboring atoms. We write the wave function of the solid as linear combination of the atomic wave functions on every lattice site **R**

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} c_{\mathbf{k},\mathbf{R}} \phi_n(\mathbf{r}-\mathbf{R}). \quad (6.53)$$

The usefulness of normalization factor $1/\sqrt{N}$ will become apparent later. The coefficients $c_{\mathbf{k},\mathbf{R}}$ are yet to be determined. They will depend on the wave vector **k**. It might not be entirely correct to use the atomic wave functions $\phi_n(\mathbf{r}-\mathbf{R})$

here because the presence of the other atoms could modify these wave functions slightly. We choose to ignore this for simplicity.

The coefficients $c_{\mathbf{k},\mathbf{R}}$ are now determined by the requirement that (6.53) must have the character of a Bloch wave if it is to be a solution of (6.49). This is achieved by choosing the coefficients such that (6.53) turns into

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} \phi_n(\mathbf{r} - \mathbf{R}), \tag{6.54}$$

where $\mathbf{k}$ takes the values permitted by the periodic boundary conditions (6.5). This wave function fulfills the Bloch condition as stated in (6.28) because

$$\begin{aligned}
\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{R}') &= \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} \phi_n(\mathbf{r} - \mathbf{R} + \mathbf{R}') \\
&= \frac{1}{\sqrt{N}} e^{i\mathbf{k}\mathbf{R}'} \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot(\mathbf{R}-\mathbf{R}')} \phi_n(\mathbf{r} - (\mathbf{R} - \mathbf{R}')) \\
&= \frac{1}{\sqrt{N}} e^{i\mathbf{k}\mathbf{R}'} \sum_{\mathbf{R}''} e^{i\mathbf{k}\cdot\mathbf{R}''} \phi_n(\mathbf{r} - \mathbf{R}'') = e^{i\mathbf{k}\cdot\mathbf{R}'} \psi_{\mathbf{k}}(\mathbf{r}),
\end{aligned} \tag{6.55}$$

where $\mathbf{R}'' = \mathbf{R} - \mathbf{R}'$.

We now use this wave function to calculate the desired band structure $E(\mathbf{k})$, using the same approach as for the hydrogen molecule (2.6). For now, we assume that the wave functions are already normalized so that

$$\begin{aligned}
E(\mathbf{k}) &= \int \psi_{\mathbf{k}}^*(\mathbf{r}) H_{\text{sol}} \psi_{\mathbf{k}}(\mathbf{r}) d\mathbf{r} \\
&= \frac{1}{N} \sum_{\mathbf{R},\mathbf{R}'} e^{i\mathbf{k}\cdot(\mathbf{R}-\mathbf{R}')} \int \phi_n^*(\mathbf{r} - \mathbf{R}') H_{\text{sol}} \phi_n(\mathbf{r} - \mathbf{R}) d\mathbf{r},
\end{aligned} \tag{6.56}$$

where both summations run over all the lattice sites and while we have a finite solid in mind, it should still be a solid in the sense of the periodic boundary conditions, that is, even if we are close to a "surface", the solid should periodically continue on the other side of this surface. Therefore, all the sums for a particular choice of $\mathbf{R}'$ are the same and we can get rid of the double summation by recognizing that we have N such sums. If we arbitrarily set $\mathbf{R}' = 0$, we obtain

$$E(\mathbf{k}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} \int \phi_n^*(\mathbf{r}) H_{\text{sol}} \phi_n(\mathbf{r} - \mathbf{R}) d\mathbf{r}. \tag{6.57}$$

Using (6.52), we can write this as

$$E(\mathbf{k}) = E_n - \beta + \sum_{\mathbf{R}\neq 0} e^{i\mathbf{k}\cdot\mathbf{R}} \int \phi_n^*(\mathbf{r}) H_{\text{sol}} \phi_n(\mathbf{r} - \mathbf{R}) d\mathbf{r}. \tag{6.58}$$

With (6.50), the integral in the above expression can now be split up into

$$\begin{aligned}
&\int \phi_n^*(\mathbf{r}) H_{\text{sol}} \phi_n(\mathbf{r} - \mathbf{R}) d\mathbf{r} \\
&\quad = E_n \int \phi_n^*(\mathbf{r}) \phi_n(\mathbf{r} - \mathbf{R}) d\mathbf{r} + \int \phi_n^*(\mathbf{r}) \nu(\mathbf{r}) \phi_n(\mathbf{r} - \mathbf{R}) d\mathbf{r}.
\end{aligned} \tag{6.59}$$

At this point, one usually neglects the first integral on the right-hand side because it contains two wave functions on different lattice sites and these have very little overlap. The second integral on the right-hand side is also small (for the same reason) but often not quite as small because the potential $v(\mathbf{r})$ falls less rapidly to zero when going away from $\mathbf{R}$ and, therefore, $v(\mathbf{r})\phi_n(\mathbf{r}-\mathbf{R})$ is increased in the region where it overlaps with $\phi_n^*(\mathbf{r})$. We introduce the abbreviation

$$\gamma(\mathbf{R}) = -\int \phi_n^*(\mathbf{r})v(\mathbf{r})\phi_n(\mathbf{r}-\mathbf{R})d\mathbf{r}, \tag{6.60}$$

and obtain the final expression for the band structure from (6.58)

$$E(\mathbf{k}) = E_n - \beta - \sum_{\mathbf{R}\neq 0} \gamma(\mathbf{R})e^{i\mathbf{k}\cdot\mathbf{R}}. \tag{6.61}$$

This describes how the atomic E_n level turns into a band when the atoms are arranged in a crystalline lattice.

We now determine this band structure for a one-dimensional chain of atoms with lattice spacing a, assuming that the band is derived from atomic s-orbitals with an energy E_s. Technically, (6.61) requires a summation over all the lattice sites. However, since the wave functions fall off very rapidly away from the site $\mathbf{R}$ they are centered on, it is sufficient to neglect all the contributions in the sum that involve lattice vectors more than one unit cell away from the origin. We thus restrict the sum (6.61) to only the nearest neighbors of an atom at $+a$ and $-a$. Moreover, since the atomic s-wave functions are spherically symmetric, we know that $\gamma_s = \gamma(-a) = \gamma(a)$ and obtain

$$E_s(\mathbf{k}) = E_s - \beta_s - \gamma_s(e^{ika} + e^{-ika}) = E_s - \beta_s - 2\gamma_s \cos ka, \tag{6.62}$$

where β_s is the value of β calculated for this s-band. This result is the lowest band plotted in Figure 6.11. The s-band has its lowest energy at $k = 0$ and its highest at $k = \pi/a$, that is, at the Brillouin zone boundary. Note that this dispersion is remarkably similar to the lowest band in the nearly free electron result in Figure 6.9d, despite of the totally different approach to calculate it. The center of the band is shifted away from the atomic energy E_s by $-\beta_s$. Usually, this shift is quite small.

The extension of this to other atomic energy levels is straightforward and the result for the next band, derived from an atomic p-level, is also shown in Figure 6.11. Again, this is very similar to the nearly free electron result in Figure 6.9(d). We also find a band gap at $k = \pi/a$. The size of this gap is given by the separation of the s- and p-levels, by difference in the shifts β_s and β_p, and the width of the two bands.

It is interesting to consider the factors influencing the absolute energy width of a band. In our one-dimensional model, the width is given by $2\gamma_s$, where the factor of 2 stems from the number of nearest neighbors and γ_s from the overlap of the wave functions and the potential. A high coordination number of the atoms, as typically present in the close-packed structures of metals, thus leads to a large band width. The value of γ_s is usually even more significant for the width of the band because of the very strong decay of the wave functions away from the nucleus. For a given

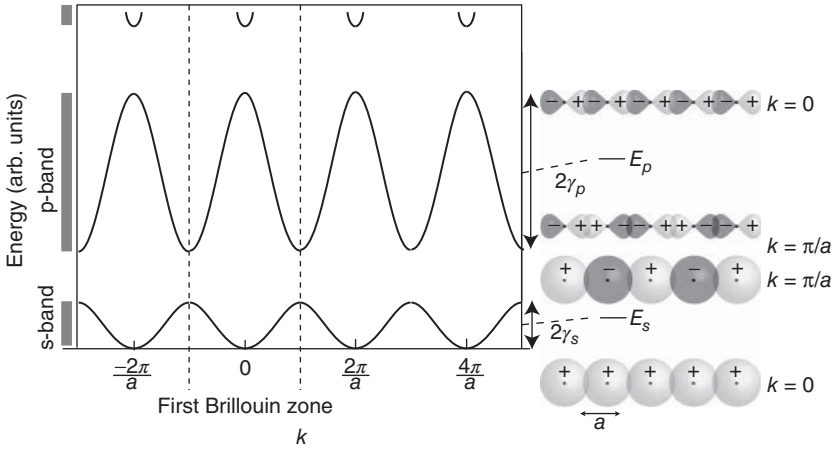


Figure 6.11 Bands for a one-dimensional solid calculated in the tight-binding approximation. The right-hand side also shows the Bloch wave functions for the s- and p-band for $k = 0$ and $k = \pi/a$. The black dots symbolize the position of the nuclei and the two different shades of gray symbolize the sign of the wave function.

structure, an atomic wave function that is stronger localized near the nucleus will lead to a significantly narrower band than a wave function that is less localized. An atomic 3d level, for instance, leads to a much narrower band than an atomic 4s level, even though the atomic levels are very similar in energy. An extreme case of a localized wave function would be the innermost (1s) level of a heavy atom. The 1s wave functions of neighboring atoms do not at all overlap in the solid and the 1s-derived band has a width approaching zero, that is, it is totally flat: It retains its atomic, localized character.

Finally, it is instructive to picture the Bloch wave functions (6.54) in the tight-binding model. Figure 6.11 shows these Bloch waves for the s-band and the p-band at $k = 0$ and at the Brillouin zone boundary $k = \pi/a$. For $k = 0$, the exponentials in (6.54) are all unity and $\psi_{\mathbf{k}}(\mathbf{r})$ is thus merely a sum over the orbitals on all the lattice sites. For the s-orbitals, this leads to an increase of probability density in between the atoms, that is, to a kind of "bonding molecular orbital." For $k = \pi/a$, the exponentials in (6.54) give rise to a sign change when moving one lattice constant a along the chain. This is symbolized by light gray (positive) and dark gray (negative) localized wave functions. This, in turn, leads to a probability density depletion in between the atoms, that is, to an "antibonding molecular orbital." This is consistent with the energies in the s-band: The energy for the bonding state at $k = 0$ is low and the energy for the antibonding state at $k = \pi/a$ is high. The opposite is true for the p-band. The sign of an atomic p-wave function changes under spatial inversion (it has an odd parity) and, therefore, adding the p-orbitals in phase for $k = 0$ leads to an antibonding state. Adding them with a sign change on every other site (for $k = \pi/a$) leads to a bonding state. Again, this is consistent with the calculated dispersion. We can also link this picture to the interpretation of the nearly free electron model wave functions near the Brillouin zone boundary in Figure 6.10.

The wave function $\psi(+)$ that has the lower energy at $k = \pi/a$ corresponds to the s wave function here, consistent with the probability density accumulation near the ion cores. The wave function $\psi(-)$ that gives rise to the higher energy state at $k = \pi/a$ with its probability density node at the ion cores corresponds to a p wave function that also has a node there. Note that this comparison is only qualitative, that is, the total probability densities in the nearly free electron model and in the tight-binding model are not at all the same, but it illustrates the consistency of the pictures.

6.6 Energy Bands in Real Solids

Now we are in a position to understand the electronic band structure in real materials and in three dimensions, at least qualitatively. In our one-dimensional models, we introduced two different ways of including a lattice-periodic potential. First, we started with free electrons and included the potential as a weak perturbation, arguing that this point of view is particularly appropriate for the nearly free electrons in metals. Then, we derived a very similar band structure by constructing Bloch wave functions from localized atomic orbitals, an approach that appears to be more natural for covalently bonded solids with more localized states. However, we have to keep in mind that both approaches are just very simple models that help us to understand the origin of band structure. In a more refined and accurate form, both should ultimately lead to the same predictions for the band structure.

In both pictures, we have seen that a lattice-periodic potential has two main effects. The first is the symmetry in the bands (6.41), which allows us to consider the dispersion in the first Brillouin zone only, because it is identical around the other points of the reciprocal lattice. This symmetry also causes a back-folding of the bands at the Brillouin zone boundary. The second effect is gap openings between bands that had been degenerate in the free electron model.

For three-dimensional materials, these effects are very similar, but the three-dimensional character of the problem makes it sometimes harder to keep the overview. The band energy now depends on a three-dimensional $\mathbf{k}$ and the Brillouin zone looks more complicated, too. As in the case of vibrational properties, we only discuss materials with an fcc Bravais lattice.

Figure 6.12 shows the energy bands of aluminum, a simple metal with only s and p electrons. The situation is still very similar to the free electron case. Consider first the dispersion in only one direction, as shown in Figure 6.12a. The dispersion is shown from the Γ to the X point at the Brillouin zone boundary and beyond into the next zone, reaching Γ again and so on. At the X point, a gap is opened and above the gap another band is dispersing back toward the Γ point. This band can easily been recognized as the band stemming from the center of the next Brillouin zone, as in the one-dimensional model in Figure 6.9. The only difference between the two figures is the energy scale. In the one-dimensional model, we have chosen the energy zero to be the bottom of the band. This appears to be the natural

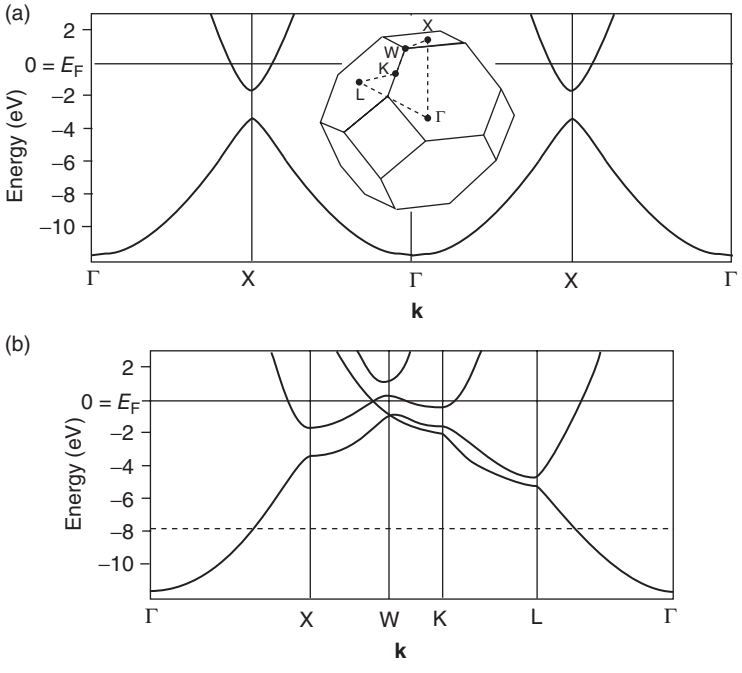


Figure 6.12 (a) Electronic energy bands in Al along the Γ – X direction only. The inset shows the first Brillouin zone. (b) Energy bands in different directions given by the dashed path between high-symmetry points of the Brillouin zone. The horizontal dashed line represents the fictitious Fermi energy for aluminum with the same structure but only one valence electron instead of three. Band structure taken from Levinson, Greuter, and Plummer (1983).

choice because it corresponds to the kinetic energy zero of the electrons. We are, however, completely free to change the origin of the energy scale. In metals, one almost always chooses the Fermi energy as $E = 0$, and this choice is also made in Figure 6.12.

The band structure of solids can be determined experimentally by angle-resolved photoemission spectroscopy. In this experiment, the sample is exposed to monochromatic ultraviolet photons and electrons are emitted because of the photoelectric effect. The emitted electrons are sorted according to their **k**-vector and energy, and from this it is possible to work back to **k** and the energy inside the sample, that is, to the band structure. The cover illustration of this book displays the experimental equivalent to Figure 6.12a.

Figure 6.12b shows the bands of Al in different high-symmetry directions in the first Brillouin zone. The continuation into the next zones, as in Figure 6.12a, is usually not shown to avoid redundancy. We can recognize the back-folding of bands from the neighboring zones and the opening of band gaps at the Brillouin zone boundaries. In fact, the band structure of aluminum can be described very well in the nearly free electron picture. It only appears complicated because of the bands appearing from the neighboring zones in three dimensions. The bands are

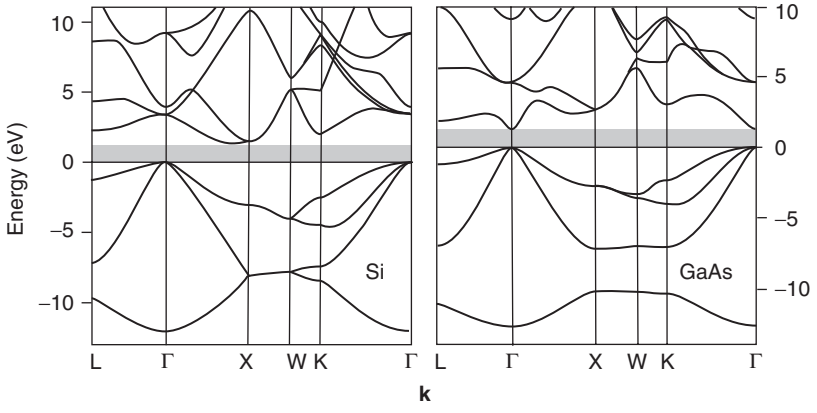


Figure 6.13 Electronic energy bands for Si and GaAs. These materials have the same Brillouin zone as Al (see Figure 6.12). The bands below the gray zone are completely filled and the bands above the gray zone are completely empty at zero temperature. The gray region represents an absolute band gap in the electronic structure. Band structures taken from Rohlfing, Krüger, and Pollmann (1993).

filled up to the Fermi energy. There are many bands crossing the Fermi energy, which means that the electrons in the occupied states just below the Fermi energy can be excited to states within the same band just above the Fermi energy. In this way, they can contribute to the transport of electric and thermal current.

Aluminum has three electrons per unit cell. These electrons have been filled into the bands, and this results in a separation of the Fermi energy from the band bottom of about 12 eV. What would happen if we had only one electron per unit cell? The Fermi energy would lie much lower, approximately at the dashed line in Figure 6.12b. In this case, the situation would be even more free-electron-like: The bands would cross the Fermi energy at the same k distance from Γ in all directions, and the electronic states at the Fermi energy would therefore constitute a sphere in k-space. This is exactly the same as in the case of the free electron model.

We can now turn back to the initial question in this chapter: What characterizes a metal as opposed to a semiconductor? From what is said above, we would define a metal as a solid where bands cross the Fermi energy, such that the energy of the electrons in these bands can be increased by a very small amount. Is this definition consistent with the band structure of typical semiconductors/insulators? To see this, we look at the band structures of Si and GaAs in Figure 6.13. Both have the same Brillouin zone shape as Al. The bands for these materials look considerably more complicated than those for the nearly free electrons or aluminum. However, we can still recognize several features. For the very lowest energies, the bands still look like parabolas. Band gap openings at the Brillouin zone boundaries and back-folded bands can also be identified. But the electronic structure of both materials differs from that of Al in two important ways. The first is the existence of an **absolute band gap**, emphasized by a gray area in the figure. "Absolute" means that this is not just a gap opening at some Brillouin zone boundary. It is a

gap in the entire Brillouin zone; there are no states at *any* **k** in the gray regions. The second remarkable difference to Al is seen when filling the states with the available electrons: When the states are occupied according to the Pauli principle, the bands below the gray band gap are exactly filled, there are no electrons left for the bands above the gap. Where this places the Fermi energy or, more precisely, the chemical potential is a subtle question that we address in the next chapter. Already now, we can say that it will be somewhere inside the gap region. The zero of the energy scale can still be set in an arbitrary way, and for semiconductors, it is often placed at the top of the occupied states. We can certainly state that these materials are not metals in the sense of the definition made above: There are no bands crossing the Fermi energy and no electrons that could increase their energy by a small amount in order to participate in electrical conduction. In fact, if the energy of an electron is to be increased, it must at least be increased by the energy corresponding to the size of the gap.

An even simpler picture for seeing the difference between metals and insulators/semiconductors emerges when we look at their density of states, as in Figure 6.14. For free electrons, we have calculated the density of states to be proportional to the square root of the energy (Figure 6.14a). For the nearly free

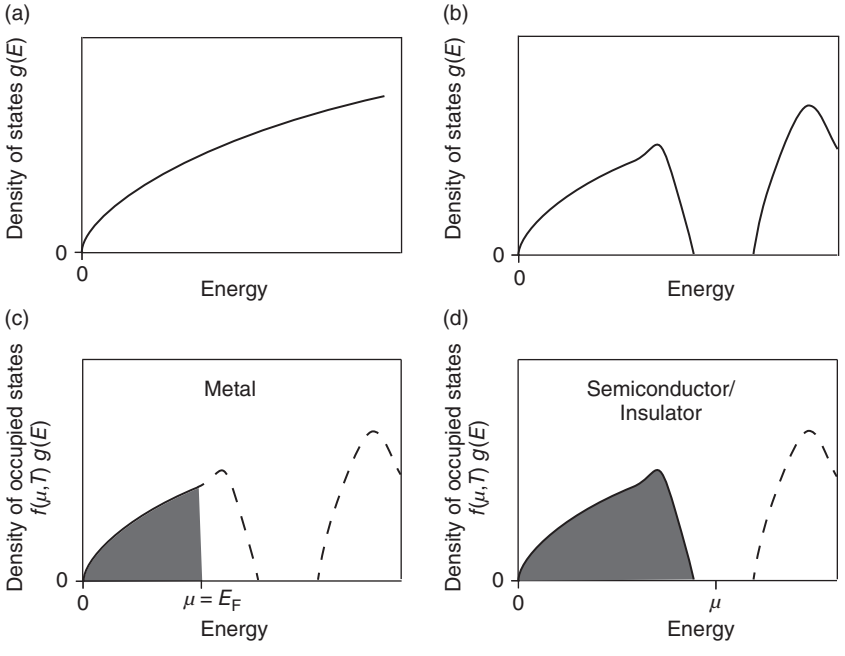


Figure 6.14 Illustration of the difference between metals and semiconductors/insulators. (a) Density of states in the free electron model. (b) Qualitative density of states in the nearly free electron model with the appearance of an absolute band gap. (c) Occupied density of states (gray area) for a metal at $T = 0$ K: The chemical potential (or the Fermi energy) cuts through a finite density of states. (d) The same for a semiconductor/insulator: The chemical potential is in a region of a vanishing density of states.

electron model or the tight-binding model, the density of states must be more complicated than this. Gaps appear in the band structure at the Brillouin zone boundaries, and this can (but need not) lead to absolute gaps in the bands. Such a more complicated density of states with an absolute band gap is shown in Figure 6.14b. Now the difference between a **metal** and a **semiconductor/insulator** depends on how many electrons we have to fill into these states. If the highest energy we fill up to lies at a finite density of states, the solid is a metal (Figure 6.14c); if we just manage to fill the states up to a gap, the solid is a semiconductor/insulator (Figure 6.14d). Note that we have so far avoided to define the difference between a semiconductor and an insulator. We will address this in the next chapter.

Can we predict if a material is a metal or a semiconductor/insulator, starting from the known crystal structure? In the picture discussed here, a material has to be a metal if a band is only partially filled with electrons. With a given number of valence electrons per unit cell, how many bands can we fill? It turns out that one band can accommodate exactly two electrons per unit cell, one for each spin direction. You can derive this important result formally in Problem 6.7, but also note that it is perfectly consistent with the very simple picture of bonding we had developed for a cluster of Na atoms in Section 6.1: There are as many states in the band as there are unit cells (or atoms in the cluster) and each state can accommodate 2 electrons. Hence, we would expect a material with an odd number of electrons per unit cell (such as Na) to be a metal. Another example is Al that crystallizes in the fcc structure. The structure has one atom per unit cell and Al has three valence electrons per atom; hence, there are three valence electrons per unit cell. Two of these can completely fill one band, leaving another band half-filled. Thus, we would expect Al to be a metal. This is of course true and also consistent with Figure 6.12. The figure also shows the position of the Fermi energy for the fictitious case of Al with only one valence electron. Following the same argument, this would also be a metal. Note, however, that the reverse argument is not valid: An even number of valence electrons per unit cell does not imply that the material is a semiconductor/insulator. The reason for this is that the electrons could be distributed into different bands in the three-dimensional band structure. Two electrons per unit cell, for example, could be either placed in one completely filled band, giving rise to a semiconductor /insulator, or in two different bands that overlap in energy. This latter case would result in two partially filled bands and thus in a metal, despite the even number of electrons.

We can use the case of graphene to illustrate these ideas further, and we will also see that graphene has some very special electronic properties, placing it at the boundary between a metal and an insulator. The bonding and electronic structure of graphene are shown in Figure 6.15. Carbon has four valence electrons (two 2s and two 2p electrons). Bonding in the honeycomb structure of graphene (see Figure 1.7a) is mainly achieved by an sp^2 hybridization between the s and $p_{x,y}$ states to form strong σ bonds. The remaining p_z orbitals stick out of the plane and form π bonds. The σ and π bonds are shown in Figure 6.15a and b, respectively, and the corresponding bands are shown in Figure 6.15c.

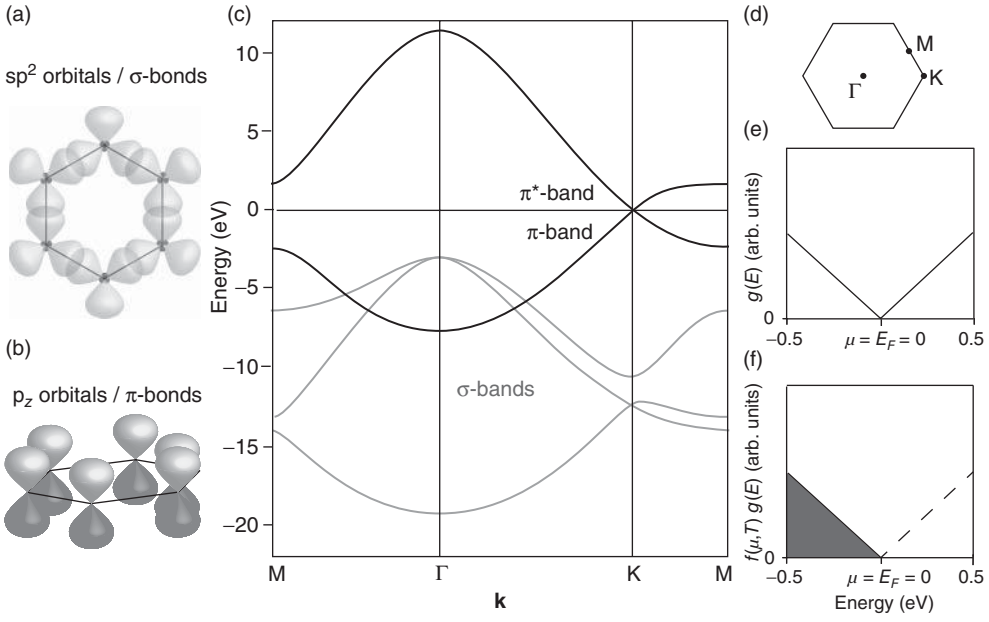


Figure 6.15 Origin of the electronic energy bands for graphene. (a) sp^2 hybrid orbitals giving rise to σ bonds. (b) p_z orbitals giving rise to π bonds. (c) Band structure taken from Kogan and Nazarov (2012) (not all the unoccupied bands are shown). (d) Two-dimensional first Brillouin zone. (e) Density of states in the immediate vicinity of Fermi energy. (f) Occupied density of states at $T = 0$ K (gray area).

Let us see if this band structure is consistent with the electron counting arguments presented above. Each carbon atom contributes with three electrons to the σ bonds and with one electron to the π bond. Graphene has two atoms per unit cell, that is, a total of six σ and two π electrons. The six σ electrons can fill three bands completely (two electrons per band) and the two π electrons can fill one band. This is also seen in the band structure. In total, we have four completely occupied bands and graphene could thus be an insulator/semiconductor. The curious thing about graphene is that there is no band gap between the occupied π band and the unoccupied π^* band. These bands meet exactly at the corner of the hexagonal, two-dimensional Brillouin zone, which is called the K point (see Figure 6.15d). The density of states and the density of occupied states in the vicinity of E_F are shown in Figure 6.15e and f, respectively. For energies close enough to E_F, the dispersion of the π and π^* bands is linear and this, together with the fact that graphene is two-dimensional, gives rise to a density of states that is linear as a function of energy (see Problem 6.2). The density of states goes exactly to zero at E_F, but there is no gap. If we define a metal as a material where the chemical potential (or Fermi energy) lies at an energy where the density of states is finite, then graphene is not a metal. On the other hand, graphene does not have band gap between the highest occupied and the lowest unoccupied states, as one finds in the case of an insulator/semiconductor.

Therefore, graphene is often called either a **semimetal** or a zero band gap semiconductor.

6.7 Transport Properties

We finally arrive at the description of transport in the quantum mechanical model, and we will confine the discussion to the transport of electrical charge. The transport of heat via electrons proceeds along similar lines, and we have some idea about the relation of electrical and thermal conductivity through the Wiedemann–Franz law. The transport properties of solids are a formidably complicated problem, and we just give some very basic ideas about what is happening. These ideas can be presented by considering a one-dimensional solid.

When inspecting the Bloch wave (6.26), we see that it describes a modulated plane wave that is delocalized over the whole crystal, very much like a free electron. In fact, introducing the periodic potential of the ions into the Schrödinger equation does not lead to any scattering. This is an extremely remarkable result: When discussing the shortcomings of the Drude model, we have asked how the electrons can manage to have a very long mean free path at low temperatures and to sneak past all the ions, completely incompatible with Drude's assumptions. Now we see why. A Bloch electron does not scatter off the lattice ions at all. Consequently, the electrical conductivity of perfectly crystalline metals should be infinite.[3] This is obviously not the case, and there must be some scattering mechanism for the Bloch electrons as well. We will discuss possible candidates at the end of this section. For now, we merely assume that there is some scattering present, which gives rise to a finite relaxation time τ. Note that the situation is very similar to the transport of heat in a harmonic crystal. Phonons, which are packets of harmonic waves, can propagate undisturbed through the crystal, and we had to invoke some effects (like defects) to obtain a finite thermal conductivity.

When discussing an electron traveling through a crystal, the Bloch waves are not an appropriate picture, precisely because they are delocalized over the whole solid. We use the same approach to describe a localized particle as in the case of lattice vibrations: We think of an electron traveling through the crystal as a wave packet, that is, as a superposition of Bloch waves within a certain Δk close to the k of interest (see Problem 6.12)[4]. Such a "particle" has a group velocity v_g given by (6.47), and we can arrive at a quasi-classical description of electrical conduction.

Consider a particle with a velocity v_g and a charge $-e$ in an electric field $\mathcal{E}$. After a short time dt, the particle has increased its kinetic energy by

$$dE = -e\mathcal{E}v_g dt. \tag{6.63}$$

3) This is actually not quite true. In a metal without scattering, one would observe the so-called **Bloch oscillations**. The Bloch oscillations are discussed in an online note on www.philiphofmann.net.

4) This is also discussed in the online note on phase velocity and group velocity on www.philiphofmann.net.

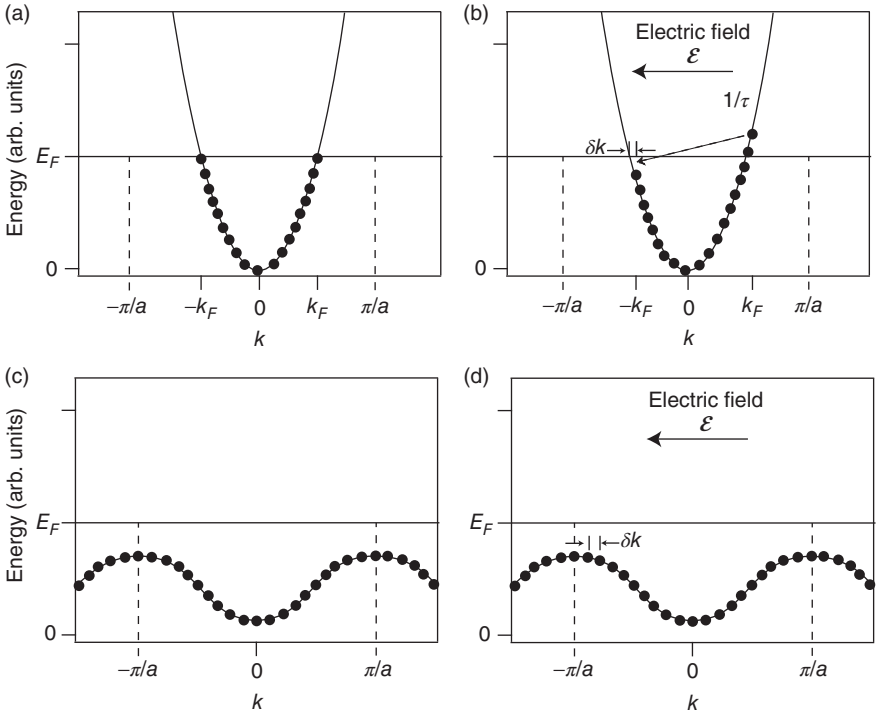


Figure 6.16 Simple picture of conduction in a metal. The circles symbolize filled electron states at the allowed k points. (a) Situation for a partially filled band without an applied field. (b) Situation for a partially filled band with an applied electric field. After some time δt, all electrons have moved by an amount δk due to the applied field. The asymmetric distribution in the electrons' group velocity gives rise to an electric current. The electrons at k_F can be scattered back to lower lying states at $-k_F$ with a probability proportional to the inverse relaxation time. (c and d) Corresponding situation for a completely filled band without and with electric field. All electrons are merely moved into states that had been occupied already in the field-free case.

On the other hand, we have

$$\frac{dE}{dt} = \frac{dE}{dk}\frac{dk}{dt}, \tag{6.64}$$

and combining this with (6.47), we get

$$\hbar\frac{dk}{dt} = -e\mathcal{E}. \tag{6.65}$$

This equation is quite plausible from the case of free electrons, where the momentum is $p = \hbar k$, but we have already shown that $\hbar k$ is not the momentum for Bloch electrons. Nevertheless, the equation is correct. It means that an electric field causes the Bloch electrons to change their k, and the rate of change is given by the field strength.

The situation for a partially filled band is shown in Figure 6.16a,b. Without an applied field, the distribution of electrons is symmetric. When the field is turned

on, the electrons will have changed their k by dk after some short time dt, giving rise to the more asymmetric distribution in Figure 6.16b. Equation (6.65) suggests that the distribution would become increasingly asymmetric with time. In reality, however, there will be an inelastic scattering mechanism with a relaxation time τ, which prevents this from happening. Such a process, shown in Figure 6.16b, brings the accelerated electrons close to k_F to unoccupied states at lower energy, close to $-k_F$. The combination of field acceleration and inelastic scattering leads to a stationary state in which all the electrons are displaced by some δk. In most cases, δk will be small compared to the size of the Brillouin zone. The inelastic scattering also leads to an energy dissipation and thus to a finite resistance.

According to (6.47), the asymmetric distribution in k corresponds to an asymmetric distribution of the group velocities of the electrons as well. While the whole distribution has been moved by δk, most electrons have ended up in states that had been occupied by other electrons before and nothing has changed. The only points where the change is important are around the Fermi energy crossings $-k_F$ and k_F. In Figure 6.16b, the asymmetry of the distribution implies that there are more electrons with a group velocity to the right than to the left, that is, there is an electric current flowing. Obviously, the size of the current depends on the group velocity of the electrons at the Fermi energy.

Figure 6.16c and d illustrate the situation for electrons in a full band. As the field is applied, these electrons are also moved by a certain δk. But this does not change the situation at all because all electrons move into states that were also occupied before the field was turned on: The full band does not contribute to the conduction. Note that this is perfectly consistent with the fact that $\hbar k$ cannot be interpreted as the momentum of the Bloch electrons. Here, we increase $\hbar k$ for all electrons but the average momentum is obviously still zero. The picture is also consistent with our previous definition of metals and insulators. In an insulator, all the bands are completely full and hence no current can be passed through it. An exotic exception to this rule is graphene. We have seen in Figure 6.15 that graphene has four filled bands. We might not expect these to contribute to a conductance, but experimentally graphene is found to be one of the best conductors there is at room temperature. The reason is the missing gap between the π band and the π^* band. The electrons from the π band can move directly into the π^* band when accelerated.

It is quite instructive to combine (6.65) with (6.47) in the following way. Consider the acceleration of an electron initially traveling with the group velocity v_g:

$$a = \frac{dv_g}{dt} = \frac{1}{\hbar}\frac{d}{dt}\frac{dE(k)}{dk} = \frac{1}{\hbar}\frac{d^2E(k)}{dk^2}\frac{dk}{dt}. \tag{6.66}$$

If we substitute (6.65) for dk/dt in the last term, we get

$$a = -\frac{1}{\hbar^2}\frac{d^2E(k)}{dk^2}e\mathcal{E}. \tag{6.67}$$

This looks exactly like a classical equation of motion, if we define that the particles have a so-called **effective mass**

$$m^* = \hbar^2 \left(\frac{d^2E(k)}{dk^2} \right)^{-1}. \tag{6.68}$$

The concept of the effective mass may appear rather artificial at first, but it allows us to describe the conduction by a classical equation of motion, as in the Drude model. The only effect of the solid is that it can change the effective mass of the electrons. This effect can be dramatic: The effective mass can be much smaller or much bigger than the free electron mass. It can also be negative (see Section 7.1.1 for a more detailed discussion of this case). For free electrons, the effective mass is of course equal to the electron mass m_e (see Problem 6.11). Again, the situation of graphene is somewhat exotic: The band structure in Figure 6.15c shows that the dispersion $E(k)$ is linear in the vicinity of the Fermi energy. If this is so, $d^2E(k)/dk^2 = 0$ and applying the definition of (6.68) would lead to a divergent effective mass, something that could perhaps suggest that graphene has a very high resistivity. Quite the opposite is the case and the reason for this confusion is a limitation of (6.68).[5)]

The concept of the effective mass brings us back to the Drude formula for the electrical conductivity (5.9). If we treat conductivity in a quantum model, starting out with an expression like (6.67), the resulting conductivity will have the same form as (5.9). We have seen that only the electrons near the Fermi energy in partially filled bands have to be considered to calculate the conductivity, so the mass in a semiclassical version of (5.9) would have to be replaced by the effective mass at the Fermi energy and the relaxation time would be that for the electrons at the Fermi energy. The electron density n would still appear but not because all the electrons contribute to the current. The reason for having n in the equation is that the number of electrons at the Fermi energy depends on the electron density. This is easily seen in the free electron model where the size of k_F and hence the size of the "Fermi sphere" depends on the electron concentration.

We have seen that Bloch waves travel through the perfect crystal without any scattering by the ions, in contrast to the electrons in the Drude model. Therefore, we still have to discuss where the relaxation time τ comes from and why metals have a finite resistivity. Ultimately, all the explanations come down to the fact that the lattice is not perfect. The most important imperfections at higher temperature are lattice vibrations that destroy the perfect translational symmetry of the lattice and cause a scattering of the Bloch electrons. This also means that our initial assumption of the Born–Oppenheimer approximation is invalid, since we have to consider the scattering of the electrons by lattice vibrations. Quite intuitively, the interaction between Bloch electrons and lattice vibrations is called the **electron–phonon interaction**. This process can be expected to be important if the temperature is not too low compared to the Debye temperature of the solid. However, even at very low temperatures, the Bloch electrons are scattered because of remaining imperfections in the crystal. These can be all kinds of defects, point

5) For a more detailed discussion of this, see online note on www.philiphofmann.net.

defects, dislocations, impurity atoms, and so on. Still, in a highly perfect crystal at low temperature, the conductivity can be several orders of magnitude higher than at room temperature.

6.8 Brief Review of Some Key Ideas

Many of the problems we had with the Drude model have been caused by not taking into account that the electrons are fermions and thus underlie the Pauli principle. This has readily been cured by the quantum mechanical free electron model. Historically, the most important result was the heat capacity of the electrons, but the fact that only the electrons close to the Fermi energy can be excited by a small amount of energy (and not all the electrons) is decisive in many properties of metals (conductivity, magnetism, screening, superconductivity, etc.). In formal terms, it is seen in the equations through the appearance of the density of states at the Fermi energy $g(E_F)$.

The free electron model can also give us some hints as to why the electrons do not seem to interact much with each other: First, as the electrons move through the solid, many scattering processes between them are impossible because the states into which they could scatter are already occupied by other electrons. In addition to this, the Coulomb interaction in metals is strongly weakened by the very efficient screening. A real understanding of why the electrons do not interact much with each other, however, is quite difficult and far beyond the scope of this book.

Despite these successes, even the free electron model has some serious limitations. Many of these were qualitatively solved by the introduction of Bloch waves as a general solution to the Schrödinger equation for the periodic lattice and by the nearly free electron model as a particularly simple solution. This could account for different band structures that depended on the Fourier coefficients in the series describing the potential (6.30). It could explain the existence of band gaps and make it plausible that some materials are metals while others are not. We have seen that we could arrive at a very similar result in the tight-binding model, even thought this starts from an entirely different construction of the wave functions.

Our discussion of electrical conduction in a metal via Bloch states has also helped to understand several problems we had with the Drude model. The increase in the length of the mean free path and the conductivity at low temperatures now follows naturally from the fact that the Bloch electrons do not scatter at all from the perfect lattice. At room temperature, they are mainly scattered by lattice vibrations, but at low temperatures, these vibrations are frozen out. We can also understand why the resistivity of alloys can be much higher than for pure metals. If the alloys are built such that the two (or more) types of ions in the alloy do not form a periodic structure, this will lead to a strongly increased scattering of the Bloch electrons. Even if they *do* form a perfectly crystalline lattice, there can still be some disorder if the two types of atoms are randomly distributed on the lattice sites.

Finally, we have encountered the concept of the effective mass, leading to the perhaps most remarkable result of the whole chapter: The electrons in the periodic solid can be treated quite similarly to free electrons by a quasi-classical theory if we replace the free electron mass by an effective mass, which contains the information about the solid's band structure.

References

Kogan, E. and Nazarov, V.U. (2012) *Phys. Rev. B*, **85**, 115418.

Levinson, H.J., Greuter, F., and Plummer, E.W. (1983) *Phys. Rev. B*, **27**, 727.

Rohlfing, M., Krüger, P., and Pollmann, J. (1993) *Phys. Rev. B*, **48**, 17791.

6.9 Further Reading

The quantum mechanical description of the electronic states is covered by all the standard texts in solid state physics. Consider, for example,

- Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Holt-Saunders. Gives a very thorough treatment of the subject.
- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.
- Omar, M.A. (1993) *Elementary Solid State Physics*, Addison-Wesley. Gives a basic and very well-written introduction.

Understanding the formation of band structure is not easy, and it can be helpful to consider alternative descriptions to those of the standard textbooks. Two of these are

- Feynman, R.P., Leighton, R.B., and Sand, M. (1966) *The Feynman Lectures on Physics*, Addison-Wesley, also available as free online version. Presents a very elegant approach to the formation of band structure in the tight-binding picture (Volume III, Chapter 13).
- Hoffmann, R. (1988) *Solids and Surfaces: A Chemist's View on Bonding in Extended Structures*, Wiley-VCH, also published in Hoffmann, R. (1987) *Angew. Chem.*, **26**, 846. A very insightful discussion of band structure formation starting from atomic orbitals.

6.10 Discussion and Problems

Discussion

1) Describe the origin of electronic energy bands in solids qualitatively.
2) Describe the free electron model. How does the energy of an electron depend on the wave vector? How does the density of states depend on the energy?

3) How do the free electrons in the quantum model contribute to the heat capacity, and how does this differ from the Drude model?
4) How does the heat capacity contribution of the electrons depend on the temperature?
5) Is it possible to measure the electronic contribution to the heat capacity despite the fact that it is usually quite small?
6) What is the form of the general solution of the Schrödinger equation for a lattice-periodic potential?
7) Do the Bloch wave functions have the periodicity of the lattice?
8) The expression for the Bloch wave function suggests that the electrons can travel through the lattice without scattering. Why is this so, and where does the observed resistance come from?
9) What causes the existence of energy gaps in the electronic bands in a quantum model (nearly free electron model or tight-binding model)?
10) What determines the width of a band (in energy) in the tight-binding model?
11) How does the electrical resistivity of a metal depend on the temperature (qualitatively) and why?
12) (*) The density of states at the Fermi energy of a metal, $g(E_F)$, appears in equations describing many different physical phenomena, such as electronic heat capacity, Pauli paramagnetism, screening, superconductivity, and others. Why is this so?
13) What is the speed of the electrons that contribute to the electrical current in the quantum model? How high is it compared to that in the Drude model, and how does it depend on the temperature?
14) What is the physical interpretation (or several interpretations) of the vector **k** for an electronic state in a solid?

Problems

1) *Free electron model:* (a) Show that the mean kinetic energy of one electron in the quantum mechanical free electron model is 3/5 E_F at $T = 0$ K. (b) Calculate the Fermi energy and the mean kinetic energy for potassium in electron volts. Use that K has a relative atomic mass of $M = 39.1$ u and a density of 856 kg m^{-3}. (c) Calculate the corresponding electron velocities. (d) Calculate the density of states at the Fermi energy. How large is the number of electrons in the "soft zone" of about $4k_BT$ around the Fermi energy at room temperature relative to the total number of electrons? (e) Estimate the Thomas–Fermi screening length in potassium.
2) *Free electron model:* We have shown that the density of states for a free electron gas in three dimensions is given by (6.13). (a) Show that the density of states for a free electron gas in two dimensions is independent of the energy. (b) How does the density of the states depend on the energy if the electronic dispersion is linear instead of quadratic, that is, if $E(k) \propto k$ instead of $E(k) \propto k^2$? Discuss this for both the three-dimensional and the two-dimensional case.

3) *Free electron model:* (a) Calculate the electronic heat capacity for 1 mol of copper at 300 K. Use that the Fermi energy of Cu is 7 eV and the molar volume 7.11 cm^3. (b) Compare the result of (a) to the Dulong–Petit value of the lattice, and explain why it is so much smaller. (c) Below which temperature is the electronic contribution to the heat capacity higher than the contribution from the lattice? Use that the Debye temperature of Cu is 343 K. (d)(*) Derive the electronic heat capacity (6.18) in a proper way by calculating the total energy and differentiating it. Hint: For this last part, assume that $k_B T \ll E_F$. In this case, the density of states $g(E)$ in the vicinity of the Fermi energy can be taken to be $g(E_F)$, independent of the energy, and the chemical potential can be taken as independent of the temperature, as usual, that is, $\mu = E_F$.
4) *Free electron model:* Show that the Wiedemann–Franz law is indeed given by (6.19).
5) *Nearly free electron model:* We have seen that (6.43) gives an approximate solution of the Schrödinger for a weak potential with Fourier components $U = U_1 = U_{-1}$. The biggest deviation from the energies of free electrons was found at the Brillouin zone boundary ($k = \pi/a$). Using (6.43), show that c_{k-g} is much larger than c_{k+g} if $k = \pi/a$. Neglect then c_{k+g} and use (6.43) to show that the size of the gap opening at the Brillouin zone boundary is $2U$.
6) *Bloch electrons:* Show that the spatial probability density for one-dimensional free electrons is constant. (b) Show that it has the periodicity of the corresponding Bravais lattice for Bloch electrons.
7) *Metals and nonmetals:* Consider a one-dimensional chain of N atoms with one atom per unit cell. Assume periodic boundary conditions and that each atom has Z valence electrons. (a) Show that you can fill exactly $Z/2$ bands with these electrons or, equivalently, that each band can accommodate $2N$ electrons. (b) Figure 6.13 shows that Si has four filled bands (for some values of $\mathbf{k}$, the energies of the bands are degenerate, but not for all). There are also four electrons per Si atom (not eight!). Explain why this is so. (c) Having an even number of electrons per unit is necessary but not sufficient for a solid to be a semiconductor/insulator. Give an example for an elemental solid that is a metal despite having an even number of electrons per unit cell.
8) *Tight-binding model:* (a) Show that the dispersion (6.62) for the s-band in the tight-binding model can be approximated by a parabolic dispersion in the vicinity of $k = 0$, as in the nearly free electron model. (b) Calculate the effective mass in this case, and discuss the result.
9) *Tight-binding model:* Figure 6.11 shows the atomic s and p orbitals in a chain of atoms and how these are combined to form the bonding and antibonding states. For the s band, the bonding state is formed with the atomic wave functions on all sites combined in phase, corresponding to a wave vector $k = 0$, and the antibonding state is associated with a sign change of the wave function on every other site, corresponding to a wave vector at the Brillouin zone boundary. Now consider the π bands in graphene, which are formed from the p_z orbitals, as shown in Figure 6.15b. (a) How should these orbitals be combined to form bonding and antibonding π states? (b) If we associate

the bonding/antibonding states with the energy extrema of the π band, it appears that both are found at the Brillouin zone center (Γ). Is this consistent with your result from (a)?

10) *Transport properties:* Show, in one dimension, that the average group velocity for a filled band is zero. Hint: For a chain with lattice constant a and macroscopic length L, the distance between the allowed k values is $2\pi/L$, that is, it is very small. It is then useful to write the sum over the allowed k-points as an integral.

11) *Transport properties:* Show that the effective mass for free electrons is equal to the free electron mass m_e.

12) *(*) Transport properties:* An electron moving through the crystal can be described by a superposition of Bloch waves. Consider the time-dependent, one-dimensional Bloch wave functions

$$\psi_k(x,t) = e^{ikx - i\omega(k)t} u_k(x) \tag{6.69}$$

and the wave packet

$$\Psi(x,t) = \int_{-\infty}^{\infty} e^{-(k-k_0)^2/b^2} e^{ikx - i\omega(k)t} u_k(x) dk. \tag{6.70}$$

If b is much smaller than the size of the Brillouin zone, the Gauss function $e^{-(k-k_0)^2/b^2}$ is strongly peaked around k_0, and we can assume that $u_k(x) = u_{k_0}(x)$ does not depend on k and that the dispersion of the electronic states $\omega(k) = E(k)/\hbar$ is linear, that is,

$$\omega(k) = v_g(k - k_0) + \omega(k_0) = v_g k + \omega_0 \tag{6.71}$$

with $\omega_0 = \omega(k_0) - v_g k_0$ with $v_g = \partial\omega/\partial k$ at k_0. Show that the maximum of the probability distribution associated with this wave packet moves with a velocity v_g.

7
Semiconductors

In the previous chapter, we have defined the difference between metals and semiconductors/insulators based on the density of states at the chemical potential at 0 K. If the density of states is finite, the solid is a metal. Otherwise, it is an insulator or a semiconductor. We have not defined the character of the solid based on its conductivity. In fact, this would have been a bad idea. The conductivity of solids spans more than 27 orders of magnitude. As a general rule, metals are good conductors and semiconductors/insulators are not, but even among the metals, the variations are big. To make matters more complicated, the conductivity of solids is strongly temperature-dependent, and this dependence is qualitatively different for different types of solids. A solid that is a poor conductor at low temperatures can be a good conductor at room temperature and vice versa.

The difference between semiconductors and insulators is not very clearly defined. However, the general idea behind a semiconductor is that the band gap between the highest occupied states and the lowest unoccupied states is sufficiently small to get thermally excited electrons at reasonable temperatures. We know that the width of the soft zone in the Fermi–Dirac distribution is about 100 meV at room temperature. Therefore, a semiconductor gap cannot be much larger than a few electron volts because otherwise the number of excited electrons would be vanishingly small. One often speaks of **semiconductors** in the case of materials with band gaps below ≈ 3 eV and of **insulators** if the gap is larger. Gap sizes for some important semiconductors and insulators are given in Table 7.1.

The most common elemental semiconductors are the group IV elements Si and Ge with their characteristic tetrahedral sp^3 bonding. Many compound semiconductors are isoelectronic to Si and Ge and show the same bonding type. Examples are SiC that also contains only group IV elements, or combinations of groups III and V such as GaAs, or II and VI such as CdSe. The latter two groups of materials are commonly referred to as III–V and II–VI semiconductors.

The conductivity of semiconductors is strongly influenced by a very small amount of impurities, and a precise control of the impurity concentration turns out to be essential for the construction of semiconductor devices. In this chapter, we shall first consider the properties of pure or **intrinsic semiconductors**, and then those of impure or **doped semiconductors**. Finally, we shall briefly discuss the basic working principles of some semiconductor devices.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

Table 7.1 Gap sizes for common semiconductors (above the horizontal line) and insulators (below the horizontal line).

Material	Gap size (eV)
InSb	0.18
InAs	0.36
Ge	0.67
Si	1.11
GaAs	1.43
CdSe	1.74
SiC	2.36
Diamond	5.5
MgF_2	11

7.1 Intrinsic Semiconductors

In this section, we discuss the properties of **pure (intrinsic)** semiconductors. Since a very small number of impurities has a strong impact on the behavior of semiconductors, this intrinsic state is hard to realize and also of little technological relevance. However, many of the concepts we explore here can be easily transferred to the case of doped semiconductors that we treat in the next section.

A semiconductor or insulator is, by definition, a solid for which the chemical potential at zero temperature is placed at an energy where the density of states is zero. In other words, there is a certain energy band that is completely filled, and the next band at higher energy is completely empty. This situation is schematically shown in Figure 7.1a. The highest occupied band and the lowest unoccupied band are called the **valence band (VB)** and the **conduction band (CB)**, respectively. At finite temperature, the electrons have to be distributed according to the Fermi–Dirac distribution shown in Figure 7.1b. This leads to some missing electrons in the VB and some extra electrons in the CB. In order to calculate the resulting electron concentrations, we have to know the position of the chemical potential.[1)] We can find this based on the argument of charge neutrality: The excited electrons in the CB must clearly come from the VB, and we must thus have as many extra electrons in the CB as missing electrons in the VB. This requirement determines the position of the chemical potential. To see this, suppose that the chemical potential is just above the **valence band maximum (VBM)**. This would result in a density of occupied states as shown in Figure 7.1c. Many states would be emptied in the VB, but only few states would be populated in the CB. In other words, the total number of electrons would have decreased and the solid would be charged positively, violating charge conservation. An equivalent argument can

1) Some authors use the term "Fermi energy" instead of "chemical potential" for semiconductors. We speak of the Fermi energy only for metals.

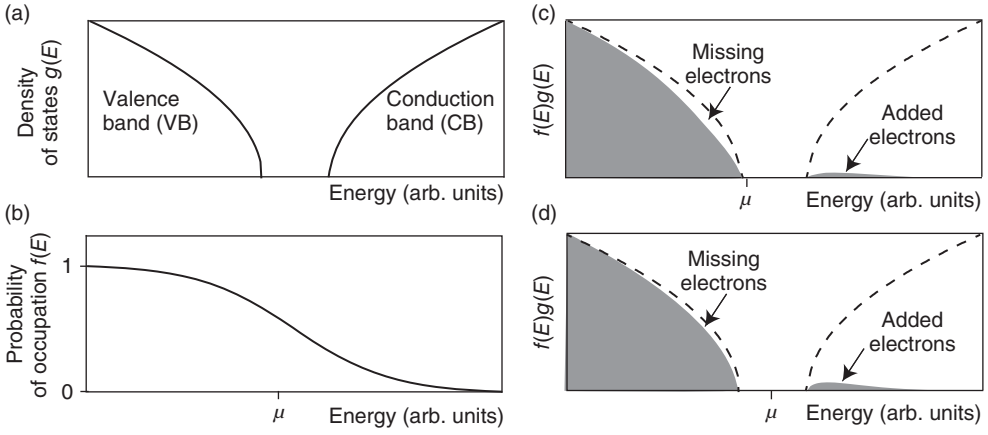


Figure 7.1 Charge neutrality and the position of the chemical potential in an intrinsic semiconductor. (a) Schematic density of states for a semiconductor. (b) Fermi–Dirac distribution at finite temperature. (c) Occupied density of states (gray area) for the chemical potential just above the valence band maximum. (d) Occupied density of states for the chemical potential close to the middle of the gap. Note that the temperature in (b–d) is much higher than room temperature to make the presence of excited carriers visible.

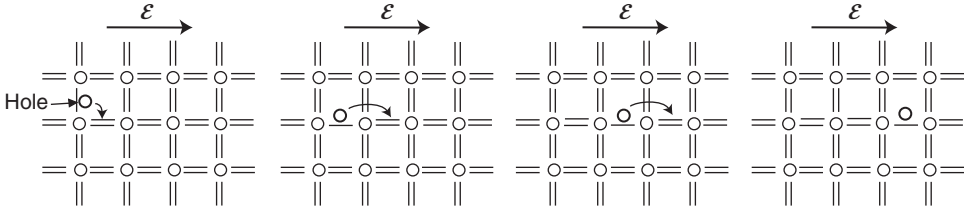


Figure 7.2 Transport of charge in an electric field $\mathcal{E}$ for a partially filled VB. The process can be interpreted as a motion of electrons in the direction opposing the field or as a motion of a hole in the field direction.

be made for a chemical potential just below the **conduction band minimum (CBM)**, which would lead to a negatively charged solid. Consequently, we see that the chemical potential must lie close to the middle of the gap between the VBM and the CBM (Figure 7.1d). At a finite temperature, this leads to an equal number of missing electrons in the VB and excited electrons in the CB. Note that the temperature for the distribution in Figure 7.1b was chosen to be *very* high in order to make the effects in Figure 7.1c and d visible.

At finite temperature, the excited electrons in the CB cause this band to be partially filled. Therefore, we can apply the same concept for electronic transport as we have used for metals. The VB is also partially filled at finite temperature and contributes to the conductivity via the remaining electrons. There is another possibility to view the conduction by the VB, which is illustrated in Figure 7.2. The semiconductor is schematically shown as atoms with valence 4 bound together.

Each bond represents one electron in the VB. At finite temperature, some of the electrons in the VB are missing and these are represented by a missing bond and a circle. When an electric field is applied, electrons can use these missing states to travel to the positive potential as shown in the figure. But there is another, simpler way of viewing this. Instead of considering the complicated motion of the electrons, one can describe the conduction in terms of the moving missing electron or so-called **hole**.

In the following two sections, we examine these ideas more quantitatively. We calculate the position of the chemical potential under several circumstances. The central idea is always that a change of temperature cannot lead to charging of the solid. We also consider the electrical transport through the VB and the CB, and we shall see that the concept of holes also emerges from the mathematical side of the problem.

7.1.1 Temperature Dependence of the Carrier Density

The current through a semiconductor can be carried by both electrons and holes and these are often discussed together as (charge) **carriers**. In order to understand the conductivity of a semiconductor, we are thus interested in the density of these carriers. It is simple to write down an expression for the electron density n in the CB. It corresponds to the integrated occupied density of states:

$$n = \frac{1}{V}\int_{E_C}^{\infty} g_C(E) f(E,T) dE, \tag{7.1}$$

where $g_C(E)$ is the density of states in the CB and E_C is the energy of the CBM. The density of missing electrons or holes in the VB p can be calculated by the analogous formula:

$$p = \frac{1}{V}\int_{-\infty}^{E_V} g_V(E)[1 - f(E,T)] dE, \tag{7.2}$$

with E_V being the energy of the VBM. For the practical calculation, we have to make some approximations. The first concerns the density of states $g_C(E)$ and $g_V(E)$. These are derived from the semiconductor's band structure and can therefore be quite complicated (see Figure 6.13). However, we have noticed that the chemical potential is roughly in the middle of the gap, quite far away from the CBM and VBM. Since the Fermi–Dirac distribution falls to zero very rapidly away from the chemical potential, most occupied electrons can be found very close to the CBM and most holes very close to the VBM. If we inspect the band dispersion close to these points, it is very nearly parabolic (see Figure 6.13). It is therefore sensible to simplify the relevant band structure of a semiconductor as in Figure 7.3a. Both the CB and the VB are represented by parabolas. These are not necessarily free-electron-like as in (6.6), especially not the VB that has the "wrong" curvature. However, they can be described using the concept of the effective mass from (6.68). The effective mass is essentially the inverse curvature of a band, such that a parabolic band has the same effective mass everywhere.

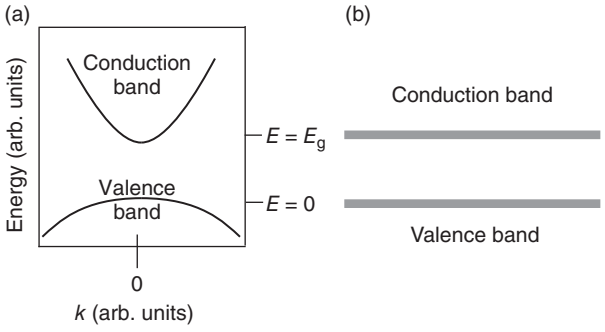


Figure 7.3 (a) Sketch of the valence band and conduction band in the vicinity of the gap. The bands are described as parabolas with different curvatures (effective masses). (b) Even simpler picture in which the valence and conduction bands are represented by single energy levels.

Hence, only the two effective masses are needed to describe the relevant part of the band structure and the density of states. Furthermore, it is common practice to define the VBM as energy zero so that we have $E_V = 0$ and $E_C = E_g$ with E_g being the gap size.

The situation in Figure 7.3 closely resembles the dispersion for GaAs in Figure 6.13, but it is somewhat different from the dispersion of Si. In GaAs, both the VBM and the CBM are found at Γ, that is, at $\mathbf{k} = (0, 0, 0)$, and the gap is called a **direct band gap**. In Si, only the VBM is found at Γ (the CBM is somewhere between Γ and X) and the gap is called an **indirect band gap**. But if we are only interested in the carrier densities, the position in $\mathbf{k}$ is not important, only the energy is. The picture shown in Figure 7.3 will therefore work for both types of band gap.

Figure 7.3a can also give us a more mathematical derivation for the concept of holes in the VB. The CB is almost empty apart from some electrons close to its minimum. When an electric field is applied, the situation is similar to that of a metal depicted in Figure 6.16b with a "Fermi energy" just above the CBM. The electrons near the CBM are accelerated in an external electric field $\mathcal{E}$ according to $a = (-e)\mathcal{E}/m_e^*$. Since m_e^* is positive, this is the "normal" motion of electrons in an electric field. In the VB, the situation is quite different. The band is almost filled, apart from some missing electrons around the VBM. In this way, the situation is very similar to conduction in a metal where the "Fermi energy" is just below the VBM. The electrons are also accelerated according to $a = (-e)\mathcal{E}/m_h^*$, but now the effective mass is negative such that we can write $a = (-e)\mathcal{E}/(-|m_h^*|)$. But this is the same as $a = e\mathcal{E}/(|m_h^*|)$, and we can therefore interpret it as the motion of holes, that is, positive carriers with a positive effective mass. In the following, we assume that m_h^* is positive, and we consider holes rather than electrons as carriers in the VB.

The dispersion of the CB can now be written as

$$E = E_g + \frac{\hbar^2 k^2}{2m_e^*}. \tag{7.3}$$

The density of states in the free electron model is given by (6.13), and consequently, the density of states in the CB is

$$g_C(E) = \frac{V}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2}(E - E_g)^{1/2}. \tag{7.4}$$

For the VB, we get analogous expressions:

$$E = -\frac{\hbar^2 k^2}{2m_h^*} \tag{7.5}$$

with the (positive) effective mass m_h^* of the holes and

$$g_V(E) = \frac{V}{2\pi^2}\left(\frac{2m_h^*}{\hbar^2}\right)^{3/2}(-E)^{1/2}. \tag{7.6}$$

The density of states between the VB and CB is of course zero. Now we can formally write down the electron density (7.1) and the hole density (7.2), but unfortunately, the integrals cannot be solved analytically.

It is therefore useful to introduce a second simplification. If the chemical potential μ is close to the middle of the gap, the gap size is at least a few hundred meV and if we are interested in the properties of the material around room temperature, then $(E - \mu) \gg k_B T$ for all the energies E in the CB. We can therefore approximate the Fermi–Dirac distribution in the CB as

$$f(E, T) = \frac{1}{e^{(E-\mu)/k_B T} + 1} \approx e^{-(E-\mu)/k_B T}. \tag{7.7}$$

For the holes in the VB, we can make the equivalent argument and obtain

$$1 - f(E, T) = 1 - \frac{1}{e^{(E-\mu)/k_B T} + 1} \approx e^{(E-\mu)/k_B T} \tag{7.8}$$

(see Problem 7.1). Basically, these approximations mean that we have replaced the Fermi–Dirac distribution with a (shifted) classical Boltzmann distribution.

Now the integrals for the electron and hole density can be solved. For (7.1) we get

$$\begin{aligned} n &= \frac{1}{V}\int_{E_g}^{\infty} \frac{V}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2}(E - E_g)^{1/2} e^{-(E-\mu)/k_B T} dE \\ &= \frac{(2m_e^*)^{3/2}}{2\pi^2\hbar^3} e^{\mu/k_B T}\int_{E_g}^{\infty}(E - E_g)^{1/2} e^{-E/k_B T} dE. \end{aligned} \tag{7.9}$$

The substitution $X_g = (E - E_g)/k_B T$ gives

$$n = \frac{(2m_e^*)^{3/2}}{2\pi^2\hbar^3}(k_B T)^{3/2} e^{-(E_g-\mu)/k_B T}\int_0^{\infty} X_g^{1/2} e^{-X_g} dX_g. \tag{7.10}$$

The last integral can be evaluated to give $\sqrt{\pi}/2$ so that the final result is

$$n = \frac{1}{\sqrt{2}}\left(\frac{m_e^* k_B T}{\pi\hbar^2}\right)^{3/2} e^{-(E_g-\mu)/k_B T} = N_{\text{eff}}^C e^{-(E_g-\mu)/k_B T}, \tag{7.11}$$

where N_{eff}^{C} can be viewed as an effective number of states per volume for the CB. The same calculation for the hole density gives

$$p = \frac{1}{\sqrt{2}}\left(\frac{m_h^* k_B T}{\pi\hbar^2}\right)^{3/2} e^{-\mu/k_B T} = N_{\text{eff}}^{V} e^{-\mu/k_B T}. \tag{7.12}$$

Equations (7.11) and (7.12) have a very intriguing and simple interpretation. Formally, they look like Boltzmann distributions for two energy levels at $E_g - \mu$ and μ. In this interpretation, the whole band character of the problem appears to be lost (it is still hidden in the effective masses), and it is quite sufficient to think of the VB and the CB as two discrete energy levels (see Figure 7.3b). Note that this is not entirely correct: The distribution is not purely Boltzmann-like because the effective numbers of states $N_{\text{eff}}^{V(C)}$are functions of the temperature themselves. Since their temperature dependence is, however, weak compared to the exponential term, it can often be neglected.

A useful relationship is derived by multiplying (7.11) by (7.12). We get

$$np = 4\left(\frac{k_B T}{2\pi\hbar^2}\right)^3 (m_e^* m_h^*)^{3/2} e^{-E_g/k_B T}, \tag{7.13}$$

meaning that the product of electron and hole concentrations is constant at any given temperature, independent of chemical potential's position. This equation is often called the **law of mass action** and particularly useful when treating doped semiconductors.

From this, we can finally calculate the carrier concentration for an intrinsic semiconductor at a given temperature. Obviously, the intrinsic electron density n_i has to be equal to the intrinsic hole density p_i and from (7.13) we get

$$n_i = p_i = \sqrt{np} = 2\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} (m_e^* m_h^*)^{3/4} e^{-E_g/2k_B T}. \tag{7.14}$$

Values for the important semiconductors Si and GaAs are given in Table 7.2. These densities are strongly temperature-dependent and much smaller than those of typical metals (Table 5.1), and we can therefore expect intrinsic semiconductors to be rather poor electrical conductors.

Again using (7.11) and (7.12), as well as the condition of charge neutrality $n = p$, we obtain an expression for the position of the chemical potential

$$\mu = \frac{E_g}{2} + \frac{3}{4} k_B T \ln\left(\frac{m_h^*}{m_e^*}\right). \tag{7.15}$$

Table 7.2 Intrinsic carrier densities for Si and GaAs.

Material	Gap size (eV)	n_i at 150 K (m^{-3})	n_i at 300 K (m^{-3})
Si	1.11	4.1×10^{6}	1.5×10^{16}
GaAs	1.43	1.8×10^{0}	5×10^{13}

For zero temperature, the chemical potential does indeed lie in the middle of the gap, as in our initial simple picture. Even at finite temperature, it remains in the middle of the gap, as long as the effective masses for holes and electrons are equal. However, in the case of different effective masses, there is a temperature-dependent correction. If, for example, the holes are heavier than the electrons, that is, the VB curvature is low and the CB curvature is high, many more holes than electrons would be generated at elevated temperature for $\mu = E_g/2$. In order to avoid this, the chemical potential has to move up as the temperature is raised.

To calculate all this, we needed essentially only three parameters: the gap size E_g and the effective electron and hole masses m_e^* and m_h^*. How can these quantities be measured? The gap size is relatively easy to measure, at least for direct gap materials such as GaAs. The basic idea is that the semiconductor cannot absorb light if the photon energy $h\nu$ is smaller than the gap size E_g. Only if $h\nu$ exceeds E_g, electrons from the VB can be excited into the CB. The gap size can therefore be measured by studying the optical absorption of a semiconductor as a function of photon energy. Strong absorption sets in as $h\nu = E_g$.

The effective masses of electrons and holes can be determined by the technique of **cyclotron resonance**. When the semiconductor is placed in a static magnetic field B, the electrons move on circular (or helical) orbits around the axis of the field (see Figure 7.4). The angular frequency of this motion is calculated by equating the Lorentz force with the centripetal force, resulting in

$$\omega_c = \frac{Be}{m_e^*}. \tag{7.16}$$

The angular frequency and thus the effective mass can be measured by shining radio frequency waves from the side into the system and measuring the transmission to the other side. Strong absorption occurs when the radio frequency exactly matches ω_c. This is because the magnetic field causes a quantization of the (nearly) free electron energy levels with a constant level spacing of $\hbar\omega_c$, similar to a harmonic oscillator, and the radio frequency wave can induce transitions between these levels. A similar argument applies for the holes.

The effective masses for some semiconductor materials are given in Table 7.3. Note that the effective masses can be quite different from the free electron mass.

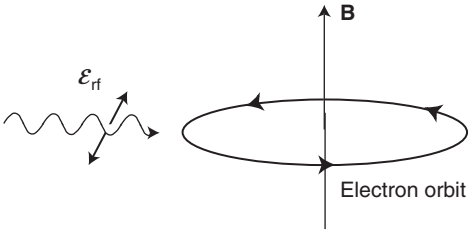


Figure 7.4 The measurement of cyclotron resonance. The electrons (or holes) are forced to move on circular orbits by a magnetic field B. Radio frequency radiation with the electric field vector $\mathcal{E}_{\mathrm{rf}}$ can induce transitions between quantized circular orbits.

Table 7.3 Effective masses for some semiconductors.

Material	m_e^*/m_e	m_h^*/m_e
Ge	0.60	0.28
Si	0.43	0.54
CdSe	0.13	0.45
InSb	0.015	0.39
InAs	0.026	0.41
GaAs	0.065	0.50

This is in contrast to the case of many simple metals. The electron masses are typically smaller than the hole masses, and they can be very small in some materials.

7.2 Doped Semiconductors

The carrier concentrations in most intrinsic semiconductors are too low to give any appreciable conductivity at room temperature. The situation can be changed by adding a very small amount of electrically active impurities, a process called doping. The dopant atoms change the conductivity by either donating (giving) electrons to the CB or accepting (taking) electrons from the VB, that is, by generating holes. They are therefore called **donors** and **acceptors**, respectively. It is evident that a very small concentration of such impurity atoms changes the carrier density drastically. Consider the case of Si. There are about 5×10^{28} Si atoms per cubic meter, but at room temperature, the intrinsic carrier concentration is only 1.5×10^{16} m^{-3}. This means that a dopant concentration in excess of 1.5×10^{16} m^{-3} would be sufficient to create more carriers than present in the intrinsic case, at least if every dopant atom gives rise to a free electron or hole. Thus, one dopant atom in 10^{12} silicon atoms would be sufficient to modify the carrier concentration significantly! It is, in fact, not possible to produce such pure samples so that there is always some amount of unintentional doping. The lowest impurity concentration that can currently be achieved is about 10^{18}m^{-3}.

7.2.1 n and p Doping

The two types of doping in a semiconductor are called **n and p doping**, for dopant atoms that give rise to free electrons in the CB (donors) and free holes in the VB (acceptors), respectively.

In silicon, n doping is achieved by putting pentavalent donor atoms such as P, As, or Sb into the lattice. These atoms have a valence configuration of s^2p^3, but only four of these five electrons are needed to form the sp^3 hybrid orbitals needed

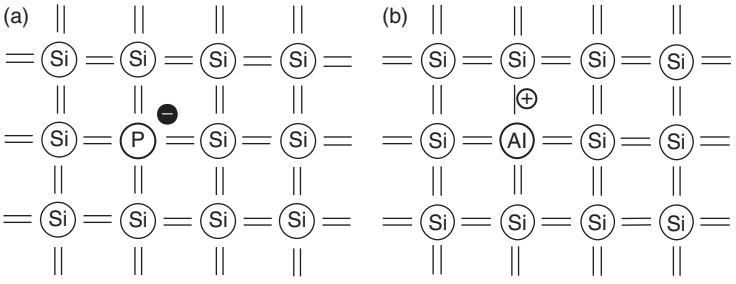


Figure 7.5 Nonionized dopant atoms in a Si lattice: (a) donor (b) acceptor.

to place the dopant atom in the lattice. The remaining electron remains loosely bound to the dopant ion, attracted by one positive net charge. This is shown in Figure 7.5a.

The important point about the n doping is that the extra electron has a very small binding energy and can therefore easily be removed by thermal excitations. We can estimate the binding energy by noticing the similarity of the problem to the hydrogen atom. The binding energies for hydrogen are

$$E_n = -\frac{m_e e^4}{8\epsilon_0^2 h^2}\frac{1}{n^2}. \tag{7.17}$$

A binding energy of zero would correspond to a free electron, that is, an electron that has left the positively charged dopant atom and moves freely in the CB. The state corresponding to $n = 1$ is the state that is most tightly bound and its binding energy thus corresponds to the ionization energy of the impurity. This energy is 13.6 eV for a free hydrogen atom. For an impurity in Si, however, we have to change (7.17) in two ways. First, we have to replace the electron mass m_e by the effective mass of a conduction electron in Si, which is $0.43m_e$. More importantly, we have to take into account that the donor "atom" is not placed in vacuum but in solid Si. This leads to a polarization of the Si atoms around the impurity, reducing the interaction. We can take this into account by replacing the dielectric constant of the vacuum ϵ_0 by $\epsilon_0\epsilon_{\mathrm{Si}}$, where $\epsilon_{\mathrm{Si}} = 11.7$. The physical origin of ϵ_{Si} is discussed in Chapter 9. The reduced effective mass and the polarization both reduce the binding energy, to the order of $E_d \approx 40$ meV or so. Since the ionized state corresponds to the electron in the CB, this binding energy refers to the CBM. The energy levels are clarified in Figure 7.6a.

The size of the nonionized donor "atom" is also affected by the reduced interaction due to the polarization of the Si atoms. We can estimate this by considering the Bohr radius, which is given by

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2}. \tag{7.18}$$

Applying the same substitutions as for the calculation of the energy, we see that this radius increases by a factor of about 30 with respect to the usual hydrogen Bohr radius.

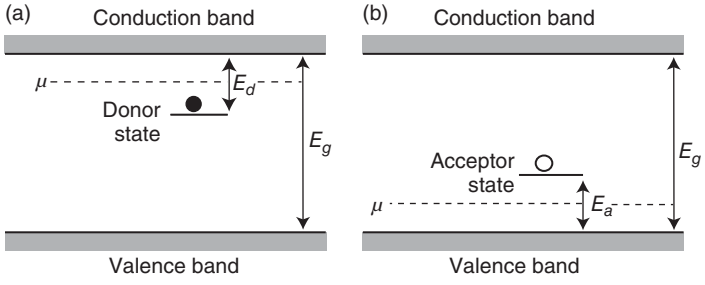


Figure 7.6 Energy levels for dopant atoms. (a) The donor ground state is placed just below the conduction band minimum. Ionization of a donor atom corresponds to the transfer of the extra electron into the conduction band. (b) The acceptor level is placed just above the valence band maximum. Ionization of an acceptor atom corresponds to accepting (taking) an electron from the valence band and thereby to generating a mobile hole.

Similar considerations apply to p doping with trivalent acceptor atoms such as B, Al, Ga, or In. This situation is shown in Figures 7.5b and 7.6b. The calculation of the binding energy is equivalent to the case of n doping, only that we now have a positively charged hole that is bound to a negatively charged ion.

7.2.2 Carrier Density

The calculation of the carrier density in a doped semiconductor is rather complicated and only sketched here. The basic principle is the same as in the intrinsic case: We have to fulfill a charge-neutrality condition to assure that a temperature change does not lead to a charging of the crystal. This can be formulated such that the number of electrons in the CB plus the number of (negatively) charged acceptor ions must be equal to the number of the holes in the VB plus the number of positively charged donor ions. The problem is further complicated by the fact that the replacement of the Fermi–Dirac statistics with the Boltzmann statistics is not necessarily appropriate because the chemical potential can be very close to the VB or CB edges.

Here, we just give a qualitative discussion for an n-doped semiconductor. The situation is shown in Figure 7.7. At temperatures much lower than E_d/k_B, essentially none of the donor atoms is ionized, they are **frozen out**. When the temperature is raised, the donors are ionized and their electrons are moved to the CB. In some intermediate temperature range, essentially all donors are ionized and the carrier concentration in the CB corresponds to the concentration of the donors. This temperature range is called the **extrinsic region**. At much high temperatures, the excitation of intrinsic carriers across the band gap becomes important, leading again to a strong increase of the carrier concentration. This happens in the **intrinsic region**. From this, it is qualitatively clear how the

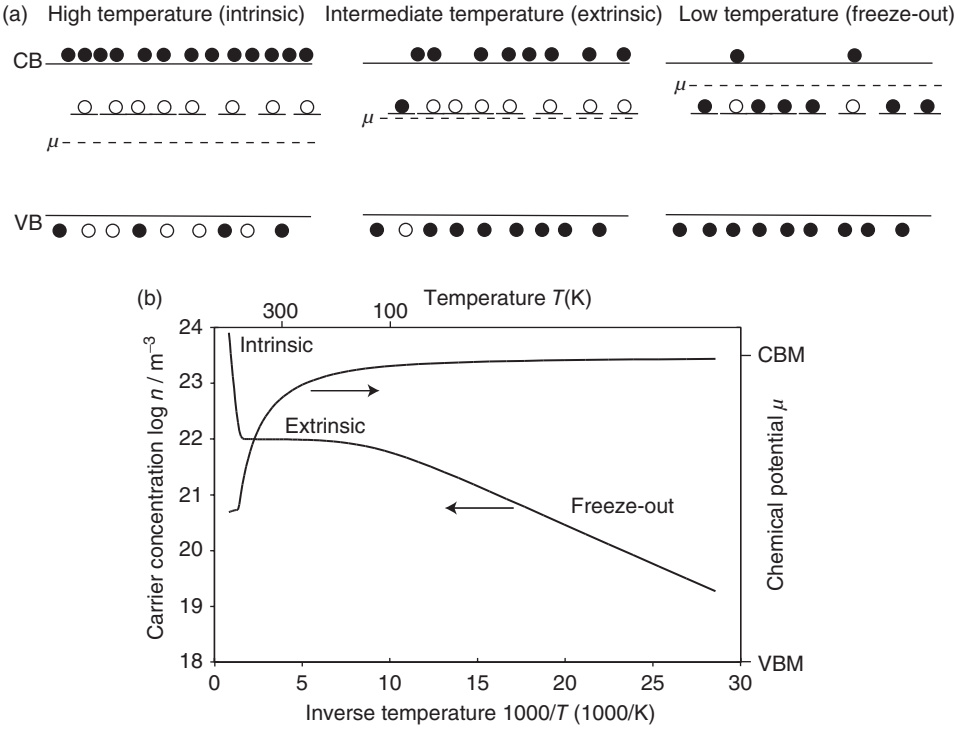


Figure 7.7 Electron density and position of the chemical potential for an n-doped semiconductor. (a) *Qualitative picture:* At very low temperatures, almost none of the donors are ionized (freeze-out region). At intermediate temperatures, almost all of the donors are ionized but very few intrinsic carriers are excited (extrinsic region). At high temperatures, all the donors are ionized and some intrinsic carriers are excited (intrinsic region). (b) *Quantitative picture:* The position of the chemical potential and the logarithmic carrier density are shown as a function of the inverse temperature.

chemical potential must behave. At very low temperatures, it must be situated between the CBM and the donor level E_d. At intermediate temperatures, it moves toward the middle of the gap. In the intrinsic region, it is close to the middle of the gap. The calculated position of the chemical potential is also shown in Figure 7.7b.

The carrier concentration in Figure 7.7b was calculated numerically for n-doped Si that would typically be used in a semiconductor device. Note that the extrinsic region is found around room temperature. There the electron concentration is much higher than in the intrinsic case but almost independent of the temperature. We shall see later that this is the key to a working semiconductor device: Via doping, we have the possibility to tune the electron or hole concentration to the desired value and it is then fairly stable for a range of operating temperatures. The construction of a working device would

hardly be possible if the electron concentration depended exponentially on the temperature near room temperature. Note also that the chemical potential in the extrinsic region is quite far away from the CBM near room temperature, so that we may even use the Boltzmann approximations to the Fermi–Dirac distribution (7.7) and (7.8) to calculate the carrier density, at least for the purpose of a semi-quantitative discussion of some semiconductor devices' working principles.

A useful expression for calculating the hole and electron densities is the law of mass action (7.13), which remains valid because it is not based on any assumptions about the position of the chemical potential. The fact that np is constant at a given temperature has some interesting consequences. If we increase the number of electrons by n doping, the number of holes in the VB has to be lower than that in the intrinsic case because the product of n and p has to remain the same. The law of mass action can thus be used to calculate the hole concentration p. The mechanism responsible for the constant np is that electrons and holes annihilate each other when they meet, since an electron in the CB can gain energy by taking the place of a hole in the VB. This process is called **carrier recombination**. If the concentration of one type of carrier is increased over the intrinsic case by doping, this also increases the chances of these extra carriers annihilating the carriers of the other type.

In a doped semiconductor, the number of one type of carriers is greatly enhanced by the doping. These carriers are called the **majority carriers**. The other type of carriers is still there, but we have just seen that its concentration is even smaller than in the intrinsic case at the same temperature. These carriers are called the **minority carriers**. One could be tempted to think that the minority carriers are utterly unimportant because there are so few. But this is actually not the case. The minority carriers are of essential importance for some semiconductor devices such as transistors.

The density of electrons and holes can be measured using the Hall effect described in Section 5.2.2. The Hall effect gives the sign and the density of the carriers. For a semiconductor, a positive R_H has the obvious interpretation as being caused by the holes in the VB. If we have purely n- or p-doped samples and the intrinsic carriers do not play a role, the interpretation of R_H is simple. If both types of doping are present, however, both holes and electrons contribute to R_H and the situation becomes more complicated (see Problem 7.6).

Finally, the results from semiconductors provide us with a plausible explanation of the fact that some metals show a positive R_H, too. The sign of R_H is determined by the details of the band structure and the position of the Fermi energy in that band structure. If the Fermi energy of a metal is placed just below a band maximum similar to the VBM in a semiconductor, the electrons in that metal will behave like holes in a Hall measurement. Similarly, the magnitude of R_H is determined by the details of the band structure. The unusual and instructive case of Bi (see Table 5.1) is discussed in an online note on www.philiphofmann.net.

7.3 Conductivity of Semiconductors

In our discussion of the conductivity of metals in the quantum model, we have seen that the simple Drude formula (5.9) can be viewed as approximately correct if we replace the free electron mass with the effective mass at the Fermi energy. The conduction of electrical current in a semiconductor is somewhat different because it is carried by both electrons and holes. We therefore have to modify the expression for the conductivity obtained in the Drude model, and we do so using the concept of the **mobility**, see (5.11) and (5.12). The total conductivity is then

$$\sigma = e(n\mu_e + p\mu_h), \tag{7.19}$$

with n and p representing the electron and hole densities and μ_e and μ_h their mobilities. For semiconductors, using the mobilities in the expression for σ is extremely convenient because the electron and hole concentrations can change over many orders of magnitudes whereas the mobilities are (approximately) constant for a given material. Note that the mobility contains both the relaxation time and the effective mass of the carriers. For most semiconductors, the electron effective mass is smaller than the hole effective mass and, therefore, a higher conductivity can be achieved by n doping than by p doping (assuming that the relaxation time is the same for both conduction mechanisms).

One of the most important characteristics of a semiconductor is the temperature dependence of its conductivity. This is so different from that of a metal, that it is sometimes also used to define semiconducting behavior. For a metal, we have seen that conductivity decreases as the temperature is raised. This is because of the increasing probability for electrons to be scattered by phonons and the accompanying reduction of the relaxation time. In a semiconductor, the same effect also leads to a shorter relaxation time and to a decreased mobility at higher temperatures. However, much more important is that the carrier densities (n or p or both) in (7.19) are typically increasing at higher temperature. This increase is much stronger than the decrease of the mobility and (7.19) therefore predicts an increased conductivity at higher temperatures. Following Figure 7.7, we see that this effect can be dramatic with the carrier concentration increasing over many orders of magnitude in a small temperature window, especially in the intrinsic regime.

However, Figure 7.7 also tells us that there are temperature ranges where the carrier concentration of a doped semiconductor can be almost constant. In the figure, this is the case around room temperature. If we measure the temperature-dependent conductivity in a small temperature interval around room temperature, it will therefore not change very much. It could even appear "metallic," showing a decrease of σ for higher T, if the effect of increased phonon scattering is more important than the additional carriers at higher temperatures. In fact, for a semiconductor device, it is very important the conductivity is relatively constant near the operating temperature. We therefore refrain from defining semiconducting behavior via the temperature-dependent conductivity.

7.4 Semiconductor Devices

Semiconductor devices are arguably the most important application of solid state physics to date. Most semiconductor devices are based on the physical phenomena that appear in inhomogeneous semiconductors, that is, semiconductors for which the doping concentration and type depend on the location. The simplest example of an inhomogeneous semiconductor is that of a pn junction. We treat pn junctions semi-quantitatively and the more complicated transistors and optoelectronic devices just qualitatively.

7.4.1 The pn Junction

The simplest inhomogeneous semiconductor device is the junction between an n- and a p-doped semiconductor. Such a junction can be used as a check valve for the current, a so-called **diode**. In real devices, such junctions are not fabricated by joining bits of semiconductors but by starting with an undoped material and doping it inhomogeneously (e.g., through diffusion of impurities into the material). In this way, no absolutely sharp boundary can be achieved, but we assume the existence of such a boundary here for simplicity.

Figure 7.8 shows what happens when we join p- (left) and n- (right) doped semiconductors. In Figure 7.8a, the two semiconductors are still separated. At not too high temperatures, the chemical potential of the p-doped semiconductor is close to the VB, and in the n-doped semiconductor, it is close to the CB. In the p-doped sample, most of the acceptors are negatively charged and the majority carriers are holes. There are also some minority electrons. In the n-type sample, most donors are ionized, the majority carriers are electrons, and there are also some minority holes.

Figure 7.8b shows what happens when the two semiconductors are joined. Electrons from the n side diffuse into the p side and holes from the p side diffuse into the n side. When mobile electrons and holes meet, they recombine. What is left then is a region of immobile ionized donors and acceptors without the compensating charge of the generated carriers. Because of the absence of mobile carriers, this region is called the **depletion layer**. The ionized donors and acceptors give rise to an electric field between the two sides, marked by $\mathcal{E}$ in the figure. This field represents an obstacle for the p holes to move into the n part and for the n electrons to move into the p part. The field increases as the depletion layer widens, and the process goes on until equilibrium is reached.

The same phenomenon can be described more formally by stating that in thermal equilibrium, the chemical potential has to be constant in the whole system. This implies a situation as in Figure 7.8c. The chemical potential is aligned in the whole system, but this can only be achieved by a macroscopic potential in the depletion region, which shifts the energy levels and "bends" the bands. The size of

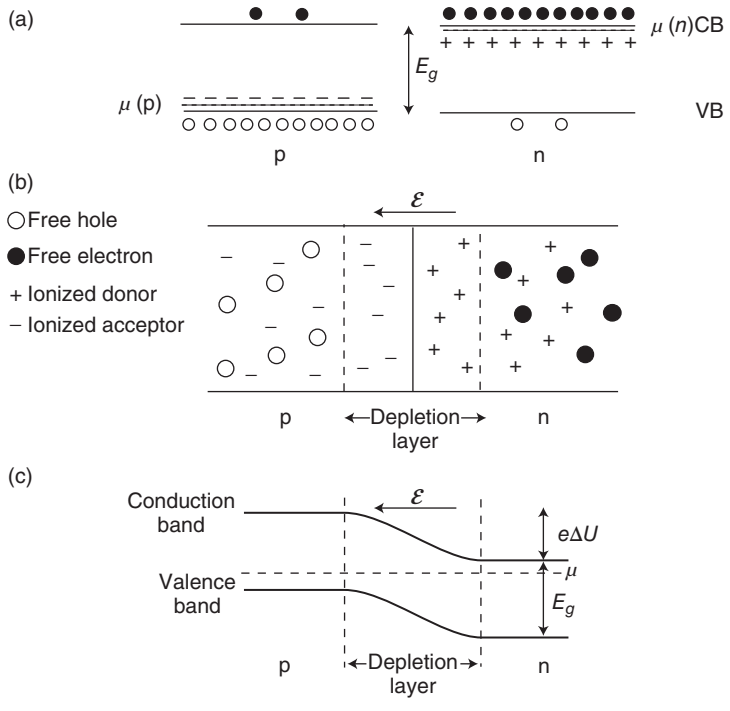


Figure 7.8 The pn junction. (a) Energy levels and carrier densities in separate p and n semiconductors. Ionized donors (acceptors) are symbolized by plus (minus) signs. The chemical potential is very close to the VB and the CB for p and n doping, respectively. (b) Formation of a depletion layer with an electric field $\mathcal{E}$ when the junction is established. (c) A position-independent chemical potential requires band bending over the depletion zone.

the potential change over the junction and the width of the depletion layer are the central quantities to be determined in the treatment of the pn junction.

The potential in the depletion region can be calculated by assuming an instantaneous transition between the depletion region and the nondepleted regions as shown in Figure 7.9a. The charge densities in the p and n part of the depletion region are $\rho_p = -eN_a$ and $\rho_n = +eN_d$, where N_a and N_d are the acceptor and donor concentrations, respectively. Charge neutrality furthermore requires that $N_a d_p = N_d d_n$, where d_p and d_n are the depths of the depletion layer in the p and n sides, respectively. If x is the direction perpendicular to the interface, it is possible to calculate the macroscopic potential $U(x)$ by solving the Poisson equation

$$\frac{d^2U}{dx^2} = -\frac{\rho}{\epsilon\epsilon_0}, \tag{7.20}$$

with ϵ being the dielectric constant of the semiconductor and the boundary conditions that both $U(x)$ and $dU(x)/dx$ have to be continuous at $x = 0$, as well as

$$\left.\frac{dU}{dx}\right|_{x=-d_p,d_n} = 0. \tag{7.21}$$

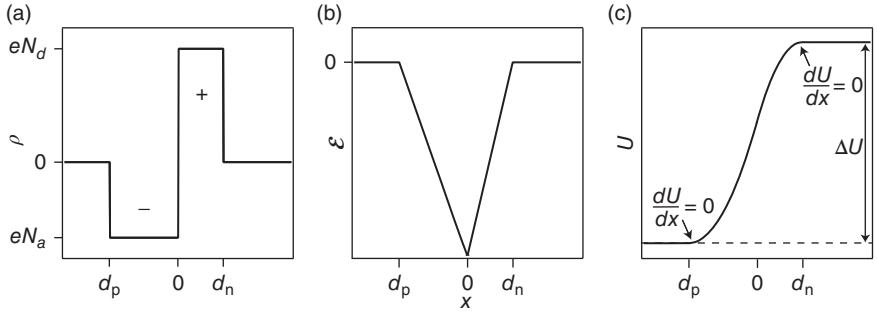


Figure 7.9 Idealized model of the depletion zone solved using the Poisson equation. (a) Charge density in the depletion zone. (b) Electric field. (c) Electrostatic potential.

The solutions for the electric field and the potential are shown in Figure 7.9b and c (see Problem 7.8 for the calculation). The total potential difference across the space charge layer is found to be

$$\Delta U = \frac{e}{2\epsilon\epsilon_0}\left(N_d d_n^2 + N_a d_p^2\right). \tag{7.22}$$

From this and the charge-neutrality conditions, the depth of the depletion zones can be calculated as

$$d_p = \left(\frac{\Delta U 2\epsilon\epsilon_0}{eN_a}\frac{N_d}{N_a + N_d}\right)^{1/2}, \qquad d_n = \left(\frac{\Delta U 2\epsilon\epsilon_0}{eN_d}\frac{N_a}{N_a + N_d}\right)^{1/2}. \tag{7.23}$$

The problem is that we know neither ΔU nor d_p and d_n. However, we do know that the chemical potential at low temperatures must be close to the VBM in the p-doped region and close to the CBM in the n-doped region. Consequently, $\Delta U \approx E_g/e$. From this, we can estimate the thickness of the depletion layer to be between 0.1 and 1 μm. This is very big on the atomic scale, justifying our macroscopic approach when using the Poisson equation.

We proceed with a simplified semi-quantitative treatment of the pn junction, illustrating how it can be used as a check valve for the current. First we write down the approximate carrier densities in the CB and VB on both sides of the junction. We use the notation that n_n and p_n are the electron and hole densities on the n side and the corresponding notation for the p side. With the energy diagram in Figure 7.10 as well as the expressions for the carrier densities in the Boltzmann approximation (7.11) and (7.12), we can write down the densities.

$$n_p = N_{\text{eff}}^C e^{(\mu - E_g - e\Delta U)/k_B T}, \qquad p_p = N_{\text{eff}}^V e^{(e\Delta U - \mu)/k_B T}, \tag{7.24}$$

$$n_n = N_{\text{eff}}^C e^{(\mu - E_g)/k_B T}, \qquad p_n = N_{\text{eff}}^V e^{-\mu/k_B T}. \tag{7.25}$$

The equilibrium of the pn junction is not static in the sense that there are no currents across the junction. In the following, we discuss the existing currents for electrons. Equivalent arguments can be made for hole currents. The majority electrons on the n side can diffuse into the p side if their energy is high enough to overcome the electric field in the depletion layer. This gives rise to a so-called

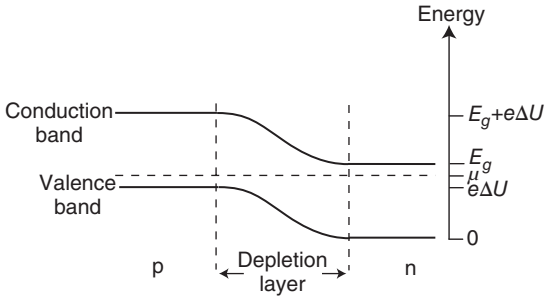


Figure 7.10 Definition of the energies in the pn junction. The VBM on the n side is taken as the energy zero.

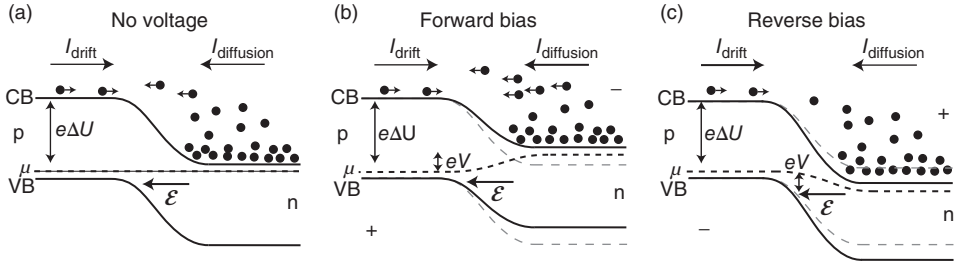


Figure 7.11 The pn junction as a diode (only considering the electrons, not the holes). (a) Without any bias voltage. (b) Forward bias. (c) Reverse bias. The gray dashed lines show the position of the n-side VB and CB without an applied voltage.

diffusion current. In equilibrium, there is an equal current of electrons from the p side to the n side. It is caused by minority electrons on the p side that enter the depletion layer. These electrons are pulled to the n side by the electric field. This gives rise to the so-called **drift current**. The two currents are illustrated in Figure 7.11a.

In order to find an expression for the drift and diffusion currents, let us assume that these are proportional to the electron density on the p side (7.24) or, what is the same, to the electron density on the n side for an energy just high enough to overcome the barrier. We can thus write

$$|I_{\mathrm{diffusion}}| = |I_{\mathrm{drift}}| = |I_0| = Ce^{(\mu - E_g - e\Delta U)/k_B T}, \tag{7.26}$$

where C is a proportionality constant.

Now it is easy to see how the pn junction with an applied external voltage acts as a kind of check valve for the current. If we apply a voltage V across the junction, we can assume that the whole voltage drop occurs over the depletion zone, that is, we can rigidly shift the bands on one side by an amount eV against the other side. This assumption is quite sensible because the resistance of the pn junction is dominated by the depletion zone where there are no free carriers.

The situation for an external field opposing the field in the depletion zone (forward bias) is shown in Figure 7.11b. As an energy zero, we keep the position of the VBM on the n side without an applied voltage. The electron drift current is unchanged, but the diffusion current is modified because the chemical potential on the n side is shifted by eV:

$$|I_{\text{diffusion}}| = Ce^{((\mu+eV)-E_g-e\Delta U)/k_BT}. \tag{7.27}$$

The net current across the junction is the difference between drift and diffusion currents:

$$I = I_{\text{diffusion}} - I_{\text{drift}} = I_0\left(e^{eV/k_BT} - 1\right). \tag{7.28}$$

This current is zero without a bias voltage and increases exponentially as the bias voltage is raised. Qualitatively, this is easy to understand because the external field lowers the barrier that has to be overcome by the majority electrons on the n side to cross to the p side.

Figure 7.11c shows the situation for the voltage applied in the opposite direction (reverse-bias). Now the chemical potential on the n side is lowered by eV; the diffusion current is

$$|I_{\text{diffusion}}| = Ce^{((\mu-eV)-E_g-e\Delta U)/k_BT}, \tag{7.29}$$

and the resulting current is

$$I = I_{\text{diffusion}} - I_{\text{drift}} = I_0\left(e^{-eV/k_BT} - 1\right). \tag{7.30}$$

For this polarity of the external voltage, the current is thus always small. It can never exceed I_0, which is about 100 nA. The resulting current–voltage ($I(V)$) curve of the pn junction is plotted in Figure 7.12. It shows the check valve behavior of a diode.

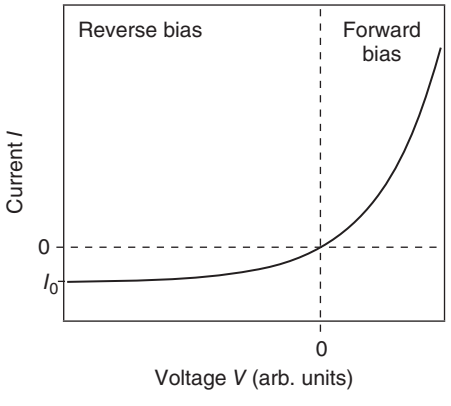


Figure 7.12 Characteristic $I(V)$ curve for a pn junction operated as diode.

7.4.2 Transistors

Transistors form essential building blocks for microelectronic devices because they can act as electrical amplifiers and switches. Here, there is not even space for a superficial discussion of the various transistor types and applications. In order to give you the flavor of transistor physics, we just pick out one specific type, a silicon **metal oxide field effect transistor (MOSFET)**, and discuss how this can be used as a switch.

The MOSFET is typically a part of an integrated circuit such as a computer memory or processor chip. A sketch of one transistor is shown in Figure 7.13a. The transistor is built onto a p-doped substrate. It consists of two n-doped regions called source and drain and an oxide layer (usually SiO_2) between them on the top of the substrate. The top of the oxide, the so-called gate, is contacted with a metal electrode and so are source and drain.

The switching behavior of the MOSFET is shown in Figure 7.13b and c. The basic idea is to use the gate voltage in order to control the current between the source and the drain. If no voltage is applied to the gate, no appreciable current of electrons can flow from the source the drain, independent of the voltage between them. This is because one of the two pn junctions in the MOSFET is always biased in the reverse direction. If, however, a positive voltage is applied to the gate, two things happen: The majority holes in the p-doped material are repelled from the oxide layer, and the minority electrons are attracted. If the voltage at the gate is larger than a threshold voltage, the minority electrons in a small channel below the oxide layer actually become the majority carriers, and an effective electron conduction is possible.

We can see how this switching is possible in Figure 7.14. For no applied voltage, the chemical potential inside the semiconductor is equal to the Fermi energy of the gate metal. When a positive voltage is applied, the electric field caused by this bends the semiconductor bands close to the interface. For a sufficiently large gate voltage, the band bending is so strong that the CB moves very close to the chemical potential, as in an n-doped semiconductor, even though the material is p-doped. Therefore, many states in the CB become populated by electrons and one speaks of an **inversion layer**.

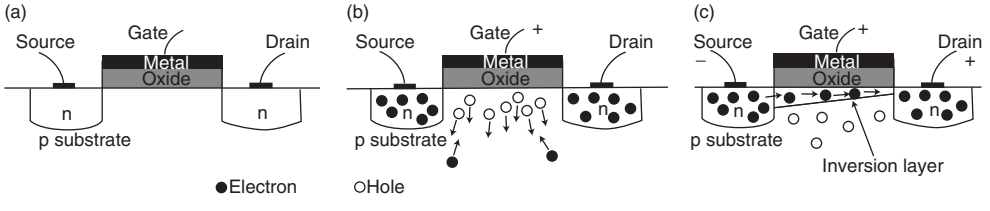


Figure 7.13 Design and working principle of a MOSFET: (a) without applied voltage; (b) with a small positive gate voltage; (c) with an applied voltage between source and drain and a gate voltage large enough to generate an inversion layer.

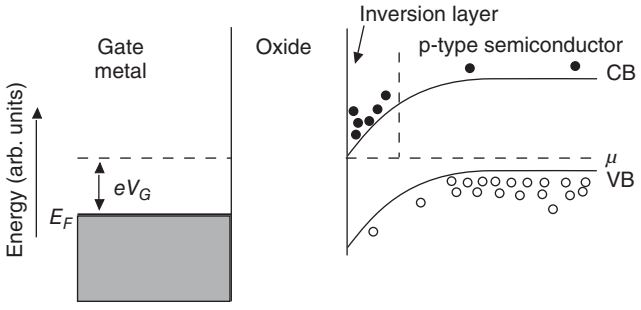


Figure 7.14 Generation of an inversion layer in the MOSFET. The positive gate voltage leads to a band bending that is strong enough to turn electrons from being minority carriers into being majority carriers.

7.4.3
Optoelectronic Devices

As for transistors, we can only give the main ideas for some optoelectronic devices here. There are two important classes: devices that turn electrical power into light and those that turn light into electrical power. Their principles of operation are closely related and sketched in Figure 7.15.

The **light emitting diode** shown in Figure 7.15a is essentially a normal diode operated in forward bias. The current is transported by transferring majority electrons from the n side to the p side and majority holes from the p side to the n side. Once these carriers have reached the depletion zone or the other side, there is a high probability for them to recombine with the other type of carrier that is in the majority. When this happens, the resulting energy can be emitted as light and the recombination process is called radiative. Since most electrons are close to the CBM and most holes close to the VBM, the energy of the emitted light is equal to the gap energy of the semiconductor. Different light colors can therefore be chosen via the appropriate semiconductor material.

Another very important design criterion is that the semiconductor material of choice actually favors radiative recombination between electrons in the CBM and

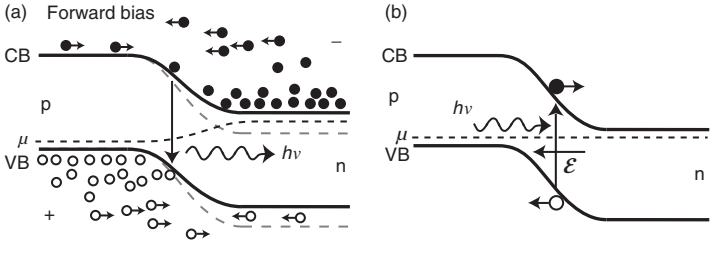


Figure 7.15 Optoelectronic devices. (a) A light emitting diode works because of radiative carrier recombination. (b) A photodetector or solar cell is based on carrier separation in the depletion zone.

holes in the VBM. This is not so for Si, the most important semiconductor by far, because of its indirect band gap: For the radiative recombination process, energy and crystal momentum have to be conserved. Energy conservation requires that the emitted photon has the energy of the band gap $h\nu = E_g$. Crystal momentum conservation, on the other hand, requires that the wave vectors of the electron and the hole can only differ by the wave vector of the generated photon (plus a reciprocal lattice vector). The wave vector of a photon is $k_{h\nu} = 2\pi\nu/c$, where c is the speed of light. For photons with energies $h\nu$ in the visible part of the spectrum, $k_{h\nu}$ is negligibly small. Therefore, the transition is only possible if the electron and the hole have the same crystal momentum $\hbar\mathbf{k}$, that is, if the transition is vertical in a band structure diagram. An inspection of the band structure of Si in Figure 6.13 shows that such a vertical transition is not possible between the VBM and CBM because they lie at different values of $\mathbf{k}$, the defining property of an indirect band gap. But this is not the case for all semiconductor materials. The band structure of GaAs, also shown in Figure 6.13, reveals that this material has a direct band gap with the CBM and the VBM at the same $\mathbf{k}$, and it is therefore better suited for optoelectronic applications. Light emitting diodes are exclusively made from materials with direct band gaps.

Power generation by light is realized in **solar cells** (Figure 7.15b), in a process that is the reverse of that used for light generation. Again, the device is set up as a pn junction. Photons of sufficient energy that hit the device can excite electrons from the VB into the CB and thereby generate electron–hole pairs. If an electron–hole pair is created in the depletion region, the built-in electric field sweeps the electron out to the n side and the hole to the p side. This gives rise to a voltage difference that can be used to drive a current. As for light emitting diodes, Si is not an ideal material for the construction of solar cells because of its indirect band gap. However, the problem is less severe: First of all, transitions across the indirect band gap are still possible when the required crystal momentum is provided by another particle, a phonon, for example. While such transitions are unlikely, this can be compensated by choosing a thicker layer of absorbing Si. Also, light with higher photon energies can still be absorbed directly, also leading to the generation of electron–hole pairs. Today, Si is the most common material for solar cells and relatively high efficiencies can be reached.

It is interesting to consider the actual design of a solar cell. A schematic illustration is given in Figure 7.16. The first step for power generation is the creation of electron–hole pairs by light absorption. For this to work, the photon energy must by higher than E_g. The incoming sunlight has a broad spectrum from the infrared to the ultraviolet. It can be described as black body radiation for a temperature of about 6000 K (the temperature of the sun's surface). This puts the maximum flux in the spectrum at ≈ 2.6 eV with a rapid decrease for higher energies. At the surface of the earth, however, certain light frequencies are strongly suppressed because of absorption in the atmosphere.

The solar cell in the figure resembles a typical silicon-based device. The cell is covered by an anti-reflective coating in order to make sure that most of the light hitting the surface also enters the cell. This can be achieved by an anti-reflective

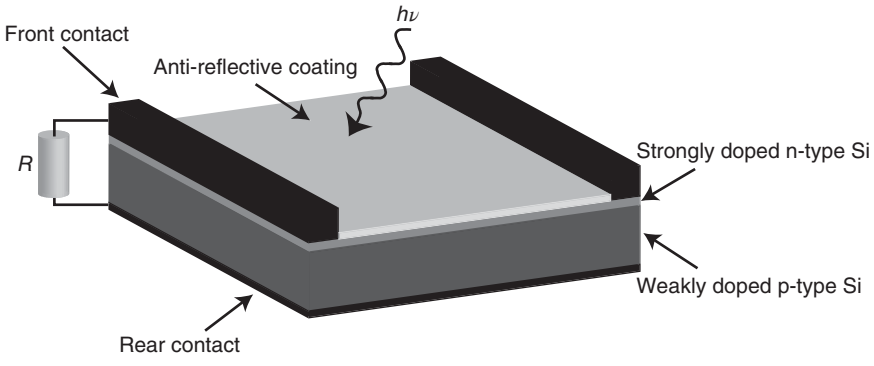


Figure 7.16 Sketch of a silicon solar cell and the electrical contact to an external load *R*. The light enters the cell through an anti-reflective coating and electron–hole pairs are generated in the silicon. These are separated by the junction between the n-type and p-type material.

transparent material, a special texture of the surface or both. The pn junction is realized between a thin strongly n-doped Si layer on the top and a more weakly p-doped layer below. When electron–hole pairs are created and separated in the device, they need to be collected by metal contacts on the front and the back of the solar cell, such that they can be used to drive a current through an external load, symbolized by the resistor *R*. For the back-contact, this is not a problem because the whole back side of the device can be covered by a metal contact. For the front side, this approach does not work since the light has to enter there, too. For Si-based devices, the n-doped layer in the front is sufficiently conductive to transport the generated electrons to metal bars that are positioned on the surface in a regular array. Alternatively, a thin, conductive, and nearly transparent material has to be placed on the top semiconductor layer to conduct the electrons to the outer contact. Clearly, the presence of such a layer or of the metal bars prevents part of the light from entering the cell.

When electron–hole pairs are generated by light absorption within the depletion layer, they are readily separated, as shown in Figure 7.15b. However, the separation can even work when the electron–hole pair is generated in the n or p type material further away from the depletion layer, where there is no electric field. For this to happen, it is sufficient that the minority carrier (the hole in the n layer or the electron in the p layer) diffuses to the space charge layer and is swept across it by the electric field. The likelihood of this decreases with the distance from the space charge layer.

A crucial issue for the generation of electric energy in the device is carrier recombination: The lifetime of the minority carrier of the electron–hole pair (e.g., the hole in the n-doped layer) is quite short because the risk of recombination with a majority carrier (an electron in the n-doped layer) is very high. Any such recombination prevents the photo-excited carriers from doing electric

work in the external circuit. Recombination can take place in different ways. We have already discussed radiative recombination that involves the emission of a photon, that is, the inverse process of electron–hole pair generation the device is intended to perform. It is also possible to have a three-body process, in which an electron recombines with a hole and gives the excess energy to another electron (a so-called Auger process). While these processes cannot be avoided, there are additional pathways for recombination that depend on the design of the solar cell: Notably, recombination can happen near structural defects, such as crystalline grain boundaries, atomic contaminants, or even dopant atom sites. Domain boundaries between small crystallites are present in solar cells based on polycrystalline Si, but they are avoided when using single crystal Si. Doing so does in fact lead to a higher efficiency, but the production cost is also higher. Carrier recombination is also favored at irregular surfaces and interfaces, but this can be limited by growing atomically flat interfaces or by passivating broken bonds at the surfaces.

There is a theoretical limit to the highest achievable efficiency of a solar cell, the efficiency being defined as the electrical energy provided by the cell divided by the total solar energy it is exposed to. This highest efficiency depends on the size of the band gap. Imagine the generation and separation of an electron–hole pair as in Figure 7.15b, and consider what happens to the electron on the n side. After the excitation, the electron will have an energy high above the CBM but due to scattering with other electrons, it will quickly loose this energy and end up near the CBM. This happens very quickly, within a hundred femtoseconds or so, much faster than possible recombination processes. The same happens to the hole on the p side. The energy difference between the hole on the p side and the electron on the n side can thus be no bigger than the size of the gap in the material (in Figure 7.15b, it is much smaller because of the strong doping difference on both sides). Only this energy difference can do work in the external circuit. If we build a solar cell from a semiconductor with a small band gap, we will thus only end up with usable electron–hole pairs that have (at best) the energy of this gap. This energy is increased for a material with a wider gap, but then we harvest fewer photons because only those with energies larger than E_g can at all excite electron–hole pairs. In both extremes, the efficiency drops and there must therefore be a maximum somewhere in the middle. It turns out that the theoretical maximum efficiency is around 30%, reached for $E_g \approx 1.2$ eV, quite close to the gap value for Si. This is known as the **Shockley–Queisser limit**. It does not take the detrimental effect of defect-induced recombination into account. In real Si solar cell devices, efficiencies of around 25% are achievable but for most commercial products, the value is lower. Using a more complex solar cell design based on GaAs with multiple junctions and external light concentration, one can beat the Shockley–Queisser limit, which is only valid for a single junction. Efficiencies of more than 40% have been reached in the laboratory but at a very considerable cost.

7.5 Further Reading

A detailed discussion on semiconductor physics is given in the general solid state physics book

- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.

More detailed information on semiconductor devices can be found in

- Sze, S.M. (1982) *Physics of Semiconductor Devices*, 2nd edn, John Wiley & Sons, Inc.
- Van Zeghbroeck, B. (2011) *Principles of Semiconductor Devices*, online book under *http://ecee.colorado.edu/~bart* (accessed 13 November 2014).

and on solar cells in

- Honsberg, C. and Bowden, S. online book on pveducation.org, *http://www.pveducation.org/pvcdrom* (accessed 13 November 2014).
- Nelson, J. (2003) *The Physics of Solar Cells (Properties of Semiconductor Materials)*, Imperial College Press.

7.6 Discussion and Problems

Discussion

1) What is the difference between a metal, a semiconductor, and an insulator?
2) How does the conductivity of a semiconductor (typically) change as a function of temperature and why?
3) In an intrinsic semiconductor, the chemical potential lies in the middle of the gap at low temperatures. Why?
4) Explain the difference between "electrons" and "holes."
5) What is the concentration of mobile electrons and holes as a function of temperature for an intrinsic semiconductor?
6) What is the meaning of the effective mass, and how can it be related to the electronic band dispersion?
7) If you have a semiconductor with heavy holes and light electrons around the VBM and CBM, respectively, can you draw the bands schematically?
8) How can you measure the effective mass of the carriers?
9) How can you measure the concentration of electrons or holes?
10) Why does a very small number of donor or acceptor atoms in a semiconductor have a big impact on the concentration of free carriers?
11) The interaction between a donor (acceptor) ion and its extra electron (hole) can be described in a way very similar to the Bohr model for the hydrogen atom, but the binding energy is much smaller and the radius much bigger. Why?

12) Where does the chemical potential in an n-doped semiconductor lie at low and high temperatures?
13) Explain why a pn junction works as a check valve for the current.
14) Explain why GaAs is a more appropriate material for optoelectronic applications than Si.

Problems

1) *Nondegenerate semiconductors:* Derive the simplified expressions for the Fermi function (7.7) and (7.8), assuming that the chemical potential is situated approximately in the middle of the gap.
2) *Intrinsic semiconductors:* The chemical potential in an intrinsic semiconductor is given by (7.15). Using the band structure in Figure 7.3a, explain qualitatively why it depends on the effective masses in this way.
3) *The chemical potential in metals:* In Chapter 6, we have simply identified the chemical potential μ in a metal with the Fermi energy E_F and stated that this is a good approximation for all temperatures, that is, that the position of the chemical potential is independent of the temperature. Use the charge-neutrality arguments introduced in this chapter to show that this is indeed a justified approximation.
4) *Doped semiconductors:* Consider a nonionized phosphorus donor atom in a Si crystal. (a) What is the "Bohr radius" of the resulting "atom?" (b) Estimate how many Si atoms are contained within a sphere of this "Bohr radius"? (c) Estimate how high the concentration of impurities would have to be for the "Bohr radii" to overlap, and what would you expect to happen in this case?
5) *Doped semiconductors:* (a) Calculate the effective number of states per volume N^C_{eff} for the conduction band of silicon at T =150 K and T =300 K. (b) In the intermediate temperature (extrinsic) case of Figure 7.7a, it appears that almost all the electrons in the donor atoms have been excited into the conduction band. On the other hand, the chemical potential in this situation must be close to the middle of the gap, which means that the Fermi–Dirac distribution is small at the donor levels but even smaller in the conduction band. So, why are the electrons in the conduction band and not in the donor levels? (Hint: Keep in mind that the doping concentration is on the order 10^{20} m^{-3}).
6) *Doped semiconductors:* It is technically very difficult to produce semiconductor crystals showing truly intrinsic behavior. For the same reason, it is hard to fabricate a truly n- or p-doped semiconductor, and in general, both types of dopant atoms will be present (although the concentration of one will generally be much higher). Show that in the case of simultaneous n and p doping, the Hall coefficient is given by

$$R_H = \frac{\mathcal{E}_H}{Bj_x} = \frac{p\mu_h^2 - n\mu_e^2}{e(p\mu_h + n\mu_e)^2}. \tag{7.31}$$

7) *Optical properties:* For an indirect gap semiconductor such as silicon, the difference in wave vector between the VBM and the CBM is on the order of

$\delta k = \pi/a$, where a is the lattice constant of the material. (a) Explain why this is so. (b) Estimate δk for Si, where the lattice constant is 0.543 nm. (c) For an optical transition to take place, the photon needs to have at least the gap energy (1.11 eV for Si) and it has to have a wave vector corresponding to δk. What is the actual size of the wave vector of 1.11 eV photons? For which type of electromagnetic radiation does the modulus of the wave vector become comparable to the reciprocal lattice distances in a solid?

8) *The pn junction:* Solve the Poisson equation (7.20) to calculate the potential within the pn junction and the absolute potential difference (7.22) across the junction.

8
Magnetism

In this chapter, we are concerned with the magnetic behavior of solids. We can divide this into two categories. We shall first discuss how solids react to an external magnetic field. For most materials, not much happens: The magnetic effects are weak and can largely be understood by the properties of the atoms making up the solid. We then inspect the more interesting case of a spontaneous magnetic ordering in the absence of an applied magnetic field. This is obviously a genuine effect of the solid that cannot be derived from atomic properties. As it turns out, it is rather difficult to describe magnetic ordering by a simple model.

It is often appropriate to consider magnetic ordering between just the spin magnetic moments of the electrons. We could then choose a model based on local spins with some interaction between them or a model with completely delocalized electrons, but the possibility of one spin (magnetization) direction to prevail. The problem is that an accurate description lies somewhere in between these extremes. Another difficulty is that it is no longer a good approximation to consider one electron in the mean potential of all the other electrons: When describing the interaction between spins, the electrons in the immediate vicinity of a given electron are more important than those that are further away, and this cannot be captured by a mean potential. We will anyway attempt to describe magnetic interactions using an averaged interaction between electrons, and we will at least be able to account for the basic phenomenon of ordering.

8.1
Macroscopic Description

Before we start to explore the magnetic properties of solids, it is quite useful to review the basics of magnetostatics. In general, we assume that Gauss' law for magnetostatics

$$\oint \mathbf{B} d\mathbf{a} = 0, \qquad \mathrm{div}\mathbf{B} = 0 \tag{8.1}$$

is obeyed: There are no magnetic monopoles. The sources of the **magnetic induction B** are magnetic dipoles. In vacuum, the magnetic induction is related to the **magnetic field H** by

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

$$\mathbf{B} = \mu_0 \mathbf{H}, \tag{8.2}$$

with the **permeability of vacuum** μ_0 ($\mu_0 = 4\pi \times 10^{-7}$ V s A^{-1} m^{-1} = $4\pi \times 10^{-7}$ T^2 m^3 J^{-1}). Note that the SI unit of $\mathbf{B}$ is the Tesla = kg s^{-2} A^{-1} and that one Tesla is a rather strong magnetic field. The Earth's magnetic field is typically on the order of 5×10^{-5} T, and strong magnetic fields in medical magnetic resonance scanners are only a few Tesla.

In matter, we have

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) = \mathbf{B}_0 + \mu_0 \mathbf{M}, \tag{8.3}$$

where $\mathbf{M}$ is the macroscopic magnetization of the solid and it is useful to regard $\mathbf{B}_0 = \mu_0 \mathbf{H}$ as the "external field." The unit of this field is T, like for $\mathbf{B}$. The magnetization $\mathbf{M}$ can be viewed as the number N of magnetic dipole moments $\boldsymbol{\mu}$ per volume V,

$$\mathbf{M} = \boldsymbol{\mu} \frac{N}{V}. \tag{8.4}$$

The unit of $\mathbf{M}$ is J T^{-1} m^{-3}. In many cases, there is a linear relation between the external field and the magnetization:

$$\mu_0 \mathbf{M} = \chi_\mathrm{m} \mathbf{B}_0, \tag{8.5}$$

where χ_m is called the **magnetic susceptibility**.[1] If the susceptibility is negative, the solid is called **diamagnetic**, and if it is positive, the solid is called **paramagnetic**. Instead of using the susceptibility, one can describe the magnetic properties of matter by the **relative permeability** $\mu = 1 + \chi_\mathrm{m}$. We do not use the relative permeability in this chapter but be aware of the possible confusion that can arise because the relative permeability and the magnetic moments are both denoted by μ. Note that the linear relation (8.5) does not always hold. In some cases, a nonlinear description must be used or the relation might even depend on the history of the piece of material at hand. We will encounter this in the case of ferromagnetism.

Like for an electric dipole in an electric field, the potential energy U of a magnetic dipole $\boldsymbol{\mu}$ in a field $\mathbf{B}_0$ is $U = -\boldsymbol{\mu} \cdot \mathbf{B}_0$. One could thus think that the energy of a macroscopic object of volume V and magnetization $\mathbf{M}$ is simply $U = -V\mathbf{M} \cdot \mathbf{B}_0$. This is also correct, but only if $\mathbf{M}$ does not depend on the field. In the case that it does, as in (8.5), one has to take into account that the energy change for a small increase of the field dB_0 depends on the already induced magnetization, giving an energy change of $dU = -V\mathbf{M}d\mathbf{B}_0$. When the field is turned on from zero to B_0, we thus get an energy of

$$U = -V \int_0^{B_0} M dB_0' = -V \int_0^{B_0} \frac{\chi_\mathrm{m}}{\mu_0} B_0' dB_0' = -V \frac{\chi_\mathrm{m}}{2\mu_0} B_0^2, \tag{8.6}$$

where we have ignored the vectorial character of $\mathbf{M}$ and $\mathbf{B}_0$ since the magnetization and the field are either parallel or antiparallel. For a **paramagnetic** solid,

1) Note that χ_m is dimensionless here. In the literature, other units for χ_m can be found, depending on the definition of M as magnetization per unit volume, unit mass, or per one mol of substance. Under some circumstances, χ_m is also defined as $\mu_0 \partial \mathbf{M}/\partial \mathbf{B}_0$.

U is thus negative, corresponding to an energy decrease for higher fields. Paramagnetic solids therefore experience a force toward locations of higher magnetic fields, that is, they are attracted to either pole of a permanent magnet. For **diamagnetic** solids, the opposite is true; they are expelled from regions of high magnetic fields. As we shall see later, χ_m is usually very small, so that these effects are not noticeable when you play with a permanent magnet and diamagnetic or paramagnetic solids.

It is tempting to explain these magnetic phenomena in a simple classical picture. In the case of diamagnetism, such an explanation comes straight from Lenz's law: An increasing external magnetic field is experienced by all the electrons in the atoms making up the solid, and this leads to the induction of microscopic currents. According to Lenz's law, the magnetic moment arising from these currents opposes the external field and hence one observes diamagnetic behavior. While diamagnetism is therefore always present, paramagnetism can only be observed when the atoms in the solid show a net magnetic moment, already without an external field. Such magnetic moments can align with the external field, leading to an energy gain. Atoms do not necessarily have a net magnetic moment because all of the electrons' orbital and spin magnetic moments may cancel out, but when such a moment is present, it usually dominates the diamagnetism.

While this picture is intuitive, it is also misleading. It turns out that a classical treatment does not give rise to any magnetism, even though some of the solid's magnetic properties are predicted correctly by classical arguments.[2] The failure of classical physics to account for magnetism is known as the **Bohr–van Leeuwen theorem**.

8.2 Quantum Mechanical Description of Magnetism

In view of the failure of classical physics to account for magnetism, a quantum mechanical treatment is called for. We approach this here from a rather general point of view, asking how the energy of an electron changes when a weak magnetic field is included as a small perturbation in the Schröding equation. In principle, this treatment can be applied to both atoms and solids. However, already in isolated atoms, the situation is complicated because the magnetic moments of the many electrons in the atom have to be added in the right way. This is treated in the next section. It does not change the physical principles illustrated here.

Before we can use perturbation theory to see how a magnetic field changes the energy of an electron, we have to discuss how the Schrödinger equation changes in the presence of an electromagnetic field. To describe this in a convenient way, we need the concept of the so-called **vector potential**, which you may not be familiar with. The key idea is as follows: In electrostatics, the electric field $\boldsymbol{\mathcal{E}}(\mathbf{r})$ can be generated by means of a potential $\phi(\mathbf{r})$ such that $\boldsymbol{\mathcal{E}}(\mathbf{r}) = -\text{grad}\phi(\mathbf{r})$. The

2) See online note on www.philiphofmann.net.

introduction of this potential greatly simplifies many calculations as we only have to find the (scalar) potential instead of the (vectorial) field. Due to the nonexistence of magnetic monopoles, it is impossible to define a similar scalar potential for the magnetic field, but one can define a so-called vector potential such that

$$\mathbf{B} = \mathrm{curl}\mathbf{A}. \tag{8.7}$$

As we shall see in a moment and later in Chapter 10, the introduction of $\mathbf{A}$ greatly simplifies the coupling of an external electromagnetic field into the Schrödinger equation. Using $\mathbf{A}$ also simplifies the notation in many other situations, for example, for obtaining the wave equation of the electromagnetic field from the Maxwell equations.

Having the vector potential $\mathbf{A}(\mathbf{r})$, and the scalar potential $\phi(\mathbf{r})$, the rules for coupling an external electromagnetic field to the Schrödinger equation are quite simple: The scalar potential obviously only acts as an addition to the potential that is already present, so we have to multiply it with the charge q of the particle described by the Schrödinger equation and add it to the Hamiltonian. The magnetic field is included by substituting the momentum operator $\mathbf{p} = -i\hbar\nabla$ by

$$\mathbf{p} \to \mathbf{p} - q\mathbf{A}. \tag{8.8}$$

We now come back to the original problem to find out how a weak external magnetic field changes the energy of an electron in an atom. Let us say that we have a magnetic field of strength B_0 only in the z direction, that is, $\mathbf{B}_0 = (0, 0, B_0)$. A vector potential generating this field is

$$\mathbf{A} = -\frac{1}{2}\mathbf{r} \times \mathbf{B}_0, \tag{8.9}$$

which is easily verified by an explicit calculation of curl$\mathbf{A}$. We know that the vector potential only affects the kinetic energy term of the electron, so we do not have to bother with the potential energy here. The original kinetic energy part of the Hamiltonian is now modified such that

$$\begin{aligned} &H_{\mathrm{kin}} \to H'_{\mathrm{kin}}, \\ &\frac{\mathbf{p}^2}{2m_e} \to \frac{1}{2m_e}(\mathbf{p} + e\mathbf{A})^2 = \frac{1}{2m_e}\left(\mathbf{p} - e\frac{\mathbf{r} \times \mathbf{B}_0}{2}\right)^2. \end{aligned} \tag{8.10}$$

Evaluating this expression gives

$$H'_{\mathrm{kin}} = \frac{1}{2m_e}\left(\mathbf{p}^2 + e\mathbf{B}_0 \cdot (\mathbf{r} \times \mathbf{p}) + \frac{e^2}{4}(\mathbf{r} \times \mathbf{B}_0)^2\right), \tag{8.11}$$

where we have used that $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = -\mathbf{c} \cdot (\mathbf{b} \times \mathbf{a})$. Now we exploit that $\mathbf{B}_0$ has only a component in the z direction, that is, $\mathbf{B}_0 = (0, 0, B_0)$.

$$H'_{\mathrm{kin}} = H_{\mathrm{kin}} + H' = \frac{\mathbf{p}^2}{2m_e} + \frac{e}{2m_e}B_0(\mathbf{r} \times \mathbf{p})_z + \frac{e^2}{8m_e}B_0^2(x^2 + y^2). \tag{8.12}$$

The first term in this expression is the original kinetic energy H_{kin}. The second and third terms represent the perturbation caused by the magnetic field. The energy

change due to the perturbation is $E' = \langle\psi|H'|\psi\rangle$ and thus

$$E' = \frac{e}{2m_e}B_0\langle\psi|(\mathbf{r}\times\mathbf{p})_z|\psi\rangle + \frac{e^2}{8m_e}B_0^2\langle\psi|(x^2+y^2)|\psi\rangle. \tag{8.13}$$

The second term in this expression represents the diamagnetism. We can see this because the term is always positive and, therefore, a higher magnetic field is always accompanied by an energy increase. The operator $x^2 + y^2$ determines the expectation value of the electron's squared distance from the origin in the plane perpendicular to the field. In the case of an atom, this origin would be the nucleus.

The first term in (8.13) contains the angular momentum of the electron projected onto the direction of the field (z). This is the paramagnetic term that gives an energy-lowering when the electron's magnetic moment aligns with the field. This happens when the z-component of the angular momentum $(\mathbf{r}\times\mathbf{p})_z$ is negative, that is, when the projection of the angular momentum is pointing in the opposite direction from the field. This is the usual situation for an electron: The orbital angular momentum and the associated magnetic moment are antiparallel because of the electron's negative charge.

Finally, the electron also has a spin and an associated magnetic moment. This is a relativistic effect and therefore not present in the nonrelativistic Schrödinger equation. We could add it to (8.13) as an additional perturbation of the energy. It is given by

$$g_e m_s \frac{e\hbar}{2m_e}B_0 = g_e m_s \mu_B B_0, \tag{8.14}$$

where μ_B is the **Bohr magneton** with a value of $9.274\times10^{-24}\mathrm{J\,T^{-1}} = 5.788\times10^{-5}\mathrm{eV\,T^{-1}}$, m_s is the spin magnetic quantum number that can take the values $-1/2$ and $1/2$, and $g_e \approx 2$ is the gyromagnetic ratio for the electron.

8.3 Paramagnetism and Diamagnetism in Atoms

The paramagnetism in atoms is a bit more complicated than described above because the spin and orbital magnetic moments interact with each other and the contributions from different electrons can cancel each other. The diamagnetism in atoms, on the other hand, can be treated by summing up the contributions from all the electrons. We describe both phenomena here, but we only treat the paramagnetism for the simplest case. For more detailed explanations, the reader is referred to the literature about quantum mechanics or atomic physics.

The magnetic moment of an atom is caused by the orbital and spin angular momenta. In a hydrogen atom, the orbital angular momentum $\mathbf{L}$ of the single electron (measured in units of $\hbar$) is accompanied by a magnetic moment $\boldsymbol{\mu}$ with

$$\boldsymbol{\mu} = -\frac{e\hbar}{2m_e}\mathbf{L} = -\mu_B\mathbf{L}. \tag{8.15}$$

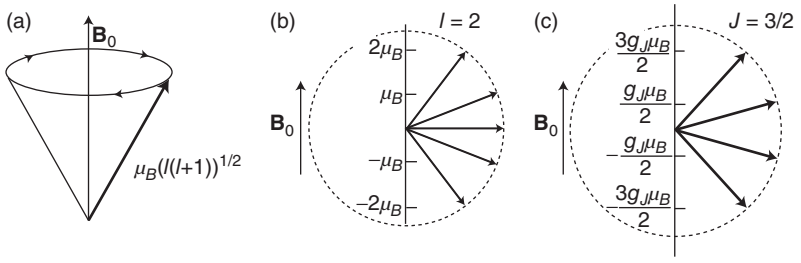


Figure 8.1 (a) Precession of an atomic magnetic moment in an external field. (b) Possible orientations for the magnetic moment in field direction for hydrogen with $l = 2$. (c) Possible orientations for the magnetic moment in field direction for Cr^{3+} ($J = 3/2$).

This magnetic moment precesses around the direction of an applied field as shown in Figure 8.1a. The component of the magnetic moment μ_l in the direction of the field is quantized and given by the magnetic quantum number m_l:

$$\mu_l = -\frac{em_l\hbar}{2m_e} = -m_l\mu_B. \tag{8.16}$$

If the orbital quantum number is l, then m_l takes the $2l + 1$ values $-l, \ldots, 0, \ldots, l$ This is illustrated for $l = 2$ in Figure 8.1b.

The situation for the spin $\mathbf{S}$ (also measured in units of $\hbar$) is very similar. It also leads to a magnetic moment

$$\boldsymbol{\mu} = -g_e\mu_B\mathbf{S}, \tag{8.17}$$

and possible magnetic moments in the field direction of

$$\mu_s = -g_e m_s\mu_B, \tag{8.18}$$

as already discussed in the previous section. Since $g_e \approx 2$, this means that $\mu_s \approx \pm\mu_B$. The hydrogen atom in the ground state has $n = 1$ and $l = 0$ and, therefore, only the spin magnetic moment matters.

What about the more complex atoms with many electrons and an interaction between spin and orbital magnetic moments? This problem is greatly simplified by the observation that for a filled shell, that is, for a set of n, l, which is completely occupied, the total orbital magnetic moment is zero because the components in field direction and opposite to the field direction are equally strong. The same is true for the total spin magnetic moment because there are equally many electrons with spin $+1/2$ and spin $-1/2$. So, we only have to worry about nonfilled shells. For these, we proceed in two steps. First, we have to find the total angular momentum that is described by the quantum number J. Then, we have to calculate the magnetic moment associated with J. Similar to the case of the orbital magnetic moment, there are $2J + 1$ possibilities for the orientation of the angular momentum with respect to a magnetic field with magnetic moments in the field direction of

$$\mu_J = -gm_J\mu_B, \tag{8.19}$$

where g is the so-called **Landé splitting factor** and m_J is the magnetic quantum number belonging to J. The total angular momentum of the electrons can be calculated from a vector sum of the spin and orbital momenta. For light atoms (with weak spin–orbit coupling), these are independent and the so-called L–S coupling scheme can be applied. With this, one obtains the quantum numbers for the total orbital and spin momenta by

$$L = \sum m_l \quad S = \sum m_s. \tag{8.20}$$

Again, we can see that L and S are zero for filled shells because all angular momenta compensate each other. For nonfilled shells the procedure for calculating L and S is given by **Hund's rules**:

1) The spins of the electrons are arranged such that the maximum value of S consistent with the Pauli principle is achieved.
2) With the given S, the quantum numbers m_l are chosen such that the maximum value of L is achieved.
3) J in the ground state is now calculated as $J = L - S$ when the shell is less than half full, as $J = L + S$ if the shell is more than half full, and as $L = 0, J = S$ if the shell is half full.

The first rule arises from an effect similar to the exchange interaction, which we have encountered for the hydrogen molecule in Chapter 2: If we require the spins to be parallel, the electrons have to be distributed into orbitals with different m_l. Since the spatial wave functions of these orbitals are mutually orthogonal, the electrons keep out of each other's way and this reduces the Coulomb repulsion. The origin of the second rule is a bit less obvious, but it is qualitatively related to the fact that the Coulomb repulsion is also lowered when the electrons "revolve around the nucleus in the same direction." The third rule minimizes the energy in the presence of spin–orbit coupling. The rules can be justified by both experiment and theory and shall not be discussed in further detail here.

As an example, consider the ion Cr^{3+} that has the electronic configuration $[Ar]3d^3$. The first of Hund's rules requires $S = 3/2$. The possible m_l values for the 3d shell are $-2, -1, 0, 1, 2$. Hund's second rule requires the largest possible value of L, that is, we have to choose $m_l = 0, 1, 2$ and therefore $L = 3$. Finally, Hund's third rule states that for less than half-filled shells, $J = L - S = 3 - 3/2 = 3/2$. The magnetic quantum number m_J therefore takes the values $-3/2, -1/2, 1/2, 3/2$.[3)]

For the calculation of the possible magnetic moments, we only lack the Landé splitting factor that is given by

$$g_J = \frac{3J(J+1) + S(S+1) - L(L+1)}{2J(J+1)}. \tag{8.21}$$

The resulting possible orientations of the magnetic moment for Cr^{3+} are shown in Figure 8.1c.

3) Note that even for partially filled shells, one can obtain $J = 0$ if $S = L$. Equation (8.19) would therefore suggest that this does not lead to any magnetic moment. This is not entirely true but we do not treat this case here.

We thus see that atoms or ions show paramagnetic behavior only when they have open shells. This is different for diamagnetism. It is always there because it is caused by all the electrons in the atoms and their reaction to the magnetic field. We will see that diamagnetic effects are usually very weak and that paramagnetism, if present, dominates.

We have already calculated the energy correction due to diamagnetism in (8.13). We can now use this expression to estimate the size of the corresponding magnetic moment in an atom. As in the derivation of (8.6), we have to take into account that the microscopic magnetic moment is induced by the field and is therefore field-dependent. We thus obtain

$$\mu = -\frac{\partial E'}{\partial B_0} = -\frac{e^2}{4m_e} B_0 \langle \psi | (x^2 + y^2) | \psi \rangle. \tag{8.22}$$

This can be calculated if the atomic wave functions are known (see Problem 8.1). In order to merely estimate the magnitude of the magnetic moment, we introduce some approximations: For a spherically symmetric electron distribution, the mean square distance of an electron from the nucleus is $r^2 = x^2 + y^2 + z^2$, hence $x^2 + y^2 = (2/3)r^2$, and we take r to be the atomic radius r_a. In addition to this, an atom contains not only one electron but Z electrons that we all take to have probability distributions with a sharp maximum on the radius r_a. With this, we arrive at an expression for the diamagnetic moment of an atom, which is

$$\mu = -\frac{Ze^2}{6m_e} r_a^2 B_0. \tag{8.23}$$

The approximation of putting all the electrons into an "orbital" with the atomic radius is rather crude and, together with the fact that the expression contains r_a^2, it will certainly lead to an overestimation of μ. We shall see, however, that the resulting estimate is very small despite of this.

8.4 Weak Magnetism in Solids

As we shall see below, the magnetic effects in solids tend to be quite weak except for the cases of magnetic ordering and superconductivity (see Chapter 10). Here, we discuss the sources of this weak magnetism and find them directly related to the results for atoms in the previous section.

What do we expect for the magnetism in solids related to that in atoms? The diamagnetism discussed above arises from all electrons in the atom, the valence electrons, and the core electrons. Therefore, it is not going to change much as we form the solid. In fact, we can view the solid simply as a dense cloud of atoms and calculate the diamagnetic susceptibility for this cloud. The only correction to this picture is the diamagnetism of the itinerant electrons in metals that we will take into account separately.

The situation is more difficult in the case of paramagnetism. Even though many atoms with open outer shells should have a nonzero J and therefore a permanent

magnetic moment, not so many solids actually show this behavior. It appears as if the magnetic moment disappears upon the formation of the solid. The reason for this is particularly easy to understand for ionic solids. Even though the contributing atoms generally have an atomic magnetic moment because of their open shells, the ionic solid does not, because it basically consists only of ions with closed shells. A similar situation is found for covalent bonds. Consider, for example, our discussion of the H_2 molecule in Chapter 2. Even though the electrons in the individual hydrogen atoms have a net spin of 1/2, the molecular ground state has zero spin and it does not have any magnetic moment either.

For having a "good" paramagnetic solid, we need atoms with open shells, that do not participate in the bonding and therefore do not change their properties much upon the formation of a solid. Possible candidates could be the relatively localized d states in the 3d and 4d transition metals, but in these the d electrons still participate in the bonding to a large extent. The best examples for quasiatomic paramagnetism in solids are therefore found in compounds of the 4f rare earth elements because the 4f electrons are very localized indeed.

8.4.1 Diamagnetic Contributions

8.4.1.1 Contribution from the Atoms

The atomic contribution to the diamagnetic susceptibility of a solid can be estimated directly from (8.23):

$$\chi_{\mathrm{m}} = \mu_0 \frac{M}{B_0} = -\frac{\mu_0 Z N e^2}{6 V m_e} r_a^2 \tag{8.24}$$

and this is always very small, 10^{-5} or so, much smaller than 1, that is, the magnetization in the sample is much weaker than the external magnetic field. Since it is purely an atomic effect, it is also independent of the temperature.

8.4.1.2 Contribution from the Free Electrons

The (nearly) free electrons in metals also show a diamagnetic contribution to the susceptibility in a quantum mechanical picture. This contribution is given here without further derivation. It is

$$\chi_{\mathrm{m}} = -\frac{1}{3V} \mu_B^2 \mu_0 g(E_F) \left(\frac{m_e}{m^*} \right)^2 . \tag{8.25}$$

The main ingredients of this contribution are quite intuitive. First of all, there is the ubiquitous density of states at the Fermi energy, which stems from the fact that only the electrons close to the Fermi energy can respond to a magnetic field (or perform any other low-energy excitation, for that matter). In addition, the susceptibility depends on the ratio of electron mass and the effective mass. The smaller the effective mass, the stronger the diamagnetic contribution. In total, the diamagnetic contribution of the free electrons is very small, of the same order as the contribution from the atoms.

8.4.2 Paramagnetic Contributions

As for diamagnetism, we treat two contributions to paramagnetism. One is the alignment of existing atomic magnetic moments and the other stems from the free electrons in metals. We will see that the first, when present at all, is usually much stronger than the second and than the diamagnetic response, and it is therefore the dominant contribution to the magnetic properties. The second is of the same order as the diamagnetic contribution of free electrons, but it is easier to understand and the discussion is useful for our later description of spontaneous magnetic ordering.

8.4.2.1 Curie Paramagnetism

Consider a solid with a unit cell that contains an atom with a localized magnetic moment. Such a moment could stem from an ion with a partially filled 4f shell, for example. Treating such a system of independent, distinguishable, and localized magnetic moments is a standard example in statistical physics and details can be found in the literature. Here we only sketch how the mean magnetization and the susceptibility can be found.

We know that the possible energy levels of the magnetic moment in an external field are given by $g_J \mu_B m_J B_0$ (see Figure 8.1c). The lowest energy level is $-g_J \mu_B J B_0$ and this is reached for the situation where the magnetic moment's z-component $g_J \mu_B J$ is aligned parallel to the field B_0. We can thus calculate the mean moment in the direction of the external field by weighting all possible moments by a Boltzmann factor for their individual probabilities

$$\overline{\mu} = \frac{1}{Z} \sum_{m_J=-J}^{J} g_J \mu_B m_J e^{g_J \mu_B m_J B_0 / k_B T}, \tag{8.26}$$

while normalizing the sum with the total sum of probabilities, the so-called partition function

$$Z = \sum_{m_J=-J}^{J} e^{-g_J \mu_B m_J B_0 / k_B T}. \tag{8.27}$$

The details of this are quite involved except for a spin 1/2 system with two states, a case we shall return to later. However, once $\overline{\mu}$ has been determined, one can calculate the total magnetization of the sample according to (8.4). The result is shown in Figure 8.2 as a function of $g_J \mu_B B_0 / k_B T$.

We can distinguish two limiting cases. For $g_J \mu_B B_0 \gg k_B T$, the magnetic field is strong enough and the temperature low enough to achieve the highest possible alignment of the magnetic moments in the field direction. This corresponds to a strong and saturated magnetization of the sample, but it is hard to realize experimentally, even for strong magnetic fields and the lowest reachable temperatures. The much more important limiting case is that $g_J \mu_B B_0 \ll k_B T$. Then, the magnetization is found to be proportional to the magnetic field such that we can

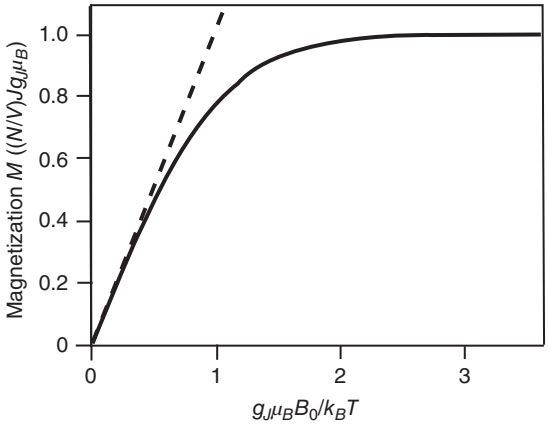


Figure 8.2 Paramagnetic susceptibility of a solid with localized magnetic moments. The limit of Curie's law is indicated as a dashed line.

define a susceptibility χ_m according to (8.5). χ_m is inversely proportional to the temperature, a result that is known as **Curie's law** (after P. Curie)

$$\chi_\mathrm{m} = \frac{C}{T}, \tag{8.28}$$

with the **Curie constant** C given by

$$C = \frac{\mu_0 N g_J^2 \mu_B^2 J(J+1)}{3Vk_B}. \tag{8.29}$$

The limit of Curie's law is indicated by the dashed line in Figure 8.2.

The Curie constant can be calculated when J, g_J, and the density of magnetic atoms are known. The calculated values compare very well to the experimental data for solids containing rare earth ions, as expected. The comparison is much less favorable for the 3d transition metal compounds (see Problem 8.2). We have already addressed this issue above: The 3d electrons participate in the bonding and the states have a different character from the atomic orbitals assumed in the derivation of Curie's law. In fact, the potential in which these electrons move is far from the spherical potential that is assumed for atoms and much more dictated by the crystal symmetry. For the elements of the iron group (Fe, Co, Ni), one observes values of C suggesting that the 3d electrons have $J = S$, that is, only spin and no orbital angular momentum at all. This effect is known as the **quenching of the orbital angular momentum**. If present, it allows us to think about the magnetization as the alignment of spin moments only.

The Curie paramagnetism is much stronger than the diamagnetism discussed previously, but it is still weak. Typical values of χ_m at room temperature are on the order of 10^{-3}–10^{-2}. Note that χ_m is also temperature-dependent, in contrast to the diamagnetic susceptibility discussed above.

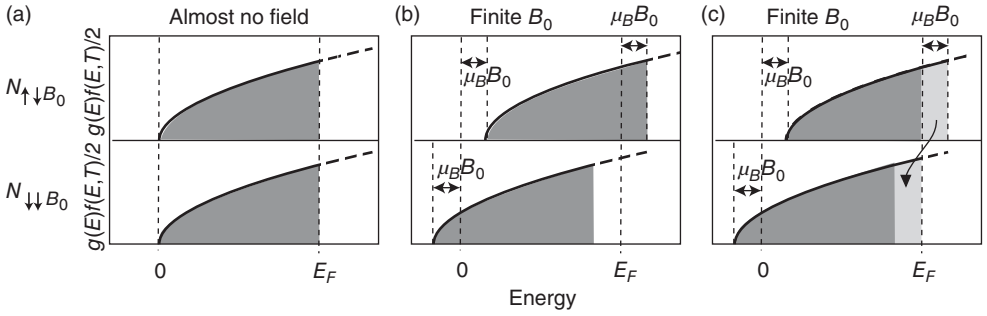


Figure 8.3 (a) Density of occupied states for free electrons at $T = 0$ K, split up into electrons having their magnetic moment antiparallel ($N_{\uparrow\downarrow B_0}$) or parallel ($N_{\downarrow\downarrow B_0}$) to an external field, but the field is almost zero. (b) When B_0 is no longer small, the energy of the electrons is increased or reduced by $\mu_B B_0$, depending on the orientation of their magnetic moments. (c) The electrons with a magnetic moment antiparallel to the field can reach a lower energy state by flipping their spin. In this way, a stable situation with a constant Fermi energy is reached. Note that the energy shift induced by B_0 is not drawn to scale.

8.4.2.2 **Pauli Paramagnetism**

Free electrons also exhibit paramagnetic behavior. If every free electron has a spin of 1/2 and a magnetic moment of μ_B, one could expect that they contribute to the saturation magnetization of the solid with μ_B times the density of the electrons. This saturation is achieved when all the magnetic moments align parallel to the field (or the spins align antiparallel to the field).[4] But this is not the case at all and the paramagnetic susceptibility of free electrons is actually very small.

The paramagnetic susceptibility of free electrons can be understood and even calculated using the picture given in Figure 8.3. In Figure 8.3a, the density of occupied states for free electrons (6.13) is divided into two parts: one with the orientation of the magnetic moments antiparallel to an external field B_0 and one with the orientation parallel. This external field is assumed to be almost zero. Figure 8.3b shows what happens when B_0 is increased to a finite value. The electrons raise or lower their energy by $\mu_B B_0$, depending on the orientation of their magnetic moments with respect to the field. Since μ_B is so small, this energy change is tiny for any achievable field, only 10^{-5} eV or so, much smaller than the distance from the bottom of the band to the Fermi energy. Once this shift happens, the electrons that have moved above the Fermi energy can lower their energy by flipping their spin and becoming electrons with a magnetic moment parallel to the field, as shown in Figure 8.3c. This gives rise to more electrons with a magnetic moment parallel than antiparallel to the field, that is, to a paramagnetic response.

4) Here and in the following, we speak loosely of spins aligning parallel or antiparallel to a field or to each other but of course they do not. Only the z component of the spin can align with the field. The actual spin precesses around the field direction.

In order to calculate χ_m, we must figure out how many electrons flip their spin in order to have a magnetic moment parallel to the field. These electrons are represented by the light gray areas in Figure 8.3c. The size of each area is $g(E_F)\mu_B B_0/2$. Therefore, we have a difference between electrons with their magnetic moment parallel and antiparallel to the field, which is

$$N_{\downarrow\downarrow B_0} - N_{\downarrow\uparrow B_0} = g(E_F)\mu_B B_0, \tag{8.30}$$

where the arrows denote whether the moments and the field are parallel or antiparallel. The net magnetization is

$$M = \frac{1}{V}(N_{\downarrow\downarrow B_0} - N_{\downarrow\uparrow B_0})\mu_B = \frac{1}{V}g(E_F)\mu_B^2 B_0, \tag{8.31}$$

and the susceptibility becomes

$$\chi_m = \frac{1}{V}\mu_0\mu_B^2 g(E_F), \tag{8.32}$$

which is small, in the same order as the diamagnetic susceptibilities. The small susceptibility is hard to understand from a semiclassical point of view. As in the case of the electronic heat capacity, it has to do with the Fermi–Dirac statistics. The field required to align all the spins would not only have to be such that $\mu_B B_0 \gg k_B T$ but such that $\mu_B B_0 > E_F$, which is huge. In fact, the Pauli paramagnetism is quite different from the Curie paramagnetism. In the latter, a strong magnetization can actually be achieved because the magnetic energy only needs to be much higher than the thermal energy. For Pauli paramagnetism, it would need to exceed the Fermi energy, something that is not possible.

8.5 Magnetic Ordering

So far, we have studied the diamagnetic and the paramagnetic behavior of solids. Neither leads to magnetic effects of appreciable size. In this section, we look at a much more spectacular phenomenon: long-range magnetic ordering without any applied field. The detailed understanding of this is very difficult, and we will not attempt it. We just concentrate on some main ideas behind the mechanism of magnetic ordering.

Different types of magnetic ordering are shown in Figure 8.4. The ordering you are probably most familiar with is the ferromagnetic type that is observed for the iron group (Fe, Co, Ni). It is also found for the rare earth elements gadolinium and dysprosium and for several alloys. Ferromagnetic ordering originates from the parallel alignment of the magnetic moments in the crystal. It gives rise to a macroscopically observable magnetization. A rather different case is antiferromagnetic ordering, which also entails a long-range ordering of magnetic moments, but the orientation of the moments on neighboring sites is opposite, such that no net magnetization is observed. Many transition metal oxide insulators show antiferromagnetic ordering. A mixture between the two cases is ferrimagnetic ordering

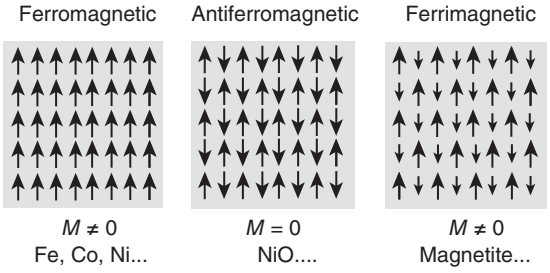


Figure 8.4 Types of magnetic ordering. The arrows denote the direction and size of the localized magnetic moments.

where there is antiferromagnetic ordering between moments of different sizes in one unit cell, but ferromagnetic ordering between the unit cells, such that a net magnetization remains. An example for a material showing ferrimagnetic ordering is magnetite (Fe_3O_4).

How can we know that there is a phenomenon such as antiferromagnetism when it does not produce an observable macroscopic field? As discussed in Chapter 1, one can determine the microscopic magnetic ordering by neutron diffraction. Whereas X-rays only "see" the structure of the material, neutrons carry a magnetic moment and are therefore sensitive to the magnetic structure. The difference between geometric and magnetic structure is pronounced in antiferromagnetic crystals because the magnetic unit cell is bigger than the geometric unit cell (see Problem 8.8).

8.5.1 Magnetic Ordering and the Exchange Interaction

Now we try to understand, at least qualitatively, where the magnetic ordering comes from. It is obviously related to some interaction between the magnetic moments. This is holding the moments aligned, despite the disordering effect of entropy (temperature). Knowing that ferromagnetic ordering exists even far above room temperature, we see that the energies needed to destroy it are at least of the order of $k_B T$ at room temperature, that is, 25 meV.

Before we describe what causes the ordering, it is a good idea to mention a mechanism that is *not* causing it. This is the direct dipole–dipole interaction of the localized magnetic moments. It is a common misconception that the "little magnets" in the solid align themselves like an array of compass needles due to their magnetic interaction, but in fact they do not. We can easily see this by estimating the strength of the magnetic dipole–dipole interaction. It is very weak and the energy difference corresponding to the parallel and antiparallel alignment of two magnetic dipoles at typical atomic distances corresponds to temperatures on the order of 1 K (see Problem 8.4).

If the magnetic interaction is not the cause of the alignment, what is? The responsible interaction is the **exchange interaction**, a funny type of energy

that stems from a combination of Coulomb interaction and the Pauli principle, that is, the need to have antisymmetric wave functions for fermions. We have encountered the exchange interaction already in our discussion of the hydrogen molecule in Chapter 2, where we have seen that the difference between the singlet and the triplet states is approximately twice the value of the exchange energy X[5] or

$$E_{\uparrow\uparrow} - E_{\uparrow\downarrow} = -2X. \tag{8.33}$$

For the hydrogen molecule, the exchange energy X is always negative (see Figure 2.2), which means that the triplet state has a higher energy than the singlet state and the ground-state ordering is therefore "antiferromagnetic." An inspection of Figure 2.2 shows that the exchange energy X (i.e., the separation between the singlet and triplet state) is by no means small. Even for large distances, it amounts to a substantial fraction of an electron volt.

This already gives the basic ingredient for the coupling in solids as well. The magnetic dipole moments order because of the exchange interaction that favors either a parallel (positive X) or an antiparallel (negative X) alignment. The exchange energy is typically smaller than in the hydrogen molecule, but it is still on the order of 100 meV for the ferromagnetic elements. It is, however, very difficult to make qualitative predictions about the size or even the sign of X, as the following considerations illustrate.

In contrast to what we see in the hydrogen molecule, a positive exchange energy X could be expected for a multielectron system on very general grounds. The reason is the Pauli principle: For two electrons, the triplet wave function (2.5) vanishes when the electrons have the same spatial coordinates, meaning that they will never be at the same place. This reduces their Coulomb repulsion, leading to an energy lowering compared to the singlet state and thus to a positive exchange energy. The most prominent example for this is the He atom for which the triplet states are found to have a lower energy than the corresponding singlet states. The same idea is also the basis of Hund's first rule that the electron states have to be occupied such that the highest possible value of S is realized. We have discussed the example of Cr^{3+}, which has a partially filled d shell with three electrons in it. The electrons are placed into different subshells to achieve the highest possible S, which also means that the electrons are kept apart from each other and the total potential energy is lowered. The same principle applies to free electrons in a metal. Electrons with the same spin direction will not be at the same place. If there is a majority of spin-up electrons, each of these electrons feels the presence of the other electrons less and it is attracted more strongly to the ions in the crystal. This leads to an energy gain.

5) In the literature on magnetism, the exchange energy is very often called J instead of X. We stick to X here, as in the Heitler–London model, and in order to avoid confusion with the total angular momentum. Actually, identifying X in the Heitler–London model with the energy difference between $E_{\uparrow\uparrow}$ and $E_{\uparrow\downarrow}$ is only approximately correct (see online note on www.philiphofmann.net) but we ignore this here.

However, the situation is not that simple because other energy considerations apply as well, even if we neglect the disordering effect of finite temperature. For a free electron gas, a complete spin polarization would lower the Coulomb repulsion between the electrons, but it would increase the kinetic energy by a very large amount (on the order of the Fermi energy), as we have already seen in our discussion of Pauli paramagnetism. In fact, spontaneous magnetization is never observed for free-electron-like metals.

We now want to explain how the exchange interaction can lead to (ferro) magnetically ordered states, even without an external field. There are two complimentary ways to describe this. We can either view a system of localized magnetic moments that interact with each other via the exchange interaction or we can inspect how the electronic band structure for completely delocalized Bloch electrons changes when we make a certain direction of the magnetic moment energetically more favorable than the other direction. The first approach works well for describing the magnetism of the rare earth metals because their 4f electrons are indeed very localized. The second description is adequate to describe the magnetism in the 3d transition metals.

8.5.2 Magnetic Ordering for Localized Spins

For a system of localized magnetic moments that interact via the exchange energy, the quantum mechanical description of the magnetization was formulated by W. Heisenberg. For simplicity, we assume that the orbital magnetic moment of the states under consideration is quenched and that we only deal with spins. Heisenberg's formulation is based on the Heitler–London model for the H_2 molecule presented in Chapter 2. There we have seen that the energy of the molecule (2.9) has three contributions and the last of these depends only on the relative spin directions of the electrons. This is very remarkable since the spin did not explicitly appear in the calculation. For the magnetism, this last contribution to the energy is the only relevant part and, therefore, Heisenberg suggested that magnetism could be studied by a Hamiltonian that only includes this spin contribution. For two spins, the **Heisenberg Hamiltonian** is

$$H = -2X\mathbf{S}_1 \cdot \mathbf{S}_2. \tag{8.34}$$

The $\mathbf{S}_i$ are the spin operators. For our present discussion, it is acceptable to view them simply as the spin directions on a certain site. The factor of 2 is introduced to obtain an energy difference of $2X$ between the singlet and triplet states, as in (8.33) for the hydrogen molecule.

We now try to capture the essence of spontaneous ferromagnetic ordering based on the Heisenberg Hamiltonian. If we extend this to the solid, the spin $\mathbf{S}_i$ on every lattice site would interact with the spin on every other lattice site. If we also include the possibility of an external magnetic field, the resulting Hamiltonian is

$$H = -\sum_{i \neq j} \sum X_{ij} \mathbf{S}_i \cdot \mathbf{S}_j + g_e \mu_B \mathbf{B}_0 \cdot \sum_i \mathbf{S}_i, \tag{8.35}$$

where i, j run over all the atoms in the solid and X_{ij} is the exchange interaction between spins on the lattice sites i and j. The second term represents the effect of an external magnetic field on all the spins. Equation (8.35) can be simplified because the exchange interaction very rapidly decreases for longer distances. It is therefore a good approximation to assume that a spin on site i only interacts with the spins on the nearest neighbor atoms. We also take X_{ij} to be the same for all of these neighbors, such that we can simply call it X. Then, we obtain

$$H = -X \sum_i \sum_{nn} \mathbf{S}_i \cdot \mathbf{S}_{nn} + g_e \mu_B \mathbf{B}_0 \cdot \sum_i \mathbf{S}_i, \tag{8.36}$$

where the second sum runs over the nearest neighbors nn of each atom.

It is quite difficult to formally find the solutions to the Hamiltonian (8.36), but it is easy to guess how the ground state looks like: For a positive exchange energy X, the state with the lowest energy must be the one in which all the spins are aligned parallel to each other and opposite to the external field (such that their magnetic moment is parallel to this field).

We would now like to show that (8.36) permits a spontaneous magnetization even without an applied field. The biggest hurdle for doing this is that the first term in (8.36) explicitly contains the local interaction between the spins on the nearest neighbor sites. Progress can be made by a so-called **mean field approximation** in which all the spins on the neighboring sites are replaced by the average spin direction in the solid $\langle \mathbf{S} \rangle$, so that

$$H = \sum_i \mathbf{S}_i \cdot \left(- \sum_{nn} X \langle \mathbf{S} \rangle + g_e \mu_B \mathbf{B}_0 \right) = \sum_i \mathbf{S}_i \cdot \left(-n_{nn} X \langle \mathbf{S} \rangle + g_e \mu_B \mathbf{B}_0 \right), \tag{8.37}$$

where n_{nn} is the number of nearest neighbors. Now we can exploit that $\langle \mathbf{S} \rangle$ is directly related to the macroscopic magnetization we are interested in since according to (8.4) and (8.17)

$$\mathbf{M} = -g_e \mu_B \langle \mathbf{S} \rangle \frac{N}{V}, \tag{8.38}$$

and it thus follows that

$$H = \sum_i \mathbf{S}_i \cdot \left(\frac{n_{nn} X V}{g_e \mu_B N} \mathbf{M} + g_e \mu_B \mathbf{B}_0 \right) = g_e \mu_B \sum_i \mathbf{S}_i \cdot \left(\mathbf{B}_W + \mathbf{B}_0 \right), \tag{8.39}$$

with

$$\mathbf{B}_W = \mathbf{M} \frac{n_{nn} X V}{g_e^2 \mu_B^2 N}. \tag{8.40}$$

Formally (8.39) describes a system of localized and independent spins that are exposed to a sum of two magnetic fields. The first field $\mathbf{B}_W$ is caused by the magnetization $\mathbf{M}$, which acts on the individual spins $\mathbf{S}_i$ and the second is the external field. $\mathbf{B}_W$ is also called the **Weiss field**, after a phenomenological treatment by P. Weiss in 1907, that is, before the advent of quantum mechanics. We emphasize that $\mathbf{B}_W$ is not a "real" magnetic field created by the magnetization of the sample but merely a clever way to cast the consequence of the exchange interaction into

something that can be treated like a magnetic field. Indeed, the crucial ingredient of (8.40) is the exchange energy X, which is needed to obtain a finite $\mathbf{B}_W$ when a magnetization is present.

The situation described by (8.39) is identical to what we have encountered in Curie paramagnetism in the sense that localized magnetic moments are exposed to a magnetic field $\mathbf{B}_W + \mathbf{B}_0$. We can exploit this to calculate the temperature-dependent magnetization in the same way as in the Curie model. The important difference to the Curie model is that a part of the magnetic field, $\mathbf{B}_W$, is not an external magnetic field but the Weiss field that arises because of the magnetization. If $\mathbf{B}_W$ is strong enough, this gives us the desired possibility to sustain a magnetization without any external field $\mathbf{B}_0$.

We assume that $B_0 = 0$ and calculate the temperature-dependent magnetization purely in the presence of the Weiss field, using (8.26) for a two-state system. We introduce the abbreviation $x = g_e|m_s|\mu_B B_W/k_B T$. If we only deal with spins, we can also approximate $g_e|m_s| \approx 1$ and $x \approx \mu_B B_W/k_B T$ and obtain

$$M(T) = \frac{\mu_B N}{V} \frac{e^x - e^{-x}}{e^{-x} + e^x} = M(0)\tanh(x), \tag{8.41}$$

where $M(0) = \mu_B N/V$ is the highest possible magnetization that can be reached at 0 K. Equation (8.41) now gives the magnetization as a function of temperature but with the difficulty that x on the right-hand side also depends on this magnetization, as it appears in the Weiss field (8.40). This is most clearly seen when rewriting (8.41) as

$$\frac{M(T)}{M(0)} = \tanh\left(\frac{M(T)}{M(0)}\frac{\Theta_C}{T}\right), \tag{8.42}$$

where we have introduced the so-called Curie temperature Θ_C as

$$\Theta_C = \frac{n_{nn}X}{g_e^2 k_B} \tag{8.43}$$

For a given temperature T, we must thus seek an $M(T)$ such that (8.42) is fulfilled. Clearly, this is always possible for a nonmagnetized sample with $M(T) = 0$. However, it turns out that nontrivial solutions also exist, as long as the temperature is lower than Θ_C. These solutions can be found numerically and are shown in Figure 8.5. The temperature-dependent magnetization in this simple model appears to agree quite well with the measured result for the ferromagnetic 3d transition metals Fe, Co, and Ni, even though the valence electrons giving rise to magnetism in these elements are not particularly localized. Data for these elements are shown along with the model.

We can estimate the magnitude of the Weiss field B_W from the measured Curie temperature (see Problem 8.5). One finds that B_W is huge, on the order of hundreds to thousands of Tesla, much stronger than any field that can be generated in the laboratory. This explains why the Weiss field can sustain a spontaneous magnetization in the sample but it should be emphasized again that the Weiss field is not an ordinary magnetic field. It is caused by the combination of magnetization and exchange, as clearly seen in (8.40).

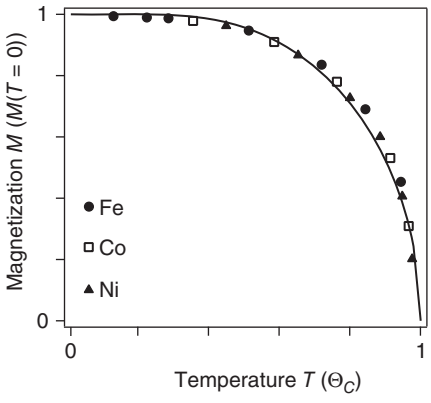


Figure 8.5 Temperature-dependent magnetization of Fe, Co, and Ni below the Curie temperature Θ_C. The line is the prediction according to (8.42) for a spin 1/2 system. The data points are taken from Tyler (1931).

Above the Curie temperature, the spontaneous magnetization is lost. This implies a lack of long-range order but the magnetic moments are of course still present. We can therefore expect to find paramagnetic behavior. To see this, we can proceed in exactly the same way as for the Curie paramagnetism and calculate (8.41) in the high-temperature limit where we can replace the tanh function by its argument. Making also use of the (8.5), we obtain the so-called **Curie–Weiss law** for the susceptibility:

$$\chi_{\mathrm{m}} = \frac{C}{T - \Theta_c}. \tag{8.44}$$

Note that this is very similar to Curie's law (8.28), only the origin is shifted by the Curie temperature. The derivation of the Curie–Weiss law and the Curie Constant C is the subject of Problem 8.6. The law suggests that the susceptibility diverges as we approach the Curie temperature and this also appears to make sense: As we go into the regime of ferromagnetism, a very small external field can cause a very strong response. We have to remember, however, that the Curie–Weiss law is a high-temperature limit, and it does therefore not necessarily describe the behavior near the Curie temperature accurately.

Overall, the description of ferromagnetism by the Heisenberg model is thus fairly successful. It can explain the existence of spontaneous magnetization, the temperature-dependent strength of the magnetization below the Curie temperature, and the paramagnetism above the Curie temperature. We have only treated it for a spin 1/2 system with two states but it can be extended to any desired magnetic moment. However, it is important to remember that the model assumes magnetic moments on the lattice sites. Therefore, we would expect it to work best for situations that come close to this idealization. The rare earth metals Gd and Dy with the 4f electrons are very close to this ideal and therefore well-described by the Heisenberg model. For the transition metals Fe, Co, and Ni, in which the magnetism is caused by the more delocalized 3d electrons, the description of ferromagnetism

by the Heisenberg model gives rise to some problems, despite the apparently good agreement in Figure 8.5, and we come back to this below.

We were only able to calculate the magnetic properties in the Heisenberg model because of the mean-field approximation, that is, because we replaced the spins $\mathbf{S}_{nn}$ on the nearest neighbors of a certain atom i by the averaged value over the whole sample $\langle \mathbf{S} \rangle$ in (8.37). This is not a very good approximation, especially not near Θ_C. Imagine what happens when we cool the sample, starting from just above Θ_C. As Θ_C is reached, certain spins will find themselves surrounded by spins of the same orientation and this local magnetization will quickly spread out. What is therefore important is the local spin environment of an atom, not the global average spin. Indeed, it is found that the Heisenberg model with the mean-field approximation does not give a very accurate description of the temperature-dependent magnetization just below Θ_C.

The Heisenberg model of localized spins can also be used to describe antiferromagnetic materials and there it often works very well, even for materials containing 3d electrons. This is because typical antiferromagnetic materials are oxides for which the oxygen atoms act as "spacers" between the magnetic atoms and the 3d electrons thus remain fairly localized (see Problem 8.8 for the structure of antiferromagnetic NiO). In fact, the 3d orbitals are so localized that there is no direct exchange interaction between these electrons on different atoms, and the exchange leading to magnetism has to be "mediated" by the oxygen atoms in between, a phenomenon called **superexchange**. The predictions for antiferromagnetism are quite similar to those for ferromagnetism. Antiferromagnetic ordering is also only possible below a certain temperature and this temperature is called the **Néel temperature**.

8.5.3 Magnetic Ordering in a Band Picture

We have assumed that the maximum possible magnetization of a ferromagnetic sample can be calculated from the density and the size of the magnetic moments. If we only had spin magnetic moments, for instance, the highest possible magnetization would be $M(0) \approx \mu_B N/V$ and this would be reached at $T = 0$ K. If the angular momentum on the ions was J instead, we would have $M(0) = \mu_B g J N/V$. The measured highest magnetizations for the 4f metals are not too far off this expectation but for the 3d transition metals, the agreement is quite poor.[6] This could have several reasons, such as the (partial) quenching of the orbital magnetic moment. It turns out, however, that the description of localized moments is inadequate because the 3d states are delocalized and thus have band character. Bands can be only partially filled and this can explain why not all the electrons participate in the magnetism but just a fraction of them.

6) Note that such problems are not apparent in Figure 8.5 because there the magnetization is normalized to the *experimental* value of $M(0)$. When plotting the data with respect to the calculated value of $M(0)$, the agreement would be much less impressive (see Problem 8.3).

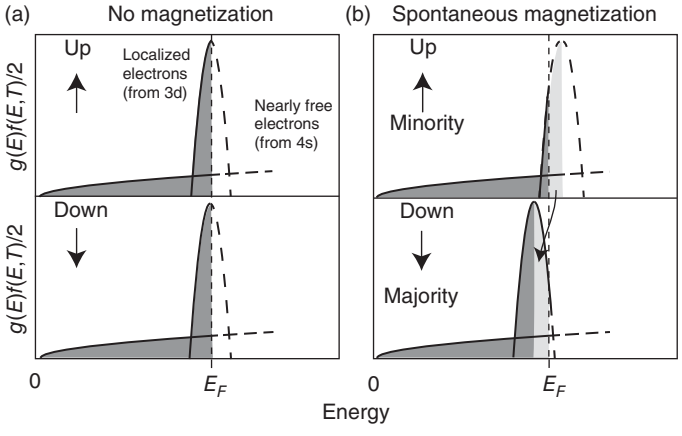


Figure 8.6 (a) Occupied density of states in a 3d transition metal, separated into two spin directions but without a net magnetization. (b) A spontaneous magnetization of the d electrons occurs when many electrons change their spin direction (here from "up" to "down"). This corresponds having more occupied states in the "down" band and this is achieved by moving the bright area in the spin "up" density of states to the spin "down" density of states.

The description of magnetism for electronic states with band character goes back to E. C. Stoner and E. P. Wohlfarth. Figure 8.6a shows a sketch of the density of states in a transition metal. The s (and sometimes also p) electrons form the familiar free electron density of states with $g(E) \propto \sqrt{E}$. The density of states looks quite different for the d electrons. The d band only exists in a small energy range, its density of states is relatively high, and it is almost centered on the Fermi energy. This can be understood by considering the character of the d electrons: The band is narrow in energy because the localized nature of the d states gives rise to a smaller overlap of the wave functions and a smaller splitting in energy. We have encountered this in the tight-binding model in Chapter 6: Equation (6.60) shows that the parameter γ that determines the band width depends on how much overlap there is between wave functions on neighboring sites. For localized d electrons this overlap, and hence the band width, is small. The density of states is high because the narrow band has to accommodate 10 d states per atom in the crystal. Finally, it is almost centered on the Fermi energy because it is only partially filled. For Fe, it is occupied by 6 out of 10 d electrons. In the sketch in Figure 8.6a, the d band is chosen to be exactly half-filled.

Now imagine a situation with a spontaneous magnetization of the d electrons (the exchange energy for the s electrons is so small that it can be ignored). Suppose we have more electrons with spin "down" than with spin "up". For obvious reasons, the "down" spin is then called the majority spin and the "up" spin the minority spin. We know that the magnetic state is stabilized by the fact that the "down" spin electrons have gained an energy on the order of the exchange energy and the "up" spin electrons have lost this energy. The two corresponding densities of states are thus shifted against each other in energy, as shown in Figure 8.6b. In order to

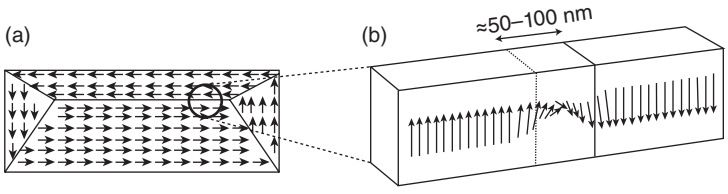


Figure 8.7 (a) Domains of different magnetization in a ferromagnetic solid. (b) Detailed picture of the magnetization rotation in a Bloch wall between two domains.

maintain a constant Fermi energy, the electrons in the bright shaded area of the occupied density of states change their spin from "up" to "down" and the desired magnetization with more "down" spin electrons is reached.

Figure 8.6b clearly illustrates that many electrons near the sharp maximum of the d density of states have moved to lower energies and we can see the magnetization directly as an energy gain in this picture. We can also understand why the highest magnetization at zero temperature (as in our picture here) can correspond to a fractional number of magnetic moments per unit cell: A complete magnetization of the d electrons would correspond to completely emptying the "up" d band but with the continuous shift here, any fractional magnetization is possible.

8.5.4 Ferromagnetic Domains

Not all ferromagnetic materials appear to show a macroscopic magnetization below the Curie temperature. Sometimes it is necessary to "magnetize" them by an external magnetic field. The reason for this is the existence of **magnetic domains**, as suggested by P. Weiss in 1907 and shown in Figure 8.7a. The domains have different magnetization directions such that the total average magnetization of the sample is zero or small. The domains are separated by so-called **Bloch walls** in which the magnetization rotates from one direction to another. The Bloch walls are typically around 50–100 nm thick. Magnetic domains can be made visible by very fine iron powder on the magnet (the so-called **Bitter method**), or optically by the so-called **Kerr effect**, or by **spin-polarized scanning tunneling microscopy**.[7)]

The existence of domains can be understood by considering what happens when the material is cooled below the Curie temperature. Ferromagnetic ordering sets in spontaneously at different places in the sample and the Bloch walls are formed where the domains meet. There is also another way to explain the origin of the domains. Consider the single-domain magnet in Figure 8.8a. It leads to a strong magnetic field outside the material with a certain energy density. By introducing a few domains as in Figure 8.8b, the external field is strongly reduced, leading to an energy gain, but the cost for this is the formation energy of the domain walls. If the latter is not too high, the state with a few domains will be favorable. If now

7) See online note on www.philiphofmann.net.

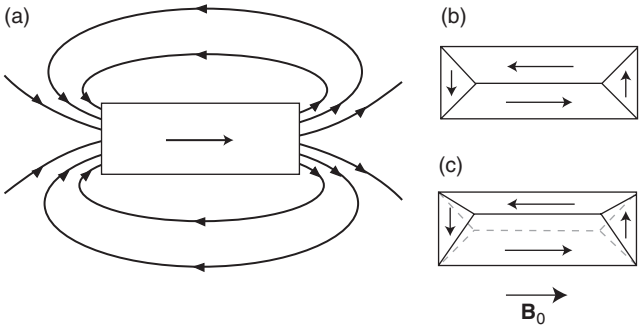


Figure 8.8 (a) Magnetic material with a single domain leading to a strong external field. (b) The introduction of a few domains greatly reduces the external field. (c) The material can be magnetized by the movement of domain walls caused by an external field $\mathbf{B}_0$. The dashed gray lines correspond to the situation in (b) before the exposure to $\mathbf{B}_0$.

an external field is applied, the sample can be magnetized by moving the domain walls relative to each other, as in Figure 8.8c. The movement of the domains walls is clearly the most efficient way to change the magnetization of a sample. An alternative mechanism would be to flip single magnetic moments in the middle of a given domain and eventually change the magnetization in the entire domain. However, the energy barrier for flipping single moments is very high because such a process would have to act against the huge local Weiss field.

8.5.5 Hysteresis

By moving the domain walls in a piece of ferromagnetic material via an external field B_0, different magnetizations can be achieved. Figure 8.9 shows the situation for an initially unmagnetized sample in a slowly oscillating field B_0. In the

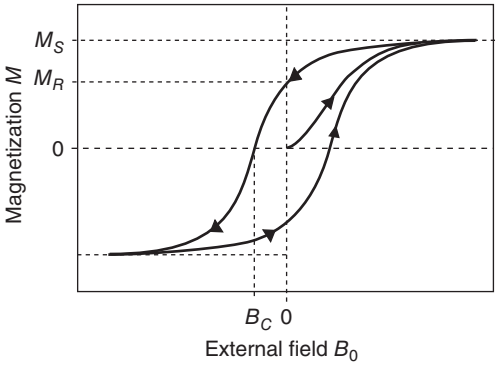


Figure 8.9 Magnetization of a ferromagnetic sample as a function of externally applied field B_0. The starting point of the curve is the origin.

beginning, the magnetization M is zero, and it increases as the field is turned on. For a certain field strength, the sample is completely magnetized in one direction, that is, the saturation magnetization M_S is reached. When the field is lowered again, the magnetization decreases, but for $B_0 = 0$, it has not reached zero but the so-called **remanent magnetization** M_R. A magnetization of $M = 0$ is first reached at the **coercive field** B_C in the opposite direction. For an even stronger field in the opposite direction, saturation is reached again.

The cause for this **hysteresis** is partly that the movement of Bloch walls through the sample is not a simple reversible process. Sometimes the Bloch wall has to pass defects and this costs energy. The energy dissipation for one closed loop of the hysteresis curve can be read directly from the curve if it is displayed in a slightly different way as $B(H)$ instead if of $M(B_0)$. Then, the dissipation is simply $\oint BdH$, that is, the area enclosed by the hysteresis curve. The $B(H)$ curve looks quite similar to the $M(B_0)$ curve, but it does not show saturation in the same sense because B still increases as H increases even if M is saturated (see (8.3)).

Hysteresis can also be obtained in a simple model for a defect-free crystal. Consider a one-domain ferromagnet at a very low temperature. All the spins are aligned with an external magnetic field. The alignment will not be lost when the field is turned to zero. In fact, a field in the opposite direction with considerable strength is required to reverse the magnetization because all the spins have to be flipped over. Once this is achieved, the external field can again be turned to zero with little further changes in the magnetization.

The exact shape of the hysteresis curve depends strongly on the type and structure of the material. It can be tailored to meet the needs of specific applications. If the goal is to have a good permanent magnet, both a large remanent magnetization and a high coercive field are desirable. These properties characterize so-called **hard magnets**. One can play certain tricks to achieve them. For example, one can make the grain size of the material smaller than the typical size of a magnetic domain. Then, the grains cannot change their magnetization by moving domain walls. The magnetization is forced to flip over as a whole, which is an expensive process.

In the opposite extreme, one needs so-called **soft magnets**, for example, for applications in transformers. From what we have seen above, the energy dissipated in each magnetization circle is the area enclosed by the hysteresis loop. In a typical transformer, this energy is lost 50 (or 60) times a second, so the area should be small. This means that both a small coercive field and a small remanent magnetization are desirable. At the same time, one is looking for a high saturation field and a material with a high resistance in order to minimize eddy currents.

References

Tyler, R. (1931) *Philos. Mag.*, **11**, 596 and references therein.

8.6 Further Reading

Magnetism is treated in all standard textbooks on solid state physics, consider in particular

- Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Holt-Saunders.
- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.
- Kittel, C. (2005) *Introduction to Solid State Physics*, 8th edn, John Wiley & Sons, Inc.
- Myers, H.P. (1990) *Introductory Solid State Physics*, 2nd edn, Taylor & Francis Ltd.
- Omar, M.A. (1993) *Elementary Solid State Physics*, Addison-Wesley.
- Rosenberg, H.M. (1988) *The Solid State*, 3rd edn, Oxford University Press.

More in-depth texts on magnetism are:

- Blundell, S. (2001) *Magnetism in Condensed Matter*, Oxford University Press.
- Buschow, K.H.J. and de Boer, F.R. (2003) *Physics of Magnetism and Magnetic Materials*, Kluwer Academic Publishers.
- Himpsel, F.J., Ortega, J.E., Mankey, G.J., and Willis, R.F. (1998) Magnetic Nanostructures, *Adv. Phys.*, **47**, 511. Journal article with special emphasis on the nanoscale.
- Stöhr, J. and Siegmann, H.C. (2006) *Magnetism. From Fundamentals to Nanoscale Dynamics*, Springer.

The diamagnetism of free electrons is discussed in

- Peierls, R.E. (1955) *Quantum Theory of Solids*, Oxford University Press.

The basics of orbital and spin angular momentum, the corresponding magnetic moments, and the coupling to magnetic fields are treated in standard quantum mechanics text, for example

- Griffiths, D.J. (2004) *Introduction to Quantum Mechanics*, Pearson Prentice Hall.

For a basic text on statistical physics and its application to magnetism, see

- Mandl, F. (1988) *Statistical Physics*, 2nd edn, John Wiley & Sons.

8.7 Discussion and Problems

Discussion

1) What is the difference between diamagnetism and paramagnetism? What are the basic physical causes for these phenomena? How do diamagnetic/paramagnetic materials react when placed into an inhomogeneous magnetic field?

2) Can you give an example where the relation between the external field and the magnetization is not linear, that is, where it deviates from (8.5)?
3) Describe the most important magnetic properties of atoms.
4) How do the free electrons in metals contribute to the magnetic properties?
5) What solids are likely to express Curie-type paramagnetism?
6) In a Curie-type paramagnet, all the ions with magnetic moments contribute to the total susceptibility, but in a metal, only very few of the free electrons contribute to Pauli paramagnetism. Why?
7) The Curie formula for the paramagnetic susceptibility $\chi_C = C/T$ is only valid under certain conditions. What are these conditions, and why is this so?
8) What interaction is responsible for the magnetic ordering in ferromagnets and antiferromagnets and how strong is it?
9) Explain the origin of the Weiss field and its significance for magnetic ordering.
10) How does a ferromagnet behave magnetically above its Curie temperature?
11) Most ferromagnetic chunks of material do not show any macroscopic magnetization. Why?
12) When a chunk of ferromagnetic material is magnetized in an external field and then the field is switched off, why does it keep a macroscopic magnetization?
13) What is the difference between a magnetically soft and a magnetically hard material?

Problems

1) *Diamagnetic susceptibility of atoms:* (a) Estimate the upper limit for the (dia) magnetic moment of an atom. Use an atom with many electrons such as Bi. From the magnetic moment, calculate the diamagnetic susceptibility and compare it to the experimental value, which is $\chi_m = -1.7 \times 10^{-4}$. (b)(*) Calculate the magnetic moment for a hydrogen atom in an external field exactly, starting from (8.13). You can use that the ground-state wave function of hydrogen is

$$\psi(r, \theta, \phi) = (\pi a_0^3)^{-1/2} e^{-r/a_0}. \tag{8.45}$$

You can also use that

$$\int_0^\infty x^n e^{-x} dx = n! \tag{8.46}$$

2) *Curie paramagnetism:* One often compares not the measured and calculated Curie constants but rather the so-called effective magneton number p with $p^2 = g_J^2 J(J+1)$. (a) Calculate p for the following ions and compare your result to the experimental value given in brackets: Nd^{3+} with three 4f electrons (3.5), Gd^{3+} with seven 4f electrons (8.0), Cr^{2+} with four 3d electrons (4.9), and Fe^{2+} with six 3d electrons (5.4). (b) Can the agreement for the 3d transition metals

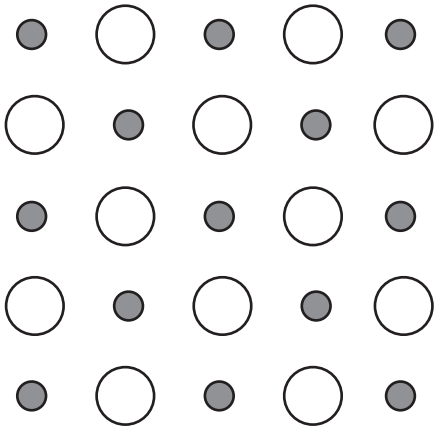


Figure 8.10 A two-dimensional version of NiO. The large circles represent Ni, the small ones O.

be improved by assuming that the orbital angular momentum is quenched, that is, that there is just the total spin?

3) *Ferromagnetic ordering:* For the ferromagnetic elements Gd^{3+} (with seven 4f electrons) and Fe^{2+} (with six 3d electrons), calculate the highest possible magnetization per atom and compare it to the experimental values taken at ≈ 0 K, which are $7.6\mu_B$ and $2.2\mu_B$, respectively. Discuss the results.
4) *Ferromagnetic ordering:* We have argued that ferromagnetic ordering is caused by the exchange interaction. It is a common misconception that it is caused by the alignment of spins due to their mutual magnetic interaction. To show this, estimate the magnetic interaction energy for two spins separated by a typical atomic distance. (Hint: If you want to estimate the magnetic field due to a microscopic moment, you can use an expression for the on-axis field of a circular wire loop and assume that the radius of the loop is very small). For what temperatures would you expect ordering due to this effect?
5) *Ferromagnetic ordering:* Estimate the size of Weiss field based on the typical Curie temperature of a ferromagnet (e.g., $\Theta_C = 1043$ K for Fe).
6) *Curie–Weiss law:* Derive the Curie–Weiss law (8.44) and show that the Curie constant C corresponds to that in the Curie law (8.29).
7) *Ferromagnetic ordering:* What would be the qualitative difference between the density of states for a 3d transition metal shown in Figure 8.6a and that of a 4f transition metal?
8) *Antiferromagnetic ordering:* NiO is antiferromagnetic with a Néel temperature of approximately 500 K. The magnetic moments are localized on the Ni ions only. A two-dimensional version of NiO is shown in Figure 8.10. (a) How would the magnetic moments be oriented in the antiferromagnetic state? What would the unit cell and the reciprocal lattice look like, when

the magnetic ordering is taken into account and when it is not taken into account? (b) What experimental technique would reveal the magnetic ordering directly (and why) and thus allow you to study the transition between the ordered and nonordered state?

9
Dielectrics

In the previous chapter, we have discussed the response of a solid to a magnetic field. In this one, we do the same thing for an electric field. We already know what an electric field does to metals: It causes a current to flow. Here, we are mainly concerned with insulators, called dielectrics in this context, so that the present chapter can also be seen as an extension of our discussion of conductivity in materials (metals, semiconductors, and now insulators). At first glance, one could suspect that nothing interesting happens when an insulator is exposed to an electric field because no current can flow. It turns out that there are other important phenomena such as the dielectric polarization and the piezoelectric effect. We will also discuss the behavior of dielectrics in time-dependent electric fields, such as their interaction with electromagnetic waves.

The formal description is in close analogy with that of magnetism and many phenomena can be described using the same ideas. But there are also some important differences. The interaction of magnetic fields with solids is almost always very weak (except in ferromagnetism) but for the electric field, this is not so. This has some advantages and disadvantages for the formal treatment. For example, the interaction of matter with electromagnetic waves is greatly simplified because we can almost always neglect the magnetic part of the interaction. On the other hand, the calculation of the electric field inside an insulator becomes difficult because we have to take into account not only the external field but also the field created by the polarized solid itself. This is something we could ignore for the description of paramagnetism and diamagnetism (but not for ferromagnetism).

9.1
Macroscopic Description

The macroscopic description of dielectric effects is similar to that for magnetism, but it is not quite the same. Again, we start by a few formal definitions. An **electric field $\mathcal{E}$** leads to a **dielectric polarization P** of the solid of the form

$$\mathbf{P} = \chi_e \epsilon_0 \mathcal{E}, \tag{9.1}$$

where χ_e is the **electric susceptibility** and $\epsilon_0 = 8.854 \times 10^{-12}\ \mathrm{C^2\,J^{-1}\,m^{-1}}$ is the **vacuum permittivity**. In vacuum, there is nothing to be polarized so $\chi_e = 0$.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

The microscopic mechanism for the polarization will be discussed in the next section. It mainly stems from the alignment of microscopic electrical dipoles that are either already there or induced by the field. We write microscopic electric dipole moments as $\mathbf{p} = q\boldsymbol{\delta}$, where q is the magnitude of the charges and $\boldsymbol{\delta}$ their separation vector. As usual for electric dipoles, this vector is defined as pointing from the negative to the positive charge. The macroscopic polarization can then be expressed in terms of the microscopic dipoles as

$$\mathbf{P} = \frac{N}{V}\mathbf{p} = \frac{N}{V}q\boldsymbol{\delta}. \tag{9.2}$$

From this, it is evident that $\mathbf{P}$ has the dimension of a surface charge density.

A material constant closely related to χ_e is the **relative permittivity** or **dielectric constant** ϵ, which is given by

$$\chi_e = \epsilon - 1. \tag{9.3}$$

Defined in this way, both χ_e and ϵ are dimensionless. When dealing with the dielectric properties of solids, it is much more common to use ϵ rather than χ_e to describe the material's polarization. This is different from the case of magnetism where we have used the susceptibility χ_m, rather than the relative permeability μ.

As in the case of magnetism, the linear relation (9.1) represents a limit for weak fields. The trouble is that the electric interaction with matter is not necessarily weak and nonlinear effects are often encountered when using strong electric fields, for example, from laser light. This is a very interesting research field in its own right, but we do not discuss it any further here.

A standard example used to illustrate the electric polarization is the plane plate capacitor shown in Figure 9.1. For the empty capacitor in Figure 9.1a, Gauss's law can be used to calculate that the electric field between the plates of size A and distance d has the constant value $|\boldsymbol{\mathcal{E}}| = \sigma/\epsilon_0$, where σ is the surface charge density on the plates. With this it immediately follows that the capacitance is $C = A\epsilon_0/d$. When a dielectric material is placed between the plates, it is polarized leading to the macroscopic polarization $\mathbf{P}$. We can imagine this as arising from a very high density of microscopically small electric dipoles (see Figure 9.1b). The most important effect of these dipoles is that they lead

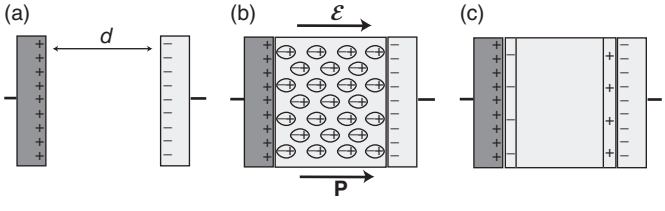


Figure 9.1 A plane plate capacitor. (a) Charges on the plates of the capacitor with no dielectric present between the plates. (b) Polarization of the dielectric material between the plates. (c) The net effect of the polarization is surface charge densities on the dielectric at the plate–dielectric interface.

to a net surface charge density where the dielectric meets the plates. When this is taken into account, the resulting average electric field in the capacitor is $|\boldsymbol{\mathcal{E}}| = (\sigma - |\mathbf{P}|)/\epsilon_0 = \sigma/\epsilon\epsilon_0$, that is, it is reduced by a factor of ϵ. The capacity, in contrast, is increased by the same factor. This is often used to determine the value of ϵ experimentally.

The seemingly simple problem of changing the capacitance of a plane plate capacitor is currently a big challenge in semiconductor device design. Consider the MOSFET in Figure 7.13. The oxide below the gate is in effect the dielectric material in a plane plate capacitor. This capacitor has to be able to store enough charge to make the MOSFET work (without needing too high a gate voltage), that is, it needs a sufficiently large capacitance. The technological goal is to design ever smaller transistors and here the problem sets in. The capacitance is $C = A\epsilon\epsilon_0/d$, where ϵ is the dielectric constant of the gate oxide SiO_2. For a decrease in the size of A, C also decreases. This can be compensated by decreasing the thickness of the oxide layer d and this is what the semiconductor industry has been doing for the last 30 years. But now it does not work anymore because d is at its limit (a few nanometers). For an even thinner oxide, the film becomes "leaky" due to tunneling. Therefore, a current field of research is to find a material with a much higher ϵ than SiO_2 to act as the gate oxide. Then, one could have the same capacitance without the need for a very thin gate oxide.

9.2 Microscopic Polarization

There are several mechanisms giving rise to the microscopic electric dipole moments leading to a macroscopic polarization. They are shown in Figure 9.2. Figure 9.2a illustrates a mechanism that is really an atomic effect and has little to do with the fact that the atoms are placed in a solid. In an electric field, the

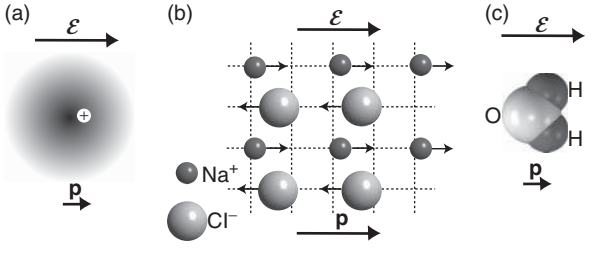


Figure 9.2 Mechanisms leading to microscopic electric polarization. (a) The electric field polarizes all the atoms in the solid. (b) In ionic solids, like NaCl, the lattice can be polarized, giving rise to local electric dipoles. The dashed grid gives the position of the ions without an applied field. (c) If there are permanent dipoles in the solid and these are free to rotate, they orient themselves parallel to the field. A molecule with a permanent dipole is, for example, water.

spherical symmetry of an atom is lifted and the negative and positive charges are displaced, giving rise to an electric dipole

$$\mathbf{p} = \alpha \boldsymbol{\mathcal{E}}, \tag{9.4}$$

where α is the **atomic polarizability**. This effect is called **electronic polarization** and present in all solids, of course. For a simple estimate of α see Problem 9.1.

The next polarization mechanism is relevant to ionic solids, where something very similar happens on a larger scale, as shown in Figure 9.2b. In the electric field, the lattice itself gets polarized since the positive ions are displaced in the direction of the external field and the negative ions opposite to the field. This effect is called **ionic polarization**.

Finally, there is the possibility that already existing dipoles are oriented in the field. Such permanent dipoles could be molecules such as water or HCl. This polarization mechanism is called **orientational polarization**. However, it is much more common in liquids or gases than in solids because the dipoles have to be free to rotate.

With (9.1) and (9.3) in mind, the presence of different polarization mechanisms should give rise to different dielectric constants ϵ. Values of ϵ for a range of materials are listed in Table 9.1. Some trends are immediately clear. Under normal conditions, air has a dielectric constant close to 1 (an electric susceptibility close to 0), simply because the density of air is very low (see (9.2)). The dielectric constant is markedly higher for solids. One should expect a noticeably higher dielectric constant for crystals with the possibility of ionic polarization (NaCl or $SrTiO_3$) than for crystals with only electronic polarization (diamond), but this is not necessarily the case. NaCl has a very similar ϵ to diamond, whereas the value is much higher for $SrTiO_3$. We will discuss this special case later on.

Table 9.1 Dielectric constant ϵ of selected materials at room temperature.

Material	Dielectric constant, ϵ
Vacuum	1
Air	1.000 573 (283 K, 1013 hPa)
Rubber	2.5–3.5
Glass	5–10
Diamond	5.7
Si	11.7
SiO_2	3.9
CdSe	10.2
NaCl	6.1
$SrTiO_3$	350
Ethanol (liquid)	25.8
Water (liquid)	81.1

The microscopic mechanisms for the polarization of the solid are somewhat reminiscent of the mechanisms leading to a magnetization. There is, however, one important difference. All the mechanisms described above lead to a polarization in the direction of the external field, that is, to "para-electric" behavior, not to "dia-electric" behavior.

9.3 The Local Field

Suppose that we know the dielectric constant ϵ of some solid and we want to calculate the microscopic polarizability α of the atoms making up the solid, knowing that there are no other mechanisms of polarization. We can use (9.1)–(9.4) to write

$$\mathbf{P} = (\epsilon - 1)\epsilon_0 \boldsymbol{\mathcal{E}} = \frac{N}{V}\mathbf{p} = \frac{N}{V}\alpha \boldsymbol{\mathcal{E}}, \tag{9.5}$$

where $\boldsymbol{\mathcal{E}}$ is the average electric field in the dielectric material, that is, the sum of the external field and the average internal field due to the polarization. In the case of a dielectric material inside a plane plate capacitor, we know this field $\boldsymbol{\mathcal{E}}$. It is the external field reduced by a factor ϵ. Rearranging (9.5) gives the desired relation:

$$\alpha = \frac{(\epsilon - 1)\epsilon_0 V}{N}. \tag{9.6}$$

Unfortunately, this is not correct because a microscopic dipole in the solid does not feel the average field $\boldsymbol{\mathcal{E}}$ but the **local field** $\boldsymbol{\mathcal{E}}_{\text{loc}}$ at its site and this can be quite different. We can think of this local field by simply taking out the dipole under consideration and inspect the effect of all the other charges in its neighborhood. This is shown in Figure 9.3. It becomes clear that the effect of the neighboring dipoles gives rise to an additional field $\boldsymbol{\mathcal{E}}'$ that is parallel with the average field, that is, the local field felt by every dipole is stronger than the average field in the solid.

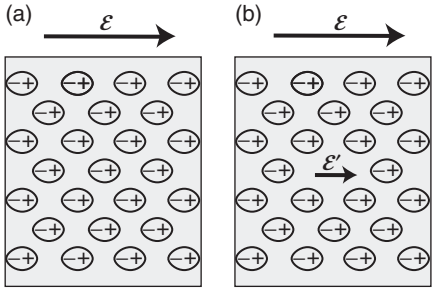


Figure 9.3 The local field on microscopic polarizable units. (a) Microscopic dipoles in a dielectric placed in an external field. The electric field $\boldsymbol{\mathcal{E}}$ is the average internal field in the dielectric. (b) The local field felt by every single dipole is not just $\boldsymbol{\mathcal{E}}$ but $\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{E}}'$ because the surrounding charges lead to a field increase.

We do not derive the strength of the local field here,[1] we merely give the result, which is

$$\boldsymbol{\mathcal{E}}_{\text{loc}} = \frac{1}{3}(\epsilon + 2)\boldsymbol{\mathcal{E}}, \tag{9.7}$$

so we get

$$\mathbf{P} = \frac{N}{V}\alpha\boldsymbol{\mathcal{E}}_{\text{loc}} = \frac{N\alpha}{3V}(\epsilon + 2)\boldsymbol{\mathcal{E}}. \tag{9.8}$$

On the other hand, we have (9.1) with (9.3) or the left side of (9.5) as an expression for **P** and by setting the two equal, we obtain the so-called **Clausius–Mossotti relation**, which relates the atomic polarizability to the dielectric constant:

$$\alpha = \frac{\epsilon - 1}{\epsilon + 2}\frac{3\epsilon_0 V}{N}. \tag{9.9}$$

Experimentally, the relation can best be tested for gases where the density can be varied.

9.4 Frequency Dependence of the Dielectric Constant

9.4.1 Excitation of Lattice Vibrations

So far, we have only studied electrostatic behavior. Far more interesting is the dynamic behavior for fields with a time dependence, in particular, in the region of optical frequencies. We know from optics that ϵ is actually a complex number,[2] and in the Drude model we have already encountered a considerable frequency dependence of ϵ (see Section 5.2.3 and in particular (5.28)). We have seen that this can explain why metals become transparent for light with a frequency above the plasma frequency ω_P. The frequency-dependent $\epsilon(\omega)$ is usually called the **dielectric function**.

For insulators, it is also found that $\epsilon(\omega)$ is complex and frequency-dependent and that energy can be resonantly transferred to the solid at some frequencies. The frequency dependence of ϵ implies a frequency dependence of the refractive index N through (5.20) and this effect is well known as dispersion in optical materials such as glass. In the following, we discuss some simple models to explain the frequency dependence of ϵ.

In a static electric field, all types of microscopic polarization are important: electronic polarization as well as ionic polarization and orientational polarization (the latter two obviously only if they are possible). At very high frequencies, on the other hand, the ions move too slowly to follow the changes of the electric field

1) For the derivation, see online note on www.philiphofmann.net.
2) For anisotropic materials, ϵ is not even a scalar quantity but a complex second rank tensor. As everywhere else in this book, we ignore the effect of anisotropy, unless it is strictly needed for a specific phenomenon.

Table 9.2 Dielectric constants ϵ of selected materials in the electrostatic case and at optical frequencies.

Material	Static ϵ	Optical ϵ
Diamond	5.68	5.66
NaCl	6.1	2.34
LiF	11.95	2.78
TiO_2	94	6.8

and the rotation of permanent dipoles is even slower. Only the very fast electronic polarization remains and, therefore, the total polarization of the solid decreases greatly. Consequently, $\epsilon(\omega)$ should be lower at high frequencies than in the electrostatic case.

In case of the ionic polarization, we have a good idea about the timescale involved. In Chapter 4, we have seen that lattice vibrations have a frequency on the order of 10^{13} Hz. So, for optical frequencies ($\approx 10^{14}$–10^{15} Hz), the lattice ions will not be able to follow the field anymore. This is confirmed by the data in Table 9.2 in which the electrostatic and the optical ϵ are given for different materials. The static ϵ is significantly larger than the optical ϵ, apart from the case of diamond, for which only electronic polarization can play a role.

We can describe the frequency dependence of ϵ more quantitatively for a simple but instructive model. We have discussed that light can only couple to optical phonons very close to the center of the Brillouin zone at $\mathbf{k} = 0$. These phonons correspond to an out-of-phase vibration of the positive and negative ions in the unit cell, whereas the motion in between unit cells is in phase (see Figure 4.5 and Problem 4.2). The mode at $\mathbf{k} = 0$ is thus the only relevant vibration for the interaction with light, and we therefore approximate the crystal as independent harmonic oscillators with one such optical oscillator per unit cell. Each oscillator shall be driven by an external field[3] of the form $\mathcal{E}_0 \exp(-i\omega t)$ and have a resonance frequency $\omega_0 = (2\gamma/M)^{1/2}$, where γ is the force constant and M the reduced mass of the two ions (see Problem 4.2). We also include a damping term $\eta dx/dt$ that is proportional to the velocity, that is, the rate at which the interatomic distance changes. The physical meaning of this is that if this particular motion gets very strong, it is likely to excite other vibrations and thus be damped. The resulting equation of motion is that of a damped and driven harmonic oscillator:

$$\frac{d^2x}{dt^2} + \eta\frac{dx}{dt} + \omega_0^2 x = \frac{e\mathcal{E}_0}{M}e^{-i\omega t}. \tag{9.10}$$

Note the similarity to the problem of a free electron in an external field where we had the equation of motion (5.22), which has no restoring force and no damping. As in that case, a good ansatz for the solution is

3) The electric field used here should actually be the local field, not the average internal field. This is not important for the conclusions from this simple model but it is important for the phenomenon of ferroelectricity, which is discussed later on.

$$x(t) = Ae^{-i\omega t}, \tag{9.11}$$

resulting in an expression for the amplitude A:

$$A = \frac{e\mathcal{E}_0}{M}\frac{1}{\omega_0^2 - \omega^2 - i\eta\omega}. \tag{9.12}$$

It will later be useful to split A into its real and imaginary parts by expanding the fraction with the complex conjugate of the denominator to give

$$A = \frac{e\mathcal{E}_0}{M}\left(\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \eta^2\omega^2} + \frac{i\eta\omega}{(\omega_0^2 - \omega^2)^2 + \eta^2\omega^2}\right). \tag{9.13}$$

The oscillation of the ions will be accompanied by an ionic polarization of $eAe^{-i\omega t}$ for every unit cell. From this, we can calculate the total polarization for a crystal with N unit cells and volume V. Apart from the ionic polarization $P_i(t)$, we consider the electronic polarization of the ions $P_e(t)$. This gives

$$P(t) = P_i(t) + P_e(t) = \frac{N}{V}eAe^{-i\omega t} + \frac{N}{V}\alpha\mathcal{E}_0 e^{-i\omega t}. \tag{9.14}$$

For simplicity, we have assumed that there is only one type of ions with a density of N/V and an effective atomic polarizability α. The two different ions in the crystal and their different polarizability can be taken care of by a suitable definition of α. Now we can calculate the dielectric function

$$\epsilon = \frac{P(t)}{\epsilon_0\mathcal{E}_0 e^{-i\omega t}} + 1 = \frac{NeA}{V\epsilon_0\mathcal{E}_0} + \frac{N\alpha}{V\epsilon_0} + 1. \tag{9.15}$$

At sufficiently high frequencies, we know that $P_i = 0$ so that the optical limit must be

$$\epsilon_{\text{opt}} = \frac{N\alpha}{V\epsilon_0} + 1, \tag{9.16}$$

and therefore

$$\epsilon(\omega) = \frac{NeA}{V\epsilon_0\mathcal{E}_0} + \epsilon_{\text{opt}}. \tag{9.17}$$

Combining this with (9.13), we get the final expression for the dielectric function $\epsilon(\omega) = \epsilon_r(\omega) + i\epsilon_i(\omega)$ with the real and imaginary parts:

$$\epsilon_r(\omega) = \frac{Ne^2}{V\epsilon_0 M}\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \eta^2\omega^2} + \epsilon_{\text{opt}} \tag{9.18}$$

and

$$\epsilon_i(\omega) = \frac{Ne^2}{V\epsilon_0 M}\frac{\eta\omega}{(\omega_0^2 - \omega^2)^2 + \eta^2\omega^2}. \tag{9.19}$$

Both parts of $\epsilon(\omega)$ are plotted in Figure 9.4. The real part of $\epsilon(\omega)$ is almost constant much below and much above ω_0 but its value is higher at lower frequencies, consistent with what we have said above. The imaginary part of $\epsilon(\omega)$ is zero almost

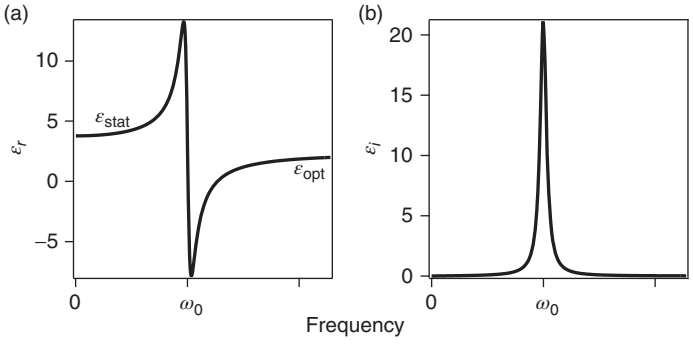


Figure 9.4 Dielectric function for the damped, driven harmonic oscillator close to the resonance frequency ω_0. (a) Real part and (b) imaginary part of ϵ. ϵ_{stat} and ϵ_{opt} stand for the static and optical value of the real part of ϵ, respectively.

everywhere apart from the immediate vicinity of ω_0 where it shows a peak with a width given mainly by η.

What is the meaning of ϵ_i? To see this, consider the energy dissipation in the system. The instantaneous electrical power dissipated per unit volume is given by

$$p(t) = j(t)\mathcal{E}(t), \tag{9.20}$$

where $j(t)$ is the (AC) current density and $\mathcal{E}(t)$ the electric field.[4] As usual, we write $\mathcal{E}(t) = \mathcal{E}_0 \exp(-i\omega t)$. In an insulator, there are no free currents and, we assume here, no magnetic fields. The only currents are polarization currents and using Ampère's law in matter (A.22) gives

$$j(t) = -\frac{\partial D}{\partial t} = -\frac{\partial}{\partial t}\epsilon\epsilon_0\mathcal{E}(t) = \epsilon_0\mathcal{E}(t)\left(i\omega\epsilon_r - \omega\epsilon_i\right). \tag{9.21}$$

The average dissipated power per cycle can now be calculated by

$$\overline{p} = \frac{1}{T}\int_0^T \mathcal{E}(t)j(t)dt, \tag{9.22}$$

where $T = 2\pi/\omega$ is the period of one oscillation. We can easily see what happens in two limiting cases. If the dielectric function is purely imaginary, $j(t)$ is out of phase with $\mathcal{E}(t)$, and the product of the two is always negative. In this case, the integral will give a nonzero value and the dissipated power per cycle is

$$\frac{1}{2}\epsilon_0\epsilon_i\omega\mathcal{E}_0^2. \tag{9.23}$$

If, however, ϵ is purely real, there will be a phase shift of $\pi/2$ between $\mathcal{E}(t)$ and $j(t)$, the integrand oscillates around zero, and the integration gives $\overline{p} = 0$. We therefore see that ϵ_i measures the degree of power dissipation in the solid. It is obviously highest at the resonance frequency where the vibrational amplitude is highest. This leads to the excitation of other vibrations via the friction term in (9.10) and to an accompanying power dissipation.

4) We work with scalar quantities here because we assume $j(t)$ and $\mathcal{E}(t)$ to be in the same direction but not necessarily with the same phase.

9.4.2 Electronic Transitions

We have now seen how light can excite optical phonons. This happens for infrared light with a photon energy far below the band gap energy E_g. For photons with $h\nu > E_g$, electronic excitations across the gap become possible, as we have already discussed in connection with solar cells (see Figure 7.15b). In the following, we explore this in a little more detail. What we describe is not limited to dielectrics. It is equally valid for semiconductors when $h\nu > E_g$ and the general idea can also be used for metals. The question we ask is how the band structure of a material influences the absorption of light, and if there are specific light frequencies where the absorption is particularly strong. This question can be answered in the framework of a complex dielectric function that we have just developed.

At frequencies higher than the optical phonon frequency, we have so far assumed that there is only electronic polarization, an atomic effect that gives rise to a high-frequency ϵ_{opt}. As we see in Figure 9.4, the resulting high-frequency dielectric function would be $\epsilon(\omega) = \epsilon_{\text{opt}} + i0$. The absence of an imaginary part implies that no energy could be absorbed. For $h\nu > E_g$, this concept needs to be revised. It turns out that $\epsilon(\omega)$ is not constant but has some pronounced structures for most materials. This is already evident from the different colors materials have, even metals. This frequency dependence of $\epsilon(\omega)$ in the visible and ultraviolet region is due to the excitation of electrons from occupied to unoccupied states, and we need to explore the band structure of a material to see what excitations are actually possible. However, the key idea is already clear from atomic physics. Atoms have discrete energy levels and when a transition between an occupied level and an unoccupied level is allowed by the optical selection rules, this will lead to a strong absorption of light when the photon energy is equal to the energy difference between these two levels. For solids, the atomic energy levels are broadened into bands, but strong absorption also takes place for photon energies that allow many transitions from occupied states to unoccupied states.

Consider the simplified insulator/semiconductor band structure in Figure 9.5a. We assume that the chemical potential lies somewhere between the valence band (VB) and the conduction band (CB). Absorption of photons can occur when $h\nu > E_g$ and it leads to the excitation of electrons from the VB to the CB. As we have discussed previously (in Section 7.4.3), the wave vector for photons with energies in the visible or ultraviolet range is very short and crystal momentum conservation therefore requires that an electron's $\mathbf{k}$ vector remains unchanged in such a transition (plus a reciprocal lattice vector). Possible photon-induced transitions are indicated by gray arrows in Figure 9.5a. These transitions correspond to the absorption of energy by the solid and we have seen that such an absorption can be described by ϵ_i. In the spirit of the simple model for the dielectric function, we can write that

$$\epsilon_i(h\nu) \propto \sum_{\mathbf{k}} M^2 \delta(E_C(\mathbf{k}) - E_V(\mathbf{k}) - h\nu), \tag{9.24}$$

where the sum runs over all the permitted $\mathbf{k}$ values in the first Brillouin zone, M is a matrix element determining the transition probability, δ stands for the

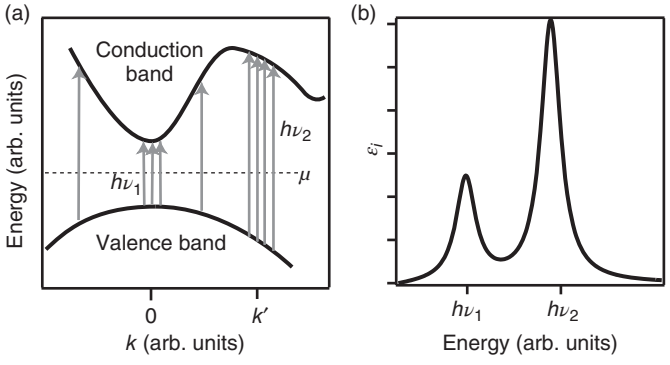


Figure 9.5 (a) Contributions to the imaginary part of $\epsilon(\omega)$ by transitions between occupied and unoccupied states. The gray lines denote possible transitions. The density of such transitions is high at $\mathbf{k} = 0$ and $\mathbf{k} = \mathbf{k}'$ because the VB and CB are parallel there. (b) ϵ_i resulting from the possible transitions in this band structure.

Dirac delta function and $E_C(\mathbf{k})$ and $E_V(\mathbf{k})$ are the dispersions of the CB and VB, respectively. If we ignore the matrix element, this expression basically counts the possible transitions for a certain $h\nu$: The δ function is only 1 if the energy difference $E_C(\mathbf{k}) - E_V(\mathbf{k})$ is exactly equal to $h\nu$. Therefore, $\epsilon_i(h\nu)$ has maxima at $h\nu$ values for which many different transitions between the VB and CB are possible. In Figure 9.5a, this is the case for the onset of the absorption at $h\nu_1 = E_g(\mathbf{k} = 0)$ and around $\mathbf{k} = \mathbf{k}'$, a region of the Brillouin zone where the VB and CB are parallel. Maxima in ϵ_i can thus be expected at $h\nu_1$ and $h\nu_2$, as illustrated in Figure 9.5b. As these maxima correspond to a strong absorption, they determine the color of the solid. While we have thereby made (9.24) plausible, this equation can also be strictly derived using Fermi's Golden Rule for time-dependent perturbations.

The key to strong absorption is thus to have the VB and CB states disperse in parallel over large fractions of the Brillouin zone. This condition is obviously fulfilled for very flat bands, arising from strongly localized states. Then, the situation is reminiscent of the absorption of light by transitions between atomic levels.

9.5 Other Effects

9.5.1 Impurities in Dielectrics

A small amount of impurities can have a pronounced effect on the properties of dielectrics, very much like in the case of semiconductors. We give two examples here. The most "visible" case, as it were, is the change of optical properties that can be caused by impurities. Due to their large gap, most insulators/dielectrics tend to be transparent, for example, NaCl, diamond, and sapphire. If there are impurities in the material with electronic states inside the gap, this can lead to a

strong change in the optical properties because now transitions from or into the impurity states are possible, and ϵ_i will show a resonance at the corresponding energy (see (9.24)). Good examples of such impurity states exist in sapphire. Depending on the type of impurity, sapphire can have many different colors (and names), for example, topaz (yellow), amethyst (purple), ruby (red), emerald (green), or sapphire (blue).

Impurities can also be used to make insulators conductive, exactly like doped semiconductors. Donor or acceptor levels have to lie close to the CB and VB, respectively, in order to give rise to a appreciable density of electrons and holes and thereby to an increased conductivity. The important advantage over a usual semiconductor material is that high-temperature applications are possible. In a narrow-gap semiconductor, high temperatures are a problem because of the exponentially increasing number of intrinsic carriers. In a doped insulator, this is not an issue for practically relevant temperatures.

9.5.2 Ferroelectricity

Ferroelectric materials are solids that exhibit a spontaneous electric dipole moment, very much like the spontaneous magnetic moment in ferromagnets. Otherwise, the term "ferroelectric" is quite misleading because the typical ferroelectric does not contain any iron and the mechanism leading to the polarization can be quite different from ferromagnetism as well. A typical ferroelectric material is barium titanate ($BaTiO_3$) (Figure 9.6), in which the electric dipole moment stems from ionic polarization: The (negatively charged) oxygen lattice is shifted against the positively charged Ba and Ti ions. What is special about this

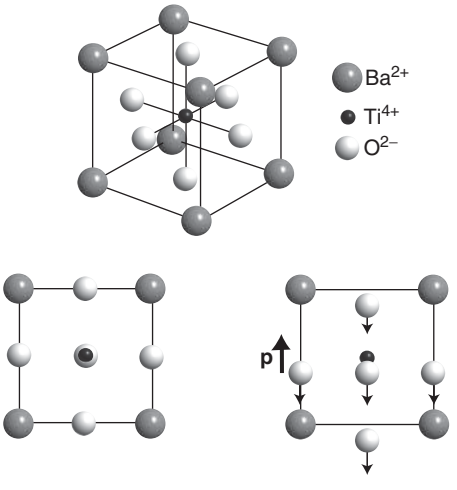


Figure 9.6 *Upper part:* The unit cell of barium titanate $BaTiO_3$ with the charges of the ions. *Lower part:* In the ferroelectric state, the (negative) oxygen sublattice is displaced from the sublattice containing the (positive) Ba and Ti ions.

polarization is that it is stable without an external electric field. If an external electric field is applied, the orientation of a ferroelectric's polarization can be reversed and there is hysteresis, very much like in the case of ferromagnetism.

The basic idea behind the ferroelectric effect is as follows: We have discussed the equation of motion for two ions in a unit cell, which leads to ionic polarization, by using a simple harmonic oscillator as a model (see (9.10)). In this framework, we have considered the motion of the ions under the influence of the average electric field in the solid. We should have used the local field, but we have argued that this does not change the course of the argument. In the case of ferroelectric materials, this distinction does matter because when we move an ion out of its equilibrium position, the local field force can pull it even further if it is stronger than the harmonic restoring force. Eventually, force equilibrium is reached but in this way, a distortion and a permanent electric dipole are generated. At a certain temperature, the thermal fluctuations become strong enough to destroy the ferroelectric state. As in the case of ferromagnetism, this temperature is called the **Curie temperature**. For barium titanate, the Curie temperature is about 130 °C.

The context of ferroelectricity allows us to understand the very high dielectric constant of strontium titanate (see Table 9.1). $SrTiO_3$ has the same crystal structure as $BaTiO_3$ and it is almost ferroelectric – but not quite (it can be made ferroelectric in thin films and under strain). It remains very easily polarizable and has a high dielectric constant. Materials on the verge of ferroelectricity could thus have interesting applications as gate insulators in MOSFETs.

9.5.3 Piezoelectricity

Piezoelectricity is an effect in which applying stress to a material leads to a macroscopic electric polarization. This, in turn, gives rise to net surface polarization charges, and these can be detected by measuring the voltage across the sample. (see Figure 9.7a). The converse effect also exists. When a voltage is applied across the material, this leads to a macroscopic strain (Figure 9.7b). The figure also shows a possible microscopic structure that can give rise to this effect. The structure contains three dipoles that are arranged such that the resulting dipole moment is zero. A deformation of the unit gives rise to a net microscopic dipole moment. Equivalently, the unit will deform in an applied electric field. It is important to note that ferroelectric materials also show piezoelectricity but the opposite is not necessarily true. What is special about ferroelectricity is the spontaneous electric polarization of the solid.

Piezoelectric materials have many applications. Some examples are sensors (like in microphones), high voltage sources (cigarette lighters), or actuators (loudspeakers). Piezocrystal-based actuators are especially important for nanotechnology because they permit positioning with unrivaled precision. They are used for positioning the tip in a scanning tunneling microscope, for example.

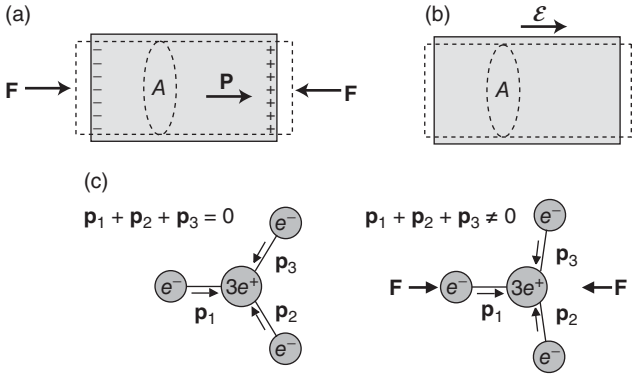


Figure 9.7 (a) Exposing a piezoelectric material to mechanical stress results in a macroscopic electric polarization. (b) Conversely, an electric field across the sample leads to a mechanical strain. (c) This is caused by the deformation of microscopic units in the crystal. Such a unit is shown without applied field or stress on the left. The unit consists of three dipole moments that sum up to a total dipole moment of zero. On the right, stress is applied. This results in a distortion of the unit and a net dipole moment that is no longer zero.

9.5.4 Dielectric Breakdown

If the electric field across an insulator is too high, the insulator will start to conduct a current. This phenomenon is known as dielectric breakdown. The mechanism for the breakdown is that some free carriers (e.g., caused by impurities) are accelerated in the field, so much that they can ionize other atoms and generate more free carriers. Then, the breakdown proceeds like an avalanche. The breakdown can be facilitated by operating the material close to a resonance frequency where much energy is dissipated, the material is heated, and the probability of having free carriers is increased.

9.6 Further Reading

Dielectrics are discussed in most general solid state physics books, see in particular

- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.

More detailed information on optical properties and ferroelectricity is given in

- Fox, M. (2010) *Optical Properties of Solids*, 2nd edn, Oxford University Press.

9.7 Discussion and Problems

Discussion

1) What are the physical mechanisms that can lead to a dielectric polarization of solids?
2) If we want to calculate the polarization of a dielectric from the polarizability of the atoms, we have to use the local electric field, which is different from the average internal electric field and from the external electric field. Explain all three fields. Why could we ignore the corresponding complications in the case of Curie-type paramagnetism?
3) Qualitatively describe the frequency dependence of the dielectric function for a material with both electronic and ionic polarization.
4) What is the physical meaning of the imaginary part of the dielectric function?
5) Describe the role of impurities in dielectrics. Why are sapphire and diamond transparent and how can impurities change this?
6) What is the difference between ferroelectricity and piezoelectricity?

Problems

1) *Electronic polarization:* (a) Assume that an atom consists of a uniform sphere of negative charge with radius R surrounding a positive point charge. Show that the atomic polarizability is equal to $4\pi\epsilon_0 R^3$. The negative charge in the sphere should be taken to remain uniform in an applied field. (b) Use this to calculate the atomic polarizability for Ne and compare it to the experimental value of 4.3×10^{-41} Fm^2. Take the atomic radius of Ne to be 0.51 Å. (c) Estimate the index of refraction for Ne gas under normal conditions.
2) *Dielectric function:* Estimate the static dielectric constant of NaCl. Use that NaCl has a density of $\rho = 2170$ kg m^{-3}, a lattice constant of $a = 5.6$ Å, and a Young's modulus of $Y = 40$ GPa. The optical ϵ can be taken from Table 9.2 and the result can be compared to the static ϵ from the same table.
3) *Dielectric function:* We have argued that the imaginary part of the dielectric function $\epsilon_i(\omega)$ is responsible for the dissipation of energy as an electromagnetic wave travels through the solid. If this is so, one would also expect that such a wave is strongly damped in regions where $\epsilon_i(\omega)$ is high. We also know that the damping of the wave is closely linked to the imaginary part of the index of refraction $N = n + i\kappa$. (a) Show that for a given complex ϵ, n, and κ can be calculated as

$$n = \sqrt{\frac{|\epsilon| + \epsilon_r}{2}} \quad \kappa = \sqrt{\frac{|\epsilon| - \epsilon_r}{2}}, \tag{9.25}$$

(b) Plot $n(\omega)$ and $\kappa(\omega)$ for one harmonic oscillator described by (9.18) and (9.19) and discuss the result. (c) Is a strong damping of the waves (or a high κ) always associated with a high imaginary part of the dielectric function?

4) *Reflectivity of Metals:* After having studied the dielectric properties of insulators more closely, we can go back to the reflectivity of a metal, which we have discussed in connection with the Drude model. In Chapter 5, we have merely argued that the a low-frequency wave cannot enter a metal (because of the high κ) but that it cannot be absorbed either (because of the lack of damping in (5.22)) and so energy conservation dictates that it should be reflected. Now we can calculate the reflectivity explicitly. The combination of the Maxwell equations with the appropriate boundary conditions leads to the so-called Fresnel equations, which describe the reflection and transmission through an interface. The normal-incidence reflectivity of a solid in vacuum, for instance, is given by

$$R = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2} \tag{9.26}$$

Calculate and plot the reflectivity $R(\omega)$ for a free electron metal. To do this, derive a general expression for n and κ for a given ϵ (see Problem 9.3a) and use (9.26) in order to obtain the reflectivity.

10
Superconductivity

Superconductivity, that is, the fact that a current can flow in materials with zero resistance, was discovered in 1911 by H. Kamerlingh Onnes, shortly after it became possible to liquefy helium and thereby to reach the required low temperatures. The discovery was made when Kamerlingh Onnes wanted to study the low-temperature resistivity of mercury. At a temperature of 4.2 K, the resistivity dropped to an unmeasurably small value. The discovery is a good example of a spectacular, totally unexpected, and practically important result from very basic research. It should be followed by many other surprises in the behavior of solids at low temperatures. Superconductivity implies several phenomena in addition to zero resistivity. These will be discussed in this chapter. It turns out that the change from the normal state to the superconducting state is a phase transition of the metal and that several properties change due to this transition.

Even though a lot of experimental and theoretical effort was made in order to arrive at a microscopic explanation of superconductivity, it took more than 40 years until the fundamental aspects of such a theory were laid out by J. Bardeen, L. N. Cooper, and J. R. Schrieffer. Their theory is now known as the BCS theory of superconductivity. The basic idea behind the theory is that charged carriers in the superconductor condense into a single ground state, forming a coherent and macroscopic matter wave, that is, a quantum mechanical wave function that exists on a macroscopic scale. Today, macroscopic wave functions are known from other branches of physics. One example is laser light, where many photons are in the same quantum state and macroscopic coherence is achieved. Other examples are superfluidity, a low-temperature quantum state of liquid ^{4}He, which allows flow without any friction, or Bose–Einstein condensation, in which many atoms can condense into a single quantum state at very low temperatures. The particles forming these macroscopic quantum states are bosons. Bosons do not underlie the Pauli exclusion principle and can all condense in the same quantum state. The obvious problem is that electrons in a metal should not be able to do this because they are fermions. It turns out that the way around this problem is to form two-electron pairs that have an integer spin and therefore behave as bosons.

In this chapter, we will first look at some basic experimental observations from superconductors before we turn to the theory and physical principles behind the phenomenon. After this, we will discuss some more advanced experimental

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

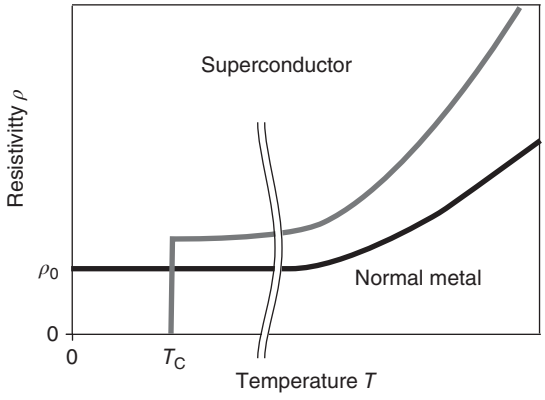


Figure 10.1 Typical temperature-dependent resistivity for a normal metal and for a superconductor. Note that the temperature axis is broken to indicate that there is a broad temperature interval between the superconducting transition and a strong temperature-induced resistivity increase.

observations and their explanation as well as at the so-called high-temperature superconductors. We conclude the chapter with a few comments on the richness of the field, which we can touch only very briefly here.

10.1 Basic Experimental Facts

10.1.1 Zero Resistivity

The most prominent phenomenon associated with superconductivity is of course the vanishing resistivity below a certain **critical temperature** T_C. The temperature-dependent resistivities for a superconductor and a "normal" metal that does not show superconductivity are shown in Figure 10.1. The resistivity of a normal metal decreases as the temperature is lowered and eventually levels off to a constant value. We have discussed the physical origin of this behavior already. The resistivity in a metal is caused by imperfections in the lattice such as impurity atoms, lattice defects, and thermal vibrations. For temperatures much lower than the Debye temperature, the vibrations are effectively "frozen in" but even at zero temperature, impurities and defects are present and hence the resistivity is expected to remain finite.

For a superconductor, the picture is entirely different. Above the critical temperature T_C, the same behavior as for a normal metal is observed. At T_C, however, the resistivity drops to an unmeasurably small value. The temperature interval for the transition is very small, usually below 10^{-3} K. The width of the transition range depends somewhat on the quality of the sample, but the transition temperature is a characteristic constant of the material.

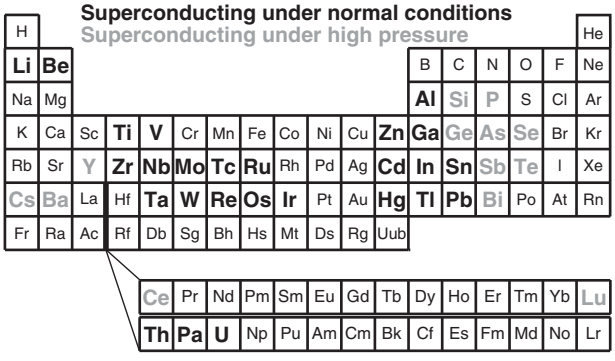


Figure 10.2 Periodic table of the elements with the superconducting elements in bold black letters. Gray letters indicate that the elements do only become superconducting in a high-pressure modification.

Superconductivity appears to be a common phenomenon as seen from Figure 10.2, which is a periodic table of the elements in which superconducting elements are highlighted. The table does of course not give any certainty that the other elements are not superconducting. It is possible that the necessary low temperature has just not been reached yet. In fact, superconductivity in Li at normal pressure was discovered as recent as 2007 at a temperature below 0.4 mK. In any case, a few trends emerge quite clearly from the table. The first is that metals that are good conductors are not necessarily also superconductors. In fact, the very best conductors such as Ag and Cu have not been found to be superconducting. Another is that the ferromagnetic elements are not superconducting either, suggesting that ferromagnetism and superconductivity are somehow mutually exclusive. Indeed, it also turns out that contaminating a superconducting sample with a very small amount of magnetic impurities can destroy the superconductivity.

The critical temperature T_C is quite low for all elements. It is highest for Nb, which has $T_C = 9.2$ K. T_C can be higher for some metallic alloys, up to around 40 K, and it can be much higher for some transition metal oxides, the so-called high-temperature superconductors.

Some elements do not undergo a transition to a superconducting state at ambient pressure, but they can be made superconducting by exerting pressure. These elements are indicated in light gray in Figure 10.2. High pressure could possibly change the structure of a crystal in order to turn it into a superconductor, but it has to be clear that the structure as such is not important in the superconducting phase transition. It is the electronic structure and the vibrations that count, as we will see later. However, if the elements do crystallize in different structures, for example, due to high pressure, these structures also have different electronic structures and, to a lesser degree, different vibrational properties.

Another possibility to achieve superconductivity in materials that are normally not superconductors is to grow them in ultrathin films or as small clusters of atoms. The films or clusters may still have the same structure as the bulk material under normal conditions but their electronic structure and the superconducting

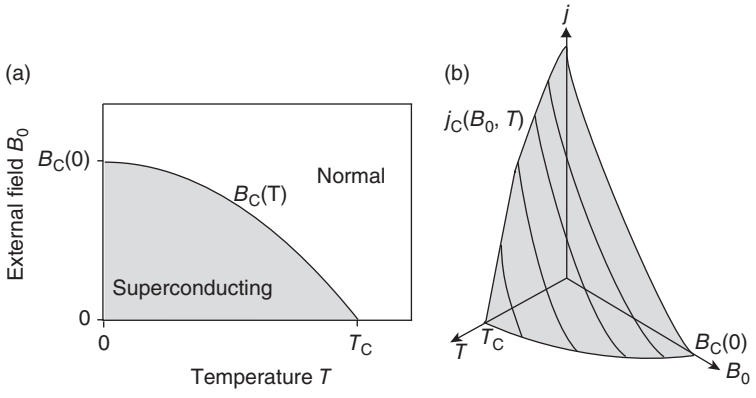


Figure 10.3 (a) Combined effect of a magnetic field and a finite temperature on a superconductor. In the region below the curve, that is, for low temperatures and weak external fields, the solid is superconducting. Above the curve, it is in its normal state. (b) The same but including the effect of a finite current density.

behavior can be different (see Chapter 11). Finally, superconductivity can also be realized in solids that do not have any long-range crystalline order, the so-called amorphous solids. In some cases, this may actually favor the transition. One example is the semimetal Bi, which is not superconducting in its normal crystalline bulk structure but which is superconducting as an amorphous film.

How does one know that the resistivity in the superconducting state is really zero and not just very small? In fact, one does not know this and it is not possible to answer this question from any experiment. One can, however, try to find an upper limit for the resistivity. The experimental approach to this is to induce a current in a superconducting ring using a magnetic field. One should then expect this current to go on forever without any decay. Observing the decay (or rather the lack of it) over a long time (years!) gives the possibility to put an upper limit on the resistivity. Currently, this upper limit is thought to be around 10^{-25} Ωm.

Apart from the temperature, there are two other important factors that can destroy the superconducting state. These are a magnetic field and a current through the sample. The combined effect of a magnetic field and the temperature is shown in Figure 10.3a. For low temperatures and weak magnetic fields, in the region below the curve, the solid is in its superconducting state. For too high a magnetic field or too high a temperature, the superconductivity is lost, corresponding to the region above the curve. For a given temperature $T < T_C$, we can assign a **critical magnetic field** $B_C(T)$, which turns out to be

$$B_C(T) = B_C(0)\left[1 - \left(\frac{T}{T_c}\right)^2\right]. \tag{10.1}$$

A current through the sample, as expressed by a current density j, has a similar effect as a magnetic field. For too high a current density, the superconductivity breaks down. Again, this **critical current density** j_C is a function of the

temperature. It is also a function of the applied magnetic field B_0, as shown in Figure 10.3b. The superconducting state is only reached for low temperatures, current densities, and magnetic fields. In a similar way, the critical magnetic field can be viewed as a function of temperature and current density, of course.

Unfortunately, the critical current densities and magnetic fields for most elemental superconductors are too low to permit any meaningful technical applications. Take, for example, the electromagnet in a medical magnetic resonance scanner. This magnet has to produce a high field of more than 1 T. In order to do this, very high currents are necessary. It is not possible to build such magnets out of normal conductors because of the large amount of heat that would be produced by ohmic losses. To build them from superconducting materials is not straightforward either, because of the high magnetic field and the high current densities. Rather than elemental superconductors, one often uses superconducting alloys that can have critical fields of up to 50 T and a critical current density of up to 10^{11} Am^{-2} at liquid He temperature.

10.1.2 The Meissner Effect

The so-called Meissner effect, discovered in 1933 by W. Meissner and R. Ochsenfeld, is another fundamental property of a superconductor. It is the fact that a superconductor shows perfect diamagnetism in its superconducting state, that is, it has $\chi_\mathrm{m} = -1$ and therefore a magnetization $\mathbf{M} = -\mathbf{B}_0/\mu_0$, which totally cancels the external field $\mathbf{B}_0$ inside the superconductor. The origin of the macroscopic magnetization $\mathbf{M}$ is, however, very different from that in normal diamagnetic materials. In the latter, diamagnetism is caused by microscopic magnetic moments that are induced by the external field throughout the entire solid. In a superconductor, this is not the case. The magnetization $\mathbf{M}$ of a superconductor is rather caused by macroscopic supercurrents that flow close to the surface of the specimen and keep the inside field-free.

The magnetic susceptibility of -1 implies that superconductors are strongly expelled from magnetic fields. You are probably familiar with the Meissner effect from diamagnetic levitation experiments, in which a small high-temperature superconductor is levitated in an inhomogeneous magnetic field (see our discussion of (8.6)).

It is very important to understand that the Meissner effect is a genuinely new effect and not simply a consequence of the vanishing resistance, which can give rise to a supercurrent to keep the inner specimen field-free. There is a difference between a (hypothetical) material that becomes merely a perfect conductor with $\rho = 0$ below T_C and a true superconductor that also displays the Meissner effect. For a perfect conductor, the entire physics is given by the Faraday's law which, in its integral form, is

$$\oint \mathcal{E}\, d\mathbf{l} = -\frac{d\Phi_B}{dt}, \tag{10.2}$$

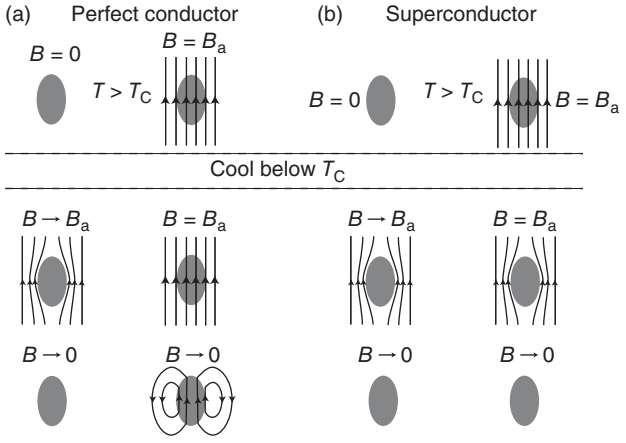


Figure 10.4 The Meissner effect is not merely a consequence of zero resistivity. (a) The behavior of a "perfect conductor," that is, a material that merely has zero resistivity below T_C. For this material, the magnetic field in the specimen below T_C depends on the presence and size of a magnetic field before cooling down below T_C. (b) The situation for a genuine superconductor that displays the Meissner effect. The interior of the specimen is field-free below T_C, independent of the sample's history.

where the integral is taken along a closed path in the conductor. For a perfect conductor, no electric field can exist along such a closed path, which means that the entire integral vanishes and the magnetic flux through the loop of integration is constant.

Figure 10.4a illustrates this behavior of a perfect conductor in a magnetic field. We start with two specimens above T_C, one in an applied magnetic field $\mathbf{B}_a$ and one in a field-free region. Both are cooled below T_C, where the resistivity of the conductor drops to zero. Now the magnetic field is also turned on for the conductor that was in the field-free region before. Because of Faraday's law, it expels the field entirely and the inner part remains field-free. For the other specimen, nothing changes because the field had been penetrating already before the transition below T_C. In the end, the field is taken to zero for both specimens. Since it cannot change inside the perfect conductor, one remains field-free, and in the other one a current is induced, which keeps the field in the specimen as it was before. In short, we can have any field we like below T_C. It only depends on the history and the situation is thus far from perfect diamagnetism.

Now consider a real superconductor that exhibits the Meissner effect as shown in Figure 10.4b. Again, we start with two specimens in the normal state, one without a penetrating magnetic field and one with such a field. Once the specimens are cooled below T_C, the magnetic field is expelled in *both* cases, that is, the interior of the superconductor is field-free, independent of the history of the sample. Once the field is turned off, the interior of the specimen remains of course field-free. The experimental observation that a superconductor does indeed behave in this way was a very important step for the understanding of

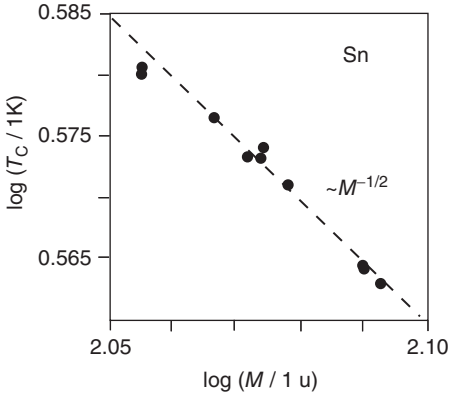


Figure 10.5 Illustration of the isotope effect. The graph shows the critical temperature as a function of isotope mass as a log–log plot. The data points lie on a straight line suggesting a power law behavior with $T_C \propto M^{-1/2}$. Data taken from Maxwell (1952), Serin, Reynolds, and Lohman (1952).

superconductivity because it permits the description of the superconducting state as a single thermodynamic phase, which can be described by a few macroscopic variables. A history-dependent magnetization of the sample would not permit such a view.

The bottom line is that the Meissner effect is a genuine effect in itself, which is not implied by zero resistivity. Any theory of superconductivity must be able to explain not only resistance-free current flow but also the Meissner effect.

10.1.3 The Isotope Effect

There are several more experimental observations associated with superconductivity and some will be discussed in a later section. The so-called isotope effect, however, is presented already here because it gives an important clue for the development of a microscopic theory.

The effect is illustrated in Figure 10.5 that shows the critical temperature of Sn as a function of the atomic mass M, which can be varied by using different isotopes of Sn. The figure is actually a log–log plot on which the data appear as a straight line, equivalent to a power law behavior with $T_C \propto M^{-1/2}$.

The fact that T_C depends at all on M is remarkable. As a consequence of the Born–Oppenheimer approximation, the electronic structure of a solid should not depend on the mass of the ions, just on their chemical nature. The vibrational properties, in contrast, do depend on the mass of the ions. For a simple harmonic oscillator, we know that $\omega = (\gamma/M)^{1/2}$, that is, the vibrational frequency depends on the mass in the same way as T_C in a superconductor. This suggests that the lattice vibrations of the solid play some role in superconductivity.

10.2 Some Theoretical Aspects

10.2.1 Phenomenological Theory

Early theories of superconductivity were not atomistic but merely sought a macroscopic formalism that explains both the resistance-free transport and the Meissner effect. This was achieved by F. London and H. London in 1935. Their theory can be summarized by the two so-called **London equations** that we now discuss.

Achieving infinite conductivity for a gas of charged particles is very easy in a macroscopic theory. In fact, in the Drude model, the equation of motion for the electrons was taken to be (5.2) and this does already lead to an infinite conductivity. In order to prevent this from happening, we had to introduce the concept of the relaxation time. The first London equation is identical to (5.2), only that the equation is now written using a current density $\mathbf{j}$ instead of the velocity $\mathbf{v}$ of a single electron:

$$\frac{\partial \mathbf{j}}{\partial t} = \frac{n_s q^2}{m_s} \boldsymbol{\mathcal{E}}, \tag{10.3}$$

where n_s is the density of superconducting particles, m_s is their mass and q their charge.[1)] If we take the curl of this equation and combine it with Faraday's law (A.17), we get

$$\frac{\partial}{\partial t}\left(\frac{m_s}{n_s q^2} \mathrm{curl}\mathbf{j} + \mathbf{B} \right) = 0. \tag{10.4}$$

What this means becomes clear if we integrate it over a cross-sectional area $\mathbf{A}$ of the solid and then use Stokes' integral theorem (A.11), giving

$$\frac{\partial}{\partial t}\left(\int \frac{m_s}{n_s q^2} \mathrm{curl}\mathbf{j} d\mathbf{A} + \int \mathbf{B} d\mathbf{A} \right) = \frac{\partial}{\partial t}\left(\oint \frac{m_s}{n_s q^2} \mathbf{j} dl + \int \mathbf{B} d\mathbf{A} \right) = 0. \tag{10.5}$$

The result contains two types of magnetic flux through the area $\mathbf{A}$. The first is caused by an electric current density $\mathbf{j}$ on the perimeter of the area and the second is the actual flux from the external field $\mathbf{B}$. We do not know the size of the first contribution but the equation tells us that the sum is constant in time. So, if we change the external field $\mathbf{B}$, this change is exactly compensated by a change in current density $\mathbf{j}$ on the solid's surface such that the total magnetic flux does not change. The result expresses exactly what we already know to be true for a perfect conductor (see Figure 10.4a).

The second London equation is obtained by not only requiring that the partial derivative with respect to time in (10.4) is zero, but that the term in brackets vanishes altogether, that is,

$$\frac{m_s}{n_s q^2} \mathrm{curl}\mathbf{j} + \mathbf{B} = 0. \tag{10.6}$$

1) If the particles carrying the supercurrent turn out to be different from electrons, we just have to substitute the corresponding mass, charge, and density into the equations.

Using the same procedure as for (10.5), we can see that this equation describes the Meissner effect correctly. Now the current density **j** can be calculated from the external field **B** and the internal field created by this current density exactly compensates the external field.

Expelling the magnetic field from the inside of the superconductor is thus achieved by a current density on the surface of the superconductor. But how far do these currents penetrate into the material? The London equations permit a more quantitative approach to this. Inside the superconductor where $\mathbf{D} = \epsilon\epsilon_0\boldsymbol{\mathcal{E}} = 0$, we have Ampère's law (A.22) as

$$\mathrm{curl}\mathbf{B} = \mu_0\mathbf{j}. \tag{10.7}$$

We combine this with (10.6) and get

$$\mathrm{curl}\,\mathrm{curl}\mathbf{B} = \mu_0\mathrm{curl}\mathbf{j} = -\frac{\mu_0 n_s q^2}{m_s}\mathbf{B}. \tag{10.8}$$

Now we can use that in general $\mathrm{curl}\,\mathrm{curl}\mathbf{B} = \mathrm{grad}\,\mathrm{div}\mathbf{B} - \triangle\mathbf{B}$, so that[2)]

$$\triangle\mathbf{B} = \frac{\mu_0 n_s q^2}{m_s}\mathbf{B}. \tag{10.9}$$

With this we have a differential equation for the penetrating magnetic field and a very similar equation can be derived for the current density, using the same technique. It is

$$\triangle\mathbf{j} = \frac{\mu_0 n_s q^2}{m_s}\mathbf{j}. \tag{10.10}$$

Both equations can be solved by an exponentially decreasing field and current density, respectively. The situation for the field is shown in Figure 10.6. The characteristic length of the exponential decrease is the so-called **London penetration depth** λ_L, which is given by

$$\lambda_\mathrm{L} = \sqrt{m_s/\mu_0 n_s q^2}. \tag{10.11}$$

Assuming that all the electrons are carriers of supercurrents and that they have a free electron mass and charge, one can estimate the London penetration depth to be on the order of 30 nm. The agreement with the measured penetration depth is not perfect but the order of magnitude is correct. The quality of this agreement is somewhat fortuitous: We will see that in a microscopic theory only a very small fraction of the electrons participates in the superconductivity. Also, the density of the superconducting particles is not constant. It varies both with position and temperature.

A considerably more sophisticated phenomenological theory of superconductivity was presented in 1950 by V. L. Ginzburg and L. D. Landau. This **Ginzburg–Landau theory** defines a so-called order parameter $\Psi(\mathbf{r})$, which can be derived

2) You may not be familiar with the notation in which the Laplace operator $\triangle$ is applied to a vector field. This notation just means that the Laplace operator is applied individually to each single component of the original field, giving again a vector field.

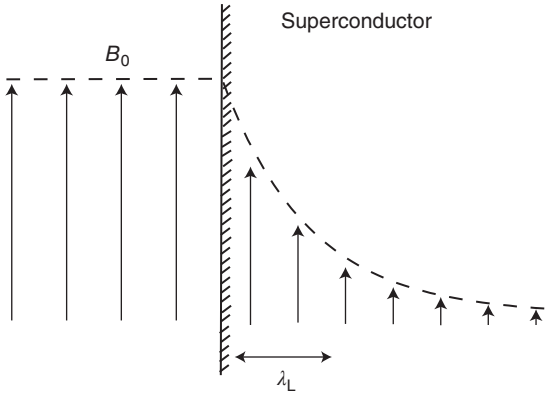


Figure 10.6 Exponential damping of an external magnetic field near the surface of a superconductor.

from an equation similar to the Schrödinger equation and can be written as the complex function

$$\Psi(\mathbf{r}) = \Psi_0(\mathbf{r})e^{i\phi(\mathbf{r})}, \tag{10.12}$$

in which $\phi(\mathbf{r})$ is a macroscopically changing phase and $|\Psi^*\Psi| = \Psi_0^2$ is equal to the density of superconducting carriers n_s. The spatial character of $\Psi(\mathbf{r})$ means that n_s does no longer need to be constant but it cannot change instantaneously either, for example, from zero to a high value at the surface of a superconductor. Appreciable changes of $\Psi(\mathbf{r})$, and thereby n_s, have to happen over a length scale ξ, the so-called **coherence length** of the superconductor. The Ginzburg–Landau theory is remarkably successful in describing many phenomena associated with superconductivity even though it is not a microscopic theory. An especially notable feature of the theory is that the order parameter already has the character of a macroscopic quantum mechanical wave function.

10.2.2
Microscopic BCS Theory

The microscopic theory of superconductivity was formulated in 1957, more than 40 years after the discovery of the effect, by J. Bardeen, L. N. Cooper, and J. R. Schrieffer. It is commonly referred to as the BCS theory, after the names of its inventors. Several requirements for a microscopic theory were clear for a long time before the theory was actually developed: It should of course be possible to explain the vanishing resistivity and the Meissner effect. The theory should probably involve lattice vibrations in some way in order to account for the isotope effect. Finally, it should not be specific to any particular material since we have seen that superconductivity is a rather common phenomenon.

Why did it take such a long time to come up with a microscopic picture of superconductivity? Part of the reason is that two of the most fundamental

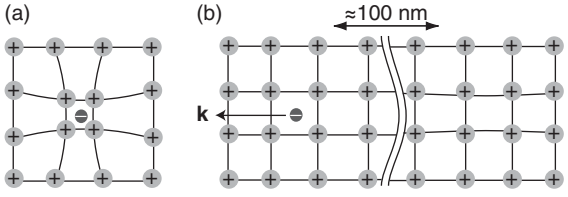


Figure 10.7 Local deformation of the lattice via the electrostatic interaction between the electrons and the ions. Situation for (a) a very slow or static electron and (b) an electron near the Fermi energy of a metal.

assumptions for the treatment of solids are not valid in the case of superconductivity and their breakdown is even essential for the theory. The first assumption is the Born–Oppenheimer approximation that allowed us to treat the electronic and the vibrational properties separately. The second is the independent electron approximation in which we considered the properties of one electron in the average potential of all others.[3)]

As already mentioned in the introduction, the superconducting state represents a quantum phenomenon on a macroscopic scale. We will see some direct experimental proof of this in the later sections. Like in a laser or in a Bose–Einstein condensate, a macroscopic wave function is realized by very many particles occupying the same quantum state. While we have to be careful in comparing the condensing particles in the solid with noninteracting bosons, it is clear that these particles cannot be the free electrons because these are fermions and have to obey the Pauli principle.

It was L. N. Cooper who realized that the formation of electron *pairs* due to a net attraction, no matter how weak, would resolve this problem and open the possibility for a new ground state of the electron gas with entirely new properties and a slightly lower energy than the original ground state discussed in Chapter 6. Since the energy of the new state is lower, one would expect a phase transition to this state, at least for very low temperatures where the entropy is not important.

A possible mechanism providing a weak attraction between electrons is the interaction between the electrons and lattice vibrations, the so-called electron–phonon interaction. This interaction is weak and usually ignored when we use the Born–Oppenheimer approximation. But for superconductivity it is vital. Figure 10.7a illustrates its origin. When an electron is placed somewhere in the ionic lattice, it slightly deforms the bonds around it due to the electrostatic interaction with the ions. Such a localized lattice distortion can be imagined as wave packets made from phonons, very similar to a localized electron that can be viewed as a packet from Bloch waves. The deformation is such that the lattice ions are attracted toward the position of the electron. This leads to a local polarization of the lattice, which in turn is attractive for other electrons. In this way, an attractive interaction can exist between two electrons.

3) We have, of course, already abandoned this approximation in the case of magnetic ordering.

The static picture given in Figure 10.7a does, however, not really work because another electron that moves to the polarized part of the lattice would still feel the strong electrostatic repulsion from the first electron. Also, we shall see that the electrons relevant for superconductivity have kinetic energies close to the Fermi energy and are therefore not static. Figure 10.7b shows a more appropriate dynamic picture. An electron moves through the lattice with a high speed, on the order of 10^6 ms^{-1} and attracts the positive lattice ions on its way. The lattice is polarized, but since the ions move much slower than the electrons, the polarization is not established instantaneously. We can estimate that the time for polarizing the lattice has to be on the same timescale as the ion movement, that is, $2\pi/\omega_D \approx 10^{-13}$ s. By the time the maximum polarization is reached, the electron has already traveled around 10^6 ms$^{-1} \times 10^{-13}$ s = 100 nm.

Now it is possible for a second electron to lower its energy by moving in the trail of the first. Electrostatic repulsion between the two electrons will be insignificant because they are so far away from each other. If the second electron is to stay on the same track, its wave vector $\mathbf{k}$ must either be the same as for the first electron or exactly opposite. It can be shown that the energetically most favorable situation arises when the two electrons have exactly opposite $\mathbf{k}$ vectors and opposite spins. The electron pairs formed from electrons with $\mathbf{k}$ and $-\mathbf{k}$ and a total spin of zero are called **Cooper pairs**.

While these simple arguments give a good visualization of the physical origin of Cooper pair formation, we must be very careful not to take them too far. It is, for example, somewhat stretching the imagination that the wave vector of the second electron should be $-\mathbf{k}$ and not $\mathbf{k}$, that is, that the second electron moves in the opposite direction of the first. But we have to realize that the question of which electron is where and moves how has already lost its meaning when we study a wave function of just two electrons. It is not obvious either that the total spin of the pair should be zero. In fact, there are some exotic superconductors in which it is one. However, the crucial point still is that it must be integer and not half-integer, that is, that the Cooper pairs are bosons, not fermions.

In order to see why we have used the velocity of the electrons at the Fermi energy to estimate the separation between the pair of electrons, we describe the electron–phonon interaction in a more quantum mechanical picture. The interaction between two electrons via phonons can be viewed as the constant emission and absorption of "virtual" phonons of energies up to $\approx \hbar\omega_D$. The emission of a phonon of energy $\hbar\omega$ does not violate energy conservation if it is only short-lived, that is, when it is rapidly absorbed by another electron. The only electrons that can participate in the exchange of such "virtual" phonons and in the formation of Cooper pairs are those close to the Fermi energy, within a window of approximately $\hbar\omega_D$. None of the other electrons can emit or absorb "virtual" phonons because they are trapped by the Fermi–Dirac distribution. All the reachable states around them are already occupied. This picture also explains why the interaction can be so important at low temperatures when essentially all the phonons are "frozen in" and the lattice only performs zero point motions.

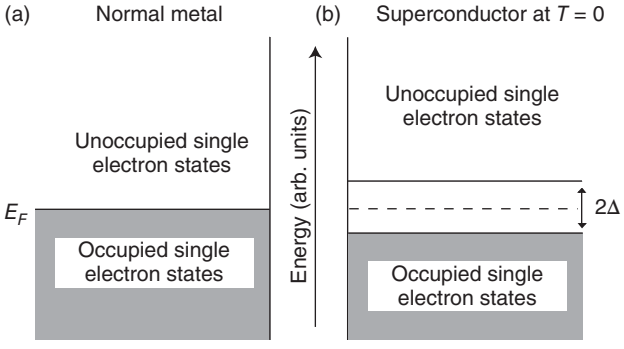


Figure 10.8 Occupation of single-electron levels at zero temperature in (a) a normal metal and (b) a superconductor. For the superconductor, the electrons close to E_F are bound in Cooper pairs and these occupy a many-body state, the BCS ground state, which cannot be shown in this figure of single-particle levels. In order to excite single electrons out of this ground state, a Cooper pair has to be broken. This costs an energy of Δ per electron and creates two unpaired electrons in the lowest possible unoccupied single-particle states.

The formation of Cooper pairs is accompanied by their condensation into a common ground state, something that is possible due to their bosonic character. This state is called the **BCS ground state** and it represents an energy gain compared to the conventional metallic grounds state. We first discuss this state at zero temperature. The energy levels for a normal metal at zero temperature have the familiar form shown in Figure 10.8a. They are completely filled up to E_F and empty above. In the BCS ground state, on the other hand, the electrons close to the Fermi energy are all bound in Cooper pairs and they have gained an average energy Δ per electron by doing so.

The Cooper pairs do not appear in a single-particle energy diagram such as Figure 10.8. Their state is a many-body state with one total energy for all the electrons bound in Cooper pairs. This is similar to the hydrogen molecule energy levels shown in Figure 2.2. For H_2 there are two energy levels for every possible interatomic distance, one for the singlet and one for the triplet state. It does not make sense to split these up into single-electron energies. The energy levels are genuine two-electron states.

For the superconductor, we can look at the remaining electrons that are not bound in Cooper pairs. Their energy levels are drastically modified as shown in Figure 10.8b: A gap of size 2Δ is opened around E_F, that is, there are no single-particle states close to the Fermi energy any more. We can understand this qualitatively by the fact that the electron states just below E_F have been removed to form the Cooper pairs. Above E_F, the lowest possible energy for a single electron is Δ. This is consistent with the energy cost of breaking up a Cooper pair, which is 2Δ and produces two single electrons in the lowest possible energy state, just above the gap.

The gap in the single-electron energies is a very characteristic and central feature of the BCS model. It is frequently referred to as the **gap in the single-particle excitation spectrum**, because now the minimal excitation energy for an unpaired electron is not zero (like in a normal metal) but 2Δ. In this respect, the situation is similar to a semiconductor that also has a gap, but the gap here is very much smaller and, as we will see in a moment, strongly dependent on the temperature.

What is the size of the superconducting gap? Only the electrons close to the Fermi energy can take part in the exchange of "virtual" phonons, which leads to the formation of Cooper pairs. One could therefore expect electrons with energies of up to $\hbar\omega_D$ below E_F to participate in the pairing, so that $\Delta \approx \hbar\omega_D$. This is not too far off the mark: We shall see below that Δ is indeed proportional to $\hbar\omega_D$ but it is usually a lot smaller. The BCS theory predicts that Δ at zero temperature is related to the critical temperature T_C as $\Delta = 3.53 k_B T_C$. Since we know T_C from experiments, we can calculate that Δ is very small, usually only a few meV.

Rather than taking the experimental value, T_C can also be obtained from the BCS theory as

$$T_C = 1.13\Theta_D \exp \frac{-1}{g(E_F)V}, \tag{10.13}$$

where Θ_D is the Debye temperature, $g(E_F)$ the electronic density of states at the Fermi energy, and V a parameter measuring the electron–phonon coupling strength. This confirms not only that $\Delta = 3.53 k_B T_C \propto \hbar\omega_D$ but it also tells us a lot about BCS-type superconductivity. First of all, the transition temperature is proportional to the Debye temperature. This readily explains the isotope effect since $\Theta_D \propto \omega_D \propto M^{-1/2}$. Equation (10.13) also shows that the Debye temperature sets the temperature scale of the phenomenon. The exponential cannot become larger than 1, so that one cannot hope to get a transition temperature higher than $1.13\Theta_D$. In reality, the exponential is much smaller than 1. Both the density of states and the interaction strength enter the expression in the same way, such that both increase T_C. It is clear that only the density of states at the Fermi energy is relevant since only the electrons at this energy participate in the formation of Cooper pairs. With these results from the BCS theory, we estimate the number of electrons in Cooper pairs (per volume) to be $g(E_F)\Delta$ and the associated total energy gain to be $g(E_F)\Delta^2$.

At finite temperatures, not all the electrons close to E_F are bound in Cooper pairs and matters are made complicated by a typical phenomenon for the condensation of bosons: The energy gain for the condensation of a particle into the ground state depends on the number of particles already in this state. In other words, the formation energy 2Δ for a Cooper pair depends on the number of pairs already in the ground state and thus Δ becomes a function of temperature. The predicted temperature dependence of Δ is shown in Figure 10.9. When the sample is cooled below T_C, Δ assumes nonzero values and some Cooper pairs start to form. This, in turn, increases the energy gain for the formation of further Cooper pairs. At zero temperature, all the electrons close to E_F are bound in Cooper pairs. As we have seen above, the size of the gap can also be taken as a measure for how many electrons are condensed in Cooper pairs.

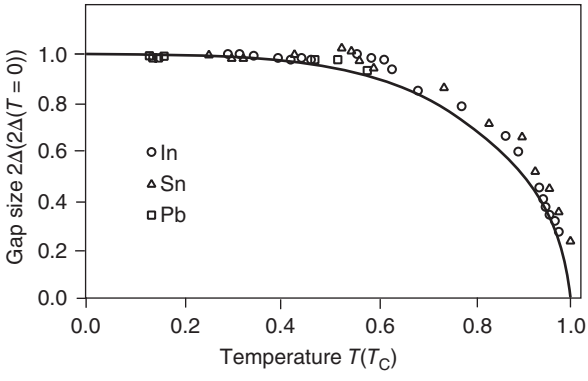


Figure 10.9 Gap size for a superconductor in the BCS model as a function of temperature and comparison to experimental data from In, Sn and Pb. At the transition temperature T_C, the gap is closed. Data from Giaever and Megerle (1961).

This intricate behavior is reminiscent of magnetic ordering in the solid where the energy gain for the orientation of a spin depends on the number of spins already present in that orientation. In magnetic ordering as in superconductivity, a self-amplifying process sets in once the sample is cooled under the transition temperature (note the similarity between Figures 8.5 and 10.9). Upon lowering the temperature further, the magnetic alignment and the number of Cooper pairs are increased in the magnet and superconductor, respectively. With this, we also understand the fact that the transition to the superconducting state is so sharp in temperature. Figure 10.9 also shows a comparison of the predicted gap size to the experimental values for some elemental superconductors. The rapidly changing gap size near T_C is clearly confirmed experimentally.

The existence of Cooper pairs and the gap in the singe-particle spectrum can be used to explain all the experimental observations associated with superconductivity. Here, we merely discuss the resistance-free transport and the existence of a critical current density and a critical magnetic field. We do not show how the Meissner effect can be explained but the basic idea is that an equation similar to the second London equation (10.6) can be derived in the BCS model.

As already mentioned in connection with the London equations, a resistance-free transport is readily obtained in both the classical and the quantum model of electrons in a solid if we do not introduce a scattering mechanism to keep the resistance finite. Exactly the same applies for Cooper pairs: The whole condensate of pairs is accelerated when an electric field is applied to the sample. Before the field is applied, the Cooper pairs are all in the same quantum state with same total $\mathbf{k} = 0$ per pair. When a current passes through the sample, all the pairs are still in the same state but with a different $\mathbf{k}' \neq 0$. The only thing we have to explain is the absence of scattering processes that can lead to a decay of this current. In order to see the difference between the Cooper pairs and unpaired electrons, it is useful to look back at Figure 6.16b. In a normal metal, the current decays because the electrons are scattered back to lower lying states by defects or thermal vibrations.

This scattering is an individual process, affecting one electron at a time. Such a process does not work for a condensate of Cooper pairs because it is not possible to change **k** for one pair without changing it for all the others at the same time, a process that is exceedingly unlikely, unless it is done by applying the same force on all the pairs at the same time, as by an electric field.

We also need to consider scattering processes that split a Cooper pair into two individual electrons. This is, in fact, possible but only if the kinetic energy of the pairs is high enough to provide the necessary 2Δ for the destruction of the Cooper pair. The process will thus become important at high current densities for which the number of Cooper pairs in the solid gradually decreases until superconductivity breaks up altogether, when the critical current density $\mathbf{j}_C$ is reached. The existence of a critical magnetic field B_C can be explained in a very similar way. This critical field is reached when the magnetic energy density exceeds the value needed for breaking up the Cooper pairs. Finally, it is clear from Figure 10.9 why the critical magnetic field and the critical current density depend on the temperature. They decrease as T_C is approached because of the shrinking gap size Δ and the lower density of Cooper pairs.

10.3 Experimental Detection of the Gap

The concept of a gap in the excitation spectrum is central to the BCS theory. Therefore, experimental tests of the existence and (temperature-dependent) size of the gap have been important for confirming the suggested mechanism for superconductivity. In fact, the comparison shown in Figure 10.9 is very compelling in this respect. Here, we discuss three approaches to actually measuring the size of the gap: single-electron tunneling, optical reflectivity, and the low-temperature heat capacity in the superconducting state.

A tunneling experiment between a superconductor below T_C and a normal metal is illustrated in Figure 10.10. The metal and the superconductor are separated by an insulating barrier, such as an oxide, which is only a few nanometers thick. The wave functions of the metal and the superconductor are not cut off at the interfaces to the oxide but leak out into the oxide. The small overlap between them permits the tunneling of electrons from one side to the other. Here, we consider only elastic tunneling of single electrons.

If no external voltage is applied, the Fermi energy in the metal is aligned with the chemical potential in the superconductor, that is, the middle of the gap (Figure 10.10a). For a small positive tunneling voltage on the superconductor, no current can flow because the electrons from the metal do not find empty states in the superconductor to tunnel into (Figure 10.10b). As the tunneling voltage U reaches a value of $U = \Delta/e$, this situation changes. Now the electrons close to the Fermi energy of the metal find empty states in the superconductor and tunneling becomes possible (Figure 10.10c). This causes a sudden rise of the tunneling current at Δ/e (Figure 10.10d) that permits the determination of the gap size.

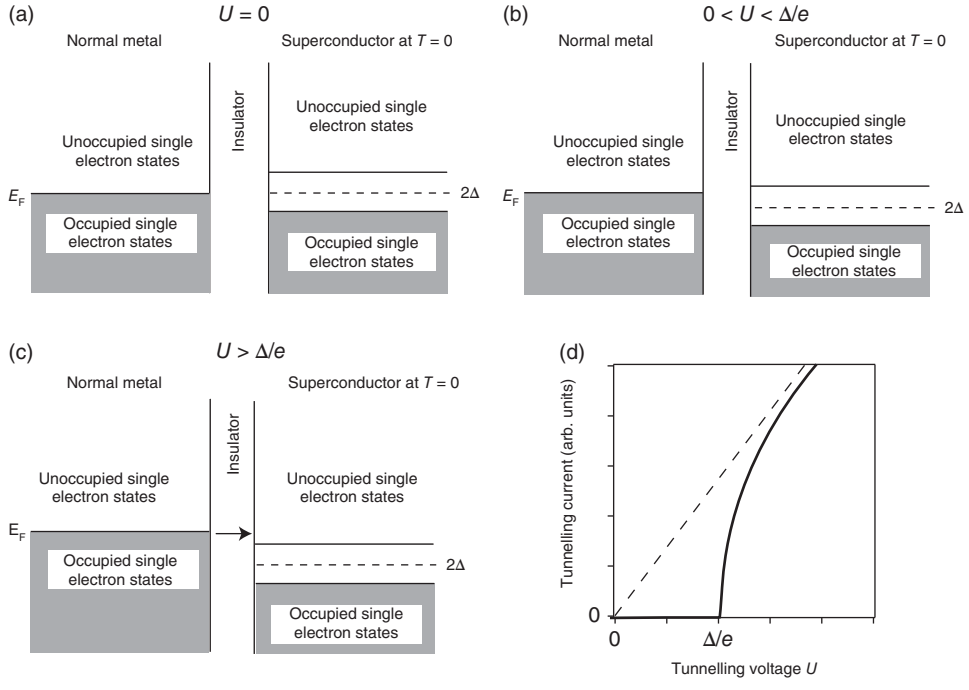


Figure 10.10 Tunneling experiment between a superconductor and a normal metal. The two are separated by a thin insulating oxide. Only elastic single-electron tunneling is considered. (a) Situation without applied voltage. (b) For a small applied voltage no tunneling is possible because of the lack of available states in the gap. (c) As the tunneling voltage exceeds Δ/e, single-electron tunneling becomes possible. (d) Thick line: tunneling current vs. voltage for the present junction; dashed line: corresponding curve for tunneling between two metals.

Note that the situation would be entirely different for tunneling between two metals: Tunneling would already be possible at very small voltages, and the number of electrons able to tunnel would increase continuously with the voltage. Such a behavior is indicated by the dashed line in Figure 10.10d.

Another possibility for detecting the existence of the gap is to measure the absorption of electromagnetic radiation as it passes through a superconductor. For electromagnetic radiation with an energy of $h\nu < 2\Delta$, electronic transitions across the gap are not possible and hence no absorption is observed. As the photon energy reaches 2Δ, absorption sets in and the transmitted intensity is strongly reduced. Since the gap size is at best a few meV, the electromagnetic radiation required for this experiment lies in the microwave or far infrared region.

Yet another experiment that points toward the existence of a gap is the low-temperature heat capacity of a superconductor. Figure 10.11 shows a superconductor's heat capacity both in the superconducting state and in the normal state. Keeping the material in the normal state below T_C can be achieved by applying a weak magnetic field. Upon cooling below T_C, the heat capacity shows

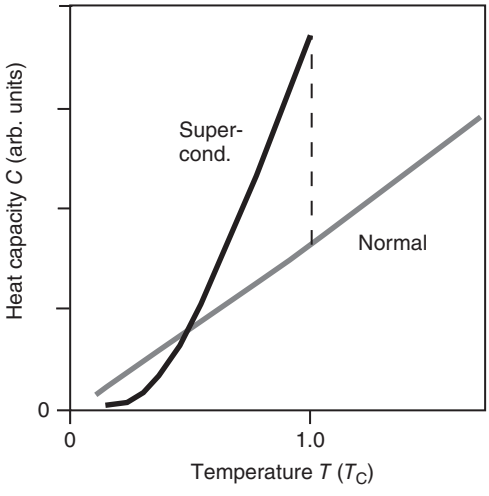


Figure 10.11 Qualitative low-temperature heat capacity of a superconductor in both the superconducting and the normal state. A normal state below T_C can be realized by applying a magnetic field.

a discontinuous change. It also shows a qualitatively different behavior at very low temperatures, where it does not exhibit the characteristic linear temperature dependence of a metal but an exponential decrease. This is another indication of an excitation gap. We have already encountered such exponential behavior in the case of the Einstein model for the phonon heat capacity where it is also caused by an "energy gap" between the ground state and the first excited state of the Einstein oscillators.

10.4 Coherence of the Superconducting State

We have mentioned several times that the superconducting state is associated with a macroscopically coherent wave function, which is built from the Cooper pairs in their common ground state. This macroscopic coherence is, in fact, directly observable in many experiments.

An immediate consequence of the macroscopically coherent wave function is the quantization of magnetic flux through a superconducting ring, as the one shown in Figure 10.12a. If we assume that the wave function of the superconductor is coherent in the entire ring, we can apply the Bohr quantization condition, like for the hydrogen atom but on a macroscopic scale. The Bohr condition states that an integer number of de Broglie wavelengths $\lambda = h/p$ has to fit in the circumference of the ring in Figure 10.12a or that

$$\oint \frac{\mathbf{p}}{h} d\mathbf{r} = n, \tag{10.14}$$

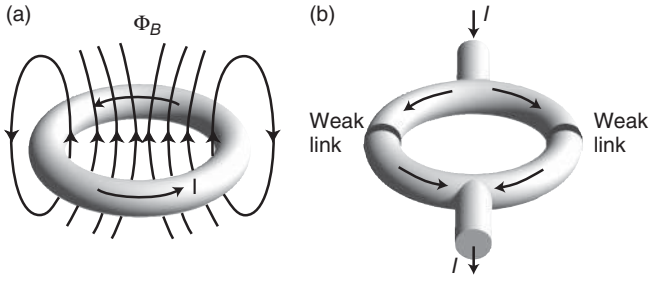


Figure 10.12 (a) A superconducting ring enclosing a magnetic flux. The magnetic flux through such a ring is quantized in multiples of $h/2e$. (b) A superconducting quantum interference device.

where the integration is carried out around the inner circumference of the ring and n is an integer. If we want to see what happens in the presence of a magnetic field, we can include this field in the equation using the vector potential and the same rules as laid out in Section 8.2 for the quantum mechanical case. Following (8.8), we write

$$\oint \mathbf{p} - q\mathbf{A}d\mathbf{r} = nh. \tag{10.15}$$

For particles of mass m_s, density n_s and charge q, we have that $\mathbf{p} = m_s\mathbf{v}$; and the current density associated with these particles is $\mathbf{j} = n_s q\mathbf{v}$ such that we can write

$$\frac{m_s}{n_s q}\oint \mathbf{j}d\mathbf{r} - q\oint \mathbf{A}d\mathbf{r} = nh. \tag{10.16}$$

We now re-write the second integral using Stoke's integral theorem (A.11), in order to change the line integral into a surface integral over the area in the middle of the ring.

$$\oint \mathbf{A}d\mathbf{r} = \int \mathrm{curl}\mathbf{A}d\mathbf{a} = \int \mathbf{B}d\mathbf{a} = \Phi_B, \tag{10.17}$$

where Φ_B is just the magnetic flux through the hole in the ring. With this, (10.16) becomes

$$\frac{m_s}{n_s q^2}\oint \mathbf{j}d\mathbf{r} - \Phi_B = n\frac{h}{q}. \tag{10.18}$$

This implies that the magnetic flux through the ring can only change in units of h/q if the current density in the first integral is constant. We can go one step further and lay the integration path of the first integral a little deeper inside the ring, just outside the screening currents that penetrate up to the London penetration depth λ_L. In many cases, the current density is then vanishingly small and we get

$$\Phi_B = n\frac{h}{q}. \tag{10.19}$$

This means that the magnetic flux through a superconducting ring is quantized in units of h/q. This was indeed shown experimentally in 1961. It was also found

that $q = -2e$, beautifully confirming the existence of Cooper pairs predicted by the BCS theory. The so-called **flux quantum** $h/2e$ is very small, only 2.067×10^{-15} T m^2. To set this into context, consider the Earth's magnetic field that is on the order of 10^{-5} T, implying that an area of 1 mm^2 contains $\approx 10^4$ flux quanta.

The coherence of the superconducting state can be exploited in so-called **superconducting quantum interference devices** (SQUIDs), shown in Figure 10.12b. We cannot describe in detail how these devices work but the key idea is as follows. The device consists of two superconducting "forks" and thin oxide layers or point contacts between them (the so-called weak links). The Cooper pairs from one superconductor can tunnel through the weak links into the other superconductor and there is a definite relation between the phase of the superconducting state in one fork and that in the other fork. A magnetic field entering perpendicular through the hole in the middle gives rise to an additional supercurrent to keep the flux through the ring an integer multiple of the flux quantum. This gives rise to a phase difference for the current going through one weak link and the current going through the other weak link. The maximum current through the entire device is given by the interference of the "left" and "right" currents and one observes oscillations as a function of the applied magnetic field. These correspond to single flux quanta entering the ring. In this way, very small magnetic field changes can be measured.

10.5 Type I and Type II Superconductors

We have so far assumed that a magnetic field cannot enter a material in its superconducting state. It would be completely expelled by supercurrents near the surface of the specimen. This is actually not always true. There is a class of superconductors for which a magnetic field can enter the bulk sample while the material still remains in its superconducting state. These superconductors are called type II superconductors, as opposed to the type I superconductors in which the field cannot enter.

Figure 10.13 shows the behavior of a type I and a type II superconductor in an external magnetic field at zero temperature. For the type I material, the superconductivity breaks down above a certain critical field B_C. For lower fields, the material is superconducting and its magnetization exactly compensates the external field such that the inner of the material is field-free. For higher fields, the magnetization vanishes and the external field completely penetrates the sample. For a type II superconductor, the situation is different. There are two critical fields B_{C1} and B_{C2}. Below B_{C1}, the material behaves exactly as a type I superconductor and above B_{C2} the superconductivity is destroyed. Between these two fields, however, the magnetic field partly enters the material. There is a finite magnetization but it is not large enough to compensate the external field. The remarkable fact is that the resistivity is still zero between these two critical fields.

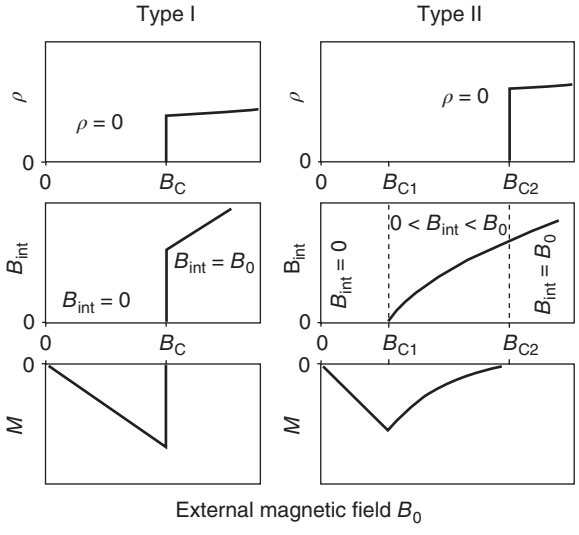


Figure 10.13 Resistivity ρ, internal magnetic field B_{int}, and magnetization M as a function of an external magnetic field for type I and type II superconductors. The temperature is assumed to be 0 K, such that the materials are in their superconducting state when no magnetic field is applied. The type I material has only one critical field B_C, whereas the type II material has two different critical fields B_{C1} and B_{C2}.

If we compare the critical field B_C for a type I superconductor to the two fields for a type II material, it is typically much closer to B_{C1} than to B_{C2}, meaning that type II superconductors can tolerate a higher field than type I materials and still remain superconducting. It is this feature that makes type II superconductors very important for technical applications. One example is the coils for superconducting electromagnets. In these, type II materials such as NbTi are used, which have critical fields above 10 T.

How can the superconductivity survive a magnetic field entering the type II material? The answer is shown in Figure 10.14. The field penetrates through very thin filaments of normal-state material while the rest of the specimen remains superconducting. The filaments of normal-state material are enclosed by vortices of supercurrent such that the rest of the material remains field-free and in its superconducting state. It turns out that the filaments contain only one magnetic flux quantum, so for higher external fields more vortices must enter the sample. Eventually, the density of vortices becomes so large that the superconducting regions of the sample disappear.

The existence of vortices in the superconductor represents a problem to the superconductivity itself. As a current passes through the material, a force similar to the Lorentz force is exerted on the vortices causing them to move perpendicular to the current and to the magnetic field. This movement leads to an energy dissipation and hence to a finite resistance even in the superconducting state, resulting in the destruction of superconductivity. However, there is one possibility to avoid

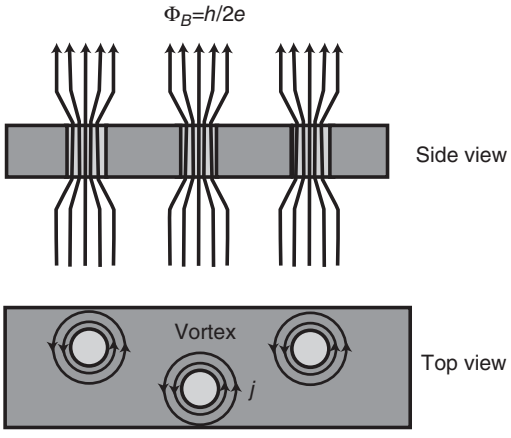


Figure 10.14 Magnetic flux in a type II superconductor. The field penetrates through thin filaments of material in the normal state (light gray) while the rest of the sample remains superconducting (dark gray). The filaments are surrounded by vortices of supercurrent that keep the rest of the sample field-free.

this phenomenon and this is to "pin" the vortices by a sufficient number of defects in the crystal. Then a certain energy will be associated with "unpinning" the vortices and as long as the current density is small enough, this energy will not be available. Type II superconductors in technical applications are therefore far from being perfect single crystals. Quite the opposite is desirable. The materials should have a sufficient number of defects such as grain boundaries in order to pin vortices efficiently.

The difference between a material being type I or type II depends on the ratio between the two characteristic length scales, the London penetration depth λ_L and the coherence length ξ in the Ginzburg–Landau theory. We have seen that ξ sets the length scale on which the density of Cooper pairs can change appreciably. For our simple treatment here, it is appropriate to think of the coherence length as the distance between the two electrons in a Cooper pair, which we estimated to be on the order of 100 nm or so. When ξ is smaller than λ_L, the formation of normal-state filaments surrounded by supercurrents becomes favorable and the material tends to be a type II superconductor. If ξ is longer than λ_L, the formation of the filaments is not favorable and the material is a type I superconductor. Disorder tends to reduce the coherence length ξ and, therefore, many intermetallic alloys are type II superconductors.

10.6 High-Temperature Superconductivity

What would we have to do to design a superconductor with a very high T_C, preferably above room temperature? We have already discussed the BCS equation

for the transition temperature (10.13) and seen that the scale for T_C is set by the Debye temperature of the solid. Debye temperatures for materials made of light elements with strong bonds can be very high (several hundred Kelvin), so that this is not the limiting factor. The limitation is in the exponential term in (10.13) where one should try to obtain both a high density of states at the Fermi energy and a strong electron–phonon coupling. Estimates of these quantities show that one cannot hope to get a T_C much higher than 30 K or so and, indeed, the progress in designing materials with increasing T_C was very slow initially (Figure 10.15).

A disruptive change of this development came in 1986 when J. G. Bednorz and K. A. Müller discovered superconductivity in the cuprate-perovskite ceramic materials. T_C for the particular material found by Bednorz and Müller was not especially high, only 30 K, but higher than for any compound found before. Moreover, superconductivity had been discovered in a new and unexpected class of materials. A short time later, superconductors of a similar kind but with a slightly different composition were found, which had a T_C above the boiling point of nitrogen. This opened (in principle) the possibility for new technological applications because cooling with liquid nitrogen is considerably cheaper than with liquid helium. It is in this sense one has to interpret the term "high-temperature superconductivity."

The cuprate-perovskite materials are actually insulators that can be doped to become (poor) conductors and, at lower temperatures, superconductors. They have a fairly complicated structure but for us the only important point is that they consist of two-dimensional CuO_2 layers that are separated by other building blocks.

The mechanism of high-temperature superconductivity is not understood yet, but there is agreement on a number of points. Most importantly, the supercurrent

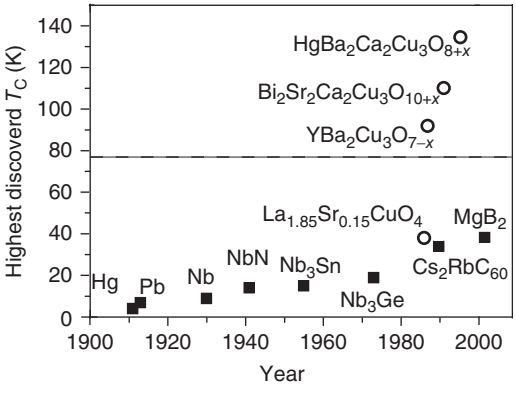


Figure 10.15 Increase of the highest critical temperature T_C of known superconductors as a function of time. The dashed horizontal line is the boiling temperature of liquid nitrogen. "Conventional" BCS-type superconductors are denoted by squares and novel "high-temperature" superconductors by circles. The x denotes a nonstoichiometric composition.

is also carried by Cooper pairs. It is unclear, however, what mechanism binds the electrons together. Some theories suggest that the electron–phonon interaction can play a role, much like in BCS superconductors, but most favor other mechanisms, such as certain magnetic excitations. The behavior of the high-temperature superconductors in the normal state is not understood either. It is very different from that of a normal metal.

The Cooper pairs in high-temperature superconductors are mainly localized in the CuO_2 planes of the perovskite structure, meaning that the material has very anisotropic properties. In fact, one can view the high-temperature superconductivity as a process that practically happens in two dimensions. Unfortunately, the anisotropy also brings about very low critical current densities for polycrystalline samples, in which the planes in one domain do not match with the planes in the neighboring domain. Nevertheless, many practical hurdles have been overcome and high-temperature superconductors have started to play a role in commercial applications.

10.7 Concluding Remarks

As already mentioned in the introduction of this chapter, we were merely able to introduce some of the very basic ideas about superconductivity and we could not even explain those in any detail. Superconductivity is an extremely rich subject, and there are many additional phenomena that were not included here. You can also find exceptions to almost every "rule" we have discussed. There are, for example, many superconductors that are well described by the BCS theory but fail to show the $M^{-1/2}$ behavior in the isotope effect. Some even show no dependence of T_C on the isotope combination or one that goes in the "wrong" direction. The high-temperature superconductors have a superconducting gap but it is of a very different nature than for the "traditional" BCS superconductors. There is, in fact, no absolute gap in the sense of Figure 10.8, that is, no energy interval without any single-particle states around the Fermi energy. The gap in these materials appears only in certain regions of the first Brillouin zone.

This list goes on but we conclude the chapter with a more general question: What is the connection between superconductivity and magnetic ordering? At first it seems that the answer is that there is none – they are mutually exclusive, as we have seen several times in this chapter. But it is really more interesting than that. Magnetic ordering and superconductivity have in common that the electrons with the highest energy, those at the Fermi energy, have a possibility to slightly lower their energy, either by establishing magnetic order or by condensing into Cooper pairs. The energy lowering appears small relative to their absolute kinetic energy but it is highly relevant compared to the average thermal energy. In some materials, there is even a close competition between magnetism and superconductivity. Often antiferromagnetic ordering occurs in the same material as

superconductivity or the material becomes superconductive at some low temperature and normal, but ferromagnetically ordered, at an even lower temperature.

There are other mechanisms apart from magnetic ordering and superconductivity that give the possibility of an energy lowering for the electrons near the Fermi energy. Particularly complex and interesting situations arise when several of these mechanisms compete with each other and this complexity is a very active research area in current solid state physics.

References

Giaever, I. and Megerle, K. (1961) *Phys. Rev.*, **122**, 1101.
Maxwell, E. (1952) *Phys. Rev.*, **86**, 235.
Serin, B., Reynolds, C.A., and Lohman, A. (1952) *Phys. Rev.*, **86**, 162.

10.8
Further Reading

Superconductivity is discussed in most more recent treatments of solid state physics. Consider in particular

- Ibach, H. and Lüth, H. (2009) *Solid State Physics*, 4th edn, Springer.

A specialized in-depth text is

- Buckel, W. and Kleiner, R. (2004) *Superconductivity*, 2nd edn, Wiley-VCH Verlag GmbH.

10.9
Discussion and Problems

Discussion

1) How can you be sure that a superconductor has zero resistance? How can you measure its resistance?
2) What can be the cause for the superconductivity of a material breaking down, even if the temperature is below the critical temperature?
3) Describe the Meissner effect. Why is the Meissner effect "more" than just a consequence of zero resistance?
4) Is it really such that magnetic fields do not penetrate at all into a superconductor?
5) Describe the key ideas behind the microscopic BCS theory for superconductivity.
6) What experimental evidence supports the BCS theory?
7) What is the difference between a type I and a type II superconductor?

Problems

1) *Zero resistivity:* We have seen that the decay of the superconducting current can be measured by inducing a current in a superconducting ring and measuring how the current decays. (a) How would you induce a current in such a ring? (b) How would you measure if it decays or not? (c) Show that the decay (if any) would be exponential in time. (d) Assume that the ring has a diameter of 0.5 cm and is made of a very thin wire. If you observe that the current decays by less than 1% in a year, estimate the maximum resistance of the wire. Hint: For this last part, you need to know the strength of the B field through the ring. For simplicity, assume that the field throughout the ring has the same strength as in the center.
2) *(*) Isotope effect:* Calculate the absolute temperature range of T_C for the T_C values in Figure 10.5. Experimentally, it can be difficult to scan the temperature through the superconducting transition for all the samples. Can you suggest an alternative experimental approach that avoids this problem?
3) *London penetration depth:* In connection with Figure 10.6 and Equation (10.9), it was argued that the **B** field decays exponentially inside the superconductor. Show this explicitly for a situation as in Figure 10.6, where the surface is in the $x-y$ plane and the field has only an x component, that is, $\mathbf{B} = (B_x, 0, 0)$.
4) *BCS theory:* The BCS theory predicts an exponential temperature dependence of the heat capacity at very low temperatures, instead of a linear heat capacity usually observed for metals. (a) Explain qualitatively why this is so. (b)(*) Explain qualitatively why the exponential behavior is only observed at very low temperatures, that is, much lower than T_C.
5) *BCS theory:* We have seen that both the typical dimension of the Cooper pairs and the London penetration depth are on the order of 100 nm and we have assumed that the Ginzburg–Landau coherence length ξ has the same order of magnitude. If ξ is the length scale on which the density of Cooper pairs can change appreciably, how likely is it that the superconducting state breaks down in a volume ξ^3? (a) Give a rough estimate of the total energy gain for the condensation of Cooper pairs in a volume of $(100\ \text{nm})^3$. (b) Can you say something about the probability of the superconductivity breaking down spontaneously due to a thermal fluctuation? Assume that the rate R of such breakdowns per second follows a Boltzmann distribution

$$R = fe^{-E/k_B T}. \tag{10.20}$$

Argue that this rate is extremely small, even though you do not know the actual value of the "attempt frequency" f.
6) *(*) BCS theory:* It is experimentally found that the thermal conductivity of a superconductor well below T_C is smaller than that for the same material in the normal state. Can you explain this?

7) *(*) BCS theory:* Consider the typical temperature dependence of the resistivity $\rho(T)$ for a superconductor and a nonsuperconducting material. In the case of Figure 10.1, it appears that the resistivity of the superconductor above T_C increases more strongly than for the nonsuperconducting material. Can you give a plausible explanation?
8) *BCS theory:* How do you expect the London penetration depth λ_L to depend on the temperature (qualitatively)?

11
Finite Solids and Nanostructures

We have so far assumed that we deal with infinite solids or, more precisely, with big but finite solids without any boundaries. This was achieved by using periodic boundary conditions, and it has been a very successful concept. The perfect periodicity of the lattice has allowed us to solve many problems that we could not have solved otherwise. One example is the vibrational motion of the atoms: Instead of solving the equations of motion for a practically infinite number of atoms in a linear chain of macroscopic dimensions, we have reduced the problem to a small set of equations. Another example is the electronic structure, where the periodicity of the problem and the introduction of the reciprocal lattice permitted a solution of the Schrödinger equation.

In this chapter, we briefly treat finite solids with a specific emphasis on the surfaces of bulk crystals as well as on nanostructured materials such as clusters of a few atoms, ultrathin films, and so on. The current interest in nanotechnology is (among other factors) driven by the fact that many materials drastically change their properties on a small scale (electronic, optical, catalytic, etc.). Here, we describe why this is so but we confine our attention to the very basic physical ideas.

We focus on three aspects of the small scale: The first is that the number of possible vibrations or electronic states is drastically reduced. As a consequence, the continuum of allowed energy states in a macroscopic solid is turned into a few discrete levels. This effect is called **quantum confinement**. The second aspect is that small objects, like clusters, necessarily have a relatively large surface (or interface) area, that is, a much larger ratio of surface area to bulk volume than macroscopic objects. The atoms at the surface (or interface) find themselves in a different environment from that of the bulk atoms and can therefore have other vibrational and electronic properties. If the structure is sufficiently small, these properties will dominate the entire nano-object. In fact, the surfaces are also important for large crystals because the surface is where the solid interacts with its environment, for example, in chemical reactions. Finally, we will see how ferromagnetic ordering in a particle is affected by its size. For very small particles, the spins are still aligned with respect to each other below the Curie temperature, but the magnetization of the entire particle can be rotated by thermal fluctuations at much lower temperatures.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

Another interesting aspect that we will not discuss any further is that of reduced dimensionality. In three-dimensional crystals, we succeeded relatively well to describe the electronic states by assuming no interaction between the electrons. As the number of dimensions is reduced, to a two-dimensional sheet or a one-dimensional wire, this assumption becomes increasingly problematic and new phenomena arise because of the correlated motion of the electrons. In the case of a strictly one-dimensional system, this is intuitively clear: Neglecting the electron–electron interaction will be a bad approximation because the electrons cannot even get past each other.

11.1 Quantum Confinement

In order to see how reducing the size of a system can change its properties, we just have to look back at Figure 4.6 that shows the allowed vibrational frequencies for a finite chain of 10 atoms. The possible vibrational frequencies are denoted by black dots, and they are clearly a very small subset of the possible vibrations for an infinite chain. Formally, this is caused by the requirement of having k-values consistent with the boundary conditions. In the case of Figure 4.6, we had taken these to be (4.20). If the system is sufficiently small, the number of allowed states is thus drastically reduced; instead of having a quasicontinuum of states, we have discrete levels that are separated by a significant energy. This is the essence of quantum confinement.

A particularly good illustration of quantum confinement can be realized by a very thin metal film that is grown on a semiconductor or an insulator (see Figure 11.1a). The wave function $\Psi(\mathbf{r})$ of the electrons in the film can be separated in parts parallel (along x, y) and perpendicular (along z) to the film

$$\Psi(\mathbf{r}) = \Psi(z)\Psi(x, y). \tag{11.1}$$

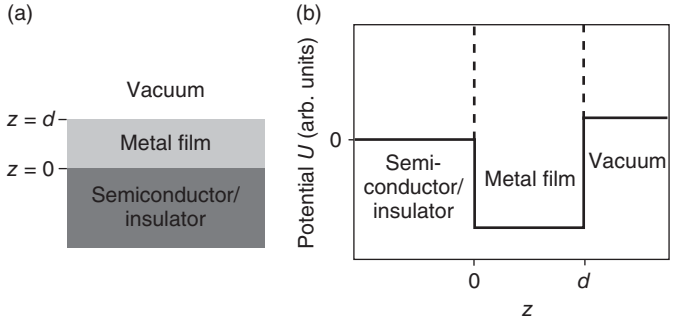


Figure 11.1 (a) A thin metal film on a semiconducting or insulating substrate. (b) Modeling the metal film as a potential well. The dashed lines indicate the simplification of an infinitely deep potential well.

For the motion parallel to the film, we assume that the electrons behave like free electrons. $\Psi(x, y)$ is then a two-dimensional free electron wave function and the energies associated with the motion parallel to the film are

$$E_{xy} = \frac{\hbar^2 \mathbf{k}_{xy}^2}{2m_e}, \tag{11.2}$$

where $\mathbf{k}_{xy}$ is the two-dimensional wave vector for the free electron motion parallel to the film. The possible values of $\mathbf{k}_{xy}$ are given by the usual periodic boundary conditions.

More interesting are the wave function and the energy levels for the motion perpendicular to the film, that is, in the z direction. An appropriate model for this is a simple **potential well** between the metal/semiconductor interface ($z = 0$) and the metal/vacuum interface ($z = d$). Such a potential is shown in Figure 11.1b. The potential inside the metal film can be taken to be zero. The potential well has a finite depth and its height on the two sides is, in general, different.

The solutions for an infinitely deep potential well are known from elementary quantum mechanics. The solution inside the well can be written as the superposition of two free electron waves, one moving to the right and the other moving to the left:

$$\Psi(z) = Ae^{ik_z z} + Be^{-ik_z z}, \tag{11.3}$$

where k_z is the component of the wave vector in the z direction and A and B are complex amplitudes. The boundary conditions are that this wave function has to vanish at $z = 0$ and at $z = d$. This immediately gives the quantization condition for k:

$$k_z d = n\pi, \qquad n = 1, 2, 3, \ldots \tag{11.4}$$

If we now realize that $k_z = 2\pi/\lambda$, then the quantization condition is $2d = n\lambda$. This has the familiar meaning that one round trip through the potential well must correspond to an integer number of wavelengths. The possible energies for the states in the z direction are

$$E_z = \frac{\hbar^2 k_z^2}{2m_e}, \tag{11.5}$$

and the total energy for a given k_z and $\mathbf{k}_{xy}$ is the sum of (11.2) and (11.5).

As mentioned above, the infinitely deep potential well is not a very appropriate model and one should really use the potential in Figure 11.1b. However, the resulting quantization condition is similar to (11.4). It turns out to be

$$2k_z d + \Phi_i + \Phi_v = 2\pi n, \qquad n = 1, 2, 3, \ldots \tag{11.6}$$

The meaning of this equation is that the total phase change for one round trip of the electron must be a multiple of 2π. This total phase change must include the phase shifts at the metal/semiconductor interface Φ_i and at the metal/vacuum interface Φ_v. This model describes the resulting k_z values for real thin metal films relatively well.

In all this, we have only considered the Schrödinger equation for free electrons for which we know the solution already. The only thing we have changed is the boundary conditions for the solution (in the z direction). The new boundary conditions strongly restrict the possible states and give rise to the discrete energy levels in the z direction.

Similar ideas can be used to describe the properties of semiconductor clusters with a radius of less than ≈100 nm, the so-called **quantum dots**. In these clusters, the possible electron and hole energies are quantized because of confinement in three dimensions. In a bulk semiconductor, the smallest energy for forming an electron–hole pair is the gap energy E_g. In quantum dots, this energy is larger and it increases as the dot's size decreases. An approximate formula for the minimum energy to separate an electron from a hole is

$$E_{\min} = E_g + \frac{\hbar^2\pi^2}{2\mu r^2} - \frac{1.8e^2}{4\pi\epsilon_0\epsilon r} \tag{11.7}$$

where r is the cluster radius and μ is the reduced mass from the electron and hole effective masses. The second term on the right-hand side is due to the confinement effect: It increases the energy when the electron and hole wave functions are "compressed" in a smaller quantum dot. Remember that in a simple potential well of size r, the possible energies are also proportional to r^{-2} (see (11.4) and (11.5)). The third term is due to the Coulomb interaction between the electron and the hole. It works in the other direction because the attraction is stronger for a smaller cluster. As the cluster size is decreased, the r^{-2} term eventually wins and becomes the most important term for really small clusters. The change of the smallest excitation energy as a function of cluster size is an extremely useful property because it allows the production of clusters with optical absorption exactly at the desired wavelength.

The small size of the clusters has also another effect on their optical properties. If the cluster is small enough, the electrons are localized and the uncertainty principle dictates that they must have a considerable momentum uncertainty, which (for a free particle) is equivalent to a k uncertainty. In other words, k is not properly defined anymore, consistent with the interpretation of k as the quantum number of infinite translational symmetry. But if k is not well defined, this means that the forbidden optical transitions across the indirect band gap in a material such as Si are no longer strictly forbidden and one can start to exploit the optical properties of Si and other materials with an indirect band gap.

11.2 Surfaces and Interfaces

Apart from the quantum confinement of bulk states, the existence of surfaces and interfaces can lead to entirely new electronic and vibrational states that are located at the surface (or the interface). This can give rise to surfaces with properties that are quite different from those of the bulk material. It is, for example, possible to

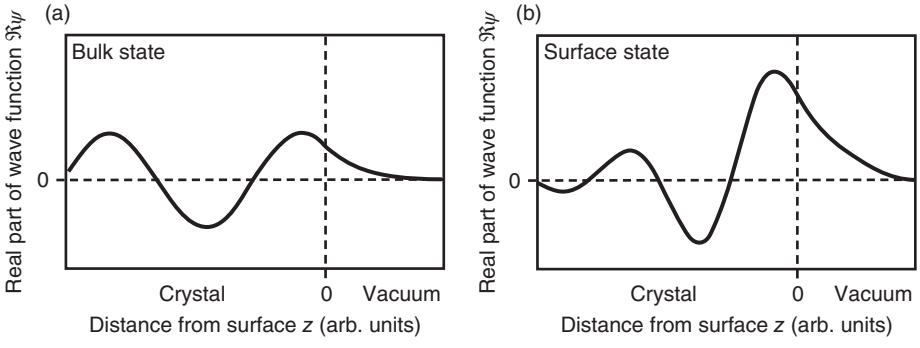


Figure 11.2 (a) Matching of a bulk electronic state (a Bloch wave) to an exponential decay outside the surface. (b) An electronic surface state that is decaying as the distance from the surface is increased, both outside and inside the solid.

have a metallic surface on an insulating bulk. We illustrate this for the electronic states in Figure 11.2. Consider first Figure 11.2a that shows the behavior of a bulk state near a surface. The bulk state is a Bloch wave

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}). \tag{11.8}$$

At the surface, this state and its derivative have to be matched continuously to an exponentially decaying wave function in the vacuum.

The wave vector $\mathbf{k}$ of the Bloch wave is a purely real quantity. Surface-localized states can be described by allowing for a complex $\mathbf{k}$. If we assume that the z component of $\mathbf{k}$ was complex, we could take the imaginary part $\Im(k_z)$ of it out of the plane wave part of the Bloch wave, leading to

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{\Im(k_z)z} e^{i\mathbf{k}'\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}), \tag{11.9}$$

where $\mathbf{k}'$ is the remaining real part of the wave vector, that is, $(k_x, k_y, \Re(k_z))$. For a bulk state, this wave function clearly has a problem because it exponentially increases in the $+z$ direction. It can therefore not be normalized and is not physically meaningful.

Near the surface, however, solutions with a complex wave vector are possible and an example is given in Figure 11.2b. The solution is exponentially increasing in the z direction but, as the crystal is terminated by the surface, it can be matched to an exponentially decreasing wave function outside the surface. Therefore, it can still be normalized and it forms a **surface-localized electronic state**. Similar arguments can be made for the interface between two solids where localized interface states can exist.

While this argument for the possible existence of surface states is based on very general principles, we also discuss an example related to the different bonding of atoms at the surface. Suppose that we have a typical semiconductor with its sp^3-type bonding in the bulk. In the bulk, each sp^3 orbital is filled with two electrons. At the surface, some of these sp^3 bonds will be broken and the broken bonds will be filled with only one electron because of the missing neighbor atom. If the

surface structure is now such that we have an odd number of broken bonds per two-dimensional surface unit cell, then this corresponds to an odd number of surface electrons in such a unit cell. We have seen in Section 6.6 that this leads to a half-filled band and thereby to a metallic surface. This is quite remarkable because we would have a metallic surface on a semiconducting bulk! The half-filled band would exist only at the surface and it would be described by wave functions such as (11.9). However, it is often found that the situation of metallic surface states is not very stable. They could be destroyed by filling a second electron into the broken sp^3 bond. This could happen via a chemical reaction of the broken bond with its environment (with hydrogen atoms, for instance) or even by pairing two bonds from different unit cells. In fact, the nonmetallic situation is often energetically more favorable and, therefore, metallic surfaces on semiconductors are rarely encountered.

An exception to this rule of thumb are the surfaces of the recently discovered so-called **topological insulators**. These materials have an insulating inside but their surfaces are metallic, independent of the surface structure and orientation, even when allowing for chemical reactions with the environment. The same is true for their interfaces to any normal (or the so-called "topologically trivial") insulator. This remarkable behavior can be derived from the special properties of the topological insulator's bulk band structure, and it is easiest to understand if we consider an interface between the topological insulator and a normal insulator. It is sufficient to use the simplest possible representation of the valence band (VB) and conduction band (CB) as single energy levels, as introduced in Section 7.1.1. The situations for a normal insulator and a topological insulator are shown in Figure 11.3a and b, respectively. The difference between the two is here the color of the bands: The normal insulator has a bright VB and a dark CB and for the topological insulator, this is the other way round. The color of the bands represents a symmetry property of the wave functions. For the topological insulator, the color ordering is inverted with respect to the normal insulator and one often speaks of an inverted band gap. The appearance of a metallic interface is now illustrated in Figure 11.3c. When an interface between the two materials is formed, only bands of the same symmetry can be joined. As a consequence, the VB (CB) of the normal insulator is connected with the CB (VB) of the topological insulator and the states have to cross the chemical potential at or near the interface. Thus, the interface becomes metallic.

The band crossing and the metallic interface do therefore result from the inner material's properties, not from the surface, and are stable against all changes at the surface. This is called a **topological** protection. An excellent analogy is the situation that arises when two countries are to be joined by a road bridge, with the difficulty that the driving rules in one country enforce right-hand traffic and in the other left-hand traffic. A possible solution to this problem is the traffic flipper bridge shown in Figure 11.3c, a proposal for a road connection between Hong Kong and Mainland China. Clearly, the detailed design of the bridge could be changed but the lane crossing cannot be avoided unless the traffic rules in one of the countries are changed.

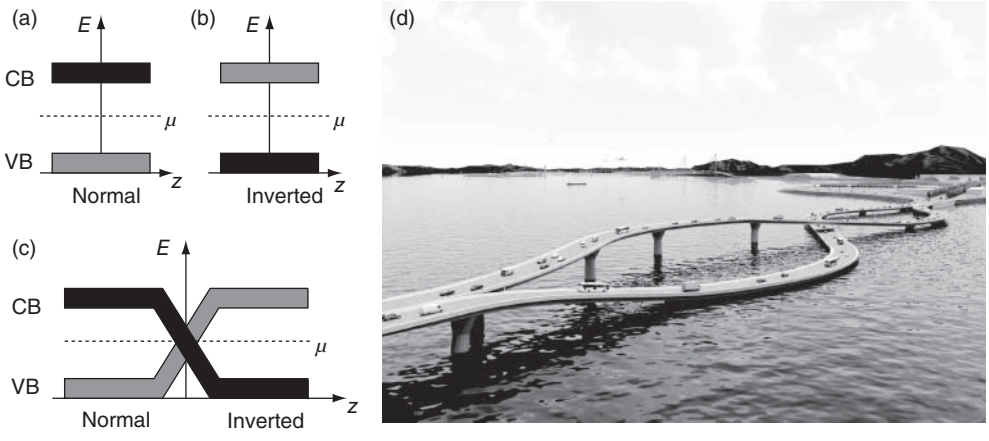


Figure 11.3 Illustration of topologically protected metallic states between two insulators. (a) Simplified band diagram for a "normal" insulator with single levels for the VB and CB energies as a function of position (i.e., the edges signify the macroscopic ends of the sample). The grayscale represents the symmetry of the bands. (b) Topological insulator with an inverted band gap. (c) Joining an insulator with a normal band gap and one with an inverted band gap while maintaining the symmetry of the states gives rise to metallic interface states. (d) "Topological" traffic lane crossing at the border of two countries with left- and right-hand traffic. Image courtesy of NL architects (www.nlarchitects.nl).

This simple picture leaves some open questions: What is the meaning of the CB and VB's color? The color represents the parity of the wave functions, describing how they transform under an inversion of the coordinates. As an example, consider the s-type and p-type wave functions for $k = 0$ in Figure 6.11. If the sites of the atoms are used as the origin of the coordinate system, the s-type wave function is not affected by an inversion of the coordinates (it has a parity of 1) but the p-wave function changes sign (it has a parity of -1). In a topological insulator, the parity ordering of the states is different from a normal insulator.

Also, the explanation may work for an interface between a topological and a normal insulator, but what about surface of a topological insulator, that is, the interface to vacuum or air? This is indeed not obvious from the simple picture given here but it is in fact possible to derive the existence of metallic surface states from the band structure of the topological insulator alone, without the need of an interface to a second material.

11.3 Magnetism on the Nanoscale

If a solid is sufficiently small, we have to rethink our approach to its magnetization and this has enormous technological consequences. For a large magnet, the energy required to flip one magnetic moment is on the order of the exchange energy, or the energy corresponding to the Curie temperature $k_B\Theta_C$. This can be much

higher than $k_B T$ at room temperature. Hence, large permanent magnets do not lose their magnetization due to some thermal fluctuations. For a small ferromagnetic particle this is still true, but now another possibility arises that can change the magnetization of the particle even if $T \ll \Theta_C$. If the particle is smaller than the typical thickness of a Bloch wall, it will only contain one magnetic domain. Instead of changing the magnetization by flipping all the moments one by one, they can then stay aligned with each other and the magnetization of the entire domain can be rotated *at the same time*. If the particle is sufficiently small, the energy cost for doing this can become small enough to change the magnetization of the entire particle at room temperature, even for ferromagnetic materials with a strong exchange interaction and a high Θ_C. This phenomenon is called **superparamagnetism**.

We give an estimate of the energy cost for rotating the entire magnetic moment of a small particle. A key factor in this is that the magnetization of such a small particle will have a preferred direction, the so-called **easy axis**. Energetically, it does not matter if **M** is parallel or antiparallel to the easy axis, but it does cost energy to rotate it out of the easy axis direction. Therefore, an energy barrier has to be overcome to reverse **M**. The height of the energy barrier depends on several factors, such as the shape of the small particle and its crystal structure. We cannot go into these effects here, but merely present a very crude estimate of the order of magnitude. Suppose that the magnetization of the particle is **M** and we assume the barrier ΔE is given by the energy it takes to move **M** out of the **B** field created by **M** (see Section 8.1). In this case, we obtain

$$\Delta E \approx V \mu_0 M^2. \tag{11.10}$$

The energy required for rotating the magnetization thus depends linearly on the volume of the particle. For iron particles with a size below a few nanometers, this energy becomes comparable to $k_B T$ at room temperature (see Problem 11.3), i.e., the particles become superparamagnetic. This is obviously a problem for the fabrication of magnetic storage devices. The magnetic particles used for storing one bit must be large enough to prevent thermally induced rotations of the magnetization. On the other hand, superparamagnetic nanoparticles do have useful applications because of their large magnetic moments, for example, in medical magnetic resonance imaging.

11.4 Further Reading

For further information on quantum confinement in nanoparticles, see

- Delerue, C. and Lannoo, M. (2004) *Nanostructures - Theory and Modelling*, Springer.
- Klimov, V. (2010) *Nanocrystal Quantum Dots*, CRC Press.

Surface states and topological insulators are discussed in

- Hofmann, Ph. (2013) *Surface Physics*, Self-Published, *www.philiphofmann.net/Philip_Hofmann/SurfacePhysics.html*.
- Moore, J. (2010) The birth of topological insulators. *Nature*, **464**, 194.

For magnetism on the nanoscale, see

- Himpsel, F.J., Ortega, J.E., Mankey, G.J., and Willis, R.F. (1998) Magnetic nanostructures. *Adv. Phys.*, **47**, 511.
- Stöhr, J. and Siegmann, H.C. (2006) *Magnetism: From Fundamentals to Nanoscale Dynamics*, Springer.

11.5 Discussion and Problems

Discussion

1) What are the main physical reasons why the properties of nanoscale solids are different from macroscopic solids of the same material?
2) How are the optical properties of semiconductors changed in nanoscale clusters?
3) Explain why it is possible that the magnetization of ferromagnetic nanoparticles can rapidly fluctuate at room temperature, even though the Curie temperature of the material is much higher.

Problems

1) *Boundary conditions:* In this chapter, we have seen that the different properties of solids on the nanoscale are largely due to new boundary conditions, which lead to fewer and discrete solutions of the Schrödinger equation. This is a bit worrying because in the previous chapters, we have just used whatever boundary conditions seemed most convenient. In particular, we have preferred the periodic boundary conditions over the fixed boundary conditions where the wave function (or the vibrational amplitude) has to vanish at the boundary of a crystal. Fortunately, the precise choice of boundary conditions is not so important if the solid is only big enough. In order to show this, calculate the density of states for the free electron model in Chapter 6 when assuming fixed boundary conditions. Compare the result to (6.13).
2) *Semiconducting nanoclusters:* (a) Calculate the minimum electron–hole separation energy for CdSe nanoclusters as a function of cluster size. Plot the two contributions (confinement and Coulomb interaction) and their sum separately. (b) Such nanoclusters can be used as fluorescent markers. When exposed to ultraviolet light, a separation of electrons and holes take place. Eventually, these recombine under the emission of light corresponding to

the minimum separation energy. How large would the clusters have to be to emit yellow light?

3) *Superparamagnetism:* (a) How small would an iron particle need to be such that the energy for rotating its entire magnetization becomes comparable to $k_B T$ at room temperature? Assume that each iron atom contributes to the magnetization with a moment of $2.2\mu_B$. (b) Given the energy barrier ΔE, assume that the rate of magnetization rotations R, that is, the number of rotations per second, follows a Boltzmann distribution:

$$R = fe^{-\Delta E/k_B T}. \tag{11.11}$$

The frequency f is of the order 10^9 s^{-1}. How large would the magnetic iron particles in a hard disk have to be if you only want to tolerate one flip every 10 years?

Appendix A

This appendix briefly deals with some basics in electromagnetism. It assumes that you are familiar with the Maxwell equations in vacuum in their integral form, as discussed in most introductory physics textbooks. What is given here are the explicit forms of vector calculus operations, the so-called differential form of the Maxwell equations, and how the Maxwell equations change in the presence of matter.

A.1
Explicit Forms of Vector Operations

In this book, several vector calculus operations are used, which are given here explicitly. They all deal with differentiating a vector field or a scalar field. Let $\boldsymbol{\mathcal{E}}(\mathbf{x})$ be a vector field and $\phi(\mathbf{x})$ a scalar field.

The gradient of the scalar field $\phi(\mathbf{x})$ is a vector with the components:

$$\mathrm{grad}\phi = \nabla\phi = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)\phi = \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right). \tag{A.1}$$

The often-used operator ∇ (called "nabla") is defined as

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right). \tag{A.2}$$

The divergence of a vector field $\boldsymbol{\mathcal{E}}(\mathbf{x})$ is a scalar field:

$$\mathrm{div}\boldsymbol{\mathcal{E}} = \nabla\cdot\boldsymbol{\mathcal{E}} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)\cdot\boldsymbol{\mathcal{E}} = \frac{\partial\mathcal{E}_x}{\partial x} + \frac{\partial\mathcal{E}_y}{\partial y} + \frac{\partial\mathcal{E}_z}{\partial z} \tag{A.3}$$

The curl of a vector field $\boldsymbol{\mathcal{E}}$ is

$$\mathrm{curl}\boldsymbol{\mathcal{E}} = \nabla\times\boldsymbol{\mathcal{E}} = \left(\frac{\partial}{\partial y}\mathcal{E}_z - \frac{\partial}{\partial z}\mathcal{E}_y, \frac{\partial}{\partial z}\mathcal{E}_x - \frac{\partial}{\partial x}\mathcal{E}_z, \frac{\partial}{\partial x}\mathcal{E}_y - \frac{\partial}{\partial y}\mathcal{E}_x\right). \tag{A.4}$$

It is always such that $\mathrm{curl}\,\mathrm{grad}\phi = 0$ and $\mathrm{div}\,\mathrm{curl}\boldsymbol{\mathcal{E}} = 0$. Finally, an important second derivative is the divergence of the gradient of a scalar field $\phi(\mathbf{x})$:

$$\mathrm{div}\,\mathrm{grad}\phi = \nabla^2\phi = \triangle\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}. \tag{A.5}$$

$\triangle$ is called the Laplace operator.

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

A.2 Differential Form of the Maxwell Equations

The familiar, integral form of the Maxwell equations is

- Gauss's law

$$\oint \boldsymbol{\mathcal{E}} d\mathbf{a} = \frac{Q_{\text{encl}}}{\epsilon_0}; \tag{A.6}$$

- Gauss's law for magnetism

$$\oint \mathbf{B} d\mathbf{a} = 0; \tag{A.7}$$

- Ampère's law

$$\oint \mathbf{B} d\mathbf{l} = \mu_0 \left(I_{\text{encl.}} + \epsilon_0 \frac{\partial \phi_E}{\partial t} \right); \tag{A.8}$$

- Faraday's law

$$\oint \boldsymbol{\mathcal{E}} d\mathbf{l} = -\frac{\partial \Phi_B}{\partial t}. \tag{A.9}$$

In order to transform these into the differential form frequently used in this book, one makes use of two fundamental theorems. The first is the divergence theorem

$$\oint \boldsymbol{\mathcal{E}} d\mathbf{a} = \int \text{div} \boldsymbol{\mathcal{E}} dV, \tag{A.10}$$

which turns a surface integral over a vector field into a volume integral over the divergence of the same field. The second is Stokes' integral theorem, which states

$$\oint \boldsymbol{\mathcal{E}} dl = \int \text{curl} \boldsymbol{\mathcal{E}} d\mathbf{a}, \tag{A.11}$$

turning an integral along a closed line into an integral over a surface enclosed by the line.

Using the divergence theorem, we can find a differential version of Gauss's law (A.10). We calculate:

$$\oint \boldsymbol{\mathcal{E}} d\mathbf{a} = \int \text{div} \boldsymbol{\mathcal{E}} dV = \frac{Q_{\text{encl}}}{\epsilon_0} = \int \frac{\rho}{\epsilon_0} dV. \tag{A.12}$$

Since this holds for an arbitrary volume, the differential version of Gauss' law is

$$\text{div} \boldsymbol{\mathcal{E}} = \frac{\rho}{\epsilon_0}. \tag{A.13}$$

Gauss's law of magnetism can be treated in the same way to give

$$\text{div} \mathbf{B} = 0. \tag{A.14}$$

Equation (A.11) can be used in a similar way to rewrite the two other equations. We start with Ampère's law:

$$\oint \mathbf{B} d\mathbf{l} = \int \text{curl} \mathbf{B} d\mathbf{a} = \mu_0 \left(I_{\text{encl.}} + \epsilon_0 \frac{\partial \Phi_E}{\partial t} \right) = \int \mu_0 \mathbf{j} + \mu_0 \epsilon_0 \frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} d\mathbf{a}, \tag{A.15}$$

which means that the differential version of Ampère's law is

$$\mathrm{curl}\mathbf{B} = \mu_0\mathbf{j} + \mu_0\epsilon_0\frac{\partial\boldsymbol{\mathcal{E}}}{\partial t}. \tag{A.16}$$

Finally, we find Faraday's law in the differential form as

$$\mathrm{curl}\boldsymbol{\mathcal{E}} = -\frac{\partial\mathbf{B}}{\partial t}. \tag{A.17}$$

A.3 Maxwell Equations in Matter

In vacuum, the Maxwell equations contain two fields: the $\boldsymbol{\mathcal{E}}$ field that is created by all charges and the $\mathbf{B}$ field that is created by all currents. In matter, it is convenient to modify the equations by the introduction of the new fields $\mathbf{D}$ and $\mathbf{H}$. Both appear in a modified version of Ampère's law. In vacuum, we have the current density $\mathbf{j}$ of free charges in (A.16) but in matter there is the possibility to have additional "bound" currents. The first, $\mathbf{j}_\mathrm{m}$ gives rise to a macroscopic magnetization $\mathbf{M}$ via $\mathbf{j}_\mathrm{m} = \mathrm{curl}\mathbf{M}$, and the second is related to a change of the dielectric polarization $\mathbf{P}$ via $\mathbf{j}_\mathrm{e} = \partial\mathbf{P}/\partial t$. If we add these to the "free" current density $\mathbf{j}$, this leads to

$$\mathbf{j} + \mathbf{j}_\mathrm{m} + \mathbf{j}_\mathrm{e} = \frac{1}{\mu_0}\mathrm{curl}\mathbf{B} - \epsilon_0\frac{\partial\boldsymbol{\mathcal{E}}}{\partial t}, \tag{A.18}$$

or

$$\mathbf{j} = \mathrm{curl}\left(\frac{1}{\mu_0}\mathbf{B} - \mathbf{M}\right) - \frac{\partial}{\partial t}\left(\epsilon_0\boldsymbol{\mathcal{E}} + \mathbf{P}\right). \tag{A.19}$$

This is then simplified by the introduction of the new fields

$$\mathbf{H} = \frac{1}{\mu_0}\mathbf{B} - \mathbf{M} \tag{A.20}$$

and

$$\mathbf{D} = \epsilon_0\boldsymbol{\mathcal{E}} + \mathbf{P}, \tag{A.21}$$

where the $\mathbf{D}$ and $\mathbf{H}$ fields correspond to the $\boldsymbol{\mathcal{E}}$ and $\mathbf{B}$ fields with the difference that the latter are caused by all charges and currents, whereas $\mathbf{D}$ and $\mathbf{H}$ are only caused by the "free" charges and currents. Ampère's law in matter is stated in terms of these new fields as

$$\mathbf{j} = \mathrm{curl}\mathbf{H} - \frac{\partial\mathbf{D}}{\partial t}. \tag{A.22}$$

Gauss's law is also modified slightly. One has to take into account that the polarization of the solid can lead to a local increase of charge density $\rho_e = -\mathrm{div}\mathbf{P}$.[1] We therefore get

$$\mathrm{div}\boldsymbol{\mathcal{E}} = \frac{\rho}{\epsilon_0} + \frac{\rho_e}{\epsilon_0} = \frac{\rho}{\epsilon_0} - \frac{\mathrm{div}\mathbf{P}}{\epsilon_0}, \tag{A.23}$$

1) The minus sign stems from the sign difference in the definition of $\boldsymbol{\mathcal{E}}$ and $\mathbf{P}$. In the former, the vector direction is from the positive to the negative charges, and in the latter, it is opposite.

or, using the usual definition of $\mathbf{D}$ (A.21),

$$\mathrm{div}\mathbf{D} = \rho. \tag{A.24}$$

Gauss's law for magnetostatics is unaffected in matter because there are still no magnetic monopoles. The same is true for Faraday's law, as there are no magnetic monopole currents either.

Index

Solid State Physics: An Introduction, Second Edition. Philip Hofmann.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.