Solid State Physics

Dimitri D. Vvedensky
The Blackett Laboratory,
Imperial College London, London SW7 2AZ

Second Term (2013–2014)

Copyright ©2014 by Dimitri D. Vvedensky
## Contents

1 Introduction to Solid State Physics  
1.1 What is Solid State Physics? .......................... 1  
1.2 Why Study Solid State Physics? ...................... 3  
1.3 Experimental Probes  ................................ 4  
  1.3.1 X-Ray Scattering  ............................... 5  
  1.3.2 Neutron Scattering  ............................ 5  
  1.3.3 Thermodynamic Measurements .................. 6  
  1.3.4 Transport Measurements ........................ 6  
1.4 Lattice Structures of Crystalline Solids .......... 7  
  1.4.1 Bravais Lattices  .............................. 7  
  1.4.2 Primitive Unit Cells  .......................... 8  
  1.4.3 The Basis .................................... 10  
1.5 What Holds Solids Together? ....................... 13  
1.6 Summary .......................................... 15  

2 Electrons in Crystalline Solids ..................... 17  
2.1 Bloch Functions .................................. 19  
2.2 The Kronig–Penney Model .......................... 21  
  2.2.1 Formulation and Solution ...................... 21  
  2.2.2 Energy Bands  ................................ 23  
  2.2.3 Wave functions* ................................ 26  
2.3 Extension to Three-Dimensional Energy Bands .... 29  
  2.3.1 Periodicity of the Energy Dispersion Relation 29  
  2.3.2 Dimensionality ................................ 31  
2.4 Energy Bands in Crystalline Materials .......... 34  
  2.4.1 Lithium ...................................... 37
Chapter 1

Introduction to Solid State Physics

1.1 What is Solid State Physics?

Solid state physics, the largest branch of condensed matter physics, which is itself the largest branch of modern physics, is the study of rigid matter which resists changes of shape or volume. This characterizes materials which are called solids, one of the four phases of matter. Solids are macroscopic materials which are typically composed of $10^{23}$–$10^{26}$ atoms and exhibit properties that depend on factors such as the atomic numbers of the constituent atoms, the structure of the solid, and the thermodynamic conditions, including temperature, pressure, and external forces and fields. Many properties of solids are fairly insensitive to the details at length scales much shorter than an Ångstrom and energy scales much higher than an electron volt. These are quite adequately described by non-relativistic quantum mechanics, i.e. Schrödinger’s equation. Relativistic effects, e.g. spin-orbit coupling, are sometimes needed for heavier atoms, but these are typically included as additional terms in the Schrödinger equation, rather than by solving the Dirac equation (the relativistic analogue of Schrödinger’s equation).

The majority of research in solid-state physics is focused on crystals, in which the atoms as arranged in a perfect periodic array extending indefinitely in all directions. This is because of Bloch’s theorem, which specifies the form of electronic wave functions and several characteristics of their energies in periodic potentials, and has had a pervading influence on the development of solid state physics as an academic discipline. Modern theories of the properties of solids rely on Bloch’s theorem to make large-scale numerical solutions of Schrödinger’s equation man-
The richness of solid state physics is revealed by the wide range of properties shown by different types of material. Broad categories of materials are metals, insulators, semiconductors, and superconductors, which are distinguished by their widely-varying optical, electronic, electrical and thermal transport, and mechanical properties. Metals are characterized by a non-zero electrical conductivity at $T = 0$, while insulators have vanishing electrical conductivity at $T = 0$. Semiconductors are insulators at $T = 0$, but can be doped with foreign atoms to control their electrical properties at $T > 0$. Superconductors have infinite conductivity below a critical temperature (i.e. there is no resistance to the flow of electrical current) and exhibit the Meissner effect, that is, they expel magnetic fields. In a ferromagnetic material, which can be a metal, an insulator, or a superconductor, all of the spins associated with individual atoms in the crystal line up in the same direction, even without a magnetic field, while, in an antiferromagnetic material, spins on neighboring atoms are aligned in opposite directions. The properties of some materials in these broad categories exhibit some quite unexpected properties. For example, diamond, which is an insulator, is the hardest known natural material, and has a thermal conductivity at room temperature which is five times higher than that of copper, an archetypal electrical conductor! We cannot begin to understand this without quantum mechanics. Another property which requires quantum mechanics is the transparency of diamond. Modern discoveries are constantly extending the realm of solid state physics with new phenomena which have the potential of fostering new and potentially revolutionary applications. Moreover, a fundamental understanding of nanoscience and nanotechnology, which are based on the manipulation of atoms to form materials and systems, requires quantum mechanical approaches.
1.2 Why Study Solid State Physics?

More than most other areas in physics, solid state physics involves the interplay between fundamental science and commercial applications. There is the intrinsic interest of understanding the behavior of electrons, lattice vibrations, and other excitations in periodic systems and their interactions with each other, which leads to superconductivity in certain materials, and with external driving forces, such as electric fields and photons. Specific examples of how interactions in solids lead to practical applications are as follows:

- **Electronic properties.** Transistors, which are semiconductor devices that amplify and switch electronic signals, are the heart of all modern electronic devices. Originally produced individually, transistors are now a key component of integrated circuits.\(^1\)

- **Electron-photon interactions.** Laser diodes, a type of laser with a doped semiconductor as the active medium, are used in applications ranging from telecommunications to bar code scanners. Photodiodes, another type of semiconductor-based device, convert light into electrical current (or a voltage). They are used, for example, in compact disk players, and as receivers in infrared remote-control devices.

- **Electron-phonon interactions.** The piezoelectric effect, discovered in 1880 by Pierre Curie and his brother Jacques in 1880, is the accumulation of charge as the result of an applied mechanical stress. This is a reversible process in that piezoelectric materials develop internal mechanical strain in response to an applied electric field. Piezoelectric materials are used in several instruments that image materials with atomic resolution, such as the scanning tunnelling microscope.

---

\(^1\)Since the early 1970s the number of transistors that semiconductor manufacturers can put on a chip has, on average, doubled every two years. This is known as Moore’s Law, named after Gordon Moore (one of the co-founders of Intel), who made this observation in 1965. This exponential increase in transistor densities has been sustained largely on the back of improvements in lithographic technology, a key component of integrated circuit technology, which has seen dramatic reductions in the feature size, from several microns to a few tenths of a micron. This increase has been accomplished without corresponding increases in production costs, resulting in processors with expanding memories and increasing operating frequencies which, for the consumer, translates into more powerful, but less expensive, computers.
• *Electron spin-electron charge interactions.* Spintronics, an amalgamation of ‘spin’ and ‘electronics,’ refers to the use of electron spin as an additional degree of freedom with which to control the motion of an electron. All spintronic devices act as follows: (1) information is written as a particular spin orientation (up or down), (2) the spins, attached to mobile electrons, carry the information along a wire, and (3) the information is read at a terminal. The spin orientation of conduction electrons survives for a relatively long time (nanoseconds), which makes spintronic devices an attractive avenue for quantum computing where electron spin would represent a bit (called qubit) of information. Room-temperature metal-based spintronics is now well established, with Albert Fert and Peter Grünberg having received the 2007 Nobel Prize in Physics for the discovery of the giant magnetoresistance (GMR) effect, which has proved to be of enormous significance with its application to disk drive sensor technology. The prototype device, already in use in industry as a read head and a memory-storage cell, is a GMR sandwich structure which consists of alternating ferromagnetic and nonmagnetic metal layers. Depending on the relative orientation of the magnetizations in the magnetic layers, the device resistance changes from small (parallel magnetizations) to large (antiparallel magnetizations). This change in resistance in the device, which is called magnetoresistance, is used to sense changes in magnetic fields.

This list is by no means exhaustive, even as a current snapshot. Discoveries of new phenomena and new materials are constantly changing the landscape of the solid state physics for both theory and experiment.

## 1.3 Experimental Probes

Solid state physics benefits from a symbiotic relationship between theory, computation, and experiment. Sometimes a new experimental measurement challenges theorists to provide an explanation, while a theory or computation may require altogether new experiments to verify a prediction. Of the various available experimental methods, the most widely used are scattering measurements to probe atomic and magnetic ordering, thermodynamic measurements for the response of a system to changes of temperature, pressure, and other control variables, and transport measurements of thermal and electrical currents in response to tempera-
ture and electric potential gradients. In this section, we briefly review the physical principles behind each type of method and discuss the information they provide about crystalline solids. We will invoke the results of such measurements throughout this course.

1.3.1 X-Ray Scattering

In 1912 Max von Laue discovered that crystalline materials act as three-dimensional diffraction gratings for X-rays whose wavelengths are similar to the spacing of the planes in the crystal. X-rays interact primarily with electrons in atoms. Diffraeted waves from different atoms can interfere with each other and the resultant intensity distribution is strongly modulated by this interaction. If the atoms are arranged in a periodic structure, as in crystals, the diffracted waves will consist of sharp interference maxima (peaks) with the same symmetry as in the distribution of atoms. Measuring the diffraction pattern therefore allows the distribution of atoms in a material to be deduced. The principle of this process is described by Bragg’s law,

$$ n\lambda = 2d \sin \theta, $$

which relates, through an integer $n$, the wavelength $\lambda$ of electromagnetic radiation to the diffraction angle $\theta$ and the lattice spacing $d$ of the diffracting planes in a crystalline sample. By changing the geometry of the incident rays, the orientation of the crystal and the detector, many diffraction directions of the lattice can be realized, from which the structure of the material can be deduced. The symmetry of crystals shows itself in the diffraction pattern by the presence or absence of characteristic reflections. The intensity of the diffracted beams depends on the arrangement and atomic number of the atoms in the unit cell. X-ray diffraction is now a standard technique for the determination of crystal structures and atomic spacings.

1.3.2 Neutron Scattering

X-rays interact primarily with the electron cloud surrounding each atom, whereas neutrons scatter mostly from the nuclei of the atoms. Thus, neutrons can provide complementary information about crystalline structure which can exploit the dependence of the scattering on the atomic numbers of the scattering centers. Although neutrons are uncharged, they carry a spin, and therefore interact with mag-
netic moments, including those arising from the electron cloud around an atom. Neutrons can detect magnetic order in solids: just as they can detect ordered planes of atoms, they can detect ordered planes of electron-spins. For example, in ferromagnets like iron oxide, new Bragg peaks appear due to the magnetic structure of these spins with which the neutron can interact. Accordingly, neutron diffraction can be used to investigate the microscopic magnetic structure of a material.

1.3.3 Thermodynamic Measurements

Measurements of thermodynamic properties of materials typically focus on the specific heat, thermal expansion, thermal conductivity, and the presence of any phase transitions. The experimental probes that are used for such measurements are photons in the visible, infrared, and X-ray regimes, and neutrons. These are microscopic probes in that they have atomic-scale resolution and, therefore, often require detailed computations for their interpretation. Macroscopic measurements of specific heat and thermal expansion, and the use of high-pressure, high-temperature experiments are also important for characterizing materials, especially for identifying solid-to-solid phase transitions. Among the best-known examples of such a phase transition is in iron. At room temperature, iron has a body-centered cubic structure and is magnetic. As the temperature is raised to 770°C (known as the Curie point), iron becomes non-magnetic, but remains body-centered cubic. Raising the temperature to 912°C, sees a phase transition to the face-centered cubic structure. Another phase transition occurs at 1394°C, back to the body-centered cubic structure. The ultimate goal of such research is to derive an equation of state for a particular material.

1.3.4 Transport Measurements

Transport properties typically focus on thermal and electrical transport. Thermal conductivity measures the ability of a material to conduct heat. Thermal conduction takes place when a temperature gradient exists in a solid, with heat flow occurring in the direction of decreasing temperature because higher temperature equates to higher kinetic energy, or more atomic movement. Energy is transferred

---

A phase is a region of a thermodynamic system within which all physical properties are uniform. Quite apart from the usual solid, liquid, and gas phases of materials, this includes the magnetic and structural states of a material.
from the more energetic to the less energetic atomic when neighboring atoms collide. Electrical transport is measured analogously, by introducing an electrical potential gradient, causing a flow of electrons from higher to lower potential. The measurement of electrical conductivity, which measures the ability of a material to conduct electricity provides information about the densities of conduction electrons, their mobilities, and their scattering mechanisms.

1.4 Lattice Structures of Crystalline Solids

Structure is one of the most important properties of a solid and is the starting point of any study of a material. Many materials of fundamental and technological interest are crystalline. Examples include metals such as iron, copper, silver, and gold, insulators such as diamond, and semiconductors such as Si, Ge, and GaAs. Information about crystalline structure is provided by experiments, usually X-ray measurements, as discussed above. Quantum mechanical computations can sometimes be used to provide some input, but this usually requires some information about the crystal lattice of a material, which is then used as a starting point to determine the lattice spacing that yields the energy minimum. Even with modern techniques and computational power, the determination of the structure of a crystal given only the atomic numbers of its constituents is an impractical task. In this section we review the basic elements of the structure of crystals and discuss how the requirement of three-dimensional periodicity restricts the allowed structures.

1.4.1 Bravais Lattices

An ideal crystal is constructed by the infinite repetition of identical structural units in space. As we will see below, this places constraints on these units. In the simplest cases the structural unit is a single atom, as in crystalline copper, silver, gold, iron, aluminum, and the alkali metals. The structures of all crystals can be described in terms of a lattice, with a group of atoms attached to every lattice point, called the basis. When the basis is repeated in space it forms the structure of the crystal containing those atoms.

The underlying structure of a crystal is described by a Bravais lattice, sometimes called a space lattice, which is defined by vectors that determine the lattice sites,

$$R(n_1, n_2, n_3) = n_1a_1 + n_2a_2 + n_3a_3,$$  \hspace{1cm} (1.1)
where $n_1$, $n_2$, and $n_3$ are integers and $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are \textit{primitive vectors} of the lattice (see below). The unit cell defined by the primitive vectors has one atom and can be used to fill all of space. In two dimensions there are five Bravais lattices (Problem Set 1), while in three dimensions there are 14 Bravais lattices, whose \textit{conventional} unit cells are shown in Fig. 1.1 and are summarized below:

- Cubic: each face is a square. Simple cubic, body-centered cubic, and face-centered cubic.
- Tetragonal: stretched in one direction. Simple tetragonal and centered tetragonal.
- Orthorhombic: sides of three different lengths at right angles to one another. Simple orthorhombic, base-centered orthorhombic, base-centered orthorhombic, and face-centered orthorhombic.
- Monoclinic: one face is a parallelogram, the other two are rectangular. Simple monoclinic, centered monoclinic and base-centered monoclinic.
- Triclinic: all faces are parallelograms. Simple monoclinic.
- Trigonal: each face is a rhombus. Simple trigonal.
- Hexagonal: two-dimensional hexagonal lattices, stacked directly above one another. Simple hexagonal.

There are several key points about primitive unit cells of Bravais lattices. Bravais lattices extend over all space, so the primitive unit cells, which are associated with a single lattice site, are space filling. Secondly, the environment of all lattice sites is \textit{identical}, i.e. apart from coordinates, there is nothing to discriminate between different lattice sites.

\subsection{1.4.2 Primitive Unit Cells}

The definition (1.1) of lattice sites in terms of primitive vectors with integer coefficients that define a cell with a single lattice site (Problem Set 1) places stringent restrictions on the choice of primitive vectors. The most straightforward case is the simple cubic lattice, for which a natural choice of primitive vectors is $\mathbf{a}_1 = a \mathbf{i}$. 


Figure 1.1: Conventional unit cells for the fourteen three-dimensional Bravais lattices.
\( a_2 = a j \), and \( a_3 = a k \), in which \( i, j, \) and \( k \) are the usual unit vectors along the \( x, y, \) and \( z \)-directions, respectively, and \( a \) is the length of an edge of the cube. Hence, the position every lattice site of a simple cubic lattice is given by
\[
R(n_1, n_2, n_3) = an_i + an_j + an_k,
\]
for some choice of \( n_1, n_2, \) and \( n_3 \). Analogous choices can be made for the simple tetragonal and simple orthorhombic lattices, and the extension to the simple monoclinic, simple triclinic, and simple trigonal lattices are straightforward.

For other Bravais lattices, however, other the relationship between the primitive and conventional unit cells is not so apparent. Consider the body-centered cubic lattice. A standard choice of primitive vectors is shown in Fig. 1.2:
\[
\begin{align*}
    a_1 &= \frac{1}{2}ai + \frac{1}{2}aj - \frac{1}{2}ak, \\
    a_2 &= -\frac{1}{2}ai + \frac{1}{2}aj + \frac{1}{2}ak, \\
    a_3 &= \frac{1}{2}ai - \frac{1}{2}aj + \frac{1}{2}ak.
\end{align*}
\]
Notice that these basis vectors are not orthogonal. Any lattice vector in the body centered-cubic system can be expressed as
\[
R(n_1, n_2, n_3) = \frac{1}{2}a(n_1 - n_2 + n_3)i + \frac{1}{2}a(n_1 + n_2 - n_3)j + \frac{1}{2}a(-n_1 + n_2 + n_3)k.
\]
If each of the \( n_i \) is odd, then each of the coefficients of the unit vectors is also odd, in which case the corresponding lattice vector yields a body-centered position in the coordinate system in Fig. 1.2. A body-centered position is also obtained if one of the coefficients is odd and the other two are even. Lattice positions at cube corners are obtained if all three coefficients are even or if only one is even. In this way, as we scan through the integers, all of the sites of the body-centered cubic lattice are accounted for. Analogous choices for primitive vectors yield the other non-simple Bravais lattices.

### 1.4.3 The Basis

Every point of a Bravais lattice is equivalent to every other point. In an elemental crystal, in which there is only one type of constituent atom, it is possible that the elemental atoms are located at the vertices of a Bravais lattice. In general, however, a crystal structure will be a Bravais lattice with a basis, which are the
Figure 1.2: The primitive lattice vectors of the body-centered cubic Bravais lattice. (C. Kittel, *Introduction to Solid State Physics* 7th edition (Wiley, New York, NY, 1996), p. 21.)

Figure 1.3: Conventional unit cells (a) crystalline Cu, with each sphere representing a Cu atom, and (b) the diamond form of C, with each sphere now representing a C atom. The structure in (a) is a Bravais lattice, while that in (b) is not.

atom(s) or molecule(s) attached to each lattice point. An important point to keep in mind is that lattice positions are abstract points, while the atoms in the Bravais lattice are physical entities.
We consider two examples. Crystalline copper has the face-centered cubic Bravais lattice, with each lattice site occupied by a copper atom. The conventional unit cell is shown in Fig. 1.3(a). Now consider the diamond form of carbon, whose conventional unit cell is shown in Fig. 1.3(b). This structure is not a Bravais lattice, but in fact is a face-centered cubic lattice with a two-atom basis (sometimes referred as two interpenetrating face-centered cubic lattices). Associated with each lattice position \( \mathbf{R}(n_1, n_2, n_3) \) are two atoms: one at \( \mathbf{R}(n_1, n_2, n_3) \) and the other at \( \mathbf{R}(n_1, n_2, n_3) + \frac{1}{2}a(\mathbf{i} + \mathbf{j} + \mathbf{k}) \). The resulting crystal is still based on face-centered cubic Bravais lattice, but the two-atom basis results in each atom having a local tetrahedral environment. However, not every tetrahedral environment is the same: those of the atoms within the cube are rotated by 90° with respect to those of the vertex and face-centered atoms (Fig. 1.4). Silicon also has this structure, with the basis now formed of two Si atoms, as does GaAs, with a Ga atom and As atom assigned to each lattice site. We will see in a later chapter how this structure impacts on the properties of materials such as Si, and GaAs.

We have now discussed the three fundamental components of the structure of crystalline materials: (1) the bravais lattice, which provides the ‘scaffolding’ to which (2) the basis is attached to yield (3) the crystal structure of the material. The relationship between these concepts can be represented symbolically as

\[
\text{Bravais lattice} + \text{basis} = \text{crystal structure}.
\]
Beginning with 14 Bravais lattices based on 7 crystal types (cubic, tetragonal, orthorhombic, monoclinic, triclinic, trigonal, and hexagonal), which are determined solely by symmetry, that is, the tiling (or filling) of space, the addition of atomic bases leads to 230 crystal structures, known as space groups. Every crystalline material has one of these structures.

1.5 What Holds Solids Together?

The basic constituents are atoms, whose electrons occupy discrete states. When atoms are brought into contact the key quantity that determines whether or nor a solid forms is the cohesive energy, which is the energy difference between the separated atoms and those in the crystalline solid:

\[
\text{cohesive energy} = \text{energy of free atoms} - \text{energy of atoms in crystal}.
\]

Figure 1.5 shows a schematic representation of the potential energy \( U \) between two atoms as they are brought closer together. The potential energy has a minimum at a value \( r = r_0 \). For \( r > r_0 \), \( U \) increases sharply, gradually approaching zero as \( r \to \infty \), while for \( r < r_0 \), \( U \) rapidly and becomes infinite at \( r = 0 \). The value \( U_0 = U(r_0) \) represents the cohesive energy. The magnitude of the cohesive energy ranges from 1 eV/atom to 10 eV/atom, except for inert gases, for which the cohesive energy is of the order of 0.1 eV/atom.

The attractive forces reflect the presence of bonds between atoms in solids, which are responsible for the stability of the crystal. The four main types are:

- **van der Waals (or molecular) bonding**, which is found in crystals of inert gases, such as neon (Ne), argon (Ar), krypton (Kr) and xenon (Xe). The atoms of these elements have filled electronic shells and spherical electronic clouds. In the crystal the inert gas atoms pack together in the cubic face-centered cubic structure. What holds these atoms together? Consider two inert gas atoms separated by distance \( r \). Although the average charge distribution in a single atom is spherically symmetric, which implies that the average dipole moment of each atom is zero, a non-zero dipole moment may form because of fluctuations of this charge distribution. According to electrostatics this dipole moment produces an electric field, which induces a
The ionic bond results from the electrostatic interaction of oppositely charged ions. Consider sodium chloride (NaCl) as an example. In the crystalline state, each Na atom loses its single valence electron to a neighboring Cl atom, producing Na$^+$ and Cl$^-$ ions which have filled electronic shells. As a result an ionic crystal is formed containing positive and negative ions coupled by a strong electrostatic interaction.

The covalent bond between two atoms is usually formed by two electrons, one from each atom. The electrons forming the bond are partly localized in the region between the two atoms joined by the bond. The covalent bond is typically strong: for example, covalent bonding between carbon atoms is responsible for the hardness and other properties of diamond. The covalent bonding is also found for silicon and germanium crystals.

The main feature of the metallic bond is the lowering of the kinetic energy of the valence electrons in a metal as compared to the free atoms. According to the Heisenberg uncertainty principle, the impreciseness in a coordinate $x$
and corresponding momentum \( p \) are related by \( \Delta x \Delta p \sim h \). In a free atom, the valence electrons are confined to a relatively small volume, so \( \Delta p \) is relatively large, as are the kinetic energies of the valence electrons. In the crystal, however, the electrons are free to move, the volume of which is large, so the kinetic energy of the electrons is greatly reduced. This lowers the total energy of the electrons in the solid.

Although the nature of the attractive potential energy is different in different solids, the origin of the repulsive potential energy is similar. The repulsive force is due mainly to the Pauli exclusion principle, the basic statement of which is that no two electrons can occupy the same orbital (including spin). As ions approach each other, the orbits of the electrons begin to overlap, i.e. some electrons attempt to occupy orbits already occupied by others. This is, however, forbidden by the Pauli exclusion principle. As a result, electrons are excited to unoccupied higher energy states of the atoms. Thus, the electron overlap increases the total energy of the system and thereby provides a repulsive contribution to the interaction.

### 1.6 Summary

This chapter has provided an introduction to solid state physics. We have focused on the crystallographic aspects of materials because that has been the basis of the historical importance of this subject. The production of high-quality single crystals of Ge and Si, and their controlled doping, was a key ingredient for the architecture of the first transistors. The translational symmetry of traditional crystals has a profound effect on their properties, as we will discover in the following chapters. New discoveries have challenged and expanded the notion of what a solid is and has provided substantial opportunities for both fundamental research as well as potentially revolutionary applications.
Chapter 2

Electrons in Crystalline Solids

There are two basic types of electrons in solids: core and valence. The core electrons are those closest to the nucleus in the atomic state and remain so in the solid. The more loosely bound valence electrons are responsible for the bonding between neighboring atoms and for properties such as conduction in metals and semiconductors. Both types of electron are important for determining the structure of crystalline solids: the bonding of the valence electrons is an attractive force that pulls atoms closer together, but the core electrons mediate this attraction with a short-range repulsion that prevents atoms from coming too close together. The equilibrium structure is obtained from the balance of these forces. The challenge for solid state physics is to use this intuitive picture of electrons to build a thorough conceptual and computational framework to understand the properties of solids.

The electronic properties of crystalline solids are determined by three factors: (i) the periodic potential of the atoms in the solid, (ii) the electrostatic interactions between the electrons and between the electrons and nuclei, and (iii) the quantum statistics of the occupancy of electronic states (the Pauli exclusion principle). The history of our understanding the behavior of electrons in crystalline solids is essentially the history of how each of these factors have been incorporated into a theoretical description that can be tested against experiment. The first quantum mechanical model of electrons in solids was due to Arnold Sommerfeld.1 The

---

1Arnold Sommerfeld (1868–1951) was a German physicist who made many contributions in the early days of quantum mechanics. He is perhaps better known as a teacher, with six of his students – Werner Heisenberg, Wolfgang Pauli, Peter Debye, Hans Bethe, Linus Pauling, and Isidor Rabi – having been awarded Nobel Prizes in Physics.
Electrons in Crystalline Solids

electrons were treated as independent and moving in a constant potential, thereby neglecting the effects of (i) and (ii) altogether. Nevertheless, despite its simplicity, the Sommerfeld model was able to explain the temperature-dependence of the heat capacity of metals, the general shape of the density of states, and several other properties, including those related to conduction, which we will describe in a later chapter.

The next important advance was due to Felix Bloch, who determined the effects of the periodicity of crystalline solids on the form of the electronic wave function – now called Bloch functions – and on the structure of energy bands. If only the electrostatic interaction between the valence electrons and the nuclei are included in the periodic potential, then the electrons are still effectively independent. The electrostatic interaction between valence electrons does not affect the periodicity within the solid, but alters the nature of the wave function dramatically and, of course, severely complicates the quantum mechanical calculation of electronic states. The desire to retain the simplicity of the independent particle picture, while acknowledging the importance of electrostatic interactions between valence electrons, has led to the popularity of ‘average’ approximations to this interaction. This has enabled many realistic calculations of crystalline materials to be carried out to the point where some are now regarded as routine. However, the search for more fundamental and practical approaches remains an active research-level activity.

This chapter is devoted to energy bands in crystalline solids. We begin with a derivation of Bloch functions and then review the salient points of the Kronig–Penney model, which reveals the effect of a one-dimensional periodic potential on the energy bands. Energy bands in two- and three-dimensional solids are then examined, followed by a discussion of how the presence of different types of atomic state gives rise to different types of energy band. We conclude with several examples of energy bands of metals and semiconductors to discuss several properties that can be explained directly from these bands.

Felix Bloch (1905–1983) was a Swiss-born American physicist who made many fundamental contributions to physics. His theory of wave functions in periodic potentials in the 1920s provided the foundation for many future advances. But Bloch is best known for his development of nuclear magnetic resonance (NMR) in the 1940s. NMR is a technique that utilizes the magnetic properties of nuclei to determine structural information about molecules and has become a standard method of medical imaging. Bloch was awarded the Nobel Prize for Physics in 1952 for this work, sharing the honor with Harvard physicist Edward Purcell, who had conducted similar experiments independently and concurrently.
2.1 Bloch Functions

The solution of the Schrödinger equation for electrons in a periodic potential is enormously simplified if we utilize the effect of this symmetry on the wave function from the outset. The time-independent Schrödinger equation has the form

$$-\frac{\hbar^2}{2m} \frac{d^2 \psi}{dx^2} + V(x) \psi = E \psi,$$

where the potential has period $a$: $V(x+a) = V(x)$. We are not concerned with the details of this potential – that will affect the solution but not its basic form.

Consider the effect of translating $\psi(x)$ by $a$, i.e. changing $x$ to $x+a$. The second derivative is unaffected because $a$ is a constant, and the potential is unchanged on account of its periodicity. The energy $E$ is a constant and is, therefore, unaffected by any transformation. Thus, the Schrödinger equation is transformed to

$$-\frac{\hbar^2}{2m} \frac{d^2 \psi(x+a)}{dx^2} + V(x) \psi(x+a) = E \psi(x+a),$$

which is the same as (2.1) for $\psi(x)$. Because $\psi(x)$ and $\psi(x+a)$ satisfy the same Schrödinger equation, any observable quantity, such as the probability density associated with $\psi(x)$, must also embody this invariance: $|\psi(x+a)|^2 = |\psi(x)|^2$. This implies that $\psi(x+a)$ can be written as

$$\psi(x+a) = e^{i\phi(a)} \psi(x),$$

that is, $\psi(x)$ and $\psi(x+a)$ can therefore differ only by a pure phase, i.e. a complex quantity that has modulus unity.

To determine the functional form of $\phi$, we first observe that

$$|\psi(x+na)|^2 = |\psi(x)|^2,$$

that is, the probability density is unaffected by any number of translations by $a$. We compute $\psi(x+na)$ in two ways. The first is by a single translation by $na$:

$$\psi(x+na) = e^{i\phi(na)} \psi(x).$$

The other way to compute $\psi(x+na)$ is by a sequence of $n$ individual translations, each by $a$, and each producing a factor of $e^{i\phi(a)}$:

$$\psi(x+na) = e^{i\phi(a)} \psi(x+(n-1)a)$$

$$= e^{2i\phi(a)} \psi(x+(n-2)a) = \cdots = e^{i(n-1)\phi(a)} \psi(x).$$
Equating our two expressions $\psi(x + na)$ yields

$$e^{i\phi(na)} = e^{in\phi(a)}.$$  \hspace{1cm} (2.4)

The arguments in these exponentials already imply that $\phi(a)$ is proportional to $a$, but we can also show this directly. We regard $n$ as a continuous variable and differentiate (2.4) with respect to $n$:

$$i \frac{d\phi}{d(na)} \frac{d(na)}{dn} e^{i\phi(na)} = i\phi(a) e^{in\phi(a)},$$

which, after setting $n = 1$ and cancelling the common factors, yields

$$\frac{d \phi}{da} = \phi.$$  \hspace{1cm} (2.5)

This equation is easily integrated, either by the trial solution or separation of variables method, to obtain the solution $\phi(a) = ka$, in which $k$ is a constant (with units of inverse length). Thus, (2.3) reduces to one form of **Bloch’s theorem**:

$$\psi(x + a) = e^{ikx} \psi(x).$$  \hspace{1cm} (2.5)

We can obtain a more explicit form for $\psi(x)$ by solving (2.5) for $\psi(x)$, then multiplying both sides of the resulting equation by $e^{-ikx}$:

$$e^{-ikx} \psi(x) = e^{-ikx} e^{-ika} \psi(x + a) = e^{-ik(x + a)} \psi(x + a),$$

which indicates that $e^{-ikx} \psi(x)$ is a periodic function with period $a$. Calling this function $u(x)$, we can write

$$\psi(x) = e^{ikx} u(x),$$  \hspace{1cm} (2.6)

where $u(x + a) = u(x)$. This is called a **Bloch function**. The exponential factor is the solution for a free particle, while the periodic function is a direct consequence of the periodic potential. This is one of the central results for describing electrons in crystalline solids and, as our derivation makes clear, is valid for any periodic potential.
2.2 The Kronig–Penney Model

The general aspects of wave functions and energy bands mandated by Bloch’s theorem are obtained for any periodic potential, but the solution of a specific problem helps to visualize the how various parameters affect the solution. In this section, we summarize the main steps in the solution of the Kronig–Penney model.

2.2.1 Formulation and Solution

The Kronig–Penney model consists of an periodic array of unit cells, or ‘atoms,’ which are infinite square wells, as shown in Fig. 2.1. When these square wells are placed adjacent to one another, the potential $V(x)$ for the electrons is a periodic array of Dirac $\delta$-functions separated by a distance $a$:

$$V(x) = V_0 \sum_{n=-\infty}^{\infty} \delta(x - na),$$

Figure 2.1: (a) An infinite square well of length $a$, which is the ‘atomic’ constituent of the Kronig–Penney model. The wave functions corresponding to the lowest three energies are shown, with positive and negative values indicated by solid and broken lines, respectively. (b) When the square wells in (a) are placed side-by-side, the resulting potential is a periodic array of Dirac $\delta$-functions, which is the Kronig–Penney model.
where $V_0$ is the strength of the $\delta$-function. In the atomic case, $V_0 \to \infty$, so there is no interaction between wells, while $V_0 = 0$ corresponds to the absence of a potential, i.e. free electrons. The solution of this problem proceeds by observing that the potential vanishes in the region between the $\delta$-functions, whereupon the time-independent Schrödinger equation reduces to

$$\frac{d^2 \psi}{dx^2} + q^2 \psi = 0,$$

where

$$q = \frac{\sqrt{2mE}}{\hbar}.$$  \hspace{1cm} (2.7)

The general solution $\psi_n$ to this equation within each region between the $\delta$-functions, $(n-1)a \leq x \leq na$ is (Fig. 2.1):

$$\psi_n(x) = A_n e^{-iq(x-na)} + B_n e^{iq(x-na)},$$

where $A_n$ and $B_n$ are constants to be determined by boundary and matching conditions. In particular, the wave function must be continuous across the $\delta$-function at $x = na$, $\psi_n(na) = \psi_{n+1}(na)$, which yields

$$A_n + B_n = A_{n+1} e^{iq} + B_{n+1} e^{-iq}.$$  \hspace{1cm} (2.8)

The cusp condition of the derivative of the wave function at the $\delta$-function (Sec. A.1),

$$\psi'_{n+1}(na) - \psi'_n(na) = \frac{2mV_0}{\hbar^2} \psi_n(na),$$

produces

$$-iqA_{n+1} e^{iq} + iqB_{n+1} e^{-iq} - (-iqA_n + iqB_n) = \frac{2mV_0}{\hbar^2} (A_n + B_n).$$  \hspace{1cm} (2.9)

By using Bloch’s theorem to write $A_{n+1} = e^{ika} A_n$ and $B_{n+1} = e^{ika} B_n$, we can represent (2.8) and (2.9) as the matrix equation

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} A_n \\ B_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$  \hspace{1cm} (2.10)
in which

\[ a_{11} = 1 - e^{i(k+q)a}, \]
\[ a_{12} = 1 - e^{i(k-q)a}, \]
\[ a_{21} = iqa \left[ 1 - e^{i(k+q)a} \right] - 2U_0, \]
\[ a_{22} = -iqa \left[ 1 - e^{i(k-q)a} \right] - 2U_0, \]

and

\[ U_0 = \frac{maV_0}{h^2} \]

is a dimensionless measure of the strength of the potential. Equation (2.10) has nontrivial solutions for \( A_n \) and \( B_n \) if and only if the determinant of the matrix of coefficients vanishes, which yields the condition

\[ \cos ka = \cos qa + U_0 \frac{\sin qa}{qa}. \]  \hspace{1cm} (2.11)

Note that, according to (2.7), \( q \) is a function of the energy \( E \), so this equation gives \( k \) as a function of \( E \). Thus, to obtain the dispersion relation \( E(k) \), this equation must be inverted, which is a straightforward numerical procedure. The next two sections discuss the dispersion relations and wave functions obtained from (2.11).

### 2.2.2 Energy Bands

Figure 2.2 shows typical energy bands for the Kronig–Penney model obtained from (2.11). This diagram provides a wealth of information and provides a basic understand of many aspects of the energy bands of all materials.

- **Band gaps.** One of the most striking feature of the energy bands for any \( U_0 > 0 \) is the appearance of gaps in the energy, that is, values of the energy for which there is no propagating solution. The physical reason for this that the band gaps occur whenever \( ka = n\pi \), for any non-zero integer \( n \), which is where the Bragg condition is satisfied for an electron moving along this
Figure 2.2: The energy dispersion $E(k)$ for free electrons, which corresponds to $U_0 = 0$ (light solid line), and for periodic potentials with $U_0 = 1$ (bold solid line), and $U_0 = 9$ (bold broken line). The broken vertical lines mark the points $ka = n\pi$, where $E(k)$ is discontinuous, and the shaded areas are the energy gaps for $U_0 = 1$, where there is no solution of 2.11.

- **Band widths.** Each continuous band results from a discrete level in the ‘atomic’ system, i.e. the infinite square well. If $U_0 \to \infty$, then an electron initially in one of the wells remains there, in which case there is no dispersion of the corresponding energy band. At the other extreme, when $U_0 = 0$, the electrons propagate through the system unimpeded by any potential. For any $U_0 > 0$, gaps in the dispersion appear whose magnitude increases as $U_0$
increases, and the band widths decrease. All of the dispersion relations for $U_0 \geq 0$ intersect at $ka = n\pi$, for non-zero integer $n$, which yields $q = n\pi$, according to (2.11), that is, the $n$th energy level of the infinite square well:

$$E_n = \frac{\hbar^2 n^2}{8ma^2}.$$  

The widths of the first four bands as a function of $U_0$ are shown in Fig. 2.3. The higher bands show a greater dispersion than the lower bands, which is essentially a manifestation of their kinetic energy. Higher-lying levels in the infinite square well have larger kinetic energies than lower lying levels (Fig. 2.1), so as decreases in their confinement (by lowering $U_0$) results correspondingly more pronounced decrease of their kinetic energy.

**Figure 2.3:** Widths of the first four bands as a function of $U_0$ as obtained from (2.11). The shaded regions indicate the widths of each band and the dotted line marks the widths for $U_0 = 9$, which is one of the cases considered in Fig. 2.2.
- **Band curvature.** At the bottom of the band, near $k = 0$, the effect of $U_0$ is to shift the energies upward from the free-electron energies and to alter the curvature. These effects can be subsumed by writing

$$E(k) \approx E_0 + \frac{\hbar^2 k^2}{2m^*},$$

where $m^*$ is the effective mass. The physical content of this equation is that, near the bottom of the band, the behavior of an electron resembles that of a free electron. However, because the curvature of the energy dispersion $E(k)$ is different from that of the free-electron parabola, the acceleration of an electron in a constant external force is not that same as that of a free electron in vacuum. The electron behaves like a free particle with an ‘effective’ mass that is greater than the inertial mass of a free electron because the potential barriers impede the motion of the electron. The association of the effective mass with the curvature of the energy dispersion leads to the following definition:

$$\frac{1}{m^*} = \frac{1}{\hbar^2} \frac{d^2E}{dk^2} \bigg|_{k=0}.$$ 

Thus, flat bands, which correspond to a large values of $U_0$ and small curvature, lead to large effective masses, while broad bands, resulting from smaller values of $U_0$ and thus a larger curvature, lead to effective masses close to free electron masses.

### 2.2.3 Wave functions*

The behavior of the energy dispersion relations obtained from (2.11) is mirrored by the wave functions, which are obtained by solving (2.10). Strictly speaking, this equation yields only the ratio $B_n/A_n$, but normalizing the solution over the unit cell alleviates any ambiguity. Figure 2.4 shows the real and imaginary parts of the Bloch functions and their absolute values for several values of $U_0$ for the state at $ka = \frac{1}{2}\pi$ in the first band, which corresponds to the $n = 1$ state in the infinite square well [Fig. 2.1(a)]. Figure 2.5 shows the analogous quantities in the second band, which corresponds to the $n = 2$ state in the infinite square well.

Notice that the squared modulus of the wave functions always has the periodicity of the lattice for $U_0 > 0$. When $U_0 = 0$, the modulus is a constant, as expected for free electrons. As $U_0$ increases, the effect of the potential becomes
Electrons in Crystalline Solids

Figure 2.4: Bloch functions for $ka = \frac{1}{2}\pi$ in the lowest band, corresponding to the $n = 1$ state of the infinite square well [Fig. 2.1(a)] for different values of $U_0$. The real and imaginary parts of the Bloch functions are plotted as solid and broken lines, respectively, in (a), (c), and (e). The corresponding absolute values are shown in (b), (d), and (f). The values of $U_0$ are (a), (b) $U_0 = 0$, (c), (d) $U_0 = 9$, and (e), (f) $U_0 \to \infty$. The positions of the $\delta$-functions are indicated in each panel.
Figure 2.5: Bloch functions for $ka = \frac{3}{2}\pi$ in the second band, corresponding to the $n = 2$ state of the infinite square well [Fig. 2.1(a)] for different values of $U_0$. The real and imaginary parts of the Bloch functions are plotted as solid and broken lines, respectively, in (a), (c), and (e). The corresponding absolute values are shown in (b), (d), and (f). The values of $U_0$ are (a), (b) $U_0 = 0$, (c), (d) $U_0 = 9$, and (e), (f) $U_0 \to \infty$. The positions of the $\delta$-functions are indicated in each panel.
evident, with the limit \( U_0 \to \infty \) forcing the modulus to zero at the \( \delta \)-functions. The same basic trends are seen for the real and imaginary parts of the wave functions at \( k = \frac{1}{2} \pi \), for the band corresponding to \( n = 1 \), and at \( k = \frac{3}{4} \pi \) for the band corresponding to \( n = 2 \), all of which have periodicity of \( 4a \) and whose relative phases are characteristic of travelling waves. The wave functions for \( U_y = 0 \) have the sinusoidal form expected of free electrons. For \( U_y = 9 \), the cusps in the wave functions at the \( \delta \)-functions are clearly evident. In the limit \( U_0 \to \infty \), the wave functions are composed of the solutions of the infinite square well for the band in question, with the relative phases determined by \( k \). An overall phase does not affect the energy of these states, so the energy of the wave function is just the corresponding ‘atomic’ energy. In this limit, there is no dispersion, so the energies are the same for all \( k \). For \( U_0 \approx 0 \), the electrons move through the lattice relatively unimpeded by the periodic potential, so a band picture is appropriate for describing the electronic states in the solid. As \( U_0 \) increases, electron motion becomes increasingly inhibited, so the band width narrows accordingly. In this case, a more localized ‘atomic’ description of the electronic states becomes more appropriate.

### 2.3 Extension to Three-Dimensional Energy Bands

The solution of the Kronig–Penney model embodies many of the fundamental features found for electronic states in all crystalline solids: continuous bands of allowed energies separated by gaps with no propagating solutions, a vanishing group velocity at band edges, and parabolic behavior near band edges whose curvature is subsumed by an effective mass. But the band diagram in Fig. 2.2 is not the most convenient for displaying the energy bands of real three-dimensional solids. In this section, we will develop the standard presentation of the energy band calculations by re-examining the Kronig–Penney solution.

#### 2.3.1 Periodicity of the Energy Dispersion Relation

The energy bands obtained from the Kronig–Penney model in Fig. 2.2 highlight the differences between the dispersion curve of free-electrons and that of electrons moving in a periodic potential. Our intuition tells us that a small periodic potential should produce a correspondingly small change in the behavior of the electrons. This is largely confirmed in both the energy dispersion curves in Fig. 2.2 and
the wave functions in Figs. 2.3 and 2.4. But, however small the value of \( U_0 \), the solutions of (2.11) shown in this diagram are not the only solutions. To understand how additional solutions arise, we define quantities \( G_n \) by

\[
G_n = \frac{2n\pi}{a},
\]

where \( n \) is any integer, and examine the effect of changing \( k \) to \( k + G_n \) on the right-hand side of (2.11). Using the standard trigonometric identity for the cosine of the sum of two angles, we obtain

\[
\cos[(k + G_n)a] = \cos(ka) \cos(G_na) - \sin(ka) \sin(G_na)
\]

\[
= \cos(ka) \cos(2n\pi) - \sin(ka) \sin(2n\pi) = \cos(ka),
\]

because \( \cos(2n\pi) = 1 \) and \( \sin(2n\pi) = 0 \). This shows that changing \( k \) to \( k + G_n \) yields the same solution of (2.11). In other words, the energy dispersion \( E(k) \) obtained from (2.11) is a periodic function of \( k \), with a fundamental period of \( G_1 \):

\[
E(k + G_n) = E(k). \tag{2.12}
\]

Although we have obtained this periodicity for a specific potential, the result is valid for any periodic potential.

The left panel in Fig. 2.6 shows the periodic solutions of the bands in Fig. 2.2. The additional periods do not provide any new information, so this diagram contains redundant information. Indeed, we need not display an entire period of any band because the left-right (i.e. reflection) symmetry of the lattice ensures that equal halves of a complete period are reflections of one another. This can be easily seen from the left-hand side of (2.11), where the substitution \( k \to -k \) yields \( \cos(-ka) = \cos(ka) \) because cosine is an even function. Thus, we have an additional symmetry in the energy dispersion solution:

\[
E(-k) = E(k). \tag{2.13}
\]

The foregoing observations point the way to an especially concise representation of energy bands that, while useful even for the Kronig–Penney model, becomes essential for energy bands in higher dimensions. The right panel of Fig. 2.6
Figure 2.6: (Left) The periodic solution for the energy bands in Fig. 2.2 over the range $-2\pi \leq ka \leq 2\pi$. (Right) Reduced presentation of the bands of the Kronig–Penney model, showing the bands only over the range $0 \leq ka \leq \pi$. Taking into account the $k \rightarrow -k$ symmetry and the periodicity of the dispersion curves, these bands contain the same information.

This figure displays the same bands as the left panel, but includes only the unique information provided by the dispersion curves. The periodicity in (2.12) alleviates the need to consider intervals of $ka$ greater than $2\pi$, while the reflection symmetry (2.13) reduces this interval by one-half. Hence, the interval $0 \leq ka \leq \pi$ is sufficient to represent all of the information provided by the solutions of (2.11). This is the type of diagram used to represent the energy bands of all two- and three-dimensional crystalline solids.

### 2.3.2 Dimensionality

The next modification of our description of the energy dispersion of electrons in solids is due to spatial dimension of the system being considered. In the Kronig–Penney model, which is based on a one-dimensional lattice, the electrons can move only along the positive or negative $x$-direction, so the energy dispersion re-
Figure 2.7: The dispersion surface for the energy states of an electron moving in a two-dimensional potential on a square lattice. Only the lowest energy band is displayed.

lation $E(k)$ is a function of a single wave number, $k$. Suppose that electrons can move in the $x$-$y$ plane of a periodic two-dimensional system. The energy dispersion relation now depends on the wave numbers $k_x$ and $k_y$ along the $x$- and $y$-directions $E(k_x, k_y)$. Thus, $E(k_x, k_y)$ is represented as a surface. In three dimensional, the realm of most materials, the energy dispersion is a function of the wave numbers along the $x$-, $y$-, and $z$-directions, $E(k_x, k_y, k_z)$.

An example of an energy dispersion surface for an electron moving in a two-dimensional potential on a square lattice is shown in Fig. 2.7. This diagram shows immediately why an alternative representation of energy states in higher dimensions is desirable. Only the lowest energy band is shown and, while the general qualitative features of the energy dispersion surface are apparent, quantitative information, such as the band width and band curvature and, if higher bands were also shown, band gaps, is difficult to discern directly from such a diagram. The conventional way of dealing with this is to provide a representation of energy bands in higher dimensions is based on plotting the intersections of this surface with planes along high-symmetry directions to obtain dispersion curves along those directions. This basic construction is shown in Fig. 2.8 along the $k_x$ and $k_y$ axes and along the lines specified by $k_x = k_y$ and $k_x = -k_y$ on a square lattice. We can again invoke various symmetries of the underlying lattice, as in the
Figure 2.8: The representation of energy bands for a two-dimensional square lattice based on the intersection of the energy dispersion surface with planes (shown shaded) along high-symmetry directions: (a) along the $k_x a$-axis, (b) along the $k_y a$-axis, (c) along the line $k_x = k_y$, and (d) along the line $k_x = -k_y$. The bold curve is the intersection of the energy dispersion surface with each plane. Note that, because of the symmetry of the square lattice, the dispersion curves in (a) and (b) and those in (c) and (d) are the same.

preceding section, to obtain a much more concise representation of the energy dispersion of this system. By plotting these curves for several such directions, the main quantitative features of the energy dispersion surface can be ascertained (Fig. 2.9). The main drawback of this approach is that we no longer have an exact representation of the entire energy dispersion surface, but the practical advantages outweigh this.
The calculation of energy bands represents a substantial activity in modern solid state physics and has provided the basis of much of our understanding of the electronic, mechanical, optical, and thermodynamic properties of materials at the atomic level. We begin this section by explaining how the Kronig–Penney model accounts for the difference between metals and insulators. This is one of the great triumphs of the quantum theory of solids and is due ultimately to the alternating sequence of allowed and forbidden regions of energy states, which gives rise to a ‘shell’ structure in crystalline solids analogous to that found in atoms.

The basic idea behind the occupancy of energy bands is as follows. As atoms are brought together to form a crystalline solid, discrete atomic levels are broadened into bands. The occupancy of these bands is determined by the occupancy of the atomic levels: completely (resp., partially) filled atomic levels lead to completely (resp., partially) filled energy bands (Fig. 2.10). The widths of the bands are determined by the ease with which electrons can move across the solid: narrow bands correspond to electrons that remain localized near their parent atoms.

---

3The further distinction between insulators and semiconductors is somewhat more subtle and will be discussed in Chapter 3.
## Figure 2.10: The Periodic Table of Elements showing the configurations of the outer-most shells of electrons. Partially filled electronic shells usually leads to partial filled bands in crystalline solids, which explains why most elemental solids are metals.
Figure 2.11: Schematic depiction of (a) a material with a partially-filled band, which is a metal, and (b) a material with completely filled bands, which is an insulator at $T = 0$.

and retain at least some of their atomic character, while wide bands to mobile electrons that can easily move between atomic sites.

A crystalline material that has a partially filled band is the microscopic definition of a metal [Fig. 2.11(a)]. Partial band occupancy has several immediate consequences for electronic properties. The fact that the Fermi energy lies within a band means that electrons can be excited to higher states with small energy differences (compared to $k_B T$). Metals are therefore characterized by large electrical and thermal conductivities and familiar optical properties, such as having a luster. Because most atoms have partially-filled outer shells, most elemental solids are metals. The alkali elements, which constitute the first column of the periodic table, exemplify many of these properties. They have an atomic structure of a single $s$ electron outside a closed shell and thereby form a half-filled band.

At the opposite end of the periodic table from the alkali metals are the noble gases. These atoms have filled electronic shells that form completely filled bands [Fig. 2.11(b)]. Electrons must now be excited across an energy gap separating the occupied and the lowest unoccupied bands. These gaps typically have energies of the order of electron volts, which strongly suppresses the thermal excitation of
Figure 2.12: The energy bands of crystalline lithium along the principal directions of $k$: $\Delta$ (100), $\Sigma$ (110), and $\Lambda$ (111). The bands along the three directions have been superimposed upon one another. [From F. S. Ham, ‘Energy bands of alkali metals. I. Calculated bands,’ Physical Review 128, 82–97 (1962).]

electrons and leads to electrical and thermal conductivities that are substantially lower than those of metals. In particular, their electrical conductivities vanish at $T = 0$. Solids with these properties are called **insulators**.

### 2.4.1 Lithium

The atomic number of lithium (chemical symbol Li, from the Greek *lithos*) is 3, with an atomic electronic configuration of $1s^22s^1$, that is, a half-filled outer $s$ shell. Lithium is in the first column of the Periodic Table, the elements of which are known as the alkali metals. The energy bands of the alkali metals have been interest since the early days of quantum mechanics because of the relative simplicity afforded by a single $s$ valence electron.

The energy bands of lithium are shown in Fig. 2.12. The structure of crys-
Electrons in Crystalline Solids

talline lithium is body-centered cubic, so the energy dispersion curves are shown along the high-symmetry directions in such a crystal: $\Delta$ (100), $\Sigma$ (110), and $\Lambda$ (111). Note, however, that the dispersion curves along these directions have been plotted on the same graph. The fact that they lie on top of one another is an indication of the isotropy of the $s$ valence electrons. The energy scale of the bands is shown in Rydbergs (1 Ry=13.6 eV), so the width of the broadest band is in excess of 8 eV, which also results from the spatial extent of the atomic $s$ orbitals. Finally, the Fermi energy lies at approximately 0.35 Ry, as expected for a half-filled band. This quantum mechanical description of the electronic states of lithium thus show this material to be a metal, which conforms to our everyday experience. Calculations for the other alkali metal yield qualitatively similar results.

2.4.2 Copper

The atomic number of copper (chemical symbol Cu, from the Latin cuprium) is 29 and its atomic electronic configuration is

\[ \text{Cu: } 1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^1 \]

where the core electrons retain much of their atomic character in the crystalline solid, while the valence electrons form metallic bonds (Sec. 1.5) that are responsible for the properties of crystalline copper, such as ductility, high electrical conductivity and corrosion resistance.

The energy bands of copper are shown in Fig. 2.13. Crystalline copper has a cubic-based structure, so the energy dispersion curves are shown along the high-symmetry directions in such a crystal: $\Delta$ (100), $\Sigma$ (110), and $\Lambda$ (111), which are the same as that for the body-centered lattice. The Fermi level intersects a wide band, which is consistent with the metallic behavior of copper. There are two types of energy bands revealed by the calculation. The wide ($\approx 8$ eV) bands that emanate from the lowest energy shown and intersect the Fermi level are derived from the 4s atomic levels of Cu. The narrower ($\approx 3$ eV) bands that lie several electron-volts below the Fermi level are formed mainly from the atomic 3d levels of atomic copper. From our discussion of band widths in the Kronig–Penney model, we can see that the 4s electrons can move easily through the lattice, which results in
Figure 2.13: The energy bands of crystalline copper along the principal directions of $k$: $\Delta$ (100), $\Sigma$ (110), and $\Lambda$ (111). The Fermi energy is indicated by the horizontal line near $-0.2$ Ry (1 Ry $\approx 13.6$ eV). [From B. Segall, ‘Fermi surface and energy bands of copper,’ *Physical Review* **125**, 109–122 (1962).]
the efficiency of absorption and re-emission is approximately equal at all optical energies, then the different colors in white light will be reflected equally well, leading to the ‘silvery’ color of polished silver and iron surfaces. In copper, the efficiency of reflection decreases with increasing energy. The reduced reflectivity at the blue end of the spectrum results in a reddish color.

### 2.4.3 Silicon

Silicon (chemical symbol Cu) has atomic number 14 with the atomic configuration

\[
\text{Si: } 1s^2 2s^2 2p^6 3s^2 3p^2. \\
\text{core } 3s^3 3p^3. \\
\text{valence }
\]

In the Si crystal, however, the effective configuration of the valence electrons is \(3s^3 3p^3\) to enable each Si atom to form covalent bonds with four nearest neighbors arranged in a tetrahedral configuration (Fig. 1.4). This is one of the key features for understanding not only the energy bands in pure silicon and other tetrahedrally-bonded semiconductors, but also how they respond to the implantation of foreign atoms. We will discuss this further in Chapter 3.

Silicon, the best known material of the computer age, forms approximately 25% of the earth’s crust (by weight), making it the second most abundant element, exceeded only by oxygen. Silicon is not found naturally in pure form, but occurs mainly as an oxide and as silicates. Sand, quartz, and rock crystal are some of the forms in which the oxide appears. The Czochralski process is commonly used to produce macroscopic single crystals of silicon used for semiconductor devices.

The energy bands of Si within several electron volts on either side of its band gap are shown in Fig. 2.14. The most apparent feature of these bands is the energy gap between the bands that are completely occupied at \(T = 0\) (the valence bands) and those which are empty at \(T = 0\) (the conduction bands). The top of the valence bands has been chosen as the zero of energy, so the conduction bands lie at positive energies. The valence and conduction bands are separated by an energy gap between the maximum energy of the valence bands and the bottom of the conduction bands, at approximately 1.2 eV. The most important property of this energy gap for applications of Si is that the valence and conduction band extrema occur at different points in \(k\)-space. In particular, the maximum of the valence bands is at the point \(\Gamma\) (the origin of \(k\)-space), but the conduction band
minimum lies along the (100) direction. For this reason, Si is said to have an **indirect** energy gap.

The indirect gap of Si has far-reaching consequences for the excitation of electrons between valence and conduction bands. A photon can provide the required energy to an electron in the valence band to overcome the energy gap, but an additional momentum is needed to carry this electron to the conduction band minimum (since the photon absorption produces a negligible momentum change). This additional momentum is provided by the lattice in the form of atomic vibrations (Chapter 4). Since this excitation involves two almost simultaneous processes (photon absorption and lattice excitation) the probability of its occurrence...
Electrons in Crystalline Solids

is considerably reduced compared to the case where only one process alone is required. Thus, the electronic structure of Si means that it is not suitable for light amplification, which includes optoelectronic devices that act as gateways between electronics and photonics.

In a pure semiconductor, there are no electron energy levels in the band gap between the valence and conduction bands. Thus, the lowest-energy light that can be absorbed corresponds to the energy required to excite an electron from the top of the valence band up to a level at the bottom of the conduction band. This corresponds to the band-gap energy, denoted as $E_g$. If the semiconductor has a large band gap, such as diamond, with $E_g = 5.4$ eV, then no light in the visible spectrum can be absorbed. Hence, single-crystal diamond transmits all incident and is, therefore, clear. Alternatively, all light is absorbed when $E_g$ is less than the 1.77 eV limit of the visible spectrum. Narrow band-gap semiconductors, such as the lead sulfide ($E_g = 0.37$ eV at room temperature), therefore absorb all light and are black.

### 2.5 Summary

The electronic energy states of the Kronig–Penney model provide a framework for understanding many of the qualitative features of the three-dimensional band structures of real materials. Even though the Kronig–Penney model contains a description only of the kinetic energies of the electrons, while real solids embody more complex effects of bond formation and Coulomb interactions, the basic idea of filled and partially filled bands already provides a distinction between metals and insulators. The analysis of the band structures of real materials requires accounting for the nature of the atomic states that participate in band formation. This can explain many of materials properties with computational methods which are now regarded as routine.
A.1 Appendix: Matching Conditions

A.1.1 Basic Considerations

Quantum mechanical potentials are often represented by idealized functions that are piecewise constant with discontinuities. The most common of these are square wells and step barriers for scattering calculations. The reason for using such potentials is mathematical convenience: solutions of Schrödinger’s equation are much easier to obtain than for “real” potentials. The solution of Schrödinger’s equation with a discontinuous potential proceeds by determining the solutions within each region where the potential is continuous then applying appropriate matching conditions at the discontinuities to determine the allowed energies and the explicit form of the wavefunctions. An extreme example of a discontinuous potential is the Dirac δ-function, which is often used to model atomic potentials in molecules and solids.

The properties of an electron moving through any potential is based on obtaining solutions of Schrödinger’s equation supplemented by two conditions:

- **Continuity of the wavefunction.** This is one of the fundamental constraints on the wavefunction of any physical system and must be observed even at points of an infinite discontinuity of the potential. Continuity is required to ensure that the probabilistic interpretation of the wavefunction is unambiguously defined at every point.

- **Continuity of the derivative of the wavefunction, except at an infinite discontinuity in the potential.** This condition is required because a finite discontinuity in the first derivative of the wavefunction implies an infinite discontinuity in the second derivative, and therefore an infinite discontinuity of the kinetic energy term in the Schrödinger equation, which is physically unrealistic.

---

4The Dirac δ-function, denoted by \( \delta(x) \), was introduced by the British physicist Paul Dirac to approximate narrow “spike” functions. Although not formally a true function, the δ-function may be regarded as a quantity that has the value of infinity at \( x = 0 \), zero elsewhere, such that the integral is unity. The δ-function is the same level of abstraction as a point mass, a point charge, and a sharp impulse. See M. Boas, Mathematical Methods in the Physical Sciences 2nd edition (Wiley, New York 1983), pp. 665–670.
A.1.2 Matching Conditions for Different Potentials

Consider Schrödinger’s equation for a particle of mass $m$ moving in a potential $V(x)$,

$$-\frac{\hbar^2}{2m} \frac{d^2 \psi}{dx^2} + V(x) \psi = E \psi,$$  \hspace{1cm} (A.1)

which we rearrange as

$$\frac{d^2 \psi}{dx^2} = \frac{2m}{\hbar^2} \left[ V(x) - E \right] \psi.$$  \hspace{1cm} (A.2)

The left-hand side of this equation is the rate of change of the slope of $\psi$, which is the curvature of the wavefunction. To characterize the behavior of $\psi$ at a point $x_0$, we integrate both sides of this equation over a small interval surrounding $x_0$, from $x_0 - \Delta x$ to $x_0 + \Delta x$:

$$\int_{x_0 - \Delta x}^{x_0 + \Delta x} \frac{d^2 \psi}{dx^2} \, dx = \frac{d\psi}{dx} \bigg|_{x_0 + \Delta x} - \frac{d\psi}{dx} \bigg|_{x_0 - \Delta x}$$

$$= \frac{2m}{\hbar^2} \int_{x_0 - \Delta x}^{x_0 + \Delta x} V(x) \psi(x) \, dx - \frac{2mE}{\hbar^2} \int_{x_0 - \Delta x}^{x_0 + \Delta x} \psi(x) \, dx.$$  \hspace{1cm} (A.3)

The left-hand side of this equation represents the difference in the derivative of $\psi$ at $x_0$ as this point is approached from the right ($x + \Delta x$) and from the left ($x - \Delta x$). For the derivative $\psi'(x_0)$ to be well-defined, these two quantities must be equal.

Since the wavefunction is always continuous, the second term on the right-hand side vanishes in this limit, so we need not consider this term further in the following discussion. The nature of the derivative at $x_0$ therefore depends on the value of the first integral on the right-hand side, which is determined by the behavior of the potential at $x_0$:

$$\left. \frac{d\psi}{dx} \right|_{x_0 + \Delta x} - \left. \frac{d\psi}{dx} \right|_{x_0 - \Delta x} = \frac{2m}{\hbar^2} \int_{x_0 - \Delta x}^{x_0 + \Delta x} V(x) \psi(x) \, dx.$$  \hspace{1cm} (A.4)

If this integral vanishes as $\Delta x \to 0$, then the left and right derivatives of $\psi$ are equal, which means that $\psi'$ is continuous at $x_0$. Alternatively, if this integral has a finite value as $\Delta x \to 0$, there is a discontinuity in the derivative of $\psi$, which is called a “cusp”. There are three cases to consider.
A.1.3 Continuous Potentials

We first consider a potential which is continuous at $x_0$, as shown in Fig. A.4. All physical potentials are continuous. Then, upon taking the limit $\Delta x \to 0$ in (A.4), the integral on the right-hand side vanishes because the integrand is finite and the range of integration is reduced to zero in this limit. Hence, the left and right derivatives of $\psi$ at $x_0$ are equal, which means that the derivative of $\psi$ is continuous at $x_0$:

$$\lim_{\Delta x \to 0} \left( \frac{d\psi}{dx} \bigg|_{x_0+\Delta x} - \frac{d\psi}{dx} \bigg|_{x_0-\Delta x} \right) = 0,$$  

(A.5)

This result implies that wavefunctions for physical potentials are continuous and differentiable.

A.1.4 Finite Discontinuous Potentials

Discontinuous potentials, such as finite steps, are used as the boundaries of finite square wells and step barriers for scattering calculations. An example of such a potential is shown in Fig. A.5, where $V = V_1$ for $x < x_0$ and $V = V_2 > V_1$ for $x > x_0$. This potential is discontinuous at $x_0$, but the magnitude of the discontinuity, $V_2 - V_1$, is finite. Upon substitution of this potential into (A.4), we obtain

$$\frac{d\psi}{dx} \bigg|_{x_0+\Delta x} - \frac{d\psi}{dx} \bigg|_{x_0-\Delta x} = \frac{2mV_1}{\hbar^2} \int_{x_0-\Delta x}^{x_0} \psi(x) \, dx + \frac{2mV_2}{\hbar^2} \int_{x_0}^{x_0+\Delta x} \psi(x) \, dx.$$

(A.6)

Figure A.4: A potential that is continuous at $x_0$.

Figure A.5: A discontinuous finite potential at $x_0$. 

Electrons in Crystalline Solids
Taking the limit $\Delta x \to 0$ yields

$$\lim_{\Delta x \to 0} \left( \frac{d\psi}{dx} \bigg|_{x_0+\Delta x} - \frac{d\psi}{dx} \bigg|_{x_0-\Delta x} \right) = 0. \quad (A.7)$$

Thus, the derivative of $\psi$ remains continuous even across a finite discontinuity of the potential. The solution of Schrödinger’s equation for such potentials proceeds by obtaining solutions within each region where the potential is constant and then matching these separate solutions for $\psi$ and its derivative at the points where $V$ is discontinuous. This yields the allowed energies of the wavefunctions and produces continuous and differentiable wavefunctions throughout the entire system.

### A.1.5 Infinite Discontinuous Potentials

Potentials with infinite discontinuities are perhaps best known as the boundaries of infinite square wells (Sec. 2.1). Our main interest in this chapter, however, is in another type of infinite discontinuity of the potential: the $\delta$-function. An example of such a potential is shown schematically in Fig. (A.7). For substitution into (A.4), we need the following integral, which is one of the key properties of the $\delta$-function:

$$\int_{x_0-a}^{x_0+a} \delta(x-x_0)f(x)\,dx = f(x_0), \quad (A.8)$$

for any $a > 0$. Then, with the potential given by a $\delta$-function of strength $V_0$,

$$V(x) = V_0\delta(x-x_0), \quad (A.9)$$

Eq. (A.3) becomes

$$\frac{d\psi}{dx} \bigg|_{x_0+\Delta x} - \frac{d\psi}{dx} \bigg|_{x_0-\Delta x} = \frac{2mV_0}{\hbar^2} \psi(x_0). \quad (A.10)$$
Taking the limit $\Delta x \to 0$ yields

$$
\lim_{\Delta x \to 0} \left( \frac{d\psi}{dx} \bigg|_{x_0 + \Delta x} - \frac{d\psi}{dx} \bigg|_{x_0 - \Delta x} \right) = \frac{2mV_0}{\hbar^2} \psi(x_0).
$$

Thus, the derivative of $\psi$ is *discontinuous* across an infinite discontinuity of the potential, i.e. the wavefunction has a *cusp* there. The wavefunction for a particle moving in a region of a $\delta$-function potential is obtained by matching the solutions of Schrödinger’s equation on either side of the $\delta$-function using the continuity of $\psi$ and the “cusp condition” in Eq. (A.11).
Chapter 3

Electrons and Holes in Semiconductors

Gaps in the energy bands of crystalline materials provide a way to distinguish between metals and insulators. In a metal, the Fermi level lies within a band, so electrons can be excited from just below to just above this energy. Good elemental conductors, such as copper, silver, and gold, can conduct electricity with little resistance because they have a small number (no more than three) of valence electrons which are only loosely bound to the atom because of metallic binding (Sec. 1.5) and can easily move to generate an electrical current in the presence of an applied electric field. On the other hand, if the energy bands are completely filled, there are no free electrons in the material. Thus, insulators typically have large band gaps, which impedes the excitation of electrons from occupied to unoccupied levels, resulting in poor electrical conductance.

As their name implies, the electrical conductivity of semiconductors lies between those of metals and insulators. At $T = 0$, a semiconductor has completely filled bands (the valence bands) separated by an energy gap from the unoccupied bands (the conduction bands), just as an insulator. Indeed, at $T = 0$, a semiconductor is an insulator. But, unlike insulators, semiconductors have much smaller band gaps, of the order of an electron volt or smaller, which enables electrons to be thermally excited as the temperature is raised. In contrast to metals, whose electrical conductivity decreases with increasing temperature because of thermal vibrations of the lattice, the electrical conductivity of semiconductors increases because of the increasing number of electrons excited into the conduction bands.
However, the difference between metals and insulators is not the size of the band gap – there is no energy cut-off that separates the two types of material. The term ‘semiconductor’ is applied to a material in which the electrical conductivity, when extremely pure, rises exponentially with temperature, as noted above, and may be increased from its low ‘intrinsic’ value by many orders of magnitude by controlled ‘doping’ with electrically active impurities. The ability to control the electrical conductivity of semiconductors is they key to understanding their impact on modern information technology.

This chapter is devoted to the electrical properties of semiconductors and, in particular, how their electrical properties can be manipulated. We begin with a discussion about the densities of electrons and holes in intrinsic semiconductors. This provides the basis for examining the effects of dopants on the concentrations of carriers. The mechanism by which dopants affect carrier densities is best understood in terms of the tetrahedral bonding in semiconductors such as Si, which provides many insights into the nature of processes such as electronic excitation, electron and hole creation, and doping. We conclude with an examination of how doping affects the temperature dependence of carrier densities in semiconductors.

### 3.1 Intrinsic Semiconductors

At $T = 0$ a semiconductor is an insulator with all valence bands occupied and all conduction bands empty. As $T$ increases from absolute zero electrons can acquire enough thermal energy from the lattice to be excited into the conduction bands, leaving an empty state – a ‘hole’ – behind in the valence band. If the semiconductor has a direct band gap, then this process depends only on the temperature. But if the gap between valence is indirect, i.e. occurs in different regions of $k$-space, then additional momentum must be available (from lattice vibrations) to the electron for this excitation to occur. Therefore, the excitation process in indirect band-gap semiconductors is a less frequent process than in direct band-gap semiconductors and focus on the latter case.

The occupation probability of the conduction bands will be seen to depend exponentially on the temperature, so the most important region of the band diagram is near the valence and conduction band extrema. Figure 3.1 shows the region of interest along with the corresponding energies. The number $N$ of electrons in the conduction band is calculated by integrating the product of the density of states
in the conduction band, $D_c(E)$, and the occupation probability, which is given by the Fermi function $f(E)$. Using the energy scale in Fig. 3.1, we have

$$N = \int_{E_c}^\infty D_c(E) f(E) \, dE = \int_{E_c}^\infty \frac{D_c(E) \, dE}{e^{(E-\mu)/k_BT} + 1}. \quad (3.1)$$

By extending the integral to infinity, we are acknowledging that, for $T > 0$, there is a probability, however small, any state can be occupied.

The number $P$ of holes in the valence band is calculated similarly, by integrating the product of the density of states in the valence band $D_v(E)$ and the probability that the state at energy $E$ is unoccupied, which is $1 - f(E)$:

$$P = \int_{E_0}^{E_v} D_v(E) \left[1 - f(E)\right] \, dE = \int_{E_0}^{E_v} \frac{D_v(E) \, dE}{e^{(\mu-E)/k_BT} + 1}, \quad (3.2)$$

where $E_0$ is the energy at the bottom of the conduction band.
In perfect intrinsic semiconductors, i.e. those without defects or dopants, every electron in the conduction band corresponds to a hole in the valence band. Thus,

\[ N = P, \]

which determines \( \mu \) and thereby allows \( N \) and \( P \) to be determined as a function of temperature. However, unless we can make some simplifications to the expressions in (3.1) and (3.2), these calculations must be done numerically.

### 3.1.1 Fermi Function for Large Band Gaps

Intrinsic semiconductors are not used in devices because the conductivity at room temperature, the typical operating temperature for semiconductors, is quite small because of the low density of electrons that have been excited into the conduction band. This observation allows us to approximate the factors in (3.1) and (3.2) to enable an analytic evaluation of the integrals.

We first consider the Fermi factors by estimating the chemical potential. The thermodynamic definition of the chemical potential,

\[ \mu = \left( \frac{\partial U}{\partial N} \right)_{S,V}, \]

measures the change in the internal energy of a system under the change of particle number at constant entropy and volume. The chemical potential has the same effect on particle motion as the temperature does for heat flow. In particular, the temperature and chemical potential must be constant in a system at thermal equilibrium. The subscripts on the partial derivatives in (3.4) are standard thermodynamic notation to indicate that the entropy \( S \) and volume \( V \) are held constant. The restriction to constant volume means that the energy levels cannot change, while the restriction to constant entropy means dictates the choice of where electrons are added or removed from the system. We will approximate the derivative of the internal energy as the average of the left- and right-hand finite differences (a common numerical procedure):

\[ \left( \frac{\partial U}{\partial N} \right)_{S,V} \approx \frac{1}{2} \left\{ [U(N) - U(N - 1)] + [U(N + 1) - U(N)] \right\} \bigg|_{S,V}, \]
where $U(N)$ is the internal energy of a semiconductor with $N$ electrons. If we confine ourselves to $T = 0$, these expressions are straightforward to evaluate. An electron removed from the system under constant $S$ and $V$ must be taken from the top of the valence band, where there is a single state at $E = E_v$. Thus,

$$U(N - 1) = U(N) - E_v.$$  

Similarly, an electron added to the system under constant $S$ and $V$ must be added to the bottom of the conduction band, where there is a single state at $E = E_c$. Thus,

$$U(N + 1) = U(N) + E_c,$$

and (3.5) yields

$$\mu = \frac{1}{2}(E_v + E_c) = E_v + \frac{1}{2}E_g,$$  \hspace{1cm} (3.6)

i.e. at $T = 0$, the chemical potential is in the center of the energy gap that separates the valence and conduction bands. If the energy gap is of the order of 1 eV, and we confine ourselves to temperatures near room temperature, where $k_B T \approx \frac{1}{40}$ eV, then

$$E_c - \mu \gg k_B T, \quad \mu - E_v \gg k_B T,$$

that is,

$$\frac{E_c - \mu}{k_B T} \gg 1, \quad \frac{\mu - E_v}{k_B T} \gg 1.$$  

Under these conditions, the arguments of the exponentials in (3.1) and (3.2) are large, so the Fermi functions may be approximated by Boltzmann functions:

$$\frac{1}{e^{(E-\mu)/k_B T} + 1} \approx e^{-(E-\mu)/k_B T},$$  \hspace{1cm} (3.7)

$$\frac{1}{e^{(\mu-E)/k_B T} + 1} \approx e^{-(\mu-E)/k_B T}.$$  \hspace{1cm} (3.8)

This implies that the number of excited electrons is so small that the Pauli exclusion principle plays no appreciable role in determining their motion. In effect, therefore, these electrons are described by classical statistics.
3.1.2 Densities of States near Band Edges

The other approximation commonly made in evaluating (3.1) and (3.2) follows from the preceding discussion of the Fermi factors. Because of the low densities of the electrons and holes, the changes to the occupancies of the valence and conduction bands are confined to a small energy region around the band extrema. As we can see from Fig. 3.1, the approximation of parabolic bands with appropriately-chosen effective masses provides an adequate account of the bands near their extrema. Accordingly, since the Boltzmann factors suppress the occupancy of energies far away from the band extrema, we can approximate the densities of states in the valence and conduction bands by that for free-electrons (with effective masses) for all energies, safe in the knowledge that the errors away from the band extrema are made small by the fact that these states are effectively excluded from the excitation process.

The density of states for a three-dimension system of electrons is

\[ D(E) = \frac{V}{2\pi^2} \left( \frac{2m^*}{\hbar^2} \right)^{3/2} E^{1/2}, \]

where \( V = L^3 \) is the volume of the system. We need to make two modifications to this expression for substitution into (3.1) and (3.2): (i) the zeros of energy, which are now at the band edges, and (ii) replacing the free-electron mass \( m \) with the effective masses \( m^*_v \) and \( m^*_c \) for the valence and conduction bands, respectively. The approximate densities of states in the valence and conduction bands are therefore given by

\[ D_v(E) = \frac{V}{2\pi^2} \left( \frac{2m^*_v}{\hbar^2} \right)^{3/2} (E_v - E)^{1/2}, \quad (3.9) \]

\[ D_c(E) = \frac{V}{2\pi^2} \left( \frac{2m^*_c}{\hbar^2} \right)^{3/2} (E - E_c)^{1/2}, \quad (3.10) \]

3.1.3 Electron and Hole Densities

We are now in a position to evaluate the numbers of electrons and holes as a function of temperature. Upon substitution of (3.7), (3.8), (3.9), and (3.10) into
Eqs. (3.1) and (3.2) we obtain

$$N = \frac{V}{2\pi^2} \left( \frac{2m^*_c}{\hbar^2} \right)^{3/2} \int_{E_c}^{\infty} (E - E_c)^{1/2} e^{-(E - \mu)/k_BT} dE , \quad (3.11)$$

$$P = \frac{V}{2\pi^2} \left( \frac{2m^*_e}{\hbar^2} \right)^{3/2} \int_{E_0}^{E_v} (E - E)^{1/2} e^{-(\mu - E)/k_BT} dE . \quad (3.12)$$

We consider the integral in (3.11) first. By changing the integration variable to

$$x = \frac{E - E_c}{k_BT} ,$$

we have that

$$E = E_c + xk_BT , \quad dE = k_BT \, dx , \quad (3.13)$$

whereupon the lower limit of integration \((E = E_c)\) becomes \(x = 0\) and the upper limit of integration \((E \to \infty)\) \(x \to \infty\). The transformed integral is, therefore,

$$\left( k_BT \right)^{3/2} e^{-(E_c - \mu)/k_BT} \int_0^\infty x^{1/2} e^{-x} \, dx . \quad (3.14)$$

The transformation of the integral in (3.12) proceeds along similar lines, with the result

$$\left( k_BT \right)^{3/2} e^{-(\mu - E_0)/k_BT} \int_0^{(E_v - E_0)/k_BT} x^{1/2} e^{-x} \, dx .$$

The upper limit of integration is the valence band width \(E_v - E_0\) divided by the thermal energy \(k_BT\). Since the band widths of semiconductors are typically several electron-volts, and, at room temperature, \(k_BT \sim \frac{1}{30} \text{eV}\), we can confidently take the upper limit of integration to infinity. The error made in doing so is strongly reduced by the negligible occupancy of levels at such high energies because of the exponential factor. Thus, the integral to be evaluated is

$$\left( k_BT \right)^{3/2} e^{-(\mu - E_0)/k_BT} \int_0^\infty x^{1/2} e^{-x} \, dx . \quad (3.15)$$

The remaining integral in (3.14) and (3.15) can be evaluated by transforming the integration variable from \(x\) to \(x = s^2\) and performing an integration by parts to leave a Gaussian integral. We leave the details as an exercise and simply quote the result:

$$\int_0^\infty x^{1/2} e^{-x} \, dx = \frac{1}{2} \pi^{1/2} .$$
Thus, the numbers of electrons and holes in (3.1) and (3.2) become

\[
N = \frac{V}{4} \left( \frac{2m_e^* k_B T}{\pi \hbar^2} \right)^{3/2} e^{-\left( E_c - \mu \right)/k_B T}, \quad (3.16)
\]

\[
P = \frac{V}{4} \left( \frac{2m_v^* k_B T}{\pi \hbar^2} \right)^{3/2} e^{-\left( \mu - E_v \right)/k_B T}. \quad (3.17)
\]

We can use the relation (3.3) to solve for \( \mu \), and therefore, obtain the numbers (or densities) of electrons and holes. Equating (3.16) and (3.17) and cancelling common factors yields

\[
m_e^{3/2} e^{-\left( E_c - \mu \right)/k_B T} = m_v^{3/2} e^{-\left( \mu - E_v \right)/k_B T}.
\]

This equation can solved for \( \mu \) to obtain

\[
\mu = E_v + \frac{1}{2} E_g + \frac{3}{4} k_B T \ln \left( \frac{m_v^*}{m_e^*} \right), \quad (3.18)
\]

which yields an expression for the densities \( N/V \) and \( P/V \) of electrons and holes:

\[
n = p = 2 \left( \frac{2\pi k_B T}{\hbar^2} \right)^{3/2} \left( m_v^* m_e^* \right)^{3/4} e^{-E_g/2k_B T}. \quad (3.19)
\]

The exponential factor, which is due to the Boltzmann distribution, provides the dominant contribution of the temperature-dependence of the carrier densities of semiconductors, with the factor involving \( T^{3/2} \), which is due in part to the density of states, making a weaker contribution. Figure 3.2 shows the intrinsic carrier densities of Si, Ge, and GaAs over a wide temperature range on a \( \ln n = \ln p \) versus \( T^{-1} \) plot. Upon taking the logarithm of our expression for \( n = p \),

\[
\ln n = \ln p = \ln 2 + 3 \left( m_v^* m_e^* \right)^{1/2} \ln \left( \frac{2\pi k_B T}{\hbar^2} \right) - \frac{E_g}{2k_B T}, \quad (3.20)
\]
we see that the dominant contribution is indeed due to the exponential term in (3.19), so that the carrier densities should appear as approximately straight lines on such a plot with a slope that becomes more steep as $E_g$ increases. Indeed, as Fig. 3.2 clearly indicates, the slopes conform to that expected from the band gaps of the three semiconductors: 1.424 eV (direct) for GaAs, 1.124 eV (indirect) for Si, and 0.664 (indirect) for Ge.

We return to consider the temperature dependence of the chemical potential in (3.18). Notice that, at $T = 0$, $\mu = E_v + \frac{1}{2}E_g$, that is, $\mu$ lies midway between the valence and conduction band edges, in agreement with (3.6). For $T > 0$, $\mu$ may ei-
Electrons and Holes in Semiconductors

\[ \mu(T) = \mu(0) \]

\[ \mu(T) = \mu(0) \]

\[ \mu(T) = \mu(0) \]

Figure 3.3: Schematic illustration of the effect of differing effective masses \( m_v^* \) and \( m_c^* \) in the valence and conduction bands of a semiconductor. (a) \( m_v^* > m_c^* \), (b) \( m_v^* < m_c^* \), and (c) \( m_v^* = m_c^* \).

ther increase or decrease, depending on the relative magnitudes of \( m_v^* \) and \( m_c^* \). We can understand this from a physical standpoint as follows. We again take the definition of the chemical potential as the difference in the electronic energy caused by the simultaneous creation of a hole in the valence band and an electron in the conduction band, as in (3.5). If \( m_v^* > m_c^* \), then the curvature of the conduction band is greater than that of the valence band (Sec. 2.2.2). Thus, for given (equal) occupancies electrons and holes, the lowest available state in the conduction band lies further away from the band edge than the highest occupied state in the valence band, and the chemical potential is increased to greater values than \( \frac{1}{2}E_g \), as shown schematically in Fig. 3.3(a). If \( m_v^* < m_c^* \), the arguments are reversed and the chemical potential is decreased to lower values than \( \frac{1}{2}E_g \) [Fig. 3.3(b)]. But, in either case, this is often not a huge effect, and the chemical potential is frequently set equal to its value at \( T = 0 \).

### 3.2 Extrinsic Semiconductors

The property of semiconductors that is most useful for electronic devices is that their conductivity may easily be modified by introducing impurities into their crystal lattices. The process of adding controlled impurities to a semiconductor is known as doping. Doping provides a method for precisely controlling the density of electrons in the valence band and/or holes in the conduction band. Semiconductors in which dopants are deliberately introduced are known as extrinsic semicon-
Electrons and Holes in Semiconductors

Dopants dictate the properties of extrinsic semiconductors. For example, a dopant concentration of 1 atom in $10^{12}$ is enough to make silicon extrinsic at room temperature! Almost all commercial semiconductors are extrinsic. Before calculating the effect of dopants on carrier densities, we discuss the conceptual basis for dopant incorporation into the lattice of semiconductors.

### 3.2.1 Hybridization

The bonding in most semiconductors is either covalent or a mixture of covalent and ionic (Sec. 1.5). For example, Si, with 14 electrons, is covalently bonded. Each silicon atom is bonded to 4 others in a tetrahedral bond, which leads to the diamond cubic crystal structure [Fig. 1.3(b)]. The electronic structure of atomic Si is $1s^22s^22p^63s^23p^2$. The angular parts of the $3s$ and $3p$ orbitals, which are obtained from the real and imaginary parts of the corresponding spherical harmonics, are shown in Fig. 3.4. The $s, p_x, p_y,$ and $p_z$ states can be combined to form directional orbitals that form covalent bonds along the directions of a tetrahedron (Fig. 1.4):

$$
\begin{align*}
\varphi_{(111)}(x) &= \frac{1}{2}(s + p_x + p_y + p_z), \\
\varphi_{(\bar{1}11)}(x) &= \frac{1}{2}(s - p_x - p_y + p_z), \\
\varphi_{(1\bar{1}1)}(x) &= \frac{1}{2}(s - p_x + p_y - p_z), \\
\varphi_{(11\bar{1})}(x) &= \frac{1}{2}(s + p_x - p_y - p_z),
\end{align*}
$$

where we have used the standard abbreviation $\bar{I} \equiv -1$. When the four Si atoms create tetrahedral covalent bonds with these orbitals (Fig. 3.5), the $3s$ and $3p$ electrons from a new set of hybrid orbitals in which the effective configuration is $3s3p^3$, which is usually abbreviated to $sp^3$. The promotion of a $3s$ electron to the $3p$ energy level costs energy, but that energy is more than compensated by the formation of strong, direction covalent bonds in crystalline silicon. Such tetrahedral bonds are highly directional and form the basis of the valence bands in covalently-bonded semiconductors. High temperatures or another energy source (such as a photon) is needed to remove an electron from this strong covalent bond. When an electron leaves the valence band to go to the conduction band, it is essentially breaking free of its covalent bond, leaving behind an empty state for another electron in the valence band energy levels. This empty state is a ‘hole,’ as dis-
Electrons and Holes in Semiconductors

Figure 3.4: The absolute values of the angular parts of the 3s and 3p orbitals of Si, obtained from the spherical harmonics, as indicated. The s state is $|Y_{0,0}(\theta, \phi)|$, $p_x$ is $\text{Re}(Y_{1,1})$, $p_y$ is $\text{Im}(Y_{1,1})$, and $p_z$ is $Y_{1,0}$.

This explains why such semiconductors have relatively low carrier densities, which leads to low electrical conductivities, unless dopants are added. Because of the directionality of the covalent bond, atoms in a covalent solid cannot be easily displaced from their equilibrium positions, making covalent solids hard and brittle.

3.2.2 Donors

The basic mechanism by which donors affect the electronic states in a semiconductor is shown in Fig. 3.6. An atom with one more electrons than an atom of an intrinsic semiconductor (e.g. As in the case of Ge) is substituted into the lattice by
a process known as implantation. The dopant can form the four bonds required of the original atom, but there is one electron remaining. This electron is still bound to the dopant, but, because of the relatively large static dielectric constants of most semiconductors ($\varepsilon \sim 10$), not as tightly as in the atomic state. This extends the wave function in space and, more importantly, reduces the binding energy to the dopant to $\sim 0.1 \, \text{eV}$ (Fig. 3.7). Thus, the dopant can be ‘ionized’ relatively easily, thereby releasing the electron into the conduction band without the concomitant creation of a hole in the valence band. Because of the ease with which this process
4fl

Electrons and Holes in Semiconductors

Figure 3.6: Schematic illustration of a donor atom in an extrinsic semiconductor, shown unshaded. The donor forms the required bonds, leaving a loosely-bound electron, whose extended wave function is indicated by shading. This electron can be released from the donor at relatively low temperatures.

occurs, at low temperatures, electrons released from donors dominate the number of electrons created by the intrinsic excitation from the valence bands.

3.2.3 Electrons in Extrinsic Semiconductors

To calculate the number of electrons in an extrinsic semiconductor with $N_d$ donors, we first observe that the probability that the donor state is occupied is $f(E_d)$, where $f$ is the Fermi function and $E_d$ is the donor energy level. We denote the number of ‘neutral’ donors by $N_d^0$, so we have that

$$N_d^0 = N_d f(E_d) = \frac{N_d}{e^{(E_d-\mu)/k_BT} + 1}.$$ 

The number of ‘ionized’ donors $N_d^+$, whose electron has been released into the conduction band, is then given by

$$N_d^+ = N_d - N_d^0 = \frac{N_d}{e^{(\mu-E_d)/k_BT} + 1}.$$ 

(3.21)
Figure 3.7: The energy $E_d$ of a donor level in relation to the band edges of an extrinsic semiconductor. This is the energy needed to release the loosely-bound electron from the donor (Fig. 3.6) into the conduction band. At low energies, this process dominates the excitation of electrons from the valence bands.

The basic relation analogous to (3.3) is obtained by observing that the electrons in the conduction band are created either by the release of an electron from a donor, or by the excitation from the valence band, thereby creating a hole. Hence, we can write

$$N = N_d^+ + P, \quad (3.22)$$

where $N_d^+$ is given by (3.21) and $N$ and $P$ are given by (3.16) and (3.17). This equation determines the chemical potential and, therefore, the densities of all the quantities involved.

We work in a regime where the excitations from the valence of the valence band are neglected, that is, $N \approx N_d^+$, but we retain the full Fermi–Dirac distribution, so the equation which determines $\mu$ is

$$N_0(T)e^{-(E_c-\mu)/k_BT} \approx \frac{N_d}{e^{(\mu-E_d)/k_BT} + 1}, \quad (3.23)$$
where

\[ N_0(T) = \frac{V}{4} \left( \frac{2m^*_e k_B T}{\pi \hbar^2} \right)^{3/2} , \]

as in (3.16). Equation (3.23) can be arranged as

\[ \left[ e^{(\mu - E_d)/2k_B T} + 1 \right] e^{-(E_c - \mu)/k_B T} = N_d , \]

which can be written as a quadratic equation for \( e^{\mu/k_B T} \):

\[ \left[ N_0(T) e^{-(E_c + E_d)/k_B T} \right] e^{2\mu/k_B T} + \left[ N_0(T) e^{-E_c/k_B T} \right] e^{\mu/k_B T} - N_d = 0 . \]

By using the quadratic formula

\[ x = \frac{1}{2a} \left( -b \pm \sqrt{b^2 - 4ac} \right) , \]

in which \( x = e^{\mu/k_B T} \), and

\[ a = N_0(T) e^{-(E_c + E_d)/k_B T} , \]

\[ b = N_0(T) e^{-E_c/k_B T} , \]

\[ c = N_d , \]

we obtain

\[ e^{\mu/k_B T} \]

\[ = \frac{1}{2N_0 e^{-(E_c + E_d)/k_B T}} \left\{ -N_0 e^{-E_c/k_B T} + \left[ N_0^2 e^{-2E_c/k_B T} + 4N_0 N_d e^{-(E_c + E_d)/k_B T} \right]^{1/2} \right\} \]

\[ = \frac{1}{2e^{-(E_c + E_d)/k_B T}} \left\{ -e^{-E_c/k_B T} + \left[ e^{-2E_c/k_B T} + 4N_d e^{-(E_c + E_d)/k_B T} \right]^{1/2} \right\} \]

\[ = \frac{1}{2} \left\{ -e^{E_d/k_B T} + \left[ e^{2E_d/k_B T} + \frac{4N_d}{N_0} e^{(E_c + E_d)/k_B T} \right]^{1/2} \right\} , \]

where we have suppressed the temperature dependence of \( N_0 \) and we have taken the positive sign of the square root because \( e^{\mu/k_B T} \) must be a positive quantity.
With this solution, the expression for the electron density is

\[ N(T) = N_0 e^{(\mu - E_c)/k_B T} \]

\[ = \frac{N_0}{2} \left\{ -e^{(E_d - E_c)/k_B T} + \left[ e^{2(E_d - E_c)/k_B T} + \frac{4N_d}{N_0} e^{(E_d - E_c)/k_B T} \right]^{1/2} \right\}. \]

In fact, we can write this result in a simpler form by factoring out the term \( e^{(E_d - E_c)/k_B T} \) from the square root:

\[ N(T) = \frac{1}{2} N_0(T) e^{-(E_c - E_d)/k_B T} \left\{ \frac{1}{2} + \frac{2N_d}{N_0(T)} \frac{e^{(E_c - E_d)/k_B T}}{e^{(E_d - E_c)/k_B T}} \right\}^{1/2} \cdot (3.24) \]

In the limit of high temperatures, we use the expansion \( \sqrt{1+x} \approx 1 + \frac{1}{2}x \) to write the square root in (3.24) as

\[ \left[ 1 + \frac{2N_d}{N_0(T)} \frac{e^{(E_c - E_d)/k_B T}}{e^{(E_d - E_c)/k_B T}} \right]^{1/2} \approx 1 + \frac{2N_d}{N_0(T)} e^{(E_c - E_d)/k_B T}. \]

Substitution of this result into (3.2) yields

\[ N(T) = \frac{1}{2} N_0(T) e^{-(E_c - E_d)/k_B T} \left[ \frac{2N_d}{N_0(T)} e^{(E_c - E_d)/k_B T} \right] = N_d, \]

so all of the donors are ionized.

In plotting \( N(T) \) in (3.24), we have used \( m = 9.11 \times 10^{-31} \) kg for the mass of the electron, \( N_d = 10^{15} \) cm\(^3\) for the donor concentration, and \( E_c - E_d = 0.06 \) eV for the energy difference between the bottom of the conduction band and the donor level. The result is shown in Fig 3.8. We have also shown the corresponding plot obtained by replacing the Fermi function in (3.23) by its classical limit, i.e.

\[ \frac{1}{e^{(\mu - E_d)/k_B T} + 1} \approx e^{-(\mu - E_d)/k_B T}, \]

which leads to the carrier density

\[ N(T) = [N_d N_0(T)]^{1/2} e^{-(E_c - E_d)/2k_B T}. \cdot (3.25) \]
Figure 3.8: Electron density (3.24) for an extrinsic semiconductor with a concentration $N_d = 10^{15}$ cm$^3$ of donors (solid line) shown with the corresponding curve (3.25) obtained with the Fermi function replaced by its classical limit. Note that the full form of the Fermi function is needed to describe the saturation of the electron density, when all of the donors have been ionized.

Both curves agree at low temperatures, but significant differences appear as the temperature is raised. In particular, the plot of (3.24) shows that the electron density saturates, as all of the donors become ionized, while the curve with the approximate Fermi function continues to rise, which is clearly an unphysical result because this theory is valid only at low temperatures where the density of electrons is low.

Figure 3.9 shows the electron density for a Si sample with a concentration of dopants of $10^{15}$ cm$^{-3}$ presented as in Fig 3.2 as a $\ln n$ versus $T^{-1}$ plot over an
If electron (hole) density is measured as a function of temperature in a doped semiconductor, one observes three regimes:

- **Freezeout**: Temperature is too small to ionize the donors (acceptors), i.e., $k_BT < E_c - E_D$ ($k_BT < E_D - E_V$).
- **Saturation**: Most of the donors (acceptors) are ionized.
- **Intrinsic**: Temperature is so high that $n_i >$ doping density.

Electron density as a function of temperature for a Si sample with donor impurity concentration of $10^{15}$ cm$^{-3}$. It is preferable to operate devices in the saturation region where the free carrier density is approximately equal to the dopant density.

It is not possible to operate devices in the intrinsic regime, since the devices always have a high carrier density that cannot be controlled by electric fields. Every semiconductor has an upper temperature beyond which it cannot be used in devices. The larger the bandgap, the higher the upper limit.

---

Figure 3.9: Electron density for a Si sample with a donor concentration of $10^{15}$ cm$^{-3}$ presented as a ln $n$ versus $T^{-1}$ plot over an extended temperature range. [From J. Singh, *Semiconductor Devices: Basic Principles* (John Wiley, New York, NY, 2001).]
temperature beyond which it cannot be used in devices: the larger the band gap, the higher this upper limiting temperature.

Comparison between Figs. 3.8 and 3.9 shows that the electron density in (3.24) reproduces all of the main features of the experimental data associated with the donors. The ‘freeze-out’ regime is seen corresponds to a range of temperatures where the Fermi function is adequately described by its classical counterpart, the Boltzmann factor, so the expression (3.25) provides an adequate account of the experimental data. However, this is a crude approximation and fails to describe the onset of ionization of the impurities. The low temperature regime through to saturation are accurately described by (3.24), which means that the full form of the Fermi function must be used, though we have still neglected the excitation of electrons from the valence to the conduction band. Reproducing the freeze-out, saturation, and intrinsic regimes necessitates using (3.22) with no approximation to the Fermi functions. This results in a cubic equation for $e^{\beta \mu}$ which must be solved numerically.

### 3.3 Summary

The main results of this chapter relate to the calculation of carrier densities in intrinsic and extrinsic semiconductors. In the intrinsic case, we obtained

$$N = 2V \left( \frac{2\pi k_B T}{\hbar^2} \right)^{3/2} (m_c^* m_v^*)^{3/4} e^{-E_g/2k_B T},$$

which is equal to the number $P$ of holes, while in the extrinsic case, in the presence of $N_d$ donors, we found

$$N(T) = \frac{1}{2} N_0(T) e^{-(E_c-E_d)/k_B T} \left\{ 1 + \frac{4N_d}{N_0(T)} e^{(E_c-E_d)/k_B T} \right\}^{1/2} - 1,$$

where

$$N_0(T) = 2V \left( \frac{2\pi m_v^* k_B T}{\hbar^2} \right)^{3/2}.$$

The most apparent difference in these two expressions are the energies in the exponential factors. In the intrinsic case, this is the energy gap between the valence and conduction bands, but in the extrinsic case, this is the binding energy of the
donor electron, which is much smaller than the band gap. This is the reason for
the dominance of extrinsic electrons in the conduction band at the operating tem-
peratures of devices based on particular doped semiconductors.

The other important result we have obtained concerns the chemical potential.
For the case of intrinsic semiconductors, at $T = 0$, the chemical potential lies
midway between the valence and conduction band edges and either increases,
decreases, or remains the same, depending on the effective masses in these bands.
For the extrinsic case, we considered a semiconductor doped with donors atoms,
which yields a chemical potential that, at $T = 0$, lies midway between the donor
level and the conduction band edge. At $T > 0$, as the donor atoms become ionized,
the chemical potential moves toward the donor level. Analogous considerations
apply for acceptor levels.
Chapter 4

The \textit{p-n} Junction and the Bipolar Junction Transistor

The \textit{p-n} junction is the fundamental ‘building block’ of almost all semiconductor electronic devices, including diodes, transistors, solar cells, light-emitting diodes (LEDs), and integrated circuits. A \textit{p-n} junction is formed at the interface between an \textit{n}-type semiconductor (i.e. one doped with donors, which have been ionized) and a \textit{p}-type semiconductor (i.e. one doped with acceptors, which have been ionized). Constructing the junction from two separate semiconductors, would introduce a physical boundary, so instead \textit{p-n} junctions are created in a single semiconductor crystal by spatially-controlled doping.

The discovery of the \textit{p-n} junction is usually attributed to American physicist Russell Ohl of Bell Telephone Laboratories in New Jersey, the same laboratory where William Shockley, John Bardeen, and Walter Brattain – the inventors of the transistor – worked.\footnote{A lively description of the innovations that came out of Bell Telephone Laboratories may be found in Jon Gertner’s book \textit{The Idea Factory: Bell Labs and the Great Age of American Innovation} (Penguin Books, London, 2012).} Prior to the advent of solid-state electronics, vacuum tubes\footnote{A vacuum tube is a glass tube surrounding a vacuum, that is, an area from which all gases have been removed. When electrical contacts are put on the ends, a current flows though the vacuum. This phenomenon, first observed by Thomas Edison in 1883, is known as the ‘Edison effect.’} were the state-of-the-art in radio equipment. The use of crystalline materials had been attempted but, as their production methods were not carefully controlled, their performance was unacceptably inconsistent. Vacuum tubes, on the other hand, were simple in design and, most importantly, reliable. But Ohl
had a different vision for the future of electronics. He thought that the erratic behavior of silicon crystals was not due to any problem with silicon itself, but was due to impurities in the crystal. Ohl believed that if crystalline silicon could be made pure enough, i.e. with minimal uncontrolled impurities, it might provide an avenue for devices that would improve radio broadcasting capabilities.3

In this chapter, we begin by describing the basic physics of the $p-n$ junction, including the formation of equilibrium when the two types of semiconductor are brought into contact, and the consequences for their current-carrying capabilities. While a single $p-n$ junction is useful, it also has limitations. It can either conduct current or not, but cannot really control how much current it will conduct. Understanding the operation of a $p-n$ junction is crucial for understanding devices based on such junctions. We will illustrate this for one of the most common types of transistor, the bipolar junction transistor, which consists of two $p-n$ junctions in series, such as $n-p-n$ or $p-n-p$.

### 4.1 Thermodynamic Equilibrium

When a $p$-type and an $n$-type semiconductors are joined together an abrupt carrier concentration gradient is created at the interface of the $n$ and $p$ regions, as shown in Fig. 4.1. This gradient leads to the diffusion of carriers across the interface because of Fick’s first law:

$$J = -D \nabla c,$$

which states that the diffusion current $J$ flows in opposition to concentration gradients, with the proportionality constant $D$ known as the diffusion constant. In the case of Fig. 4.1, diffusion occurs because the concentration of electrons in the $n$-type semiconductor is much higher than that in the $p$-type semiconductor. The opposite behavior is observed for the holes. The hole concentration in the $p$-type semiconductor is much higher than that in the $n$-type semiconductor. Consequently, holes from the $p$ side diffuse to the $n$ side and electrons from the $n$ side diffuse to the $p$ side.

Since holes leave the $p$-type region, the negative acceptor ions near the junction are left uncompensated and a space charge region is formed. This means that

---

3 A personal discourse on the road to single crystals for solid-state electronics has been given by G. K. Teal, ‘Single crystals of germanium and silicon – Basic to the transistor and integrated circuit,’ *IEEE Transactions on Electronic Devices* ED-23(7), 621–639 (1976).
The \( p-n \) Junction and the Bipolar Junction Transistor

Figure 4.1: The equilibration of a \( p-n \) junction when \( n \)- and \( p \)-type semiconductors are brought into contact. (a) Isolated, charge-neutral regions of \( n \)- and \( p \)-type semiconductors, with the energies of their valence and conduction band edges, together with their chemical potentials. (b) The \( p-n \) junction, showing the space charge in the transition region \( W \), the resulting electric field \( E \), and the coordinate \( x \) used for the calculations in the text, with the origin taken at the junction. The entire system now has a common chemical potential and is, therefore, in thermodynamic equilibrium. (c) The directions of the four components of particle currents within the transition region and the resulting charge flow. In equilibrium, the net electron and hole currents are zero, so that their diffusion and drift currents must cancel.
the acceptor ions are localized, whereas the holes are mobile in the valence band. The same applies for the electrons in the conduction band. The electrons diffuse towards the interface and the localized donor states are uncompensated in the $n$-type semiconductor. As a consequence a space charge region is formed there as well. In this case, free carriers are removed from the space charge region and the region becomes depleted.

The localized charges (acceptor and donor ions) create a potential barrier that prevents additional electrons and holes from moving towards the space charge region. As a consequence, an electric field is created that maintains drift currents of electrons and holes in opposite directions. Therefore, drift and diffusion currents\(^4\) occur simultaneously (Fig. 4.2).

\(^4\)Diffusion is the mechanism by which mass is transported in response to a gradient of concentration, as noted above. The diffusion equation is a macroscopic statement of this process. Drift, on the other hand, is the average motion of, say, an electron in an electric field, where the transport is directed field, but interrupted by various scattering events.

---

**Figure 4.2:** Schematic illustration of (a) drift and (b) diffusion. The gray circles symbolize vibrating atoms in a lattice which disrupt the periodicity of the lattice and cause scattering of the Bloch functions of electrons and holes in a semiconductor. In the case of drift in the presence of a field point from right to left, the particle maintains a component of its scattered velocity along the field direction, producing a net drift velocity. But for diffusion, the driving force is the concentration gradient of the diffusing particles. There is no force on the particles, whose motion resembles a random walk.
Thermodynamic equilibrium is attained when the chemical potential throughout the entire sample becomes constant. At this point, there is no additional excitation of carriers and there is no temperature gradient in the sample, so the overall current flow through the sample is zero, meaning electron and hole currents at each position in the $p$-$n$ junction are zero. This can be achieved only if the drift current is equal to the diffusion current for each spatial position throughout the device. Hence, the current caused by the electric field (drift) is compensated by the current caused by the carrier gradient (diffusion).

### 4.1.1 The Built-In Voltage

The two key quantities that characterize a $p$-$n$ junction are the built-in electric field $E$ and the width $W$ of the transition region. We can calculate both quantities based on ideas and methods we have already developed. In this subsection, we calculate the built-in voltage and in the next subsection, the width of the transition region.

We suppose that all dopants are ionized, so we are working in the saturation region of the $p$-type and $n$-type materials (Fig. 3.9). Consider the $n$-type material first. The basic sum rule is $N = N_d^+ = N_d$, as all donors are ionized. Therefore, we can write

$$ N = N_0(T)e^{-(E_c - \mu_n)/k_BT} = N_d, $$

where $\mu_n$ denotes the chemical potential in the $n$-type material and

$$ N_0(T) = 2V \left( \frac{2\pi m^*_e k_BT}{\hbar^2} \right)^{3/2}. $$

Solving for $\mu_n$,

$$ \mu_n = E_c - k_BT \ln \left( \frac{N_0}{N_d} \right). $$

For the $p$-type material, we have $P = N_a^+ = N_a$, which yields

$$ P = P_0(T)e^{-(\mu_p - E_v)/k_BT} = N_a, $$

where $\mu_n$ denotes the chemical potential in the $n$-type material and

$$ P_0(T) = 2V \left( \frac{2\pi m^*_h k_BT}{\hbar^2} \right)^{3/2}. $$
Solving for $\mu_p$, 

$$\mu_p = E_v + k_B T \ln \left( \frac{p_0}{N_a} \right).$$

Thus, the potential energy difference $e\phi_0$ is the difference between the chemical potentials in the $p$-type and $n$-type materials:

$$e\phi_0 = \mu_n - \mu_p = E_c - E_v - k_B T \ln \left( \frac{p_0 N_0}{N_a N_d} \right)$$

$$= E_g - k_B T \ln \left( \frac{p_0 N_0}{N_a N_d} \right).$$

A more compact way of writing this quantity is in terms of the intrinsic carrier densities:

$$N_i(T) = N_0(T) e^{-(E_c - \mu)/k_BT}, \quad P_i(T) = P_0(T) e^{-(\mu - E_v)/k_BT},$$

in terms of which we have (Problem Set 5)

$$\phi_0 = \frac{k_B T}{e} \ln \left( \frac{N_a N_d}{N_i^2} \right),$$

or, in terms of densities, which are represented by the corresponding lower case letters:

$$\phi_0 = \frac{k_B T}{e} \ln \left( \frac{n_a n_d}{n_i^2} \right), \quad (4.1)$$

### 4.1.2 The Transition Region

The width of the transition region will be obtained by solving Poisson’s equation:

$$\frac{d^2 \phi}{dx^2} = -\frac{\rho(x)}{\varepsilon \varepsilon_0},$$

in which the charge density is

$$\rho(x) = \begin{cases} 
-N_a e, & -w_p \leq x < 0; \\
N_d e, & 0 < x \leq w_n; \\
0, & \text{otherwise}.
\end{cases}$$
Thus, Poisson’s equation must be solved in the $p$-type and $n$-type regions separately:

$$\frac{d^2\phi}{dx^2} = \begin{cases} \frac{N_a e}{\varepsilon \varepsilon_0}, & -w_p \leq x < 0; \\ -\frac{N_i e}{\varepsilon \varepsilon_0}, & 0 < x \leq w_n; \\ 0, & \text{otherwise}. \end{cases}$$

We consider the $p$-type region first. For $-w_p \leq x < 0$, we have

$$\frac{d^2\phi}{dx^2} = \frac{N_a e}{\varepsilon \varepsilon_0}.$$ 

Performing two integrations yields:

$$\frac{d\phi}{dx} = \frac{N_a e}{\varepsilon \varepsilon_0} x + A,$$

$$\phi = \frac{N_a e}{2 \varepsilon \varepsilon_0} x^2 + Ax + B,$$

where $A$ and $B$ are constants to be determined by the boundary conditions. We set the zero of energy so that the potential beyond the transition region in the $p$-type region is zero: $\phi(-w_p) = 0$. Since the electric field must vanish at the boundary of the transition region, we also have that

$$\left.\frac{d\phi}{dx}\right|_{x=-w_p} = 0.$$ 

This equation yields

$$\frac{d\phi}{dx} = \frac{N_a e}{\varepsilon \varepsilon_0} (-w_p) + A = 0,$$

which determines $A$:

$$A = \frac{N_a e w_p}{\varepsilon \varepsilon_0}.$$ 

The other boundary condition,

$$\phi(-w_p) = \frac{N_a e w_p^2}{2 \varepsilon \varepsilon_0} - \frac{N_a e w_p^2}{\varepsilon \varepsilon_0} + B = 0,$$
yields $B$:

$$B = \frac{N_a e w_p^2}{2\varepsilon\varepsilon_0}.$$ 

Hence, the expression obtained for the potential $\phi_p(x)$ in the $p$-type region is:

$$
\phi_p(x) = \frac{N_a e}{2\varepsilon\varepsilon_0} x^2 + \frac{N_a e w_p}{\varepsilon_0} x + \frac{N_a e w_p^2}{2\varepsilon\varepsilon_0} \\
= \frac{N_a e}{2\varepsilon\varepsilon_0} (x^2 + 2w_p x + w_p^2) \\
= \frac{N_a e}{2\varepsilon\varepsilon_0} (x + w_p)^2.
$$

The calculation for the potential in the $n$-type region proceed in an analogous fashion. The boundary conditions are

$$
\phi(w_n) = \phi_0, \quad \frac{d\phi}{dx} \bigg|_{x=w_n} = 0,
$$

where the first condition is obtained from the built-in potential across the junction, and the second states that the electric field vanishes outside of the transition region on the $n$-side of the junction. Proceeding as above, we obtain for the potential $\phi_n(x)$ in the $n$-type region

$$
\phi_n(x) = \phi_0 - \frac{N_d e}{2\varepsilon\varepsilon_0} (x - w_n)^2.
$$

The continuity of the potential requires that $\phi_p(0) = \phi_n(0)$, i.e.

$$
\frac{N_a e}{2\varepsilon\varepsilon_0} w_p^2 = \phi_0 - \frac{N_d e}{2\varepsilon\varepsilon_0} w_n^2, \quad (4.2)
$$

while charge neutrality means that $N_a e w_p = N_d e w_n$. Solving these simultaneous equations for $w_n$ and $w_p$ yields

$$
\begin{align*}
  w_n &= \left[ \frac{2\varepsilon\varepsilon_0 N_a \phi_0}{e N_d (N_a + N_d)} \right]^{1/2}, \\
  w_p &= \left[ \frac{2\varepsilon\varepsilon_0 N_d \phi_0}{e N_a (N_a + N_d)} \right]^{1/2}.
\end{align*}
$$

\begin{equation}
\text{(4.3)}
\end{equation}
For typical doping levels, these formulas give widths of the order of a micron. Note that the continuity condition (4.1.2) enables to express the in-built potential in terms of the depletion widths as

\[
\phi_0 = \frac{e}{2\varepsilon_0} (N_{dw}^2 + N_{dp}^2).
\]  

(4.4)

4.2 Forward and Reverse Bias

One of the most important properties of \( p-n \) junctions is rectification, i.e. the ability to readily allow current flow in one direction (forward biased), but not the other (reverse biased). In this section, we first provide a qualitative description of this property before providing a more quantitative discussion leading to the Shockley equation.

4.2.1 Qualitative Analysis

Figure 4.3 shows a \( p-n \) junction in equilibrium and under forward and reverse bias. In equilibrium the diffusion and drift currents balance, and there is no net current flow across the junction (Fig. 4.1). The application of a bias voltage disrupts this equilibrium and there results a rearrangement of charge. Consider the forward bias case first, as shown in Fig. 4.3(b). The voltage draws electrons from the \( n \)-type region and holes from the \( p \)-type region toward the depletion region, where they combine, so that a continuous current can be maintained. The forward-biased voltage thus counteracts the in-built voltage, which results in a reduced depletion zone. The reduced depletion zone can also be seen by replacing the in-built voltage with \( V \) in (4.3).

Now consider the case of reverse bias, as shown in Fig. 4.3(c). The applied voltage pulls electrons and holes away from the junction, resulting in a current. However, this is a transient current, because, when the potential formed by the widened depletion zone equals the applied voltage, this current ceases, except for a small current due to small thermal fluctuations, which results in excitations from the valence to the conduction band. Hence, the reverse bias voltage reinforces the in-built voltage, again as expected from (4.3).
4.2.2 Carrier Currents under Forward and Reverse Bias

Figure 4.3: Schematic diagram of the rectification of a p-n junction, showing that junction (a) in equilibrium, (b) in forward bias, and (c) in reverse bias. Holes are indicated by open circles, electrons by closed circles, and the depletion region is indicated by shading, with the charges of immobile donors and acceptors indicated by “+” and “−” signs, respectively.

Figure 4.5 shows the electric fields and valence and conduction bands of a p-n junction in equilibrium and under forward and reverse bias. In the absence of bias, the in-built potential $\phi_0$ acts as a barrier to the flow of electrons from the n-type to the p-type region. A few electrons can cross the barrier, giving rise to a small current $I_e^0$. If there is no bias, there is no current overall, so $I_e^0$ must be balanced
by an electron current from the \( p \)-type to the \( n \)-type region. Although there are comparatively few electrons on the \( p \) side, those that drift into the depletion zone are immediately swept across to the \( n \)-side by the in-built potential.

In the case of forward bias, the applied field acts in opposition to the in-built field, thereby reducing the potential barrier across the junction. On the other hand, the effect of reverse bias is to reinforce the in-built field, leading to a much greater barrier at the junction. Understanding the effect of these barrier changes on the drift and diffusion currents is the key to understanding why the \( p-n \) junction conducts electrical current for forward bias, but much less so reverse bias.

Consider the diffusion of electrons and holes first. The application of a forward bias reduces the energy barrier for holes and electrons at the junction and causes a misalignment of the chemical potentials on the two sides of the junction. In particular, the chemical potential for electrons on the \( n \) side of the junction is greater than that on the \( p \) side. Thus, electrons with sufficient energy in the conduction band can diffuse readily to the \( p \) side. Similarly, holes on the \( p \) side of the junction with sufficient energy in the valence band can diffuse to the \( n \) side of the junction. But in the reverse bias case, the energy barrier that must be surmounted near the barrier is too large for any significant number of electrons and holes to cross the junction, so the diffusion current is substantially diminished.
Figure 4.5: The current-voltage characteristics of a p-n junction indicate conductance with forward \((V > 0)\), but not reverse \((V < 0)\) bias.

Drift current provides an altogether scenario. The direction of the electric field is the same as in equilibrium in the presence of either forward or reverse bias. Therefore, the direction of drift particle and their corresponding electrical currents are the same in the three cases. In fact, as we now argue, the magnitude of these currents is also approximately the same in the three cases. In the presence of an electric field from the \(n\)-side to the \(p\)-side of the junction, electrons drift from the \(p\)-side of the junction, where they are minority carriers, to the \(n\)-side. Similarly, holes drift from the \(n\) side of the junction, where they are minority carriers, to the \(p\) side. How are these minority carriers created? By thermal excitation of an electron from the valence band, which creates a hole in that band, to the conduction band. As we discussed in Chapter 3, carrier densities are dominated by the (majority) extrinsic carriers, but the density of these “intrinsic” carriers is still non-zero. These carriers are responsible for the current in the when the p-n junction is placed under reverse bias.

We can now understand qualitatively why the p-n junction conducts electrical current under forward, but not reverse bias (Fig. 4.5). The current in the forward bias direction is dominated by diffusion current of extrinsic carriers, but the the current in the reverse bias direction is due mainly to intrinsic carriers generated by excitation across the energy gap of the semiconductor.
4.3 The Bipolar Junction Transistor

Bipolar junction transistors were the prevalent type of transistor in the semiconductor industry until around the 1970s, when complementary metal-oxide semiconductor (CMOS) technology took over. CMOS is now the predominant technology in digital integrated circuits which are used, for example, in microprocessors. Nevertheless, bipolar junction transistors are still advantageous in many applications, such as radio frequency circuits and wireless systems.

As we have discussed at length in the preceding sections, the p-n junction blocks current in one direction. When two p-n junctions are placed back-to-back, the resulting device – called a bipolar junction transistor – allows current to be regulated. In particular, an n-p-n transistor, which we will discuss in this section (Fig. 4.6), allows the current of electrons through an n-type semiconductor to be controlled. Electrons flow through the device from one side, called the emitter, to the other side, called the collector, with the voltage at the base regulating the current between the emitter and collector. Figure 4.7(a) shows the energy levels

Figure 4.6: Three representations of an n-p-n bipolar junction transistor. (a) Simplified cross-section, (b) simplified circuit diagram, and (c) a microscopic view.
in an \(n-p-n\) transistor when there are no externally applied voltages. In each of the \(n\)-type regions conduction can take place by the free movement of electrons in the conduction band. In the \(p\)-type region, conduction can take place by the movement of the free holes in the valence band. However, in the absence of any externally applied electric field, depletion zones form at both \(p-n\) junctions, so no charge moves from one layer to another.

The transistor works on the same principles as the \(p-n\) junction, but with one twist. Consider first the flow of electrons through the device. The junction between emitter and base is operated at a forward-bias voltage difference Figure 4.7(c). Therefore, the majority electrons of the \(n\)-type emitter pour through in great numbers. By the normal logic, these electrons should produce a current between the emitter and base electrodes.

But here is the twist. The \(p\) region is extremely thin, much smaller than than the diffusion length of the electrons. Essentially all of the electrons that pour through the junction continue into the second junction, the one between base and collector. This second junction is operated at a reverse-bias voltage Figure 4.7(c). This produces a strong electric field that sweeps the electrons into the collector. (The electrons are minority carriers in the base, so they get swept through the junction by the electric field.) As a result, virtually all electrons leaving the emitter end up as an electron flow to the collector electrode instead of to the base. This explains why the base voltage can regulate the current between emitter and collector without diverting much of the current. Furthermore, as for the \(p-n\) junction, the number of electrons pouring through the junction from emitter to base varies strongly with the base voltage. Small voltage changes at the base can therefore decimate or explode the electron flow, and almost all of which goes to the collector. An \(p-n-p\) works just like an \(n-p-n\) transistor, but with holes taking the place of electrons.

### 4.4 Summary

The \(p-n\) junction is the interface between an \(n\)-type and a \(p\)-type semiconductor. The space formed at the junction provides a barrier for majority carriers, but accelerates minority carriers passing through the junction. Under normal conditions, the \(p-n\) junction will only conduct a significant current in one direction, called the forward direction, in which case both the \(n\)-side electrons and the \(p\)-side holes...
Figure 4.7: The operation of an $n-p-n$ bipolar junction transistor.
move towards the junction. The diode equation describes the current versus voltage relation of p-n junctions, but only in a limited range. The bipolar junction transistor, form from two p-n junctions allows current to be regulated.
Chapter 5

Magnetism and Magnetic Materials

The magnetism of solids originates from the electrons of atoms, due to their orbital motion and intrinsic spin. Since an electron has charge, a magnetic moment appears from the orbital motion just like magnetic field is generated when current flows in the coils of a solenoid. A magnetic moment also appears from the intrinsic spin orientation of electrons. Although in a typical solid many orbitals are occupied, so that there are many contributions from electronic orbital motions and the orientations of intrinsic spins around the nuclei, only two electrons can occupy each orbital, one with spin up and one with spin down, due to the Pauli exclusion principle. Therefore, most of the magnetic moments which originate from electronic spins cancel each other and do not contribute to the magnetic moment of a solid. However, in the case of transition elements, whose 3d orbitals are not fully occupied, magnetic moments appear due to the spins. According to Hund’s rule (see below), electron are occupied so that the total spin is maximized. In other words, the number of up-spins increases from one to five according to the atomic number, thereafter, total spin amount decreases as up-spin and down-spin cancel each other when two electrons occupy the same orbit. The reason why the actual metal never has a magnetization as large as five unpaired spins is that the 3d-electrons are shared by many atoms in crystals as delocalized electrons.

Most solids are considered to be non-magnetic, which means broadly that they become magnetized only in the presence of a magnetic field. In most cases, these effects are very weak and the magnetization is lost as soon as the magnetic field is removed. Of more interest, both from the fundamental and technological points of view, are materials that retain their magnetization even if there is no external field.
Magnetism and Magnetic Materials

Such materials, called permanent magnets, have the property their like (resp., unlike) poles repel (resp., attract) one another, which was known by the ancient Greeks and the Chinese some 2000 years ago. The importance of magnets continued to increase as more uses were realized for their remarkable properties and they play a crucial role in many modern technologies. The information technology revolution owes as much to developments in magnetic information storage as it does to the transistor and the integrated circuit.

In this chapter, we review the fundamental of magnetism in materials, beginning with why spins align. This will then allows us to discuss the different types of materials in terms of their responses to external magnetic fields. We conclude with a brief discussion of a recent suggestion for the use of permanent magnets in refrigeration cycles.

5.1 Why Spins Align

To understand why spins align, which provides the physical basis for Hund’s rule mentioned in the introduction, we focus on the case of two electrons in an atom. The two electrons are at positions \( \mathbf{r}_1 \) and \( \mathbf{r}_2 \) with spins \( \sigma_1 \) and \( \sigma_2 \). Each spin can either point up or down, that is, \( \sigma_i = \uparrow, \downarrow \), for \( i = 1, 2 \). The total wave function \( \Psi \) must be anti-symmetric under the exchange of the two electrons to conform to the Pauli exclusion principle. We write \( \Psi \) as the product of a configurational factor \( \psi \) and a spin factor \( \phi \):

\[
\Psi(\mathbf{r}_1, \sigma_1; \mathbf{r}_2, \sigma_2) = \psi(\mathbf{r}_1, \mathbf{r}_2) \phi(\sigma_1 \sigma_2).
\]

Suppose that both electrons are in the same spin state: \( \phi(\uparrow, \uparrow) \). This spin wave function is symmetric, so the configurational part of the wave function, \( \psi(\mathbf{r}_1, \mathbf{r}_2) \), must be anti-symmetric which means, in particular, that the two electrons cannot occupy the same orbital. Thus, electrons with the same spin cannot get too close to one another, which has the effect of reducing their (repulsive) Coulomb interaction.

There is an additional effect to consider. The positive charge of the nucleus seen by one electron is screened by the negative charge of the other electron. This screening reduces the binding energy of the electrons to the nucleus. However, when the two spins are aligned, the electrons repel each other and therefore screen the nucleus less effectively. In this case, the electrons see the full charge...
of the nucleus and bind more strongly, thus lowering their energies. Another way of understanding this is to realize that when the spins are not aligned, one electron sometimes gets between the other electron and the nucleus, thereby reducing the effective charge seen by the outer electron, reducing the binding energy, and increasing the total energy of the atom. However, when the electrons are spin aligned, the Pauli principle largely prevents this configuration from occurring, thereby lowering the total energy of the system.

Thus, we see that spin alignment is determined ultimately by the Coulomb interaction. Recall from the Kronig–Penney model that, in the case of a small inter-site energy barrier $U_0$, there are wide bands with delocalized electrons. For a large inter-site barrier, however, the bands are narrow, indicating that the electrons are localized around the atomic sites and retain their atomic properties. What determines which type of behavior results in a particular material? The electronic structure of a material is determined by the competition between the Coulomb energy of the atomic state and the tendency toward bond formation. The fact that a crystalline solid forms indicates that the cohesive energy is positive, which means that there is significant bond formation (Sec. 1.5). But in a transition-metal, which has delocalized $4s$ electrons in wide bands and $3d$ electrons in relatively narrow bands, the energy balance becomes more delicate. In such cases, detailed quantum calculations are required to determine the outcome.

## 5.2 Types of Magnetism

When a material is placed in a magnetic field, the magnetic forces of the electrons in that material will be affected. This is known as Faraday’s law of magnetic induction. However, behavior of materials in magnetic fields can be quite diverse and depends on several factors, such as the atomic and structure of the material, and the net magnetic field associated with the atoms. The magnetic moments associated with atoms have three origins. As discussed in the preceding section, these are the orbital motion of the electron, the change in motion caused by an external magnetic field, and the spin of the electrons.

In most atoms, electrons occur in pairs with spins in opposite directions. Thus, when electrons are paired, their opposite spins cause their magnetic fields to cancel each other, so no net magnetic field exists. Alternately, materials with some unpaired electrons will have a net magnetic field and will have a more pronounced
response to an external field. The relationship between a small applied field $H$ and the magnetization $M$ of a material is characterized by the susceptibility $\chi$:

$$M = \chi H.$$  

This is an example of a linear response relation in that, for a small external impulse (in this case the applied magnetic field), the response (the magnetization here) is proportional to the impulse. Materials are classified by their response to an external magnetic field and whether they are themselves magnetic:

**Diamagnetic** materials are those for which $\chi < 0$ and so have a weak, negative susceptibility to magnetic fields. Diamagnetic materials are slightly repelled by a magnetic field and do not retain the magnetic properties when the external field is removed. In diamagnetic materials all the electrons are paired, so there is no permanent net magnetic moment per atom. Diamagnetic properties arise from the realignment of the electronic orbits under the influence of an external magnetic field. Most elements in the periodic table, including copper, silver, and gold, are diamagnetic.

**Paramagnetic** materials are those for which $\chi > 0$ and so have a small, positive susceptibility to magnetic fields. These materials are slightly attracted by a magnetic field and do not retain the magnetic properties when the external field is removed. Paramagnetic properties are due to the presence of some unpaired electrons and from the realignment of the electronic orbits caused by the external magnetic field. Paramagnetic materials include magnesium, molybdenum, lithium, and tantalum.

**Ferromagnetic** materials have a large, positive susceptibility to an external magnetic field. They exhibit a strong attraction to magnetic fields and are able to retain their magnetic properties after the external field has been removed, that is, $M \neq 0$ even without a magnetic field. Ferromagnetic materials have some unpaired electrons so their atoms have a net magnetic moment. They get their strong magnetic properties due to the presence of magnetic domains. In these domains, the moments of large numbers of atom are aligned in parallel, so that the magnetic force within the domain is strong. Iron, nickel, and cobalt are examples of ferromagnetic materials.

Figure 8.3 shows one of the most striking phenomena associated with ferromagnetic materials, namely, the transition between magnetic and non-magnetic phases at a critical temperature $T_c$. For $T < T_c$, a ferromagnetic material exhibits
spontaneous magnetization (that is, without the application of an external field, while, for $T > T_c$, there is no such magnetization. This phase transition is accompanied by singularities in various thermodynamic functions. For example, as this figure shows, the behavior of the magnetization near $T_c$ for a two-dimensional (2D) Ising model is

$$M \sim (T_c - T)^{1/8}.$$ 

This shows that the magnetization is continuous through $T_c$, but not differentiable. However, in the presence of an external magnetic field, this singular behavior is smoothed out.

Even more striking is the behavior of the susceptibility of this 2D Ising model
near $T_c$:

$$\chi \sim \frac{1}{|T - T_c|^{7/4}}$$

which shows that the susceptibility becomes infinite at the critical temperature (with $H = 0$). This indicated that a small magnetic field elicits a macroscopic response from a ferromagnetic material near its critical point and is a characteristic feature of such phase transitions.

### 5.3 The Magnetocaloric Effect

The magnetocaloric effect is a magneto-thermodynamic process in which a change in temperature of a magnetic material is caused by exposing that material to a changing magnetic field. This effect forms the basis for the magnetic cooling cycle, which is a promising alternative to the conventional vapor-cycle refrigeration, that may soon be viable for commercial and domestic applications. In this section we will describe how the magnetocaloric effect is used in the refrigeration cycle of model for a magnetic material. The cycle consists of the following steps:

- **Step 1.** The magnetic material is initially at the ambient temperature $T_i$, which is greater than the critical temperature $T_c$ of the magnetic material, $T_i > T_c$. An external magnetic field $H$ is now applied under adiabatic conditions, which means that there is no heat transfer. There are two contributions to the total entropy change $\Delta S_{\text{total}}$: the entropy change of the spin of the electrons $\Delta S_{\text{spin}}$ and the entropy change $\Delta S_{\text{lattice}}$ of the lattice. Hence,

$$\Delta S_{\text{total}} = \Delta S_{\text{spin}} + \Delta S_{\text{lattice}}.$$

As the process is adiabatic, $\Delta S_{\text{total}} = 0$. But the application of the external magnetic field has increased the magnetization of the material, which means that the spins are now more aligned. Thus, $\Delta S_{\text{spin}} < 0$, which implies that $\Delta S_{\text{lattice}} > 0$, so the temperature of the material increases to a temperature

$$T_1 = T_i + \Delta T_{0 \rightarrow H}^{\text{adiabatic}}.$$

- **Step 2.** The magnetic material is placed in thermal contact with the environment in the presence of $H$. The temperature of the material reduces to

$$T_2 = T_i,$$
Magnetism and Magnetic Materials

Figure 5.2: Schematic illustration of the use of the magnetocaloric effect in a refrigeration cycle. The shaded regions indicate the magnetic state of the material, with the arrows indicating spin directions. The cycle begins at the top (Step 1 in the main text) and then proceeds counterclockwise to Steps 2, 3, and 4.

with the material still magnetized.

- **Step 3.** The magnetic is again adiabatically isolated and the external field is turned off. Entropy randomizes the moments, so that \( \Delta S_{\text{spin}} > 0 \) and, as the total entropy change must again be zero, \( \Delta S_{\text{lattice}} < 0 \), so the temperature of the system is reduced to a temperature

\[
T_3 = T_i - \Delta T^{H \to 0}_{\text{adiabatic}}.
\]

- **Step 4.** The system is brought back to the ambient temperature \( T_i \) by heat exchange with the material being refrigerated:

\[
T_4 = T_i,
\]

and the cycle is repeated.
5.4 Summary

We have provided in this chapter a brief summary of the atomistic origins of magnetism and of the different types of magnetic material. We then discussed a modern application of magnetic materials, the use of the magnetocaloric effect in a refrigeration cycle. There are many applications of magnetic materials and the list is constantly growing.