Jenő Sólyom

Fundamentals of the Physics of Solids

Volume 2 – Electronic Properties

Springer

Periodic table of elements

	1 IA	2 IIA	3 IIIB	4 IVB	5 VB	6 VIB	7 VIIB	8	9 VIIIB	10	11 IB	12 IIB	13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 VIIIA
1	1 H $1s^1$ Hydrogen																	2 He $1s^2$ Helium
2	3 Li [He] $2s^1$ Lithium	4 Be $1s^2$ Beryllium											5 B $2s^22p^1$ Boron	6 C $2s^22p^2$ Carbon	7 N $2s^22p^3$ Nitrogen	8 O $2s^22p^4$ Oxygen	9 F $2s^22p^5$ Fluorine	10 Ne $2s^22p^6$ Neon
3	11 Na [Ne] $3s^1$ Sodium	12 Mg $3s^2$ Magnesium											13 Al $3s^23p^1$ Aluminum	14 Si $3s^23p^2$ Silicon	15 P $3s^23p^3$ Phosphorus	16 S $3s^23p^4$ Sulfur	17 Cl $3s^23p^5$ Chlorine	18 Ar $3s^23p^6$ Argon
4	19 K [Ar] $4s^1$ Potassium	20 Ca $4s^2$ Calcium	21 Sc $3d^14s^2$ Scandium	22 Ti $3d^24s^2$ Titanium	23 V $3d^34s^2$ Vanadium	24 Cr $3d^54s^1$ Chromium	25 Mn $3d^54s^2$ Manganese	26 Fe $3d^64s^2$ Iron	27 Co $3d^74s^2$ Cobalt	28 Ni $3d^84s^2$ Nickel	29 Cu $3d^{10}4s^1$ Copper	30 Zn $3d^{10}4s^2$ Zinc	31 Ga $3d^{10}4s^24p^1$ Gallium	32 Ge $3d^{10}4s^24p^2$ Germanium	33 As $3d^{10}4s^24p^3$ Arsenic	34 Se $3d^{10}4s^24p^4$ Selenium	35 Br $3d^{10}4s^24p^5$ Bromine	36 Kr $3d^{10}4s^24p^6$ Krypton
5	37 Rb [Kr] $5s^1$ Rubidium	38 Sr $5s^2$ Strontium	39 Y $4d^15s^2$ Yttrium	40 Zr $4d^25s^2$ Zirconium	41 Nb $4d^45s^1$ Niobium	42 Mo $4d^54s^1$ Molybdenum	43 Tc $4d^55s^2$ Technetium	44 Ru $4d^75s^1$ Ruthenium	45 Rh $4d^85s^1$ Rhodium	46 Pd $4d^{10}$ Palladium	47 Ag $4d^{10}5s^1$ Silver	48 Cd $4d^{10}5s^2$ Cadmium	49 In $4d^{10}5s^25p^1$ Indium	50 Sn $4d^{10}5s^25p^2$ Tin	51 Sb $4d^{10}5s^25p^3$ Antimony	52 Te $4d^{10}5s^25p^4$ Tellurium	53 I $4d^{10}5s^25p^5$ Iodine	54 Xe $4d^{10}5s^25p^6$ Xenon
6	55 Cs [Xe] $6s^1$ Cesium	56 Ba $6s^2$ Barium	57 La $5d^16s^2$ Lanthanum	72 Hf $5d^26s^2$ Hafnium	73 Ta $5d^36s^2$ Tantalum	74 W $5d^46s^2$ Tungsten	75 Re $5d^56s^2$ Rhenium	76 Os $5d^66s^2$ Osmium	77 Ir $5d^76s^2$ Iridium	78 Pt $5d^96s^1$ Platinum	79 Au $5d^{10}6s^1$ Gold	80 Hg $5d^{10}6s^2$ Mercury	81 Tl $5d^{10}6s^26p^1$ Thallium	82 Pb $5d^{10}6s^26p^2$ Lead	83 Bi $5d^{10}6s^26p^3$ Bismuth	84 Po $5d^{10}6s^26p^4$ Polonium	85 At $5d^{10}6s^26p^5$ Astatine	86 Rn $5d^{10}6s^26p^6$ Radon
7	87 Fr [Rn] $7s^1$ Francium	88 Ra $7s^2$ Radium	89 Ac $6d^17s^2$ Actinium	104 Rf $6d^27s^2$ Rutherfordium	105 Db $6d^37s^2$ Dubnium	106 Sg $6d^47s^2$ Seaborgium	107 Bh $6d^57s^2$ Bohrium	108 Hs $6d^67s^2$ Hassium	109 Mt $6d^77s^2$ Meitnerium	110 Ds $6d^97s^1$ Darmstadtium	111 Rg $6d^{10}7s^1$ Roentgenium							

Lanthanoids	58 Ce $4f^26s^2$ Cerium	59 Pr $4f^36s^2$ Praseodymium	60 Nd $4f^46s^2$ Neodymium	61 Pm $4f^56s^2$ Promethium	62 Sm $4f^66s^2$ Samarium	63 Eu $4f^76s^2$ Europium	64 Gd $4f^75d^16s^2$ Gadolinium	65 Tb $4f^96s^2$ Terbium	66 Dy $4f^{10}6s^2$ Dysprosium	67 Ho $4f^{11}6s^2$ Holmium	68 Er $4f^{12}6s^2$ Erbium	69 Tm $4f^{13}6s^2$ Thulium	70 Yb $4f^{14}6s^2$ Ytterbium	71 Lu $4f^{14}5d^16s^2$ Lutetium
Actinoids	90 Th $6d^27s^2$ Thorium	91 Pa $5f^26d^17s^2$ Protactinium	92 U $5f^36d^17s^2$ Uranium	93 Np $5f^57s^2$ Neptunium	94 Pu $5f^67s^2$ Plutonium	95 Am $5f^77s^2$ Americium	96 Cm $5f^76d^17s^2$ Curium	97 Bk $5f^86d^17s^2$ Berkelium	98 Cf $5f^{10}7s^2$ Californium	99 Es $5f^{11}7s^2$ Einsteinium	100 Fm $5f^{12}7s^2$ Fermium	101 Md $5f^{13}7s^2$ Mendelevium	102 No $5f^{14}7s^2$ Nobelium	103 Lr $5f^{14}6d^17s^2$ Lawrencium

Fundamentals of the Physics of Solids

Jenő Sólyom

Fundamentals of the Physics of Solids

Volume II

Electronic Properties

Translated by Attila Piróth

With 238 Figures and 40 Tables

Springer

Author
Professor Dr. Jenő Sólyom
Hungarian Academy of Sciences
Research Institute for
Solid State Physics & Optics
P.O. Box 49, 1525 Budapest
Hungary

and

Department of Physics
Eötvös Loránd University
1171 Budapest
Pázmány sétány 1/A
Hungary
solyom@szfki.hu

Translator
Attila Piróth
www.pirothattila.com

Title of the Hungarian edition: *A Modern Szilárdtestfizika Alapjai* II, Elektronok a szilárd testekben, ISBN 963 463 589 X © ELTE Eötvös Kiadó 2003

ISBN: 978-3-540-85315-2 e-ISBN: 978-3-540-85316-9

Library of Congress Control Number: 2007929725

Cover design: eStudio Calamar S.L., Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To Márta, Gyöngyvér, Tünde, and Iringó

Preface

The reader is holding the second volume of a three-volume textbook on solid-state physics. This book is the outgrowth of the courses I have taught for many years at Eötvös University, Budapest, for undergraduate and graduate students under the titles *Solid-State Physics* and *Modern Solid-State Physics.*

The main motivation for the publication of my lecture notes as a book was that none of the truly numerous textbooks covered all those areas that I felt should be included in a multi-semester course. Especially, if the course strives to present solid-state physics in a unified structure, and aims at discussing not only classic chapters of the subject matter but also (in more or less detail) problems that are of great interest for today's researcher as well. Besides, the book presents a much larger material than what can be covered in a two- or three-semester course. In the first part of the first volume the analysis of crystal symmetries and structure goes into details that certainly cannot be included in a usual course on solid-state physics. The same applies, among others, to the discussion of the methods used in the determination of band structure, the properties of Fermi liquids and non-Fermi liquids, and the theory of unconventional superconductors in the present and third volumes. These parts can be assigned as supplementary reading for interested students, or can be discussed in advanced courses.

The line of development and the order of the chapters are based on the prerequisites for understanding each part. Therefore a gradual shift can be observed in the style of the book. While the intermediate steps of calculations are presented in considerable detail and explanations are also more lengthy in the first and second volumes, they are much sparser and more concise in the third one, thus that volume relies more on the individual work of the students. On account of the prerequisites, certain topics have to be revisited. This is why magnetic properties are treated in three, and superconductivity in two parts. The magnetism of individual atoms is presented in an introductory chapter of the first volume. The structure and dynamics of magnetically ordered systems built up of localized moments are best discussed after lattice vibrations, along the same lines. Magnetism is then revisited in the third volume, where the role of electron–electron interactions is discussed in more detail. Similarly, the phenomenological description of superconductivity is presented in this volume after the analysis of the transport properties of normal metals, in contrast to them, while the microscopic theory is outlined later, in the third volume, when the effects of interactions are discussed.

Separating the material into three similar-sized volumes is a necessity in view of the size of the material – but it also reflects the internal logical structure of the subject matter. At those universities where the basic course in solid-state physics runs for three semesters working through one volume per semester is a natural schedule. In this case the discussion of the electron gas – which is traditionally part of the introduction – is left for the second semester. This choice is particularly suited to curricula in which the course on solid-state physics is held parallel with quantum mechanics or statistical physics. If the lecturer feels more comfortable with the traditional approach, the discussion of the Drude model presented in this volume can be moved to the beginning of the whole course. Nevertheless the discussion of the Sommerfeld model should be postponed until students have familiarized themselves with the fundamentals of statistical physics. For the same reason the lecturer may prefer to change the order of other chapters as well. This is, to a large extent, up to the personal preferences of the lecturer.

In presenting the field of solid-state physics, special emphasis has been laid on discussing the physical phenomena that can be observed in solids. Nevertheless I have tried to give – or at least outline – the theoretical interpretation for each phenomenon, too. As is common practice for textbooks, I have omitted precise references that would give the publication data of the discussed results. I have made exceptions only for figures taken directly from published articles. At the end of each chapter I have listed textbooks and review articles only that present further details and references pertaining to the subject matter of the chapter in question. The first chapter of the first volume contains a longer list of textbooks and series on solid-state physics.

Bulky as it might be, this three-volume treatise presents only the fundamentals of solid-state physics. Today, when articles about condensed matter physics fill tens of thousands of pages every year in Physical Review alone, it would be obviously overambitious to aim at more. Therefore, building on the foundations presented in this series, students will have to acquire a substantial amount of extra knowledge before they can understand the subtleties of the topics in the forefront of today's research. Nevertheless at the end of the third volume students will also appreciate the number of open questions and the necessity of further research.

A certain knowledge of quantum mechanics is a prerequisite for studying solid-state physics. Various techniques of quantum mechanics – above all field-theoretical methods and methods employed in solving many-body problems – play an important role in present-day solid-state physics. Some essential details are listed in one of the appendices of the third volume, however, I have omitted more complicated calculations that would have required the application of the modern apparatus of many-body problems. This is especially true for the third volume, where central research topics of present-day solid-state physics are discussed, in which the theoretical interpretation of experimental results is often impossible without some extremely complex mathematical formulation.

The selection of topics obviously bears the stamp of the author's own research interest, too. This explains why the discussion of certain important fields – such as the mechanical properties of solids, surface phenomena, or amorphous systems, to name but a few – have been omitted.

I have used the International System of Units (SI), and have given the equations of electromagnetism in rationalized form. Since nonrationalized equations as well as gaussian CGS (and other) units are still widely used in the solid-state physics literature, the corresponding formulas and units are indicated at the appropriate places. In addition to the fundamental physical constants used in solid-state physics, the commonest conversion factors are also listed in Appendix A of the first volume. I deviated from the recommended notation in the case of the Boltzmann constant using k_{B} instead of k – reserving the latter for the wave number, which plays a central role in solid-state physics.

To give an impression of the usual values of the quantities occurring in solid-state physics, typical calculated values or measured data are often tabulated. To provide the most precise data available, I have relied on the Landolt–Börnstein series, the CRC Handbook of Chemistry and Physics, and other renowned sources. Since these data are for information only, I have not indicated either their error or in many cases the measurement temperature, and I have not mentioned when different measurement methods lead to slightly disparate results. As a rule of thumb, the error is usually smaller than or on the order of the last digit.

I would like to thank all my colleagues who read certain chapters and improved the text through their suggestions and criticism. Particular thanks go to professors György Mihály and Attila Virosztek for reading the whole manuscript. In spite of all efforts, some mistakes have certainly remained in the book. Obviously, the author alone bears the responsibility for them.

Special thanks are due to Károly Härtlein for his careful work in drawing the majority of the figures. The figures presenting experimental results are reproduced with the permission of the authors or the publishers. The challenge of translating the first and second volumes of the book from the Hungarian original was taken up by Attila Piróth. I acknowledge his work.

Finally, I am indebted to my family, to my wife and children, for their patience during all those years when I spent evenings and weekends with writing this book.

Budapest, August 2008 Jenő Sólyom

Contents

Contents Volume 1: Structure and Dynamics

16

Free-Electron Model of Metals

The first volume of this series was primarily concerned with the structure of condensed matter. We studied whether an order is present in the arrangement of the atoms or ions that determine the overall structure of the solid, and then analyzed the dynamics of crystal lattices in the ordered (crystalline) phase. Next, we treated a very similar problem presented by magnetically ordered materials, determining the possible ordered arrangements of localized atomic magnetic moments and the elementary excitations arising from the dynamics of the moments. Various thermodynamic and magnetic properties of solids could be explained in terms of these.

A second constituent of solids, which is perhaps even more important than the ion cores, is the system of electrons that are not tightly bound in the inner shells of the ions and that form metallic bonds or valence bonds between ion cores. In this volume we shall be concerned with the behavior of such electrons – which participate in bonding, and thus affect substantially the properties of solids. Throughout, we shall employ the one-particle approximation; the analysis of the role of electron–electron interactions will be the subject of the third volume.

Soon after the discovery of the electron,[1] a straightforward explanation was suggested for the most characteristic properties of metals (good electrical and thermal conductivity) in terms of *conduction electrons*, i.e., electrons moving freely in metals. The successful description of the behavior of nearly ideal gases at the end of the 19th century, due primarily to L. Boltzmann's contributions to classical statistical mechanics, enabled P. Drude (1900) to apply a simple formulation of the kinetic theory of gases to a gas of electrons. This model was developed further by H. A. Lorentz (1905) to give a more detailed account of conduction phenomena. This model of the classical gas of free electrons is called the *Drude model* or *Drude–Lorentz model*.

Despite its initial success, the inadequacies of the model soon transpired. These were rooted in the fact that the gas of electrons – if it can be consid-

[1] J. J. Thomson, 1897.

ered a gas of weakly interacting particles at all – can hardly ever be treated classically. Electrons, as quantum mechanical objects, obey the *Fermi–Dirac statistics*.[2] Immediately after the formulation of quantum mechanics, Sommerfeld and his co-workers gave a much better – albeit still imperfect – description of the electronic properties of metals based on the new theory. The present chapter is devoted to this free-electron model. Before turning to the quantum mechanical treatment, we shall briefly overview the classical description as its terminology and underlying physical picture are often of great use even today.

Even though the approximation of lumping all interactions into the collision time of electrons sounds extremely rough at first, the model nevertheless provides a surprisingly good explanation of those properties of metals that are determined by the motion of electrons in applied electric or magnetic fields. However, it does not explain either the existence of nonmetallic materials (insulators and semiconductors) or the properties of superconductors.

Throughout the present volume we shall deal with these problems, assuming that the interactions between electrons are almost negligible. Among others, the theories discussed in Volume 3 aim to provide an explanation for this electron-gas-like behavior, and to point out under what circumstances interactions between electrons play an important role.

16.1 Classical Drude Model

Calculated from the atomic mass, material density, and the well-known value of the Avogadro constant, the electron density is estimated to be a few times 10^{22} per cm^3 in simple metals. This is much larger than the usual densities in gases. If the ideal gas law is assumed to be valid, the pressure of the electron is found to be on the order of a thousand atmospheres. Despite this unrealistically high value we shall assume below that electrons in solids behave like particles of an almost ideal gas confined to a finite box.

16.1.1 Basic Assumptions of the Model

According to Drude's assumption, Z electrons get detached from each atom, and in metals they fill the space between atoms essentially uniformly. As they are responsible for metallic conduction, they are called conduction electrons. By making the straightforward assumption that core electrons do not contribute to the electrical conductivity, Z can be identified with the number of valence electrons, that is, electrons on the outermost open shell in the atomic configuration. The legitimacy of this assumption is supported by the resistivity data of alkali metals and noble metals that hardly increase with increasing atomic number.

[2] E. Fermi and P. A. M. Dirac, 1926.

The conduction electrons have strong attractive and repulsive Coulomb interactions with ion cores and other conduction electrons, respectively. Since the system as a whole is neutral, it can be assumed that the strong Coulomb attraction due to ion cores is essentially compensated for by the repulsion due to other conduction electrons. In this approximation electrostatic forces can be neglected, and the system of electrons can be considered as a gas of free, neutral particles. However, this rough picture requires substantial improvement.

The neglect of direct collisions with ions can be justified if the volume filled by ion cores is small compared to the volume occupied by electrons. To evaluate their ratio, each atom is associated with a sphere whose volume is one atom's share of the total volume. The radius r_{WS} of this *Wigner–Seitz sphere*[3] is determined from the formula

$$\frac{4\pi r_{\mathrm{WS}}^3}{3} = \frac{V}{N}\,, \tag{16.1.1}$$

where V is the volume of the sample and N is the number of atoms. Values of r_{WS} – either calculated from the lattice constant or obtained from the density, atomic weight, and the Avogadro constant – are listed in Table 16.1 for some simple metals.

Table 16.1. Wigner–Seitz radius for some simple metals

Element	r_{WS} (Å)	Element	r_{WS} (Å)	Element	r_{WS} (Å)	Element	r_{WS} (Å)	Element	r_{WS} (Å)
Li	1.73	Be	1.25						
Na	2.07	Mg	1.78					Al	1.59
K	2.62	Ca	2.18	Cu	1.41	Zn	1.54	Ga	1.67
Rb	2.81	Sr	2.27	Ag	1.60	Cd	1.73	In	1.83
Cs	2.99	Ba	2.47	Au	1.59	Hg	1.76	Tl	1.89

The value of the Wigner–Seitz radius is seen to be usually around 2 Å. However, in alkali metals, alkaline-earth metals, and other simple metals, which are all considered to be good metals, the radius of the ion core is around 1 Å (see Chapter 4 of Volume 1). This implies that only one-eighth to one-tenth of the total volume of the metal is occupied by ion cores. According to the model, conduction electrons can move around freely in the remaining volume, it can therefore be assumed that collisions occur relatively infrequently.

Even though the net Coulomb interaction is weak and collisions with ions are rare, the finiteness of the metallic conductivity cannot be understood without taking these scattering events into account. We shall discuss in detail the

[3] E. P. Wigner and F. Seitz, 1933.

interaction between electrons and the lattice of ions in later chapters. Below we shall use the simplification that, just like in the kinetic theory of gases, these collisions can be characterized by the collision time τ (also known as the relaxation time or the mean free time). Physically, $\mathrm{d}t/\tau$ is the probability that a randomly selected electron experiences a collision in time $\mathrm{d}t$. It also means that the probability that no collision occurs to the same electron for time t after (or before) this collision is $\mathrm{e}^{-t/\tau}$, while the probability that the time between two consecutive collisions is between t and $t + \mathrm{d}t$ is $\mathrm{e}^{-t/\tau}\mathrm{d}t/\tau$. This implies that an electron undergoes collisions once every τ on the average.

In this phenomenological treatment the relaxation time τ is a fundamental parameter of the theory. Instead of trying to derive it from first principles, we shall determine its value from experiments. Only after the discussion of scattering processes shall we turn to the quantum mechanical treatment of the relaxation time (Chapter 24).

According to the laws of classical statistical physics, the velocity distribution of particles in the thermal equilibrium arising from collisions is described by the Maxwellian velocity distribution,[4] also known as the Maxwell–Boltzmann distribution.[5] The probability that the speed of an electron of mass m_e is between v and $v + \mathrm{d}v$ in an isotropic system is

$$f_0(v) 4\pi v^2 \mathrm{d}v = \left(\frac{m_\mathrm{e}}{2\pi k_\mathrm{B} T}\right)^{3/2} \exp\left(-\frac{m_\mathrm{e} v^2}{2 k_\mathrm{B} T}\right) 4\pi v^2 \, \mathrm{d}v \,. \tag{16.1.2}$$

Denoting the number of electrons per unit volume by n_e, the density of electrons of velocity $\boldsymbol{v}$ is

$$\rho_\mathrm{e}(\boldsymbol{v}) = n_\mathrm{e} f_0(|\boldsymbol{v}|) = n_\mathrm{e} \left(\frac{m_\mathrm{e}}{2\pi k_\mathrm{B} T}\right)^{3/2} \exp\left(-\frac{m_\mathrm{e} v^2}{2 k_\mathrm{B} T}\right) . \tag{16.1.3}$$

Using (C.2.5) [see Volume 1], it can be shown that for this classical Maxwell–Boltzmann distribution

$$\langle v^2 \rangle = \int_0^\infty v^2 f_0(v) \, 4\pi v^2 \, \mathrm{d}v = 3 \frac{k_\mathrm{B} T}{m_\mathrm{e}} \,, \tag{16.1.4}$$

that is, the mean kinetic energy of electrons satisfies the equipartition theorem:

$$\langle \varepsilon \rangle = \tfrac{1}{2} m_\mathrm{e} \langle v^2 \rangle = \tfrac{3}{2} k_\mathrm{B} T \,. \tag{16.1.5}$$

It is readily seen that the heat capacity per electron is then

$$c = \tfrac{3}{2} k_\mathrm{B} \,, \tag{16.1.6}$$

[4] J. C. Maxwell, 1859.

[5] L. Boltzmann, 1871.

while the specific heat per unit volume[6] is

$$c_{\mathrm{el}} = \tfrac{3}{2} n_{\mathrm{e}} k_{\mathrm{B}} \,. \tag{16.1.7}$$

At not too low temperatures the specific heat of solids is indeed found to be independent of the temperature, however the molar heat of divalent metals is not twice as much as that of monovalent metals. It was shown in Chapter 12 that this temperature-independent contribution to the specific heat does note come from the motion of electrons but from the vibrations of the lattice. The electronic contribution to specific heat is much smaller. As mentioned in connection with Fig. 12.8, this contribution is not constant but proportional to temperature. It can only be observed at low temperatures where the phonon contribution is also smaller than the classical value and vanishes as T^3. The classical model thus badly overestimates the electronic contribution to specific heat.

Identifying the obtained value of the kinetic energy with the energy of an electron moving with an average speed of $\overline{v}$, we have

$$\overline{v} = \left(\frac{3 k_{\mathrm{B}} T}{m_{\mathrm{e}}} \right)^{1/2} . \tag{16.1.8}$$

At room temperature this thermal velocity is around 10^5 m/s. Moving at this speed, the electron travels a mean distance of

$$l = \overline{v}\, \tau = \left(\frac{3 k_{\mathrm{B}} T}{m_{\mathrm{e}}} \right)^{1/2} \tau \tag{16.1.9}$$

between two successive collisions. This is the *mean free path* of electrons. To evaluate it, the collision time has to be known. As mentioned above, this can be derived from the electrical resistivity.

16.1.2 Electrical Conductivity

When a metal is placed into a uniform electric field $\boldsymbol{E}$, electrons acquire, beyond their thermal motion, an additional *drift velocity*. Following DRUDE, we shall assume that the part of the kinetic energy that an electron acquires from the electric field is entirely lost (dissipated) in the collisions – that is, immediately after a collision the velocity of electrons is described by the Maxwell–Boltzmann distribution, which corresponds to thermal equilibrium, irrespective of the velocity distribution prior to the collision. When averaged over directions, the mean velocity is zero right after collisions; electrons then

[6] Strictly speaking, specific heat capacity (or specific heat) should be used only for the heat capacity per unit mass. However, we shall follow common practice and use specific heat in a broader sense, for heat capacity per unit volume and heat capacity per unit amount of substance (molar heat capacity).

accelerate for a short time. As we shall see, the drift velocity is much smaller than the root-mean-square (rms) velocity of thermal motion, therefore the energy loss is tiny on the thermal energy scale, consequently collisions can be considered nearly elastic. However, scattering is not strictly elastic, and it is precisely this inelasticity that leads to the establishment of thermal equilibrium.

The electrostatic force on electrons, $\boldsymbol{F} = -e\boldsymbol{E}$, would change the velocity of each electron by

$$\boldsymbol{v}_{\mathrm{dr}}(t) = -\frac{e\boldsymbol{E}t}{m_{\mathrm{e}}} \tag{16.1.10}$$

in time t. According to our previous assumptions, regardless of the choice of the particular moment, the average lapse of time since the last collision is τ, so the mean drift velocity of electrons is

$$\boldsymbol{v}_{\mathrm{dr}} = -\frac{e\boldsymbol{E}\tau}{m_{\mathrm{e}}}\,. \tag{16.1.11}$$

In what follows – and especially in the discussion of semiconductors – we shall repeatedly write the relation between drift velocity and electric field strength as

$$\boldsymbol{v}_{\mathrm{dr}} = -\mu\,\boldsymbol{E}\,, \tag{16.1.12}$$

where μ is the carrier mobility. For free electrons $\mu = e\tau/m_{\mathrm{e}}$.

The isotropic thermal motion of electrons gives a vanishing contribution to electrical conductivity, therefore it can be ignored. Conduction phenomena can then be treated as if particles of charge $-e$ were drifting with the same speed $\boldsymbol{v}_{\mathrm{dr}}$ opposite to the field direction. In reality, even when thermal motion is neglected, electrons accelerate in the electric field, stop upon collision, and then start to accelerate again. This motion can be approximated by a uniform motion at the average velocity. Note that if each electron collided at regular intervals τ with the ion cores then the average drift velocity would be only half of the value given in (16.1.11). DRUDE used this assumption in his original calculations. In a more careful calculation the time between collisions is assumed to follow a Poisson distribution. Because of acceleration, particles that collide less frequently than the average get much farther and acquire an above-average terminal velocity. That is why the mean velocity of electrons is given by (16.1.11).

The description of the average motion of electrons can also be based on a classical equation of motion that will be used repeatedly later, too. Besides the electrostatic force that accelerates electrons in the electric field, the expression contains a phenomenological damping (relaxation) term that describes how collisions hinder the free motion of electrons and how the drift velocity relaxes to zero. Since the momentum and energy acquired from the field dissipates over an average period of τ, the equation of motion reads

$$m_{\mathrm{e}}\frac{\mathrm{d}\boldsymbol{v}_{\mathrm{dr}}}{\mathrm{d}t} = -e\boldsymbol{E} - \frac{m_{\mathrm{e}}\boldsymbol{v}_{\mathrm{dr}}}{\tau}\,. \tag{16.1.13}$$

When the transients following the switch-on of the electric field have decayed and a stationary state has been established, (16.1.11) is recovered.

To determine the electric current induced by the field, the number $\mathrm{d}N_\mathrm{e}$ of particles passing through a surface element $\mathrm{d}S$ perpendicular to the propagation direction of the particles in time $\mathrm{d}t$ has to be calculated. According to the kinetic theory of gases,

$$\mathrm{d}N_\mathrm{e} = n_\mathrm{e} v_\mathrm{dr}\,\mathrm{d}S\,\mathrm{d}t\,. \tag{16.1.14}$$

The amount of charge passing through the surface in time $\mathrm{d}t$ is

$$\mathrm{d}Q = -e n_\mathrm{e} v_\mathrm{dr}\,\mathrm{d}S\,\mathrm{d}t\,; \tag{16.1.15}$$

differentiation with respect to time and surface area then gives the current density,

$$j = -e n_\mathrm{e} v_\mathrm{dr}\,. \tag{16.1.16}$$

By treating the velocity and current of electrons as vector quantities,

$$\boldsymbol{j} = -e n_\mathrm{e} \boldsymbol{v}_\mathrm{dr}\,. \tag{16.1.17}$$

Through (16.1.11), this leads to

$$\boldsymbol{j} = \frac{n_\mathrm{e} e^2 \tau}{m_\mathrm{e}} \boldsymbol{E}\,. \tag{16.1.18}$$

In an isotropic material *Ohm's law*[7] can be written in terms of the scalar resistivity ϱ or the conductivity σ:

$$\boldsymbol{E} = \varrho \boldsymbol{j}\,, \qquad \boldsymbol{j} = \sigma \boldsymbol{E}\,. \tag{16.1.19}$$

It follows from (16.1.18) that in the Drude model, conductivity is given by

$$\boxed{\sigma = \frac{n_\mathrm{e} e^2 \tau}{m_\mathrm{e}}\,.} \tag{16.1.20}$$

Note that this relation is valid in SI and CGS units alike, however the numerical value and unit of the elementary charge – and with it, those of conductivity – are different in the two systems. Apart from very low and very high temperatures (around the melting point), the typical resistivity of metals is on the order of 1–100 nΩ m. As shown in Table 16.2, for most metals this value is between 10 and 100 nΩ m at room temperature, and between 1 and 10 nΩ m at liquid-nitrogen temperature.[8]

Starting with the above values of resistivity, and using the known value of the density of conduction electrons, the relaxation (or collision) time can

[7] G. S. Ohm, 1827.

[8] The liquefaction temperature of nitrogen (77 K) is a standard reference for experiments in solid-state physics, just like the liquid-helium temperature (4.2 K).

Table 16.2. Resistivity at 77 K and 273 K, and the relaxation (collision) time calculated from it for some metallic elements (10^{-15} s = 1 fs)

Element	ϱ(77 K)	ϱ(273 K)	τ(77 K)	τ(273 K)
	(nΩ m)		(10^{-15} s)	
Ag	2.8	14.7	200	40
Al	2.3	25.0	65	8.0
Au	4.7	20.5	120	30
Ba	67	500	17	3.8
Bi	350	1068	0.72	0.23
Cu	2.0	16.8	210	27
Fe	6.4	89	32	2.4
Ga	27.5	136	8.8	1.7
Na	8	42	170	32
Pb	47	192	5.7	1.4
Sb	80	370	2.7	0.55
Zn	10.4	54.3	24	4.9

be determined. At room temperature $\tau \sim 10^{-14}$ to 10^{-15} s is found, while at liquid-nitrogen temperature the value is an order of magnitude higher, as shown in Table 16.2. Thus 10^{-14} s and 10^{-13} s can be considered as typical relaxation times in good metals. Using (16.1.11), the drift velocity for electrons can then be estimated. For $\tau = 10^{-14}$ s, in a field of 10^{-2} V/cm, which is attainable in good metals, the drift velocity is 10^{-3} m/s, which is indeed several orders of magnitude smaller than the velocity of thermal motion (10^5 m/s) obtained from (16.1.8) at room temperature.

However, more careful studies reveal the inadequacy of this classical picture. If the mean free path of electrons is calculated using $\tau = 10^{-14}$ s and $v_{\text{therm}} = 10^5$ m/s, we obtain $l \sim 10^{-9}\,\text{m} = 10\,\text{Å}$, which is comparable to the distance of atoms. This result seems to contradict our previous assumption about the relatively free propagation of electrons. Such a short electron mean free path is not compatible with experimental results, either. Resistance measurements at low temperatures show that conductivity depends sensitively on the purity of the sample, even in samples where the separation of impurities is much larger than the lattice constant. The temperature dependence is also problematic. If the mean free path were identified with the atomic spacing, the temperature dependence of the velocity of thermal motion would give rise to a factor $T^{-1/2}$ in the relaxation time – and hence in the conductivity as well. No such dependence is observed in experiments; moreover, in the $T \to 0$ limit the conductivity of metals tends to a purity-dependent saturation value – unless they become superconductors.

This means that scattering by relatively distant impurities whose arrangement lacks order plays a more important role in determining resistivity than collisions with the atoms of the regular lattice, and the mean free path of electrons can greatly exceed the lattice constant. It also questions the applicability of the Maxwell velocity distribution for electrons and the classical method for determining resistivity. In the second part of the chapter, which goes beyond the classical treatment, we shall demonstrate that the speed of those electrons that are responsible for electrical resistivity is usually an order of magnitude higher than the classical value obtained from thermal motion, and thus even at room temperature the mean free path is much larger than atomic distances. At low temperatures, where the resistivity of high-purity copper can be as low as $10^{-12}\,\Omega\,\mathrm{m}$, the mean free path may be on the order of a few mm.

In Chapter 24 devoted to the study of the transport properties of solids we shall show that the current of electrons would flow without resistance in a regular rigid lattice. Resistivity is due to elastic scattering by impurities (for which momentum conservation is no longer valid) and inelastic scattering by the regular but vibrating lattice of ions. Nevertheless we shall see that the expression obtained for the conductivity from a more precise treatment is often well approximated by (16.1.20), even though the concept of relaxation time and the electron mass used in the relation have to be refined. If the relaxation time can be determined for all relevant scattering processes by theoretical considerations, the resistivity can be obtained – at least, in principle – from the Drude formula. Quite often the reverse path is taken, just as above: the relaxation time is determined from the measured value of resistivity.

16.1.3 Heat Conduction

If there is a temperature difference between the two ends of a sample, then some of the electrons move from the hot side to the cold side via thermal diffusion, transporting energy. According to Fourier's law[9] of heat conduction, the heat current $\boldsymbol{j}_Q$ resulting from energy transport is proportional to the temperature gradient $\boldsymbol{\nabla} T$:

$$\boldsymbol{j}_Q = -\lambda \boldsymbol{\nabla} T\,, \tag{16.1.21}$$

where λ is the thermal (or heat) conductivity of the material. The justification of the choice of the negative sign is that λ is then positive, as energy (heat) flows against the direction of the temperature gradient, from the higher-temperature part to the lower-temperature one. The room-temperature thermal conductivity of a few metals, semimetals, and semiconductors are listed in Table 16.3.

It was already mentioned in Chapter 12 in connection with phononic heat conduction that according to the kinetic theory of gases, the thermal conductivity λ can be expressed in terms of the mean free path l, the average thermal speed $\overline{v}$ of particles and the specific heat c. Instead of (12.4.20), the thermal conductivity of the system of electrons is now expressed as

[9] J. B. J. Fourier, 1822.

Table 16.3. The room-temperature thermal conductivity of a few metals, semimetals, and semiconductors

Element	λ (W m^{-1} K^{-1})	Element	λ (W m^{-1} K^{-1})
Ag	429	Ge	58.6
Al	237	Sb	25.9
Au	317	Se	2.48
Bi	7.87	Si	83.7
Cu	401	Zn	121

$$\lambda = \tfrac{1}{3} l \,\overline{v}\, c_{\mathrm{el}} \,. \tag{16.1.22}$$

The mean free path is $l = \overline{v}\tau$, while $\overline{v}$ is known from (16.1.8) and c_{el} from (16.1.7). Collecting all these terms, we have

$$\lambda = \frac{3}{2}\frac{n_{\mathrm{e}}\tau}{m_{\mathrm{e}}} k_{\mathrm{B}}^2 T \,. \tag{16.1.23}$$

The thermal conductivity of an electron gas is thus proportional to temperature.

Thermal conductivity can be related to electrical conductivity in a particularly simple way. Making use of the result in (16.1.20),

$$\boxed{\frac{\lambda}{\sigma} = \frac{3}{2}\left(\frac{k_{\mathrm{B}}}{e}\right)^2 T \,.} \tag{16.1.24}$$

The ratio of thermal and electrical conductivity is proportional to temperature in the Drude model, and the constant of proportionality is independent of the material properties as it contains only k_{B} and e. This is in good agreement with an earlier experimental finding, the *Wiedemann–Franz law*[10] – which states that the ratio of λ and σ, both measured at the same temperature, is independent of material properties –, and also with L. V. Lorenz's observation (1872) that this ratio is proportional to temperature. The constant of proportionality is called the *Lorenz number* or *Lorenz coefficient.* In the Drude model its value is

$$L = \frac{3}{2}\left(\frac{k_{\mathrm{B}}}{e}\right)^2 = 1.11 \times 10^{-8}\,\mathrm{V}^2\,\mathrm{K}^{-2} \,. \tag{16.1.25}$$

As listed in Table 16.4, the experimental value of the Lorenz number is 2–$3 \times 10^{-8}\,\mathrm{V}^2\,\mathrm{K}^{-2}$ for most metals. On account of a mistake,[11] this was considered as the most compelling evidence for the correctness of the Drude model.

[10] G. Wiedemann and R. Franz, 1853.

[11] For the reasons discussed on page 6, Drude's original formula for the conductivity contained $\tau/2$ instead of τ, and thus his theoretical estimate for the Lorenz number was twice as large as the value in (16.1.25).

Table 16.4. The experimental value of the Lorenz number at 0 °C and 100 °C for some simple metals

Metal	$L(0°\mathrm{C})$	$L(100°\mathrm{C})$	Metal	$L(0°\mathrm{C})$	$L(100°\mathrm{C})$
	$(10^{-8}\,\mathrm{V}^2\,\mathrm{K}^{-2})$			$(10^{-8}\,\mathrm{V}^2\,\mathrm{K}^{-2})$	
Ag	2.31	2.38	Li	2.22	2.43
Au	2.35	2.36	Mo	2.61	2.79
Al	2.14	2.19	Pb	2.47	2.53
Cu	2.23	2.29	Pt	2.51	2.60
Fe	2.61	2.88	W	3.04	3.20
Ir	2.49	2.49	Zn	2.28	2.30

This factor aside, the correct order of magnitude of the Lorenz number is still surprising because the Drude model gives very bad estimates for each of the three quantities in the thermal conductivity (16.1.22) derived from the kinetic theory of gases – the mean free path, the mean velocity and the electronic specific heat[12] –, but the errors compensate. To obtain the correct value of the Lorenz number, a quantum mechanical treatment is necessary. Measurements at low temperatures would show serious deviations from the values given in the table. The conditions for the applicability of the Wiedemann–Franz law and the explanation of the deviations from it require more careful investigations.

16.1.4 Hall Resistance

When a sample in which an electric current is flowing is placed in a magnetic field that is perpendicular to the current flow, electrons are deviated from their rectilinear path by the well-known Lorentz force of classical electrodynamics. Choosing the x-axis along the electric field that drives the current and the z-axis along the magnetic field, an additional electric field is induced in the y-direction, as shown in Fig. 16.1. Consequently, a transverse voltage called the *Hall voltage*[13] is observed across the sample. The ratio of the transverse voltage and the longitudinal current is the *Hall resistance.*

Resistivity against a current in the x-direction can be determined from

$$\varrho(B) = \frac{E_x}{j_x}\,. \tag{16.1.26}$$

In principle, this can be different from the value obtained in the absence of a magnetic field, leading to the magnetic-field dependence of resistivity. The

[12] The Drude model predicts a temperature-independent electronic specific heat that is several orders of magnitude higher than the experimental value, while it seriously underestimates the mean electron velocity, which it predicts to be temperature dependent.

[13] E. H. Hall, 1879.

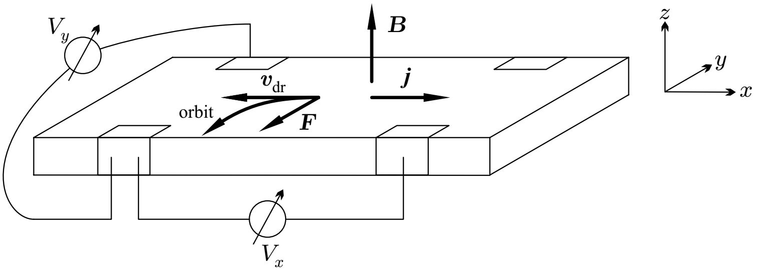


Fig. 16.1. Measurement setup for the Hall resistance. When the current is along the x-axis, the electron drift velocity is in the opposite $(-x)$ direction, and a magnetic field along the z-axis would deviate them in the $-y$-direction

change in the electrical resistance of a material upon the application of a magnetic field is called *magnetoresistance.*

The *Hall coefficient* is defined by

$$R_{\mathrm{H}} = \frac{E_y}{j_x B} \,. \tag{16.1.27}$$

To evaluate it, the drift velocity in (16.1.17) needs to be specified. In the presence of a magnetic field its value is obtained from a generalization of the equation of motion (16.1.13) that takes the Lorentz force into account, too:

$$m_{\mathrm{e}} \frac{\mathrm{d}\boldsymbol{v}_{\mathrm{dr}}}{\mathrm{d}t} = -e\left(\boldsymbol{E} + \boldsymbol{v}_{\mathrm{dr}} \times \boldsymbol{B}\right) - \frac{m_{\mathrm{e}} \boldsymbol{v}_{\mathrm{dr}}}{\tau} \,. \tag{16.1.28}$$

In the stationary state

$$-e\left(\boldsymbol{E} + \boldsymbol{v}_{\mathrm{dr}} \times \boldsymbol{B}\right) - \frac{m_{\mathrm{e}} \boldsymbol{v}_{\mathrm{dr}}}{\tau} = 0 \,. \tag{16.1.29}$$

For notational simplicity, we shall suppress the label “dr” of drift. Writing out the equation in component form,

$$v_x = -\frac{e\tau}{m_{\mathrm{e}}} E_x - \frac{eB}{m_{\mathrm{e}}} \tau v_y \,, \tag{16.1.30-a}$$

$$v_y = -\frac{e\tau}{m_{\mathrm{e}}} E_y + \frac{eB}{m_{\mathrm{e}}} \tau v_x \,, \tag{16.1.30-b}$$

$$v_z = -\frac{e\tau}{m_{\mathrm{e}}} E_z \,. \tag{16.1.30-c}$$

In measurements of the Hall effect current flows only along the x-axis, i.e., $v_y = v_z = 0$. The transverse Hall voltage must precisely compensate for the deflection due to the Lorentz force. From (16.1.30-a),

$$v_x = -\frac{e\tau}{m_{\mathrm{e}}} E_x \,, \tag{16.1.31}$$

which is the same as the form obtained in the absence of the magnetic field. The x component of the current and thus the resistivity are independent of the magnetic field applied in the perpendicular direction: $\varrho(B) = 1/\sigma_0$, where σ_0 is the conductivity given in (16.1.20). Thus the resistivity does not depend on the magnetic field – i.e., there is no magnetoresistance – in the Drude model.

However, by eliminating v_x in favor of j_x in (16.1.30-b), and making use of (16.1.17),

$$E_y = Bv_x = -B\frac{j_x}{n_e e} \tag{16.1.32}$$

is obtained. From the definition (16.1.27) of the Hall coefficient,

$$\boxed{R_\mathrm{H} = -\frac{1}{n_\mathrm{e} e}\,.} \tag{16.1.33}$$

The negative sign is the consequence of the specific choice of the measurement geometry. When the current flow is along the x-direction, electrons move in the $-x$-direction. They are then deflected in the $-y$-direction by the Lorentz force – that is, the current is deflected in the y-direction. This is compensated for by the negative field along the y-axis.

In simple, above all monovalent metals a fairly good agreement is found between the Hall coefficients determined theoretically from the number of carriers and measured in experiments: their ratio is close to unity, as can be inferred for the elements in the first column of Table 16.5. On the other hand, the fourth column contains some metals for which the agreement is poor: experimental and calculated values differ not only in magnitude but sometimes even in sign, as if carriers were positively charged. This observation does not lend itself to interpretation in the framework of the Drude model.

Table 16.5. Measured and calculated Hall coefficients of some metals in not too strong magnetic fields around room temperature

Element	$R_\mathrm{H}^\mathrm{exp}$	R_H^th	Element	$R_\mathrm{H}^\mathrm{exp}$	R_H^th
	$(10^{-10}\,\mathrm{m^3\,s^{-1}\,A^{-1}})$			$(10^{-10}\,\mathrm{m^3\,s^{-1}\,A^{-1}})$	
Li	−1.7	−1.31	Be	+2.4	−0.25
Na	−2.1	−2.36	Zn	+0.63	−0.46
K	−4.2	−4.46	Cd	+0.59	−0.65
Cu	−0.54	−0.74	Pb	+0.09	−0.47
Ag	−0.84	−1.04	As	+450	−0.50
Au	−0.71	−1.05	Sb	+270	−0.43
Al	−0.34	−0.34	Bi	−6330	−0.44

Instead of the Hall coefficient, the Hall effect is sometimes characterized by the Hall angle θ defined by

$$\tan\theta = \frac{E_y}{E_x}\,. \tag{16.1.34}$$

From (16.1.32) and the expression for the current,

$$\tan\theta = -\frac{e\tau}{m_{\mathrm{e}}}B \equiv -\mu B\,. \tag{16.1.35}$$

Thus the Hall angle is directly related to electron mobility.

In a more general geometry of the measurement setup the relation between the current and the electric field can be cast in the form

$$\boldsymbol{E} = \varrho\boldsymbol{j} + R_{\mathrm{H}}\left(\boldsymbol{B}\times\boldsymbol{j}\right) \tag{16.1.36}$$

in the presence of a magnetic field. This leads to a tensorial relation between the two quantities even in isotropic systems. When the magnetic field is along the z-axis, the resistivity tensor is

$$\hat{\varrho}(B) = \begin{pmatrix} \varrho & -R_{\mathrm{H}}B & 0 \\ R_{\mathrm{H}}B & \varrho & 0 \\ 0 & 0 & \varrho \end{pmatrix}. \tag{16.1.37}$$

Its diagonal elements are the transverse and longitudinal resistivities, while its off-diagonal elements are related to the Hall coefficient. The conductivity tensor is its inverse:

$$\hat{\sigma}(B) = \begin{pmatrix} \dfrac{\varrho}{\varrho^2+(R_{\mathrm{H}}B)^2} & \dfrac{R_{\mathrm{H}}B}{\varrho^2+(R_{\mathrm{H}}B)^2} & 0 \\ -\dfrac{R_{\mathrm{H}}B}{\varrho^2+(R_{\mathrm{H}}B)^2} & \dfrac{\varrho}{\varrho^2+(R_{\mathrm{H}}B)^2} & 0 \\ 0 & 0 & \dfrac{1}{\varrho} \end{pmatrix}. \tag{16.1.38}$$

16.1.5 AC Conductivity

The conductivity in response to an alternating electric field of angular frequency ω can also be determined easily in the Drude model. Just like in (16.1.13), we start with the equation of motion for electrons, however we are not seeking stationary solutions now but assume that once transients have decayed the current is of the same angular frequency ω as the applied electric field. Assuming the same time dependence, $\exp(-\mathrm{i}\omega t)$, for both quantities, the following equation is obtained for the frequency-dependent amplitudes:

$$-\mathrm{i}\omega m_{\mathrm{e}}\boldsymbol{v}(\omega) = -e\boldsymbol{E}(\omega) - \frac{m_{\mathrm{e}}\boldsymbol{v}(\omega)}{\tau}\,. \tag{16.1.39}$$

Substituting the solution of this equation into (16.1.17), we have

$$\boldsymbol{j}(\omega) = -en_{\mathrm{e}}\boldsymbol{v}(\omega) = \frac{n_{\mathrm{e}}e^2}{m_{\mathrm{e}}(1/\tau - \mathrm{i}\omega)}\boldsymbol{E}(\omega)\,, \tag{16.1.40}$$

which implies

$$\sigma(\omega) = \frac{n_{\mathrm{e}}e^2}{m_{\mathrm{e}}(1/\tau - \mathrm{i}\omega)} = \frac{\sigma_0}{1 - \mathrm{i}\omega\tau} \tag{16.1.41}$$

for the frequency-dependent (or AC) conductivity. Separating the complex conductivity into real and imaginary parts,

$$\operatorname{Re}\sigma(\omega) = \frac{\sigma_0}{1 + (\omega\tau)^2}\,, \qquad \operatorname{Im}\sigma(\omega) = \frac{\sigma_0\omega\tau}{1 + (\omega\tau)^2}\,. \tag{16.1.42}$$

Their variations with frequency are shown in Fig. 16.2.

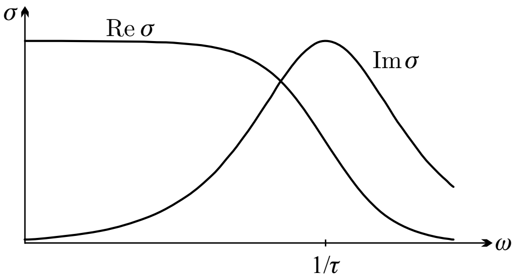


Fig. 16.2. Semi-logarithmic plot of the frequency dependence of the real and imaginary parts of the conductivity in the Drude model

The real part of conductivity is related to resistivity, i.e., the absorption of energy that gives rise to Joule heating.[14] The Lorentzian peak around $\omega = 0$ is called the Drude peak. The full width at half maximum (FWHM) of the frequency dependence is determined by the relaxation time: $\Delta\omega \sim 1/\tau$. On the other hand, the imaginary part is inductive in character because of the phase shift, and has its maximum at $\omega\tau = 1$.

In the limit where all scattering processes can be neglected, the relaxation time tends to infinity ($\tau^{-1} \to 0$). Using (C.3.1-c), it is readily seen from

$$\sigma(\omega) = \mathrm{i}\frac{n_{\mathrm{e}}e^2}{m_{\mathrm{e}}}\frac{1}{\omega + \mathrm{i}/\tau} \tag{16.1.43}$$

that the real part of the conductivity can be written as

$$\operatorname{Re}\sigma(\omega) = D_{\mathrm{c}}\,\delta(\omega)\,, \tag{16.1.44}$$

[14] J. P. Joule, 1840.

where

$$D_{\mathrm{c}} = \pi \frac{n_{\mathrm{e}} e^2}{m_{\mathrm{e}}} \tag{16.1.45}$$

is the *Drude weight*, and the imaginary part is

$$\operatorname{Im} \sigma(\omega) = \frac{n_{\mathrm{e}} e^2}{m_{\mathrm{e}} \omega} . \tag{16.1.46}$$

This form is valid for finite values of τ as well, provided the frequency satisfies the condition $\omega \gg 1/\tau$.

It should be noted for future reference that the real and imaginary parts of conductivity are not independent of one another. By extending the formula (16.1.41) of the frequency-dependent conductivity to complex values of ω, a pole is found at $\omega = -\mathrm{i}/\tau$, i.e., in the lower half-plane, while the function is analytic in the upper half-plane. As it will be shown in Appendix J of Volume 3, the Cauchy relations for analytic functions imply the Kramers–Kronig relations between the real and imaginary parts:

$$\begin{aligned} \operatorname{Re} \sigma(\omega) &= \frac{1}{\pi} \mathrm{P} \int\limits_{-\infty}^{\infty} \mathrm{d}\omega' \, \frac{\operatorname{Im} \sigma(\omega')}{\omega' - \omega} = \frac{2}{\pi} \mathrm{P} \int\limits_{0}^{\infty} \mathrm{d}\omega' \, \frac{\omega' \operatorname{Im} \sigma(\omega')}{\omega'^2 - \omega^2} \, , \\ \operatorname{Im} \sigma(\omega) &= -\frac{1}{\pi} \mathrm{P} \int\limits_{-\infty}^{\infty} \mathrm{d}\omega' \, \frac{\operatorname{Re} \sigma(\omega')}{\omega' - \omega} = -\frac{2\omega}{\pi} \mathrm{P} \int\limits_{0}^{\infty} \mathrm{d}\omega' \, \frac{\operatorname{Re} \sigma(\omega')}{\omega'^2 - \omega^2} \, , \end{aligned} \tag{16.1.47}$$

where P stands for the principal value.

16.1.6 High-Frequency Behavior of a Classical Electron Gas

The foregoing analysis is valid only at relatively low frequencies. At higher frequencies the magnetic component of the electromagnetic field cannot be neglected, and sometimes even spatial variations need to be taken into account. In such cases the complete set of Maxwell equations has to be used. In the customary notation

$$\begin{aligned} \operatorname{curl} \boldsymbol{H} &= \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{j}_{\mathrm{ext}} \, , \qquad & \operatorname{div} \boldsymbol{D} &= \rho_{\mathrm{ext}} \, , \\ \operatorname{curl} \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t} \, , \qquad & \operatorname{div} \boldsymbol{B} &= 0 \, , \end{aligned} \tag{16.1.48}$$

where ρ_{ext} denotes the density of external free charges, and $\boldsymbol{j}_{\mathrm{ext}}$ their current.

Below we shall determine the high-frequency behavior for a system without external charges in which the electromagnetic field can nevertheless induce a spatially and temporally varying charge distribution. This induced charge gives rise to the difference between the electric displacement and the electric field. The corresponding Maxwell equation can then be recast in the form

$$\epsilon_0 \operatorname{div} \boldsymbol{E} = \operatorname{div} \boldsymbol{D} + \rho_{\text{ind}} \,. \tag{16.1.49}$$

In the most general case the relationship between the electric field and the electric displacement is nonlocal but causal:

$$\boldsymbol{D}(\boldsymbol{r}, t) = \int \mathrm{d}\boldsymbol{r}' \int\limits_{-\infty}^{t} \mathrm{d}t' \, \epsilon(\boldsymbol{r} - \boldsymbol{r}', t - t') \boldsymbol{E}(\boldsymbol{r}', t') \,. \tag{16.1.50}$$

Similarly, the relationship between the magnetic field and the magnetic induction is nonlocal:

$$\boldsymbol{B}(\boldsymbol{r}, t) = \int \mathrm{d}\boldsymbol{r}' \int\limits_{-\infty}^{t} \mathrm{d}t' \, \mu(\boldsymbol{r} - \boldsymbol{r}', t - t') \boldsymbol{H}(\boldsymbol{r}', t') \,. \tag{16.1.51}$$

Since our free-electron model is isotropic, the permittivity relating $\boldsymbol{D}$ and $\boldsymbol{E}$ and the magnetic permeability relating $\boldsymbol{B}$ and $\boldsymbol{H}$ are both scalars. In fact this assumption is valid only for uniform electric and magnetic fields. For electromagnetic radiation propagating in a ponderable medium the propagation direction singles out a preferred direction, and the $\boldsymbol{D}$ and $\boldsymbol{E}$ (or $\boldsymbol{B}$ and $\boldsymbol{H}$) components parallel and perpendicular to this direction are not related by the same ϵ (μ) – that is, one has to distinguish longitudinal and transverse components of the permittivity (permeability) tensor. We shall come back to this point in Chapter 25 on optical properties. Below we shall work with a scalar permittivity and permeability.

Taking the Fourier transforms of (16.1.50) and (16.1.51), the following simple relations are obtained for the Fourier components:

$$\boldsymbol{D}(\boldsymbol{q}, \omega) = \epsilon(\boldsymbol{q}, \omega) \boldsymbol{E}(\boldsymbol{q}, \omega) \,, \qquad \boldsymbol{B}(\boldsymbol{q}, \omega) = \mu(\boldsymbol{q}, \omega) \boldsymbol{H}(\boldsymbol{q}, \omega) \,. \tag{16.1.52}$$

We shall therefore assume that the spatial and temporal variations of fields are specified by a function of the form $\exp(\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{r} - \mathrm{i}\omega t)$. The Maxwell equations then take the form

$$\begin{aligned} \mathrm{i}\boldsymbol{q} \times \boldsymbol{H} &= -\mathrm{i}\omega \boldsymbol{D} \,, & \mathrm{i}\boldsymbol{q} \cdot \boldsymbol{D} &= 0 \,, \\ \mathrm{i}\boldsymbol{q} \times \boldsymbol{E} &= \mathrm{i}\omega \boldsymbol{B} \,, & \mathrm{i}\boldsymbol{q} \cdot \boldsymbol{B} &= 0 \,. \end{aligned} \tag{16.1.53}$$

Rewriting the equation $\mathrm{i}\boldsymbol{q} \cdot \boldsymbol{D} = 0$ as

$$\mathrm{i}\epsilon(\boldsymbol{q}, \omega) \boldsymbol{q} \cdot \boldsymbol{E}(\boldsymbol{q}, \omega) = 0 \,, \tag{16.1.54}$$

a trivial solution is found immediately:

$$\boldsymbol{q} \cdot \boldsymbol{E}(\boldsymbol{q}, \omega) = 0 \,, \tag{16.1.55}$$

which implies that the electric field is perpendicular to the propagation direction: $\boldsymbol{E} \perp \boldsymbol{q}$. Then, just like for electromagnetic radiation in vacuum, transverse waves (photons) propagate in the medium. It also follows from the Maxwell equations that $\rho_{\mathrm{ind}}(\boldsymbol{q},\omega)=0$ in this case – that is, the charge distribution remains uniform. The propagation of photons in condensed matter will be discussed in more detail in Chapter 25.

Note that (16.1.54) has a longitudinal-field solution ($\boldsymbol{E}\parallel\boldsymbol{q}$) as well when charges are induced. Then the Maxwell equations imply $\boldsymbol{B}=0$, that is, this solution does not correspond to an electromagnetic wave but to a pure polarization wave. This wave appears for pairs of $\boldsymbol{q},\omega_{\mathrm{L}}(\boldsymbol{q})$ that satisfy the relation

$$\epsilon(\boldsymbol{q},\omega_{\mathrm{L}}(\boldsymbol{q}))=\epsilon_0\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega_{\mathrm{L}}(\boldsymbol{q}))=0\,, \tag{16.1.56}$$

where $\epsilon_{\mathrm{r}}=\epsilon/\epsilon_0$ is the relative permittivity or dielectric constant. Since it is a function of $\boldsymbol{q}$ and ω, the term *dielectric function* is also used. To determine the frequency ω_{L}, the dielectric function of the ideal gas of electrons has to be known. We shall start with the Fourier transform of (16.1.49),

$$\mathrm{i}\epsilon_0\boldsymbol{q}\cdot\boldsymbol{E}(\boldsymbol{q},\omega)=\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{D}(\boldsymbol{q},\omega)+\rho_{\mathrm{ind}}(\boldsymbol{q},\omega)\,. \tag{16.1.57}$$

Unless $\boldsymbol{q}$ and $\boldsymbol{E}$ are perpendicular,

$$\epsilon(\boldsymbol{q},\omega)=\epsilon_0+\mathrm{i}\frac{\rho_{\mathrm{ind}}(\boldsymbol{q},\omega)}{\boldsymbol{q}\cdot\boldsymbol{E}(\boldsymbol{q},\omega)} \tag{16.1.58}$$

and

$$\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega)=1+\mathrm{i}\frac{\rho_{\mathrm{ind}}(\boldsymbol{q},\omega)}{\epsilon_0\boldsymbol{q}\cdot\boldsymbol{E}(\boldsymbol{q},\omega)}\,. \tag{16.1.59}$$

To ensure charge conservation, the induced charge and the induced current must be related by the continuity equation

$$\operatorname{div}\boldsymbol{j}_{\mathrm{ind}}(\boldsymbol{r},t)+\frac{\partial\rho_{\mathrm{ind}}(\boldsymbol{r},t)}{\partial t}=0\,, \tag{16.1.60}$$

whose Fourier transform reads

$$\boldsymbol{q}\cdot\boldsymbol{j}_{\mathrm{ind}}(\boldsymbol{q},\omega)=\omega\rho_{\mathrm{ind}}(\boldsymbol{q},\omega)\,. \tag{16.1.61}$$

Expressing the current induced by the electric field in terms of the conductivity,

$$\boldsymbol{j}_{\mathrm{ind}}(\boldsymbol{q},\omega)=\sigma(\boldsymbol{q},\omega)\boldsymbol{E}(\boldsymbol{q},\omega)\,. \tag{16.1.62}$$

Comparison with the continuity equation gives

$$\rho_{\mathrm{ind}}(\boldsymbol{q},\omega)=\frac{\sigma(\boldsymbol{q},\omega)}{\omega}\boldsymbol{q}\cdot\boldsymbol{E}(\boldsymbol{q},\omega)\,. \tag{16.1.63}$$

Substituting this form into (16.1.58), the following relationship of the dielectric function and the conductivity is obtained:

$$\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega)=1+\mathrm{i}\frac{\sigma(\boldsymbol{q},\omega)}{\epsilon_0\omega}\,. \tag{16.1.64}$$

It is readily seen that the real part of the conductivity is related to the imaginary part of the dielectric function, while the imaginary part of the conductivity is related to the real part of the dielectric function. Just as for conductivity, the real and imaginary parts of the dielectric function are not independent of one another. The dielectric function ϵ_{r} – or more precisely, the quantity $\epsilon_{\mathrm{r}}-1$, which vanishes at infinity – also satisfies the Kramers–Kronig relations:

$$\begin{aligned}\operatorname{Re}\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega)-1&=\frac{1}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty}\mathrm{d}\omega'\,\frac{\operatorname{Im}\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega')}{\omega'-\omega}=\frac{2}{\pi}\,\mathrm{P}\int_{0}^{\infty}\mathrm{d}\omega'\,\frac{\omega'\operatorname{Im}\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega')}{\omega'^2-\omega^2}\,,\\ \operatorname{Im}\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega)&=-\frac{1}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty}\mathrm{d}\omega'\,\frac{\operatorname{Re}\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega')-1}{\omega'-\omega}=-\frac{2\omega}{\pi}\,\mathrm{P}\int_{0}^{\infty}\mathrm{d}\omega'\,\frac{\operatorname{Re}\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega')-1}{\omega'^2-\omega^2}\,.\end{aligned}\tag{16.1.65}$$

The Kramers–Kronig relations for the dielectric function are usually written in this form – however, they are valid only for insulators and semiconductors then. This is because the DC conductivity σ_{DC} of metals remains finite in the $\boldsymbol{q}\to 0$ and $\omega\to 0$ limits, and thus the imaginary part of the dielectric function exhibits singular behavior at $\omega=0$ as shown by (16.1.64); consequently the Kramers–Kronig relations cannot be written in their customary form. When the singularity is separated, the remainder,

$$\epsilon_{\mathrm{r}}-1-\mathrm{i}\frac{\sigma_{\mathrm{DC}}}{\epsilon_0\omega}\,,\tag{16.1.66}$$

is analytic in the upper half-plane, and so the Kramers–Kronig relations can be formulated. After some algebra the following relations emerge:

$$\begin{aligned}\operatorname{Re}\epsilon_{\mathrm{r}}(\omega)-1&=\frac{2}{\pi}\,\mathrm{P}\int_{0}^{\infty}\mathrm{d}\omega'\,\frac{\omega'\operatorname{Im}\epsilon_{\mathrm{r}}(\omega')}{\omega'^2-\omega^2}\,,\\ \operatorname{Im}\epsilon_{\mathrm{r}}(\omega)&=\frac{\sigma_{\mathrm{DC}}}{\epsilon_0\omega}-\frac{2\omega}{\pi}\,\mathrm{P}\int_{0}^{\infty}\mathrm{d}\omega'\,\frac{\operatorname{Re}\epsilon_{\mathrm{r}}(\omega')-1}{\omega'^2-\omega^2}\,.\end{aligned}\tag{16.1.67}$$

After this digression let us turn back to the main line of discussion, and substitute the long-wavelength formula (16.1.41) of the frequency-dependent conductivity into expression (16.1.64) of the dielectric function:

$$\begin{aligned}\epsilon_{\mathrm{r}}(\omega)=1+\frac{\mathrm{i}\sigma_0}{\epsilon_0\omega(1-\mathrm{i}\omega\tau)}&=1+\frac{\mathrm{i}n_{\mathrm{e}}e^2\tau}{\epsilon_0 m_{\mathrm{e}}\omega(1-\mathrm{i}\omega\tau)}\\ &=1-\frac{n_{\mathrm{e}}e^2}{\epsilon_0 m_{\mathrm{e}}\omega(\omega+\mathrm{i}/\tau)}\,.\end{aligned}\tag{16.1.68}$$

Introducing the notation

$$\omega_{\mathrm{p}}^2 = \frac{\sigma_0}{\epsilon_0 \tau} = \frac{n_{\mathrm{e}} e^2}{\epsilon_0 m_{\mathrm{e}}} = \frac{4\pi n_{\mathrm{e}} \tilde{e}^2}{m_{\mathrm{e}}}, \tag{16.1.69}$$

where $\tilde{e}^2 = e^2/4\pi\epsilon_0$ in line with (3.1.6), the dielectric function can be written as

$$\epsilon_{\mathrm{r}}(\omega) = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega(\omega + \mathrm{i}/\tau)}. \tag{16.1.70}$$

For sufficiently high frequencies ($\omega\tau \gg 1$)

$$\epsilon_{\mathrm{r}}(\omega) = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2}. \tag{16.1.71}$$

The dielectric function vanishes at $\omega = \omega_{\mathrm{p}}$, thus, according to our previous considerations, longitudinal vibrations (polarization waves) of this frequency can appear in the electron gas even in the absence of external excitations. Such vibrations are in fact the density oscillations of the electron system, which are similar to the oscillations of charged plasmas, and are called *plasma oscillations*, *Langmuir oscillations*, or *Langmuir waves.*[15] The angular frequency ω_{p} is called the *plasma frequency* or *Langmuir frequency.* In simple metals ω_{p} is on the order of $10^{16}\,\mathrm{s}^{-1}$, thus, in combination with the previously mentioned relaxation times, the condition $\omega_{\mathrm{p}}\tau \gg 1$ is indeed satisfied.

We shall see in the chapter on the optical properties of solids that the plasma frequency plays an important role there, too. Electromagnetic radiation in the optical region – whose frequency is lower than the plasma frequency – cannot penetrate into the solid, so it undergoes total reflection. This causes the characteristic luster of metals. On the other hand, higher-frequency (higher-energy) quanta of radiation penetrate through the metal freely: metals are transparent in the ultraviolet region.

16.1.7 Magnetic Properties

It was shown in the discussion of the magnetic properties of core electrons (see Section 3.2) that the angular frequency of electrons on closed orbits are modified by the application of a magnetic field, giving rise to Langevin (or Larmor) diamagnetism. According to the *Bohr–van Leeuwen theorem*,[16] the magnetic susceptibility of a classical electron gas is zero.[17] This can be understood most simply in the classical picture of a charged particle moving in a circular orbit

[15] I. Langmuir, 1928. Irving Langmuir (1881–1957) was awarded the Nobel Prize in Chemistry in 1932 “for his discoveries and investigations in surface chemistry”.

[16] N. Bohr, 1911, and G. van Leeuwen, 1919. Niels Henrik David Bohr (1885–1962) was awarded the Nobel Prize in 1922 “for his services in the investigation of the structure of atoms and of the radiation emanating from them”.

[17] No reference was made to the Bohr–van Leeuwen theorem in the discussion of atomic diamagnetism in Section 3.2, as electrons were assumed to be bound to the atom – and thus we tacitly went beyond the classical theory.

in a uniform magnetic field: as the particle does not take any energy from the magnetic field, magnetization – the derivative of the energy with respect to magnetic field – must vanish. No contradiction arises from Larmor's theorem, either, when proper account is taken of the orbiting motion of the electrons: even though they move on circular orbits around magnetic field lines, and thus, according to Lenz's law, induce a magnetic moment opposing the magnetic field, the resultant of these moments is precisely compensated for by the contribution of the electrons precessing in the vicinity of the sample edges. As shown in Fig. 16.3, such electrons are repeatedly reflected by the boundary surfaces, and thus traverse their orbit in the opposite sense, so that the angular momenta and magnetic moments associated with such trajectories are opposite in direction to their counterparts due to the electrons orbiting in the interior.

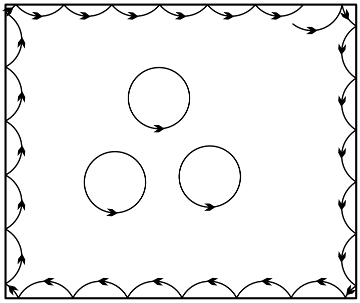


Fig. 16.3. Circular orbits of classical electrons moving in an external magnetic field inside a finite sample and the piecewise circular trajectory of electrons bouncing back from the boundary surfaces

The intrinsic angular momentum (spin) and the related intrinsic magnetic moment of electrons were not known at the time when the Drude–Lorentz model was put forth. Below we shall estimate the magnetic susceptibility of electrons obeying classical statistics but possessing spin.

Consider an electron with an intrinsic magnetic moment of one Bohr magneton μ_B. Upon the application of a magnetic field B, its energy changes by

$$\Delta\varepsilon = -\tfrac{1}{2} g_\mathrm{e} \mu_\mathrm{B} B \sigma \,, \tag{16.1.72}$$

where σ can take the values ±1, corresponding to the two quantized orientations of the spin. Following common practice, we shall speak of the spin direction rather than its opposite, the direction of the intrinsic magnetic moment. The energy of an electron of quantum number $\sigma = 1$ – i.e., whose spin is parallel to the field direction – increases because the g factor of the electron is negative, $g_\mathrm{e} \approx -2$. Similarly, the energy of an electron whose spin is antiparallel to the field direction (spin-down electron) decreases.

It follows from the Maxwell–Boltzmann distribution that the density of electrons decreases exponentially with increasing energy. For the two spin directions the number of electrons per unit volume is

$$n_\uparrow \propto \tfrac{1}{2} n_\mathrm{e} \exp\left(\frac{g_\mathrm{e}\mu_\mathrm{B}B}{2k_\mathrm{B}T}\right), \qquad n_\downarrow \propto \tfrac{1}{2} n_\mathrm{e} \exp\left(-\frac{g_\mathrm{e}\mu_\mathrm{B}B}{2k_\mathrm{B}T}\right). \tag{16.1.73}$$

It should be borne in mind that, according to the conversion formulas listed in Appendix A of Volume 1, only in a strong field of approximately 1 tesla will the magnetic energy of a magnetic moment of one Bohr magneton be of the same order as the thermal energy that corresponds to 1 kelvin. Thus, extremely low temperatures aside, $\mu_\mathrm{B}B \ll k_\mathrm{B}T$ in the customary magnetic fields of susceptibility measurements – that is, the variation of the electron energy is small on the scale of the thermal energy. Consequently,

$$n_\uparrow \approx \tfrac{1}{2} n_\mathrm{e} \left[1 + \frac{g_\mathrm{e}\mu_\mathrm{B}B}{2k_\mathrm{B}T}\right], \qquad n_\downarrow \approx \tfrac{1}{2} n_\mathrm{e} \left[1 - \frac{g_\mathrm{e}\mu_\mathrm{B}B}{2k_\mathrm{B}T}\right]. \tag{16.1.74}$$

The magnetic moment per unit volume – i.e., magnetization – is given by

$$M = \tfrac{1}{2} g_\mathrm{e}\mu_\mathrm{B} \left(n_\uparrow - n_\downarrow\right) = \tfrac{1}{2} g_\mathrm{e}\mu_\mathrm{B} n_\mathrm{e} \frac{g_\mathrm{e}\mu_\mathrm{B}B}{2k_\mathrm{B}T}. \tag{16.1.75}$$

Since magnetization is very small compared to the applied magnetic field, the equality $B = \mu_0(H + M)$ can be safely replaced by $B \approx \mu_0 H$. In this approximation, the magnetic susceptibility is

$$\chi = n_\mathrm{e} \frac{\mu_0 (g_\mathrm{e}\mu_\mathrm{B})^2}{4k_\mathrm{B}T}. \tag{16.1.76}$$

The Curie susceptibility of paramagnets is recognized in this formula. In reality, a very different behavior is observed in metals: measured susceptibilities are essentially temperature-independent and much smaller. As discussed in Chapter 14, Curie-law-like behavior is observed only in ferromagnetic metals or in substances where the atomic core itself is paramagnetic.

16.1.8 Failures of the Drude Model

The failures of the Drude model were already mentioned in connection with specific properties. Below we shall list them together.

The theory has two fundamental parameters: the density n_e of electrons participating in electrical and heat conduction, and the relaxation time τ. Both parameters are phenomenological; they have to be determined by other methods. For elements of group 1 (IA) of the periodic table – alkali metals – and of the next-to-last group of transition metals [group 11 (IB)] – noble metals –, in which the outermost incomplete shell contains a single electron, the assumption of one free electron per atom seems very promising. Good agreement is found between experimental and theoretical values for the plasma frequency, whose formula contains a single free parameter, n_e. For multivalent metals, especially for transition metals and rare-earth metals, it is no longer clear which electrons should be considered free and which ones bound to the atomic core. Nevertheless there is no reason to assume that the number of conduction electrons depends sensitively on temperature.

Unfortunately, nothing can be said about the relaxation time unless the details of the scattering mechanism are known. This is of particular importance when the temperature dependence of resistivity is considered because in the Drude model this can arise only from the temperature dependence of the relaxation time. According to estimates based on resistivity data, the mean free path is on the order of atomic dimensions, which leads naturally to the conclusion that the mean free path is temperature independent. Since the thermal velocity increases as $T^{1/2}$, the relaxation time should vary as $T^{-1/2}$. The problem is then twofold. On the one hand, no simple scattering mechanism is known that would lead to such a relaxation time. On the other hand, this model cannot account for the measured temperature dependence of resistivity, either: instead of the $T^{1/2}$ dependence, the resistivity of metals is observed to be independent of T at low temperatures and to increase linearly with T around room temperature – while a T^5 dependence is found in an intermediate region. No remedy is offered by the Drude–Lorentz model, which is based, instead of the somewhat naive calculations presented above, on a more rigorous but still classical formulation of the kinetic theory of gases due to Boltzmann. The problem is rooted in the assumptions that resistivity is due to the scattering of electrons by atoms arranged in a regular array.

The validity of the Wiedemann–Franz law was considered as a great triumph of the Drude model – but in reality the good agreement came from a calculation that was a factor of two off. Nor is it precisely true that the ratio of the thermal and the electrical conductivities is directly proportional to temperature. In most metals this proportionality is observed only at very low temperatures or above the liquid-nitrogen temperature. In the intermediate region the temperature dependence is far from linear.

It is even more difficult to interpret the behavior in a magnetic field in the classical model. Susceptibility was found to be too high and its temperature dependence too strong. It was also seen in connection with the Hall effect that for certain multivalent metals the theory cannot even predict the correct sign of the Hall coefficient. Finally the Drude model has nothing to say about the reasons why, in addition to metals and insulators, semiconductors and superconductors also exist.

All this indicates the necessity of going well beyond the Drude model if we are to provide a theoretical description for the electronic properties of solids that is in better agreement with experiments.

16.2 Quantum Mechanical Sommerfeld Model

Apart from collisions occurring once every τ on the average, electrons move freely, and their motion is described by classical equations of motion in the classical Drude model. To obtain a better description of the behavior of electrons in solids, the model needs to be improved in two aspects. Firstly, the quantum nature of electrons has to be taken into account, and secondly, the effects of the electron–ion (and electron–electron) interactions have to be treated more precisely.

In the rest of this chapter we shall be concerned with the first aspect. We shall even neglect the potential due to ions and other electrons, and examine how the behavior of free electrons is affected by using of a quantum mechanical approach. Named after its progenitor, this model is called the *Sommerfeld model.*[18]

16.2.1 Quantum Mechanical States of Free Electrons

To determine the states of a system of N_e electrons, the many-particle Schrödinger equation

$$\left(-\frac{\hbar^2}{2m_\mathrm{e}}\sum_{i=1}^{N_\mathrm{e}}\boldsymbol{\nabla}_i^2\right)\Psi(\boldsymbol{r}_1,\boldsymbol{r}_2,\dots,\boldsymbol{r}_{N_\mathrm{e}}) = E\Psi(\boldsymbol{r}_1,\boldsymbol{r}_2,\dots,\boldsymbol{r}_{N_\mathrm{e}}) \tag{16.2.1}$$

needs to be solved. Since electron–electron interactions are neglected in the present treatment, the wavefunction can be constructed from the products of one-particle states obtained from the one-particle Schrödinger equation:

$$-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}_i^2\psi(\boldsymbol{r}_i) \equiv -\frac{\hbar^2}{2m_\mathrm{e}}\left(\frac{\partial^2}{\partial x_i^2}+\frac{\partial^2}{\partial y_i^2}+\frac{\partial^2}{\partial z_i^2}\right)\psi(\boldsymbol{r}_i) = \varepsilon\psi(\boldsymbol{r}_i)\,. \tag{16.2.2}$$

Bearing in mind that the wavefunction Ψ of the many-fermion system has to be antisymmetrized on account of the Pauli exclusion principle – i.e., Ψ has to change sign upon the interchange of the spatial and spin variables of any two particles –, an antisymmetrized linear combination of the product of one-particle wavefunctions is needed. This can be written as a Slater determinant.

As is well known, in a system of infinite volume the wavefunction of a free particle can be written as a plane wave of wave vector $\boldsymbol{k}$:

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{(2\pi)^{3/2}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\,. \tag{16.2.3}$$

[18] A. Sommerfeld, 1928.

In a state of wave number $k = |\boldsymbol{k}|$ the de Broglie wavelength is $\lambda = 2\pi/k$.[19] In this state the energy of the electron is

$$\varepsilon_{\boldsymbol{k}} = \frac{\hbar^2 \boldsymbol{k}^2}{2m_\mathrm{e}} = \frac{\hbar^2}{2m_\mathrm{e}}(k_x^2 + k_y^2 + k_z^2)\,. \tag{16.2.4}$$

For simplicity, when the volume V of the sample is finite, we shall examine an electron system enclosed in a rectangular box of sides L_x, L_y, L_z. Choosing a more general form would not lead to further physical insight. To determine the eigenstates, appropriate boundary conditions have to be imposed. One possibility is to consider the walls of the box as infinitely high potential barriers into which the wavefunction cannot penetrate, i.e., $\psi(\boldsymbol{r})$ has to vanish at the boundaries. Solutions satisfying this boundary condition are stationary waves of the form

$$\psi(x, y, z) = \sin\frac{\pi n_x x}{L_x}\,\sin\frac{\pi n_y y}{L_y}\,\sin\frac{\pi n_z z}{L_z}\,, \tag{16.2.5}$$

where n_x, n_y, and n_z are integers. Changing the sign of n_α $(\alpha = x, y, z)$ leads to the same state, thus only nonnegative integers need to be considered. The energy of such a one-particle state is

$$\varepsilon = \frac{\hbar^2\pi^2}{2m_\mathrm{e}}\left[\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_y}{L_y}\right)^2 + \left(\frac{n_z}{L_z}\right)^2\right]. \tag{16.2.6}$$

According to the Pauli exclusion principle, each state can be occupied by two electrons of opposite spin. Owing to the Fermi–Dirac distribution function, electrons occupy the lowest-lying states in the ground state of the system – that is, each state is filled up to a level ε_max that depends on the number of particles, while all higher-lying states are left empty. The quantum numbers of occupied states fill the interior of one-eighth of an ellipsoid whose semiaxes are

$$a = \frac{L_x}{\pi}\sqrt{\frac{2m_\mathrm{e}\varepsilon_\mathrm{max}}{\hbar^2}}, \quad b = \frac{L_y}{\pi}\sqrt{\frac{2m_\mathrm{e}\varepsilon_\mathrm{max}}{\hbar^2}}, \quad c = \frac{L_z}{\pi}\sqrt{\frac{2m_\mathrm{e}\varepsilon_\mathrm{max}}{\hbar^2}}. \tag{16.2.7}$$

The volume of this one-eighth ellipsoid – and thus the number of states whose energy is less than or equal to ε_max – is

$$\frac{1}{8}\,\frac{4\pi}{3}\,abc = \frac{L_x L_y L_z}{6\pi^2}\left(\frac{2m_\mathrm{e}\varepsilon_\mathrm{max}}{\hbar^2}\right)^{3/2}. \tag{16.2.8}$$

[19] It would be more appropriate to call k the angular or circular wave number, since the wave number is the reciprocal of the wavelength, $1/\lambda$. It is nonetheless customary to call $2\pi/\lambda$ the wave number – just as ω is also commonly called the frequency rather than the angular frequency, although the latter would be more rigorous.

The maximum energy ε_{max} is determined by the requirement that all N_{e} electrons have to be accommodated on these states – taking into account the two values of the spin quantum number. That is,

$$N_{\text{e}} = 2\frac{L_x L_y L_z}{6\pi^2}\left(\frac{2m_{\text{e}}\varepsilon_{\text{max}}}{\hbar^2}\right)^{3/2}, \tag{16.2.9}$$

whence

$$\varepsilon_{\text{max}} = \frac{\hbar^2}{2m_{\text{e}}}(3\pi^2 n_{\text{e}})^{2/3}, \tag{16.2.10}$$

where $n_{\text{e}} = N_{\text{e}}/V$ is the electron number density.

The excited states of this system could also be examined, and the thermal properties could then be determined. However, the highly important transport phenomena in solids cannot be adequately treated in terms of such stationary waves. Therefore, instead of a system confined by infinitely high potential barriers, the periodic (Born–von Kármán) boundary condition introduced in Chapter 6 of Volume 1 is customarily used in solid-state physics. In this case the wavefunction has to take the same value on opposite faces of the box, that is,

$$\begin{aligned}\psi(L_x, y, z) &= \psi(0, y, z)\,,\\ \psi(x, L_y, z) &= \psi(x, 0, z)\,,\\ \psi(x, y, L_z) &= \psi(x, y, 0)\,.\end{aligned} \tag{16.2.11}$$

Under such boundary conditions the solutions of the Schrödinger equation can again be chosen as plane waves, however the usual normalization for a finite box is

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}}\text{e}^{\text{i}\boldsymbol{k}\cdot\boldsymbol{r}}, \tag{16.2.12}$$

and the corresponding eigenvalue is

$$\varepsilon_{\boldsymbol{k}} = \frac{\hbar^2 k^2}{2m_{\text{e}}} = \frac{\hbar^2}{2m_{\text{e}}}(k_x^2 + k_y^2 + k_z^2)\,. \tag{16.2.13}$$

The boundary condition (16.2.11) is satisfied only by a discrete set of wave vectors, thus the energy spectrum is quantized. Only those $\boldsymbol{k}$ vectors are allowed whose components satisfy the equations

$$\text{e}^{\text{i}k_x L_x} = \text{e}^{\text{i}k_y L_y} = \text{e}^{\text{i}k_z L_z} = 1\,, \tag{16.2.14}$$

that is,

$$k_x = \frac{2\pi}{L_x}n_x\,, \qquad k_y = \frac{2\pi}{L_y}n_y\,, \qquad k_z = \frac{2\pi}{L_z}n_z\,, \tag{16.2.15}$$

where n_x, n_y, and n_z are arbitrary integers. The end points of such discrete vectors

$$\boldsymbol{k} = \frac{2\pi n_x}{L_x}\hat{\boldsymbol{x}} + \frac{2\pi n_y}{L_y}\hat{\boldsymbol{y}} + \frac{2\pi n_z}{L_z}\hat{\boldsymbol{z}} \tag{16.2.16}$$

make up a regular lattice of lattice constants $2\pi/L_x$, $2\pi/L_y$, $2\pi/L_z$. This lattice is shown in Fig. 16.4.

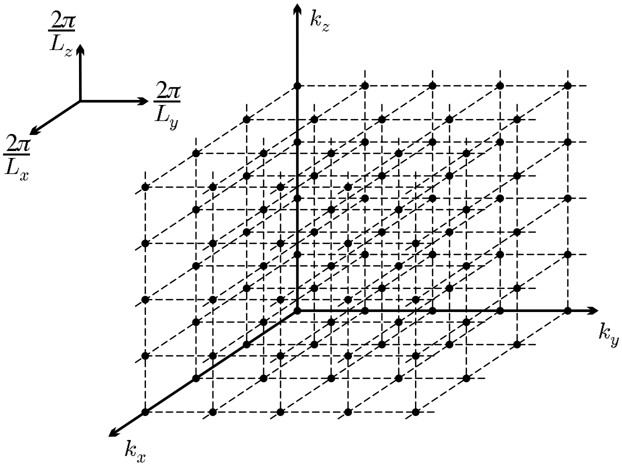


Fig. 16.4. Allowed values of the wave vector

Since each value of $\boldsymbol{k}$ is associated with a volume $(2\pi/L_x)(2\pi/L_y)(2\pi/L_z)$ in $\boldsymbol{k}$-space, the number of allowed states in volume $\mathrm{d}\boldsymbol{k}$ is[20]

$$\frac{L_x L_y L_z}{(2\pi)^3}\mathrm{d}\boldsymbol{k} = \frac{V}{(2\pi)^3}\mathrm{d}\boldsymbol{k}\,. \tag{16.2.17}$$

In addition to the Hamiltonian, plane waves are also eigenstates of the momentum operator

$$\boldsymbol{p} = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial \boldsymbol{r}} = \frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} \tag{16.2.18}$$

with eigenvalues $\boldsymbol{p} = \hbar\boldsymbol{k}$.[21] It should be noted that this relation between the momentum and wave vector of the particle is rigorously valid only for free particles.

The quantum mechanical particle-current density is

$$\boldsymbol{j}_n = \frac{\hbar}{\mathrm{i}}\frac{1}{2m_\mathrm{e}}\left(\psi^*\boldsymbol{\nabla}\psi - \psi\boldsymbol{\nabla}\psi^*\right) = \frac{1}{V}\frac{\hbar\boldsymbol{k}}{m_\mathrm{e}}\,. \tag{16.2.19}$$

[20] Using the variable $\boldsymbol{p} = \hbar\boldsymbol{k}$ instead of $\boldsymbol{k}$, the density of states in phase space is V/h^3.

[21] Theoretical works often use systems of units in which $\hbar = 1$. In this case $\boldsymbol{k}$ itself is often referred to as the momentum.

As $\boldsymbol{j}_n$ is the particle current carried by an electron, it is just the velocity of the electron (divided by a volume factor). Indeed, using

$$\boldsymbol{v} = \dot{\boldsymbol{r}} = \frac{\mathrm{i}}{\hbar}[\mathcal{H}, \boldsymbol{r}] \tag{16.2.20}$$

and the canonical commutation relation $[\boldsymbol{r}, \boldsymbol{p}] = \mathrm{i}\hbar$, the velocity operator of a free electron is

$$\boldsymbol{v} = \frac{\boldsymbol{p}}{m_\mathrm{e}} = \frac{\hbar}{\mathrm{i}m_\mathrm{e}}\boldsymbol{\nabla}\,, \tag{16.2.21}$$

thus the velocity of an electron in the state of wave number $\boldsymbol{k}$ is

$$\boldsymbol{v}_{\boldsymbol{k}} = \frac{\hbar\boldsymbol{k}}{m_\mathrm{e}}\,. \tag{16.2.22}$$

Note that the relationship

$$\boxed{\boldsymbol{v}_{\boldsymbol{k}} = \frac{1}{\hbar}\frac{\partial\varepsilon_{\boldsymbol{k}}}{\partial\boldsymbol{k}}} \tag{16.2.23}$$

between the group velocity and the energy is valid more generally, not only for free electrons.

16.2.2 Ground State of the Electron Gas

As has been mentioned, the states of a system of N_e particles can be constructed from the known one-particle states by taking the antisymmetrized combination of their products (Slater determinant form). The energy of the state is the sum of the energies of the occupied one-particle states. In the ground state electrons occupy the lowest-lying one-particle states. It follows from (16.2.13) that the wave vectors associated with the occupied states are inside a sphere in $\boldsymbol{k}$-space. The radius k_F of this *Fermi sphere* is the *Fermi wave number*, and $\hbar k_\mathrm{F}$ is the *Fermi momentum*, even though k_F is often called the Fermi momentum, too. Its value is determined by the requirement that for N_e electrons the number of allowed $\boldsymbol{k}$ states inside the Fermi sphere should be $N_\mathrm{e}/2$, as two electrons with different spin quantum numbers can have the same quantum number $\boldsymbol{k}$. Taking into account the density of $\boldsymbol{k}$-space points,

$$\frac{N_\mathrm{e}}{2} = \frac{4\pi k_\mathrm{F}^3}{3}\frac{V}{(2\pi)^3} = \frac{k_\mathrm{F}^3}{6\pi^2}V\,, \tag{16.2.24}$$

that is,

$$n_\mathrm{e} = \frac{N_\mathrm{e}}{V} = \frac{k_\mathrm{F}^3}{3\pi^2}\,. \tag{16.2.25}$$

The set of occupied states inside the Fermi sphere is often referred to as the *Fermi sea*. The energy of the highest occupied one-particle level in the ground state is the *Fermi energy* ε_F. This separates the completely filled states from the completely empty ones in the ground state. For free electrons

$$\varepsilon_{\mathrm{F}} = \frac{\hbar^2 k_{\mathrm{F}}^2}{2m_{\mathrm{e}}} \,. \tag{16.2.26}$$

It is the energy of the states on the surface of the Fermi sphere. The velocity associated with such electrons,

$$v_{\mathrm{F}} = \frac{\hbar k_{\mathrm{F}}}{m_{\mathrm{e}}} \,, \tag{16.2.27}$$

is the Fermi velocity. The Fermi energy is related to the electron density by

$$\varepsilon_{\mathrm{F}} = \frac{\hbar^2}{2m_{\mathrm{e}}} (3\pi^2 n_{\mathrm{e}})^{2/3} \,. \tag{16.2.28}$$

Note that this relation is the same as (16.2.10), the energy of the highest occupied state in a system of electrons confined to a box by infinitely high potential barriers. The result is indeed independent of the choice of the boundary condition.

Even though the quantum mechanical description is based on an extended wavefunction, the Wigner–Seitz sphere can be defined for conduction electrons as well: it is a sphere of radius r_0 whose volume is equal to the volume per conduction electron. Thus,

$$\frac{V}{N_{\mathrm{e}}} = \frac{1}{n_{\mathrm{e}}} = \frac{4\pi r_0^3}{3} \,. \tag{16.2.29}$$

Comparison with (16.2.25) gives

$$\frac{k_{\mathrm{F}}^3}{3\pi^2} = \frac{3}{4\pi r_0^3} \,, \tag{16.2.30}$$

thus k_{F} and r_0 are related by

$$k_{\mathrm{F}} = \frac{(9\pi/4)^{1/3}}{r_0} = \frac{1.919}{r_0} \,. \tag{16.2.31}$$

Instead of the parameter r_0 of dimension length, its dimensionless ratio with the Bohr radius $a_0 = 4\pi\epsilon_0\hbar^2/m_{\mathrm{e}}e^2 = \hbar^2/m_{\mathrm{e}}\tilde{e}^2$,

$$r_{\mathrm{s}} = r_0/a_0 \tag{16.2.32}$$

is often used to specify the electron density. Table 16.6 contains the calculated density of electrons, the radius of the spherical volume per electron, as well as the values of the Fermi wave number, Fermi velocity, and Fermi energy determined in the free-electron model for some simple metals.

The de Broglie wavelength of an electron with Fermi momentum is

$$\lambda_{\mathrm{F}} = \frac{2\pi}{k_{\mathrm{F}}} = \frac{2\pi r_0}{(9\pi/4)^{1/3}} = 3.27\, r_0 \,. \tag{16.2.33}$$

Table 16.6. The electron density, the radius r_0 of the spherical volume per electron, the dimensionless ratio $r_\mathrm{s} = r_0/a_0$, and the values of the Fermi wave number, Fermi velocity, and Fermi energy calculated in the free-electron model for some metallic elements

Element	Valence	n_e $(10^{22}\,\mathrm{cm}^{-3})$	r_0 (Å)	r_s	k_F $(10^8\,\mathrm{cm}^{-1})$	v_F $(10^6\,\mathrm{m/s})$	ε_F (eV)
Li	1	4.63	1.73	3.27	1.11	1.29	4.27
Na	1	2.68	2.07	3.92	0.93	1.07	3.24
K	1	1.33	2.62	4.94	0.73	0.86	2.12
Rb	1	1.08	2.81	5.31	0.68	0.81	1.85
Cs	1	0.90	2.99	5.65	0.64	0.75	1.59
Cu	1	8.47	1.41	2.67	1.36	1.57	7.00
Ag	1	5.86	1.60	3.02	1.20	1.39	5.49
Au	1	5.90	1.59	3.01	1.21	1.40	5.53
Ca	2	4.61	1.73	3.27	1.11	1.28	4.69
Zn	2	13.10	1.21	2.30	1.59	1.82	9.47
Cd	2	9.27	1.37	2.59	1.41	1.62	7.47
Hg	2	8.65	1.40	2.65	1.36	1.58	7.13
Al	3	18.06	1.10	2.07	1.75	2.02	11.63
Ga	3	15.30	1.16	2.19	1.66	1.91	10.35
Pb	4	13.20	1.22	2.30	1.58	1.82	9.47

In metals, where the number of conduction electrons per atom is usually of order unity, r_0 is on the order of atomic distances. Using this value on the right-hand side of the above equation, the wavelength of electrons that primarily determine the properties of metals is found to be of the same order. Thus the behavior of electrons in solids cannot be understood adequately without taking their wave nature into account.

In the ground state the total energy of the electron gas is the sum of the energies of independent particles inside the Fermi sphere:

$$E_0 = \sum_{|\boldsymbol{k}|\le k_\mathrm{F}} \sum_\sigma \frac{\hbar^2 k^2}{2m_\mathrm{e}} = 2 \sum_{|\boldsymbol{k}|\le k_\mathrm{F}} \frac{\hbar^2 k^2}{2m_\mathrm{e}} \,. \tag{16.2.34}$$

For a macroscopic sample the allowed values of $\boldsymbol{k}$ fill the space densely, the $\boldsymbol{k}$-space sum can therefore be replaced by an integral. Exploiting the previous result asserting that a volume $(2\pi)^3/V$ is associated with each allowed vector in $\boldsymbol{k}$-space,

$$E_0 = 2\frac{V}{(2\pi)^3} \int\limits_{|\boldsymbol{k}|\le k_\mathrm{F}} \frac{\hbar^2 k^2}{2m_\mathrm{e}}\,\mathrm{d}\boldsymbol{k} = 2\frac{V}{(2\pi)^3} \int\limits_0^{k_\mathrm{F}} \frac{\hbar^2 k^2}{2m_\mathrm{e}} 4\pi k^2\,\mathrm{d}k = \frac{V}{5\pi^2} k_\mathrm{F}^3 \frac{\hbar^2 k_\mathrm{F}^2}{2m_\mathrm{e}} \,. \tag{16.2.35}$$

Making use of the relationship between the particle number and the Fermi wave number,

$$E_0 = \frac{3}{5}\frac{\hbar^2 k_{\mathrm{F}}^2}{2m_{\mathrm{e}}} N_{\mathrm{e}} = \frac{3}{5}\varepsilon_{\mathrm{F}} N_{\mathrm{e}} \,. \qquad (16.2.36)$$

In contrast to a classical gas, the degenerate quantum mechanical electron gas has an appreciable ground-state energy.

16.2.3 Excited Electron and Hole States

To determine the properties of the electron system, over and above the ground state the excited states need to be known. With the particle number fixed, excited states can be generated in the gas of free electrons by raising individual electrons from the ground-state Fermi sea to higher-lying states whose $\boldsymbol{k}$ vector is outside the Fermi sphere. One can also say that in the excitation process holes are created in the Fermi sphere, and some electron states outside the Fermi sphere are filled.

A more formal formulation can be given most easily in the language of second-quantized operators. By introducing the creation and annihilation operators of electron states according to the prescriptions of Appendix H, the Hamiltonian of the noninteracting many-fermion system can be written in the second-quantized form

$$\mathcal{H} = \sum_{\boldsymbol{k},\sigma} \frac{\hbar^2 k^2}{2m_{\mathrm{e}}} c_{\boldsymbol{k}\sigma}^{\dagger} c_{\boldsymbol{k}\sigma} \,. \qquad (16.2.37)$$

Instead of the canonical ensemble, it is more practical to use the grand canonical ensemble, in which the particle number is not kept constant. The Hamiltonian has to be complemented by a term $-\mu N_{\mathrm{e}}$ then, where μ is the chemical potential. Writing the particle number in terms of creation and annihilation operators, too,

$$\mathcal{H} = \sum_{\boldsymbol{k},\sigma} \left(\frac{\hbar^2 k^2}{2m_{\mathrm{e}}} - \mu\right) c_{\boldsymbol{k}\sigma}^{\dagger} c_{\boldsymbol{k}\sigma} \,. \qquad (16.2.38)$$

At zero temperature the chemical potential is the same as the Fermi energy,

$$\mu(T=0) = \varepsilon_{\mathrm{F}} = \frac{\hbar^2 k_{\mathrm{F}}^2}{2m_{\mathrm{e}}} \,, \qquad (16.2.39)$$

as it separates the occupied and unoccupied levels in the ground state.

New operators can be introduced through the canonical transformation

$$d_{\boldsymbol{k}\sigma}^{\dagger} = \begin{cases} c_{\boldsymbol{k}\sigma}^{\dagger} & |\boldsymbol{k}| > k_{\mathrm{F}} \,, \\ c_{\boldsymbol{k}\sigma} & |\boldsymbol{k}| < k_{\mathrm{F}} \,. \end{cases} \qquad (16.2.40)$$

For $|\boldsymbol{k}| > k_{\mathrm{F}}$, the operator $d_{\boldsymbol{k}\sigma}^{\dagger}$ creates an excited electron state, while for $|\boldsymbol{k}| < k_{\mathrm{F}}$ it creates a hole state since by removing an electron from the Fermi sea a hole is generated. Since the Hermitian adjoint of $d_{\boldsymbol{k}\sigma}^{\dagger}$ satisfies

$$d_{\boldsymbol{k}\sigma}|\Psi_0\rangle = 0 \tag{16.2.41}$$

for any $\boldsymbol{k}$, where $|\Psi_0\rangle$ is the ground state of the fermion system, $|\Psi_0\rangle$ is the vacuum of the states created by $d^\dagger_{\boldsymbol{k}\sigma}$. (That is why the notation $|0\rangle$ is also commonly used.) In terms of these operators the Hamiltonian is

$$\begin{aligned}\mathcal{H} &= \sum_{|\boldsymbol{k}|<k_\mathrm{F},\sigma}\left(\frac{\hbar^2k^2}{2m_\mathrm{e}}-\mu\right)d_{\boldsymbol{k}\sigma}d^\dagger_{\boldsymbol{k}\sigma} + \sum_{|\boldsymbol{k}|>k_\mathrm{F},\sigma}\left(\frac{\hbar^2k^2}{2m_\mathrm{e}}-\mu\right)d^\dagger_{\boldsymbol{k}\sigma}d_{\boldsymbol{k}\sigma} \\ &= E_0 - \sum_{|\boldsymbol{k}|<k_\mathrm{F},\sigma}\left(\frac{\hbar^2k^2}{2m_\mathrm{e}}-\mu\right)d^\dagger_{\boldsymbol{k}\sigma}d_{\boldsymbol{k}\sigma} + \sum_{|\boldsymbol{k}|>k_\mathrm{F},\sigma}\left(\frac{\hbar^2k^2}{2m_\mathrm{e}}-\mu\right)d^\dagger_{\boldsymbol{k}\sigma}d_{\boldsymbol{k}\sigma}\,,\end{aligned} \tag{16.2.42}$$

where

$$E_0 = 2\sum_{|\boldsymbol{k}|<k_\mathrm{F}}\left(\frac{\hbar^2k^2}{2m_\mathrm{e}}-\mu\right) \tag{16.2.43}$$

is the energy of the filled Fermi sphere, i.e., the ground-state energy. Introducing the quantity

$$\xi_{\boldsymbol{k}} = \begin{cases}\mu - \dfrac{\hbar^2k^2}{2m_\mathrm{e}} & |\boldsymbol{k}| < k_\mathrm{F}\,, \\ \dfrac{\hbar^2k^2}{2m_\mathrm{e}} - \mu & |\boldsymbol{k}| > k_\mathrm{F}\,,\end{cases} \tag{16.2.44}$$

the Hamiltonian can be written as

$$\mathcal{H} = E_0 + \sum_{\boldsymbol{k},\sigma}\xi_{\boldsymbol{k}}d^\dagger_{\boldsymbol{k}\sigma}d_{\boldsymbol{k}\sigma}\,. \tag{16.2.45}$$

This formula also shows that $\xi_{\boldsymbol{k}}$, which is always positive, is the excitation energy of states created by the operator $d^\dagger_{\boldsymbol{k}\sigma}$. The energy of one-particle excitations is shown as a function of the wave number k in Fig. 16.5. Excitations above the Fermi energy are electron-like, while those below it are hole-like. The excitation energy vanishes at the Fermi momentum and is linear in its vicinity.

16.2.4 Density of States of the Electron Gas

To determine the macroscopically observable properties of the electron gas theoretically, by applying the methods of statistical mechanics, a sum has to be taken over all occupied electron states. In macroscopic samples the sum over $\boldsymbol{k}$-states can be replaced by an integral, just as it was done in the calculation of the ground-state energy. The substitution

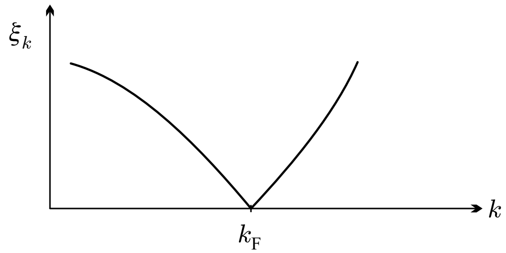


Fig. 16.5. The excitation energy as a function of the wave number for a free-electron gas. Excitations are electron-like for $k > k_{\mathrm{F}}$ and hole-like for $k < k_{\mathrm{F}}$

$$\begin{aligned}\sum_{\boldsymbol{k}} g(\boldsymbol{k}) &\to \int g(\boldsymbol{k}) \left(\frac{L_x}{2\pi}\right) \mathrm{d}k_x \left(\frac{L_y}{2\pi}\right) \mathrm{d}k_y \left(\frac{L_z}{2\pi}\right) \mathrm{d}k_z \\ &= \frac{V}{(2\pi)^3} \int g(\boldsymbol{k})\, \mathrm{d}\boldsymbol{k}\end{aligned} \tag{16.2.46}$$

can be applied to any function $g(\boldsymbol{k})$.

If the integrand depends on $\boldsymbol{k}$ only through the energy $\varepsilon_{\boldsymbol{k}}$, then the integral can be simplified further by the introduction of the electronic density of states, just as it was done for phonons. Allowing for the spin dependence of the energy, we shall use the notation $\rho_\sigma(\varepsilon)\mathrm{d}\varepsilon$ for the number per unit volume of electron states of spin σ in the energy range between ε and $\varepsilon + \mathrm{d}\varepsilon$. Then, by definition,

$$\boxed{\sum_{\boldsymbol{k}} g(\varepsilon_{\boldsymbol{k}\sigma}) = V \int g(\varepsilon)\rho_\sigma(\varepsilon)\, \mathrm{d}\varepsilon\,.} \tag{16.2.47}$$

The total density of states $\rho(\varepsilon)$ is the sum of densities of states for the two spin orientations:

$$\rho(\varepsilon) = \sum_\sigma \rho_\sigma(\varepsilon)\,. \tag{16.2.48}$$

In an ideal electron gas those particles whose energy is less than ε $(\varepsilon + \mathrm{d}\varepsilon)$ fill a sphere of radius k $(k + \mathrm{d}k)$. The volume difference of the two spheres is

$$\frac{4\pi(k + \mathrm{d}k)^3}{3} - \frac{4\pi k^3}{3} \approx 4\pi k^2\, \mathrm{d}k\,. \tag{16.2.49}$$

The number of allowed $\boldsymbol{k}$-points in this region is

$$\frac{V}{(2\pi)^3} 4\pi k^2\, \mathrm{d}k\,. \tag{16.2.50}$$

By taking into account that each point $\boldsymbol{k}$ is associated with two electron states because of the two spin orientations, the number of states in the energy range of width $\mathrm{d}\varepsilon$ per unit volume is

$$\rho(\varepsilon)\,\mathrm{d}\varepsilon = \frac{2}{(2\pi)^3} 4\pi k^2\,\mathrm{d}k\,. \tag{16.2.51}$$

On the other hand, the energy–wave number relationship implies

$$\varepsilon = \frac{\hbar^2 k^2}{2m_\mathrm{e}}\,, \qquad \varepsilon + \mathrm{d}\varepsilon = \frac{\hbar^2(k+\mathrm{d}k)^2}{2m_\mathrm{e}} \approx \frac{\hbar^2 k^2}{2m_\mathrm{e}} + \frac{\hbar^2 k}{m_\mathrm{e}}\,\mathrm{d}k\,, \tag{16.2.52}$$

that is,

$$\mathrm{d}\varepsilon = \frac{\hbar^2 k}{m_\mathrm{e}}\,\mathrm{d}k\,. \tag{16.2.53}$$

This leads to

$$\rho(\varepsilon) = \frac{2}{(2\pi)^3} 4\pi k^2 \left(\frac{\hbar^2 k}{m_\mathrm{e}}\right)^{-1} = \frac{m_\mathrm{e} k}{\hbar^2 \pi^2} = \frac{1}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon}\,. \tag{16.2.54}$$

In what follows, we shall repeatedly refer to the value of the density of states at the Fermi energy. By exploiting the connection between the particle number and the Fermi wave number, (16.2.25), the previous formula implies

$$\boxed{\rho(\varepsilon_\mathrm{F}) = \frac{m_\mathrm{e} k_\mathrm{F}}{\hbar^2 \pi^2} = \frac{1}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon_\mathrm{F}} = \frac{3n_\mathrm{e}}{2\varepsilon_\mathrm{F}}\,.} \tag{16.2.55}$$

Note that the defining equation (16.2.47) of the density of states leads formally to

$$\rho_\sigma(\varepsilon) = \frac{1}{V}\sum_{\boldsymbol{k}} \delta(\varepsilon_{\boldsymbol{k}\sigma} - \varepsilon) = \frac{1}{(2\pi)^3}\int \delta(\varepsilon_{\boldsymbol{k}\sigma} - \varepsilon)\,\mathrm{d}\boldsymbol{k}\,. \tag{16.2.56}$$

For quadratic dispersion relations the evaluation of the integral is straightforward, and the result derived above is recovered. The reason why we chose a seemingly more elaborate method is that the presented argument is more readily generalized to nonisotropic cases and to dispersion relations that are not quadratic.

16.2.5 Ideal Electron Gas at Finite Temperatures

Since electrons are fermions, they obey the Fermi–Dirac statistics, and the occupation probability of electron states in thermal equilibrium at finite temperatures is given by the Fermi–Dirac distribution function

$$f_0(\varepsilon) = \frac{1}{\exp[(\varepsilon - \mu)/k_\mathrm{B}T] + 1}\,, \tag{16.2.57}$$

where μ is the chemical potential at temperature T. The distribution function and its negative derivative with respect to energy are shown in Fig. 16.6. In the $T = 0$ limit the derivative is the Dirac delta function:

$$\lim_{T\to 0}\left(-\frac{\mathrm{d}f_0(\varepsilon)}{\mathrm{d}\varepsilon}\right) = \delta(\varepsilon-\mu)\,, \tag{16.2.58}$$

while at finite temperatures it is sharply peaked at μ with a width of a few times $k_\mathrm{B}T$, and takes practically zero value outside this peak. The Fermi edge is smeared out over this narrow energy range by the thermally created electron–hole pairs. The states are neither fully occupied nor completely empty here. At energies that are farther than a few times $k_\mathrm{B}T$ from the chemical potential, states within the Fermi sphere continue to be completely filled, as if they were frozen in, while states outside the Fermi sphere remain empty.

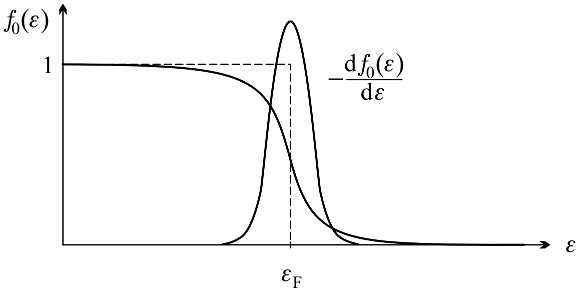


Fig. 16.6. The equilibrium distribution function of particles obeying the Fermi–Dirac statistics and its derivative with respect to energy

As mentioned above, $\mu = \varepsilon_\mathrm{F}$ at $T = 0$. At finite temperatures the value of the chemical potential can be determined from the requirement

$$N_\mathrm{e} = \sum_{\boldsymbol{k},\sigma} f_0(\varepsilon_{\boldsymbol{k}})\,, \tag{16.2.59}$$

as the number of electrons is not changed by thermal excitation. If the sum over $\boldsymbol{k}$ is replaced by an energy integral, division by the volume V leads to the following implicit equation for the chemical potential:

$$n_\mathrm{e} = \int\limits_0^\infty f_0(\varepsilon)\rho(\varepsilon)\,\mathrm{d}\varepsilon\,. \tag{16.2.60}$$

Substituting the density of states from (16.2.54),

$$\begin{aligned} n_\mathrm{e} &= \frac{1}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2}\int\limits_0^\infty \sqrt{\varepsilon}f_0(\varepsilon)\,\mathrm{d}\varepsilon \\ &= 2\left(\frac{m_\mathrm{e}k_\mathrm{B}T}{2\pi\hbar^2}\right)^{3/2} F_{1/2}(\mu/k_\mathrm{B}T)\,, \end{aligned} \tag{16.2.61}$$

where

$$F_{1/2}(x) = \frac{2}{\sqrt{\pi}} \int\limits_0^\infty \frac{y^{1/2}}{\exp(y-x)+1} \, \mathrm{d}y \tag{16.2.62}$$

is the Fermi integral of order $j = 1/2$. In general, it can be evaluated only numerically. However, in two limits, for large negative and positive values of x, closed forms are obtained. These were given in Appendix F of Volume 1. The condition for the applicability of the asymptotic form (C.2.24) for large negative values is that $\exp(\mu/k_\mathrm{B}T) \ll 1$. Then the relation

$$n_\mathrm{e} = 2 \left(\frac{m_\mathrm{e} k_\mathrm{B} T}{2\pi\hbar^2} \right)^{3/2} \mathrm{e}^{\mu/k_\mathrm{B}T} \tag{16.2.63}$$

leads to the expressions that are valid for classical gases. Obviously, this approximation is applicable if

$$\frac{n_\mathrm{e}}{2} \left(\frac{2\pi\hbar^2}{m_\mathrm{e} k_\mathrm{B} T} \right)^{3/2} \ll 1 \,. \tag{16.2.64}$$

This condition is met at room temperature by relatively low ($< 10^{19}/\mathrm{cm}^3$) electron densities. As listed in Table 16.6, in good conductors the density of electrons is three orders of magnitude higher.[22]

As we shall see, in ordinary metals the thermal energy is at least one order of magnitude smaller than μ even at the melting point – that is, we are dealing with the opposite limit, so it is certainly sufficient to keep the first correction in the asymptotic expression (C.2.25) for large positive values of $\mu/k_\mathrm{B}T$. From

$$F_{1/2}(x) \approx \frac{4}{3\sqrt{\pi}} x^{3/2} \left[1 + \frac{\pi^2}{8} \frac{1}{x^2} + \dots \right] \tag{16.2.65}$$

we get

$$n_\mathrm{e} = \frac{1}{3\pi^2} \left(\frac{2m_\mathrm{e}\mu}{\hbar^2} \right)^{3/2} \left[1 + \frac{\pi^2}{8} \left(\frac{k_\mathrm{B}T}{\mu} \right)^2 + \dots \right]. \tag{16.2.66}$$

On the other hand, making use of (16.2.28), the density of electrons can be expressed in terms of the Fermi energy as

$$n_\mathrm{e} = \frac{1}{3\pi^2} \left(\frac{2m_\mathrm{e}\varepsilon_\mathrm{F}}{\hbar^2} \right)^{3/2} , \tag{16.2.67}$$

in agreement with the assertion that ε_F is the zero-temperature limit of the chemical potential. Comparison of the two formulas leads to an implicit equation for the chemical potential:

[22] The situation is different in semiconductors, thus classical statistics may usually be applied to them.

$$\varepsilon_{\mathrm{F}} = \mu \left[1 + \frac{\pi^2}{8} \left(\frac{k_{\mathrm{B}} T}{\mu} \right)^2 + \dots \right]^{2/3} . \tag{16.2.68}$$

The first temperature correction to the ground-state value can be determined from a series expansion, which yields

$$\mu = \varepsilon_{\mathrm{F}} \left[1 - \frac{\pi^2}{12} \left(\frac{k_{\mathrm{B}} T}{\varepsilon_{\mathrm{F}}} \right)^2 + \dots \right] . \tag{16.2.69}$$

It is also customary to define the *Fermi temperature* or *degeneracy temperature* as the temperature for which the thermal energy would be equal to the Fermi energy of the electron system ($k_{\mathrm{B}} T_{\mathrm{F}} = \varepsilon_{\mathrm{F}}$).[23] In terms of the Fermi temperature,

$$\mu = \varepsilon_{\mathrm{F}} \left[1 - \frac{\pi^2}{12} \left(\frac{T}{T_{\mathrm{F}}} \right)^2 + \dots \right] . \tag{16.2.70}$$

The temperature dependence of the chemical potential is usually very weak, nevertheless it cannot be neglected completely: only by taking it into account can the specific heat of the electron system be evaluated correctly.

16.2.6 Sommerfeld Expansion

We shall often encounter integrals of the form

$$I = \int_0^\infty g(\varepsilon) f_0(\varepsilon) \, \mathrm{d}\varepsilon , \tag{16.2.71}$$

which is similar to the Fermi integral but contains some generic function $g(\varepsilon)$, and integrals of the form

$$I = \int_0^\infty G(\varepsilon) \left(- \frac{\mathrm{d} f_0(\varepsilon)}{\mathrm{d}\varepsilon} \right) \mathrm{d}\varepsilon , \tag{16.2.72}$$

which contains the derivative of the Fermi function. Note that the two are identical provided the functions $g(\varepsilon)$ and $G(\varepsilon)$ in the integrands are related by

$$g(\varepsilon) = \frac{\mathrm{d} G(\varepsilon)}{\mathrm{d}\varepsilon} , \qquad G(\varepsilon) = \int_0^\varepsilon g(\varepsilon') \, \mathrm{d}\varepsilon' . \tag{16.2.73}$$

To demonstrate this, integration by parts is performed on I in (16.2.72):

[23] At temperatures above T_{F} the electron gas would behave classically. The conditional is used because even at a Fermi energy of order 1 eV (which is lower than ε_{F} in most metals) this would occur at temperatures above 10^4 K.

$$I = -G(\varepsilon) f_0(\varepsilon) \Big|_0^\infty + \int\limits_0^\infty \frac{\mathrm{d}G(\varepsilon)}{\mathrm{d}\varepsilon} f_0(\varepsilon)\,\mathrm{d}\varepsilon\,. \tag{16.2.74}$$

The second term is indeed the integral given in (16.2.71), and the integrated part is zero as $f_0(\varepsilon)$ vanishes at the upper and $G(\varepsilon)$ at the lower limit by definition.

Expanding the function $g(\varepsilon)$ into a series, the asymptotic form (C.2.25) for Fermi integrals (F_j) can be used in each term. However, a simpler method put forward by Sommerfeld can be applied to metals.

As was shown in Fig. 16.6, the negative derivative of the Fermi function is sharply peaked around the chemical potential, and has a sharp cutoff at lower and higher energies. This implies that only a narrow (a few times $k_\mathrm{B}T$ wide) region around μ contributes to (16.2.72), therefore the integral can be formally extended to the $(-\infty, +\infty)$ range. By expanding $G(\varepsilon)$ in the integrand around $\varepsilon = \mu$ as

$$G(\varepsilon) = G(\mu) + (\varepsilon - \mu)G'(\mu) + \tfrac{1}{2}(\varepsilon - \mu)^2 G''(\mu) + \dots\,, \tag{16.2.75}$$

the integral (16.2.72) is written as the series

$$I = G(\mu) \int\limits_{-\infty}^{\infty} \left(-\frac{\mathrm{d}f_0(\varepsilon)}{\mathrm{d}\varepsilon} \right) \mathrm{d}\varepsilon + G'(\mu) \int\limits_{-\infty}^{\infty} (\varepsilon-\mu) \left(-\frac{\mathrm{d}f_0(\varepsilon)}{\mathrm{d}\varepsilon} \right) \mathrm{d}\varepsilon + \dots\,. \tag{16.2.76}$$

Since

$$\int\limits_{-\infty}^{\infty} \left(-\frac{\mathrm{d}f_0(\varepsilon)}{\mathrm{d}\varepsilon} \right) \mathrm{d}\varepsilon = 1\,, \tag{16.2.77}$$

and $\mathrm{d}f_0(\varepsilon)/\mathrm{d}\varepsilon$ is even in $\varepsilon - \mu$, the odd terms are absent from the series expansion (16.2.76):

$$I = G(\mu) + \sum_{n=1}^{\infty} \frac{1}{(2n)!} \frac{\mathrm{d}^{2n}G(\varepsilon)}{\mathrm{d}\varepsilon^{2n}} \bigg|_{\varepsilon=\mu} \int\limits_{-\infty}^{\infty} (\varepsilon - \mu)^{2n} \left(-\frac{\mathrm{d}f_0(\varepsilon)}{\mathrm{d}\varepsilon} \right) \mathrm{d}\varepsilon\,. \tag{16.2.78}$$

Let us now introduce the notation

$$c_{2n} = \int\limits_{-\infty}^{\infty} \frac{x^{2n}}{(2n)!} \left(-\frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{\mathrm{e}^x + 1} \right) \mathrm{d}x \tag{16.2.79}$$

for the constants arising from the integrals that contain the derivative of the Fermi function. By making use of (C.2.20) and the parity of the integrand,

$$c_{2n} = 2 \sum_{l=1}^{\infty} \frac{(-1)^{l+1}}{l^{2n}} = 2(1 - 2^{1-2n})\zeta(2n)\,, \tag{16.2.80}$$

where $\zeta(x)$ is the Riemann ζ function. The first few coefficients are $c_2 = \pi^2/6$, $c_4 = 7\pi^4/360$, and $c_6 = 31\pi^6/15120$.

This leads to

$$I = G(\mu) + \sum_{n=1}^{\infty} c_{2n} \, (k_{\mathrm{B}}T)^{2n} \frac{\mathrm{d}^{2n} G(\varepsilon)}{\mathrm{d}\varepsilon^{2n}} \bigg|_{\varepsilon=\mu} , \tag{16.2.81}$$

or, in terms of $g(\varepsilon)$,

$$I = \int\limits_{-\infty}^{\mu} g(\varepsilon) \, \mathrm{d}\varepsilon + \sum_{n=1}^{\infty} c_{2n} (k_{\mathrm{B}}T)^{2n} \frac{\mathrm{d}^{2n-1} g(\varepsilon)}{\mathrm{d}\varepsilon^{2n-1}} \bigg|_{\varepsilon=\mu} . \tag{16.2.82}$$

This is the *Sommerfeld expansion.* It is usually sufficient to keep the first temperature correction:

$$\boxed{\int\limits_{0}^{\infty} g(\varepsilon) f_0(\varepsilon) \, \mathrm{d}\varepsilon = \int\limits_{0}^{\mu} g(\varepsilon) \, \mathrm{d}\varepsilon + \frac{\pi^2}{6} (k_{\mathrm{B}}T)^2 g'(\mu) ,} \tag{16.2.83}$$

or, when integrals containing the derivative of the Fermi function are considered,

$$\boxed{\int\limits_{0}^{\infty} G(\varepsilon) \left(-\frac{\mathrm{d} f_0(\varepsilon)}{\mathrm{d}\varepsilon} \right) \mathrm{d}\varepsilon = G(\mu) + \frac{\pi^2}{6} (k_{\mathrm{B}}T)^2 \frac{\mathrm{d}^2 G(\varepsilon)}{\mathrm{d}\varepsilon^2} \bigg|_{\varepsilon=\mu} .} \tag{16.2.84}$$

If the method is applied to the particle number, the density of states appears in place of $g(\varepsilon)$:

$$n_{\mathrm{e}} = \int\limits_{0}^{\mu} \rho(\varepsilon) \, \mathrm{d}\varepsilon + \frac{\pi^2}{6} (k_{\mathrm{B}}T)^2 \rho'(\mu) . \tag{16.2.85}$$

Except for extremely high temperatures, the chemical potential differs little from its zero-temperature value, ε_{F}. By expanding the upper limit of the integral about ε_{F}, and keeping only the leading temperature correction,

$$n_{\mathrm{e}} = \int\limits_{0}^{\varepsilon_{\mathrm{F}}} \rho(\varepsilon) \, \mathrm{d}\varepsilon + (\mu - \varepsilon_{\mathrm{F}}) \rho(\varepsilon_{\mathrm{F}}) + \frac{\pi^2}{6} (k_{\mathrm{B}}T)^2 \rho'(\varepsilon_{\mathrm{F}}) . \tag{16.2.86}$$

As the integral in front is just the particle number,

$$\mu = \varepsilon_{\mathrm{F}} - \frac{\pi^2}{6} (k_{\mathrm{B}}T)^2 \frac{\rho'(\varepsilon_{\mathrm{F}})}{\rho(\varepsilon_{\mathrm{F}})} = \varepsilon_{\mathrm{F}} - \frac{\pi^2}{6} (k_{\mathrm{B}}T)^2 \, \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \ln \rho(\varepsilon) \bigg|_{\varepsilon=\varepsilon_{\mathrm{F}}} . \tag{16.2.87}$$

Indeed, $\lim_{T \to 0} \mu = \varepsilon_{\mathrm{F}}$, and the chemical potential shows weak temperature dependence due to the slow variations of the density of states in the vicinity of the Fermi energy. Using (16.2.54) for the density of states of the ideal electron gas, (16.2.69) is recovered for the temperature dependence of the chemical potential.

16.2.7 Specific Heat of the Electron Gas

When the thermal occupation of the quantum mechanical states of electrons are known, the thermal properties – e.g., the specific heat – and the equation of state of the electron system can be determined. To evaluate the specific heat of the ideal electron gas, consider the total internal energy of the system of electrons at finite temperatures. In terms of the density of states,

$$E = 2\sum_{\boldsymbol{k}} \varepsilon_{\boldsymbol{k}} f_0(\varepsilon_{\boldsymbol{k}}) = 2\frac{V}{(2\pi)^3}\int \varepsilon_{\boldsymbol{k}} f_0(\varepsilon_{\boldsymbol{k}})\mathrm{d}\boldsymbol{k} = V\int \varepsilon f_0(\varepsilon)\rho(\varepsilon)\,\mathrm{d}\varepsilon\,. \quad (16.2.88)$$

The last formula does not contain the customary factor of two arising from the two spin orientations, as that was already included in the density of states.

Employing the Sommerfeld expansion for the energy density E/V,

$$\frac{E}{V} = \int_0^\infty \varepsilon f_0(\varepsilon)\rho(\varepsilon)\,\mathrm{d}\varepsilon = \int_0^\mu \varepsilon\rho(\varepsilon)\,\mathrm{d}\varepsilon + \frac{\pi^2}{6}(k_\mathrm{B}T)^2\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\varepsilon\rho(\varepsilon)\right]_{\varepsilon=\mu}. \quad (16.2.89)$$

Expanding once again μ about ε_F, and using (16.2.87) for the temperature dependence of the chemical potential,

$$\begin{aligned}\frac{E}{V} &= \int_0^{\varepsilon_\mathrm{F}} \varepsilon\rho(\varepsilon)\,\mathrm{d}\varepsilon + (\mu-\varepsilon_\mathrm{F})\varepsilon_\mathrm{F}\rho(\varepsilon_\mathrm{F}) + \frac{\pi^2}{6}(k_\mathrm{B}T)^2\left[\rho(\varepsilon_\mathrm{F}) + \varepsilon_\mathrm{F}\rho'(\varepsilon_\mathrm{F})\right]\\ &= \int_0^{\varepsilon_\mathrm{F}} \varepsilon\rho(\varepsilon)\,\mathrm{d}\varepsilon + \frac{\pi^2}{6}(k_\mathrm{B}T)^2\rho(\varepsilon_\mathrm{F})\,. \end{aligned} \quad (16.2.90)$$

The specific heat is then

$$\boxed{c_\mathrm{el} = \frac{1}{V}\frac{\partial E}{\partial T} = \frac{\pi^2}{3}\rho(\varepsilon_\mathrm{F})k_\mathrm{B}^2 T\,.} \quad (16.2.91)$$

As we shall see, this formula does not apply to free electrons alone but is valid more generally – provided the appropriate expression is used for the density of states.

Now consider the density of states of free electrons at the Fermi energy, $\rho(\varepsilon_\mathrm{F}) = m_\mathrm{e}k_\mathrm{F}/(\pi\hbar)^2$. Expressing k_F in terms of the electron density through (16.2.25),

$$c_\mathrm{el} = \frac{m_\mathrm{e}}{3\hbar^2}\left(3\pi^2 n_\mathrm{e}\right)^{1/3} k_\mathrm{B}^2 T\,. \quad (16.2.92)$$

On the other hand, if the density of states is expressed in terms of n_e and ε_F as $\rho(\varepsilon_\mathrm{F}) = 3n_\mathrm{e}/2\varepsilon_\mathrm{F}$, we have

$$c_\mathrm{el} = \frac{\pi^2}{2}n_\mathrm{e}k_\mathrm{B}\frac{k_\mathrm{B}T}{\varepsilon_\mathrm{F}}. \quad (16.2.93)$$

As mentioned in connection with the Drude model, the classical result would be $c_{\mathrm{el}} = \frac{3}{2} n_{\mathrm{e}} k_{\mathrm{B}}$ because of the three translational degrees of freedom. The quantum mechanical value is smaller by a factor of $(\pi^2/3)(k_{\mathrm{B}}T/\varepsilon_{\mathrm{F}})$. This is because the majority of the electrons are frozen in states well below the Fermi energy: only electrons in a region of a few times $k_{\mathrm{B}}T$ in width around ε_{F} – i.e., about a fraction $k_{\mathrm{B}}T/\varepsilon_{\mathrm{F}}$ of all electrons – can be excited thermally, giving nonvanishing contributions to the specific heat. For most metals ε_{F} is a few eV, thus even at room temperature only 10^{-2} to 10^{-3} of the electrons can be excited, consequently the electronic contribution to the specific heat is by the same factor smaller than the phonon contribution.

Instead of the electronic heat capacity per unit volume, it is more common to specify the molar heat capacity, in the form

$$\boxed{C_{\mathrm{el}} = \gamma T\,,} \tag{16.2.94}$$

where γ is the *Sommerfeld coefficient*. The experimental values of γ are listed in Table 16.7 for some metals; for comparison, theoretical values are also indicated for some simple metals.

Table 16.7. Experimental value of the Sommerfeld coefficient (determined from the temperature dependence of the low-temperature specific heat) for some metals and so-called heavy-fermion materials, in units of $\mathrm{mJ/(mol\,K^2)}$. For simple metals the theoretical value γ_{th} obtained from the free-electron model is also listed

Metal	γ	γ_{th}	Metal	γ	Metal	γ
Li	1.63	0.749	Fe	5.0	$CeAl_3$	1600
Na	1.38	1.094	Co	4.7	$CeCu_6$	1500
K	2.08	1.668	Ni	7.1	$CeCu_2Si_2$	1100
Cu	0.69	0.505	La	10	$CeNi_2Sn_2$	600
Ag	0.64	0.645	Ce	21	UBe_{13}	1100
Au	0.69	0.642	Er	13	U_2Zn_{17}	500
Al	1.35	0.912	Pt	6.8	YbBiPt	8000
Ga	0.60	1.025	Mn	14	$PrInAg_2$	6500

Theoretical values calculated from the free-electron model are reasonably close to experimental results in monovalent metals (alkali metals and noble metals). The experimental value of γ is usually 10 to 30% above the prediction of the free-electron model, even though in one case the measured value is somewhat lower than the calculated one. Taking the number (and hence the density) of conduction electrons as given, the difference between theoretical and experimental values can be attributed, according to (16.2.92), to a change in the mass of electrons contributing to the specific heat in solids: these electrons seem to have an effective mass m^* that is different from the electron mass m_e. In the next chapter we shall see that this increase in the mass is caused – at least in part – by the potential due to ions that is neglected in the free-electron model.

In transition metals and rare-earth metals the density of states derived from specific-heat measurements is much higher than the theoretical prediction of the free-electron model. As we shall see, in addition to s- and p-electrons, substantial contributions are also due electrons that are on the d- and f-levels in the free atomic state, although these cannot by any means be considered free in solids. This is even more so for the materials listed in the last two columns. One of the big surprises of the late 1970s and 1980s was the discovery of families of compounds for which γ is two or three orders of magnitude larger than the usual value. For YbBiPt γ is as high as $8\,\mathrm{J/(mol\,K^2)}$. Converted to the effective mass, the increase compared to the free-electron value is enormous: for CeAl$_3$ and CeCu$_6$ it is 700-fold. Because of their large effective mass, these compounds are called heavy-fermion materials. To understand their behavior, in addition to the potential due to ions, electron–electron interactions need to be taken into account. We shall revisit this problem in Chapter 35 of Volume 3.

16.2.8 Equation of State for the Ideal Electron Gas

In a system of fermions each state is either empty or singly occupied when the spin quantum number σ is also taken into account. The grand canonical partition function is therefore

$$\varXi = \prod_{\boldsymbol{k},\sigma} \left(1 + \mathrm{e}^{-(\varepsilon_{\boldsymbol{k}}-\mu)/k_\mathrm{B}T}\right). \tag{16.2.95}$$

According to the general relations of thermodynamics, the grand canonical potential,

$$\varOmega = -k_\mathrm{B}T \ln \varXi = -k_\mathrm{B}T \sum_{\boldsymbol{k},\sigma} \ln\left(1 + \mathrm{e}^{-(\varepsilon_{\boldsymbol{k}}-\mu)/k_\mathrm{B}T}\right), \tag{16.2.96}$$

is equal to $-pV$, that is,

$$pV = k_\mathrm{B}T \sum_{\boldsymbol{k},\sigma} \ln\left(1 + \mathrm{e}^{-(\varepsilon_{\boldsymbol{k}}-\mu)/k_\mathrm{B}T}\right). \tag{16.2.97}$$

Replacing the $\boldsymbol{k}$-sum by an integral,

$$\begin{aligned} pV &= k_\mathrm{B}T \frac{V}{(2\pi)^3} \sum_\sigma \int \mathrm{d}\boldsymbol{k}\, \ln\left(1 + \mathrm{e}^{-(\varepsilon_{\boldsymbol{k}}-\mu)/k_\mathrm{B}T}\right) \\ &= k_\mathrm{B}T \frac{V}{(2\pi)^3} \sum_\sigma \int_0^\infty \mathrm{d}k\, 4\pi k^2\, \ln\left(1 + \mathrm{e}^{-(\varepsilon_{\boldsymbol{k}}-\mu)/k_\mathrm{B}T}\right). \end{aligned} \tag{16.2.98}$$

Integrating by parts, this can be rewritten as

$$\begin{aligned} pV &= k_{\mathrm{B}}T\frac{V}{(2\pi)^3}\sum_{\sigma}\int_0^{\infty}\mathrm{d}k\,4\pi\frac{k^3}{3}\,\frac{\hbar^2 k}{k_{\mathrm{B}}Tm_{\mathrm{e}}}\,\frac{1}{\mathrm{e}^{(\varepsilon_{\boldsymbol{k}}-\mu)/k_{\mathrm{B}}T}+1} \\ &= \frac{2}{3}\frac{V}{(2\pi)^3}\sum_{\sigma}\int_0^{\infty}\mathrm{d}k\,4\pi k^2\,\frac{\hbar^2 k^2}{2m_{\mathrm{e}}}\,\frac{1}{\mathrm{e}^{(\varepsilon_{\boldsymbol{k}}-\mu)/k_{\mathrm{B}}T}+1}\,. \end{aligned} \tag{16.2.99}$$

Note that apart from the factor 2/3, the right-hand side is just the thermal average of the energy of the electron system, so

$$pV = \frac{2}{3}E\,. \tag{16.2.100}$$

This formula could have also been easily derived from the kinetic theory of gases by determining the pressure from the change in the momentum of particles hitting the walls and bouncing back elastically.

The leading contribution to the internal energy E is the ground-state energy of the electron gas, which has already been calculated. Thus, completely at odds with classical ideal gases, quantum mechanics predicts a finite zero-temperature pressure in a degenerate electron gas:

$$p_0 = \frac{2}{3}\frac{E_0}{V} = \frac{2}{5}n_{\mathrm{e}}\varepsilon_{\mathrm{F}}\,. \tag{16.2.101}$$

Inserting (16.2.28) into (16.2.36), the ground-state energy reads

$$E_0 = \frac{3}{5}N_{\mathrm{e}}\frac{\hbar^2}{2m_{\mathrm{e}}}\left(3\pi^2\frac{N_{\mathrm{e}}}{V}\right)^{2/3}. \tag{16.2.102}$$

Substituting this into the equation of state, the pressure is readily seen to be proportional to the $-5/3$rd power of the volume. The compressibility of the electron gas is therefore

$$\frac{1}{\kappa} = -V\frac{\partial p}{\partial V} = \frac{5}{3}p\,. \tag{16.2.103}$$

Making use of the previous formulas this can be rewritten as

$$\frac{1}{\kappa} = \frac{2}{3}n_{\mathrm{e}}\varepsilon_{\mathrm{F}}\,. \tag{16.2.104}$$

The entropy of the electron gas is readily derived from the grand canonical potential (16.2.96). The thermodynamic relation

$$S = -\frac{\partial \Omega}{\partial T}\,, \tag{16.2.105}$$

implies

$$S = k_{\mathrm{B}} \sum_{\boldsymbol{k},\sigma} \ln\left(1 + \mathrm{e}^{-(\varepsilon_{\boldsymbol{k}}-\mu)/k_{\mathrm{B}}T}\right) + k_{\mathrm{B}} \sum_{\boldsymbol{k},\sigma} \frac{(\varepsilon_{\boldsymbol{k}}-\mu)/k_{\mathrm{B}}T}{\mathrm{e}^{(\varepsilon_{\boldsymbol{k}}-\mu)/k_{\mathrm{B}}T}+1}\,, \tag{16.2.106}$$

which is customarily written in the equivalent form

$$S = -k_{\mathrm{B}} \sum_{\boldsymbol{k},\sigma} \left\{ f_0(\varepsilon_{\boldsymbol{k}}) \ln f_0(\varepsilon_{\boldsymbol{k}}) + \left[1 - f_0(\varepsilon_{\boldsymbol{k}})\right] \ln\left[1 - f_0(\varepsilon_{\boldsymbol{k}})\right] \right\}. \tag{16.2.107}$$

The temperature dependence of the grand canonical potential can be obtained by applying the Sommerfeld expansion to the integral in (16.2.99). This leads to

$$S = V\frac{\pi^2}{3}k_{\mathrm{B}}^2 T \rho(\varepsilon_{\mathrm{F}})\,. \tag{16.2.108}$$

This is in agreement with the result that can be obtained by integrating the specific heat divided by the temperature.

16.2.9 Susceptibility of the Electron Gas

The magnetization of an electron gas in a magnetic field is due to the electrons' spin or orbital motion. Let us consider the spin contribution first. In a zero magnetic field the energy of the electrons is independent of the spin quantum number. States are occupied up to the same energy for both spin orientations. The presence of a magnetic field changes the energy of the electrons, and introduces spin dependence. By quantizing the spin along the magnetic field direction,

$$\varepsilon_{\boldsymbol{k}\sigma} = \varepsilon_{\boldsymbol{k}} - \tfrac{1}{2} g_{\mathrm{e}} \mu_{\mathrm{B}} B \sigma\,, \tag{16.2.109}$$

where σ takes the values ±1. In what follows, these spin states will be referred to as ↑ and ↓ states. When the magnetic field is turned on adiabatically, one would naively think that the energy shift moves the highest occupied level to different heights for the two spin orientations, as shown in Fig. 16.7. However, such a state cannot correspond to a thermal equilibrium. In an isolated electron gas the conservation of magnetization does not allow high-energy spin-↑ electrons to decay into lower-energy spin-↑ states via spin reversal. The thermalization, that is, the equalization of the chemical potential between the spin-↑ and spin-↓ subsystems requires additional interactions in which electrons flip their spin and transfer the angular momentum to another subsystem, for example the lattice. Assuming that such a spin–lattice interaction exists, the redistribution of the filled electron states continues until the ↑ and ↓ states are occupied up to the same field-dependent level $\mu(B)$, which differs from the field-free value of the chemical potential but only slightly.

Since the energy shift of the states depends only on the spin quantum number but not on the wave vector, the density of states for the two spin orientations can be expressed in a simple form:

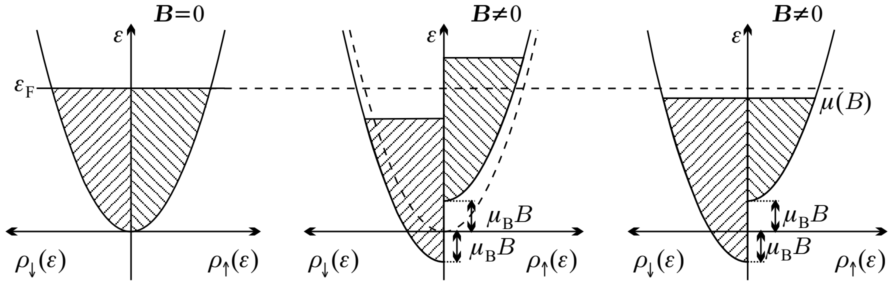


Fig. 16.7. The electronic density of states for the two spin orientations, in the absence a magnetic field (left) and in its presence, before spin-flip and spin-transfer processes occur (middle) and in thermal equilibrium (right)

$$\begin{aligned}\rho_{\uparrow}(\varepsilon) &= \tfrac{1}{2}\rho(\varepsilon + \tfrac{1}{2}g_{\mathrm{e}}\mu_{\mathrm{B}}B) \approx \tfrac{1}{2}\rho(\varepsilon) + \tfrac{1}{4}g_{\mathrm{e}}\mu_{\mathrm{B}}B\frac{\mathrm{d}\rho(\varepsilon)}{\mathrm{d}\varepsilon}\,,\\ \rho_{\downarrow}(\varepsilon) &= \tfrac{1}{2}\rho(\varepsilon - \tfrac{1}{2}g_{\mathrm{e}}\mu_{\mathrm{B}}B) \approx \tfrac{1}{2}\rho(\varepsilon) - \tfrac{1}{4}g_{\mathrm{e}}\mu_{\mathrm{B}}B\frac{\mathrm{d}\rho(\varepsilon)}{\mathrm{d}\varepsilon}\,.\end{aligned} \tag{16.2.110}$$

The number of occupied states for each orientation is

$$\begin{aligned}n_{\uparrow\downarrow} &= \int \rho_{\uparrow\downarrow}(\varepsilon)f_0(\varepsilon)\,\mathrm{d}\varepsilon\\ &= \tfrac{1}{2}\int \rho(\varepsilon)f_0(\varepsilon)\,\mathrm{d}\varepsilon \pm \tfrac{1}{4}g_{\mathrm{e}}\mu_{\mathrm{B}}B\int \frac{\mathrm{d}\rho(\varepsilon)}{\mathrm{d}\varepsilon}f_0(\varepsilon)\,\mathrm{d}\varepsilon\,.\end{aligned} \tag{16.2.111}$$

The chemical potential μ in $f_0(\varepsilon)$ can be determined from the requirement that it should lead to the correct total particle number, i.e., the equality $n_{\uparrow} + n_{\downarrow} = n_{\mathrm{e}}$ should hold. Using the previous series expansion and neglecting the weak quadratic corrections in the temperature, the chemical potential μ is found to be independent of the magnetic field up to linear order.

The magnetization

$$\begin{aligned}M &= \tfrac{1}{2}g_{\mathrm{e}}\mu_{\mathrm{B}}\,(n_{\uparrow} - n_{\downarrow}) = \tfrac{1}{4}g_{\mathrm{e}}^2\mu_{\mathrm{B}}^2 B\int \frac{\mathrm{d}\rho(\varepsilon)}{\mathrm{d}\varepsilon}f_0(\varepsilon)\,\mathrm{d}\varepsilon\\ &= \tfrac{1}{4}g_{\mathrm{e}}^2\mu_{\mathrm{B}}^2 B\int \rho(\varepsilon)\left(-\frac{\partial f_0(\varepsilon)}{\partial\varepsilon}\right)\mathrm{d}\varepsilon = \tfrac{1}{4}g_{\mathrm{e}}^2\mu_{\mathrm{B}}^2 B\rho(\varepsilon_{\mathrm{F}})\end{aligned} \tag{16.2.112}$$

is small even for fields of a few teslas, so B can be replaced by $\mu_0 H$. The susceptibility is then

$$\boxed{\chi_{\mathrm{P}} = \tfrac{1}{4}\mu_0(g_{\mathrm{e}}\mu_{\mathrm{B}})^2\rho(\varepsilon_{\mathrm{F}})\,.} \tag{16.2.113}$$

This is called the *Pauli susceptibility*.[24] In addition to the temperature-independent leading term, the T^2 correction can also be determined by employing the Sommerfeld expansion, but this correction does not have any practical importance in most metals.

[24] W. Pauli, 1926.

Taking the value of the density of states at the Fermi energy from (16.2.55),

$$\chi_{\mathrm{P}} = \frac{3}{8} n_{\mathrm{e}} \frac{\mu_0 (g_{\mathrm{e}} \mu_{\mathrm{B}})^2}{\varepsilon_{\mathrm{F}}} \,. \tag{16.2.114}$$

Expressed in terms of the Fermi temperature, the susceptibility of the electron gas then reads

$$\chi_{\mathrm{P}} = \frac{3}{2} n_{\mathrm{e}} \frac{\mu_0 (g_{\mathrm{e}} \mu_{\mathrm{B}})^2}{4 k_{\mathrm{B}} T_{\mathrm{F}}} \,. \tag{16.2.115}$$

Comparison with the expression (16.1.76) obtained for a classical electron gas shows that the susceptibility χ – which obeys the Curie law at high temperatures – gets saturated around the Fermi temperature T_{F}, where quantum effects become important, and the electron gas becomes degenerate.

Just like the Pauli susceptibility, the specific heat is also proportional to the electronic density of states at the Fermi energy. It is customary to take their ratio, known as the *Wilson ratio*[25] or *Sommerfeld–Wilson ratio*

$$R_{\mathrm{W}} = \frac{4\pi^2 k_{\mathrm{B}}^2 T}{3\mu_0 (g_{\mathrm{e}} \mu_{\mathrm{B}})^2} \frac{\chi_{\mathrm{P}}}{c_{\mathrm{el}}} \,. \tag{16.2.116}$$

It follows from the previous formulas that in a free-electron gas

$$R_{\mathrm{W}} = 1 \,. \tag{16.2.117}$$

It has been mentioned in relation to the specific heat that the measured Sommerfeld coefficient is often substantially different from the theoretical value obtained for an ideal electron gas. This difference can be interpreted in terms of the increase in the effective electron mass due to interactions. A similar tendency is observed for susceptibility. Whether or not the increase in the specific heat and susceptibility can be described by the same multiplicative factor is shown by the Wilson ratio. We shall later see that the effects of the interaction with the atoms of an ordered lattice can often be fairly well characterized by a single effective mass that appears both in the specific heat and susceptibility. However, electron–electron interactions give rise to an additional increase in the susceptibility. Therefore in systems where electron–electron correlations are important, the Wilson ratio differs significantly from the free-electron value. The converse of this statement is also true. A substantial deviation of the experimental value of the Wilson ratio from unity is an indication of strong electron–electron correlations. (In this respect, a factor of 2 should already be considered as substantially different from unity.)

In the foregoing we were concerned only with the magnetic moment arising from spins. However, in the presence of a magnetic field orbital motion can also give rise to magnetic moment. As mentioned earlier, this orbital moment vanishes for classical electrons. But the quantum treatment leads to a different

[25] K. G. Wilson, 1975.

result: as will be demonstrated in Chapter 22, a diamagnetic contribution is found. The diamagnetic susceptibility for free electrons is

$$\chi_{\text{dia}} = -\tfrac{1}{3}\chi_{\text{P}} \, . \tag{16.2.118}$$

Because of the factor $-\frac{1}{3}$, the overall behavior of the quantum mechanically treated ideal electron gas is paramagnetic.

16.3 Electric and Heat Currents in an Electron Gas

The thermodynamic equilibrium state hitherto discussed breaks down in the presence of external disturbances. Such a disturbance can be an electromagnetic field or a nonuniform temperature distribution. When they are present, electrons start to move toward lower-potential or colder places, giving rise to electric or heat currents. If the external driving force is weak, the current will be linear in it. Once set up, such a current would persist indefinitely in an ideal electron gas. In reality, electric and heat currents encounter finite resistance on account of collision processes with ions of the crystal lattice or impurities, as was discussed in connection with the Drude model. Without specifying the collision mechanism we shall, again, assume a finite relaxation time τ for electrons, and study conduction phenomena, using the Fermi–Dirac statistics this time.

16.3.1 Noninteracting Electrons in a Uniform Electric Field

Electrons in an external electromagnetic field obey simple equations of motion. This was exploited in the discussion of the Drude model. In the quantum mechanical description we have to go back to the Schrödinger equation, noting that when the electric and magnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are specified in terms of a scalar potential φ and a vector potential $\boldsymbol{A}$ as

$$\boldsymbol{E} = -\operatorname{grad}\varphi - \frac{\partial \boldsymbol{A}}{\partial t} \, , \qquad \boldsymbol{B} = \operatorname{curl}\boldsymbol{A} \, , \tag{16.3.1}$$

then the canonical momentum $\boldsymbol{p} = -\mathrm{i}\hbar\boldsymbol{\nabla}$ should be replaced by the kinetic momentum $\boldsymbol{p} - q\boldsymbol{A}$ in the kinetic energy, where q is the charge of the particle, and the energy contribution of the scalar potential is $q\varphi$. Since the charge of the electron is $-e$, the gauge-invariant Hamiltonian of free electrons in an electromagnetic field reads

$$\mathcal{H} = \frac{1}{2m_{\text{e}}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e\boldsymbol{A}\right)^2 - e\varphi \tag{16.3.2}$$

if the interactions with the spin of the electrons can be neglected.

One possibility to describe a uniform electric field is to choose a gauge in which the scalar potential vanishes ($\varphi = 0$) and the vector potential is time dependent:

$$\boldsymbol{A} = -\boldsymbol{E}\, t\,. \tag{16.3.3}$$

An obvious advantage of this choice is that it preserves translation invariance, and the wave vector remains a good quantum number. With this choice of gauge the Hamiltonian (16.3.2) takes the form

$$\mathcal{H} = \frac{1}{2m_\mathrm{e}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} - e\,\boldsymbol{E}\, t\right)^2 . \tag{16.3.4}$$

On account of the explicit time dependence, the time-dependent Schrödinger equation

$$\frac{1}{2m_\mathrm{e}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} - e\,\boldsymbol{E}\, t\right)^2 \psi(\boldsymbol{r},t) = -\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}\psi(\boldsymbol{r},t) \tag{16.3.5}$$

has to be solved. It is easily shown that the function

$$\psi(\boldsymbol{r},t) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \exp\left\{ -\frac{\mathrm{i}}{\hbar}\int\limits_0^t \frac{1}{2m_\mathrm{e}}(\hbar\boldsymbol{k} - e\boldsymbol{E}t')^2\,\mathrm{d}t' \right\} \tag{16.3.6}$$

satisfies the above equation. The wave vector $\boldsymbol{k}$ characterizing the spatial variations of the wavefunction is time independent, however the energy of the state changes with time as

$$\varepsilon_{\boldsymbol{k}}(t) = \bigl\langle\psi(\boldsymbol{r},t)\bigl|\mathcal{H}\bigr|\psi(\boldsymbol{r},t)\bigr\rangle = \frac{\hbar^2}{2m_\mathrm{e}}\left(\boldsymbol{k} - \frac{e}{\hbar}\boldsymbol{E}t\right)^2 . \tag{16.3.7}$$

The same time dependence would be obtained if the wave vector of the particle were changing as

$$\boldsymbol{k}(t) = \boldsymbol{k} - \frac{e}{\hbar}\boldsymbol{E}t\,. \tag{16.3.8}$$

We shall soon see that, in addition to the energy, other physical quantities of the electron also behave in such a way as if its wave vector were $\boldsymbol{k}(t)$.

Another possibility is to use a scalar potential, which can then be chosen as $\varphi = -\boldsymbol{E}\cdot\boldsymbol{r}$. In a macroscopic sample, where the discrete (quantized) character of the wave vector can be ignored, solutions of the equation

$$\left[\frac{1}{2m_\mathrm{e}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}\right)^2 + e\boldsymbol{E}\cdot\boldsymbol{r}\right]\psi(\boldsymbol{r},t) = -\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}\psi(\boldsymbol{r},t) \tag{16.3.9}$$

can be given in terms of a vector $\boldsymbol{k}$ that changes continuously with time:

$$\psi(\boldsymbol{r},t) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}(t)\cdot\boldsymbol{r}} \exp\left\{ -\frac{\mathrm{i}}{\hbar}\int\limits_0^t \frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{k}^2(t')\,\mathrm{d}t' \right\}, \tag{16.3.10}$$

where the time dependence of $\boldsymbol{k}(t)$ is similar to (16.3.8):

$$\boldsymbol{k}(t) = \boldsymbol{k}(0) - \frac{e}{\hbar}\boldsymbol{E}t\,. \tag{16.3.11}$$

Translations are now characterized by an explicitly time-dependent $\boldsymbol{k}$ vector. Along with the variations of the wave vector, the energy of the electron also changes. The time dependence of the energy is the same as above.

To obtain the current carried by the particle, the quantum mechanical current (16.2.19) has to be complemented by the contribution of the vector potential:

$$\begin{aligned}\boldsymbol{j}_n &= \frac{1}{2m_\mathrm{e}}\left[\psi^*(\boldsymbol{r})\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e\boldsymbol{A}\right)\psi(\boldsymbol{r}) + \psi(\boldsymbol{r})\left(-\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e\boldsymbol{A}\right)\psi^*(\boldsymbol{r})\right]\\ &= \frac{\hbar}{\mathrm{i}}\frac{1}{2m_\mathrm{e}}\left[\psi^*(\boldsymbol{r})\boldsymbol{\nabla}\psi(\boldsymbol{r}) - \psi(\boldsymbol{r})\boldsymbol{\nabla}\psi^*(\boldsymbol{r})\right] + \frac{e}{m_\mathrm{e}}\boldsymbol{A}|\psi(\boldsymbol{r})|^2\,. \end{aligned} \tag{16.3.12}$$

Whether the wavefunction obtained with one or the other gauge choice is used, the same expression arises for the electric current:

$$\boldsymbol{j}(t) = -e\frac{1}{V}\frac{\hbar\boldsymbol{k}(t)}{m_\mathrm{e}}\,. \tag{16.3.13}$$

Using the expression derived for $\boldsymbol{k}(t)$,

$$\boldsymbol{j}(t) = \boldsymbol{j}_0 + \frac{1}{V}\frac{e^2}{m_\mathrm{e}}\boldsymbol{E}t\,. \tag{16.3.14}$$

The electric-field-dependent part is common to all electrons: the current generated by freely accelerating electrons increases linearly with time.

This would indeed be the case if there were no collisions. However, in real crystals electrons are scattered by impurities and lattice vibrations, and they can also collide with each other. In metals at room temperature the average time between two collisions is $\tau \sim 10^{-14}\,\mathrm{s}$, as was determined from the resistivity in the Drude model. At the same time, assuming an electric field of $E \sim 10^{-2}\,\mathrm{V/cm}$, the wave vector $\boldsymbol{k}$ changes little between two subsequent collisions, and therefore the energy increases only slightly. In the collision process the electron can lose this small excess energy. Thus, even though electrons are not in equilibrium, a stationary distribution may arise, and a stationary current may flow. Its magnitude is basically determined by the collision processes. To calculate the current, the nonequilibrium distribution function has to be specified first.

16.3.2 Stationary Distribution Function

The occupation probability of electron states in thermal equilibrium is known to be given by the Fermi–Dirac distribution function

$$f_0(\boldsymbol{k}) = \frac{1}{\exp\left[\left(\varepsilon_{\boldsymbol{k}} - \mu\right)/k_{\mathrm{B}}T\right] + 1} \,. \tag{16.3.15}$$

The easiest way to provide an approximate treatment for its variations under external perturbations is to assume that electrons move classically under the influence of external disturbances, just like in the Drude model, but are required to obey the quantum mechanical Fermi–Dirac statistics rather than the classical statistics. We shall therefore assume again that the collision processes are such that electrons fly an average time τ between collisions, and in each collision they lose the energy gained from the field. Thermal equilibrium is thus restored in the collision, nevertheless when a snapshot is taken of the electron gas at any particular moment, the occupation probability of states is not given by the equilibrium Fermi–Dirac distribution function but bears the stamp of the changes that have occurred since the last collision.

Suppose now that a uniform electric field $\boldsymbol{E}$ acts on the system of electrons, and that the spatial variations of temperature are such that the temperature gradient $\boldsymbol{\nabla} T(\boldsymbol{r})$ is also uniform in space. In a uniform electric field the momentum of the electrons is changed, by $-e\boldsymbol{E}t$ in time t. Considering collisions occurring at intervals τ on the average, the momentum $\hbar\boldsymbol{k}$ of the electron changes by $-e\boldsymbol{E}\tau$ per collision. The occupation probability of the state of wave number $\boldsymbol{k}$ thus depends on whether the state $\boldsymbol{k} + e\boldsymbol{E}\tau/\hbar$ was occupied in equilibrium. Consequently, in a first approximation the nonequilibrium distribution function may be considered to take the same value at momentum $\hbar\boldsymbol{k}$ as the equilibrium distribution function at that momentum $\hbar\boldsymbol{k}'$ which is transformed into $\hbar\boldsymbol{k}$ by the field in time τ:

$$f(\boldsymbol{k}) = f_0(\boldsymbol{k}') = f_0\Big(\boldsymbol{k} + \frac{e\tau}{\hbar}\boldsymbol{E}\Big) \,. \tag{16.3.16}$$

As shown in Fig. 16.8, the wave vectors of the occupied states fill once again a Fermi sphere, which is nonetheless displaced with respect to the original one. The figure also shows the variation of the distribution function along the electric field direction and its deviation from the equilibrium distribution. Variations are restricted to two regions. The occupation is reduced in the neighborhood of the surface of the Fermi sphere in the direction of the field, while it is increased on the opposite side. The direction of the resultant particle current is opposite to the field direction, while the electric current is along the field direction on account of the negative charge of the electron.

Exploiting the fact that the equilibrium distribution function depends on the momentum only through the energy, in weak fields the expansion to linear order in $\boldsymbol{E}$ leads to

$$\begin{aligned} f(\boldsymbol{k}) &= f_0(\boldsymbol{k}) + \frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \frac{\partial \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}} \frac{e\tau}{\hbar}\boldsymbol{E} \\ &= f_0(\boldsymbol{k}) + \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right) \boldsymbol{v}_{\boldsymbol{k}} \cdot (-e\boldsymbol{E})\,\tau \,, \end{aligned} \tag{16.3.17}$$

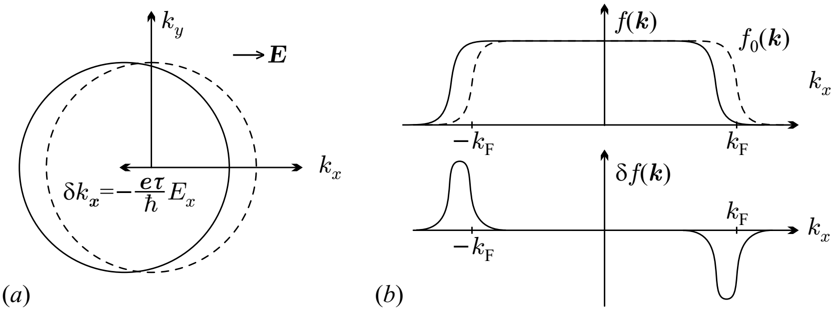


Fig. 16.8. (*a*) The Fermi sphere in the absence (dashed line) and presence (solid line) of an electric field along the x-direction. (*b*) The finite-temperature distribution function along the k_x-axis in the two cases and its variation upon turning on the field

where we made use of the relation (16.2.23) between the group velocity of electrons and the derivative of the energy with respect to $\boldsymbol{k}$.

A similar argument can be applied when a temperature gradient is present. However, owing to the spatial variations of the temperature, the distribution function must then be allowed to depend on the spatial variable $\boldsymbol{r}$. Assuming that electrons become thermalized in their collisions – that is, immediately after a collision their distribution corresponds to thermal equilibrium –, and that they can fly freely for an average time τ between collisions with a velocity $\boldsymbol{v}_{\boldsymbol{k}}$, the distribution function at position $\boldsymbol{r}$ will be the same as the equilibrium distribution function at position $\boldsymbol{r} - \boldsymbol{v}_{\boldsymbol{k}}\tau$:

$$f(\boldsymbol{r},\boldsymbol{k}) = f_0(\boldsymbol{r} - \boldsymbol{v}_{\boldsymbol{k}}\tau, \boldsymbol{k}) \,. \tag{16.3.18}$$

We shall use the form

$$f_0(\boldsymbol{r},\boldsymbol{k}) = \frac{1}{\mathrm{e}^{[\varepsilon_{\boldsymbol{k}} - \mu(\boldsymbol{r})]/k_{\rm B}T(\boldsymbol{r})} + 1} \tag{16.3.19}$$

for the equilibrium distribution function, which takes into account the spatial dependence of the chemical potential arising from the spatial variations of the temperature. If the temperature varies little over a mean free path, which, in turn, is small compared to the dimensions of the sample, then, performing an expansion about the equilibrium distribution function and exploiting the relation

$$\boldsymbol{\nabla}\mu = \frac{\partial\mu}{\partial T}\boldsymbol{\nabla}T \,, \tag{16.3.20}$$

the rules of implicit differentiation lead to

$$\begin{aligned} f(\boldsymbol{r},\boldsymbol{k}) &= f_0(\boldsymbol{k}) + k_{\rm B}T\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\frac{\partial}{\partial\boldsymbol{r}}\frac{\varepsilon_{\boldsymbol{k}} - \mu(\boldsymbol{r})}{k_{\rm B}T(\boldsymbol{r})} \cdot (-\boldsymbol{v}_{\boldsymbol{k}}\tau) \\ &= f_0(\boldsymbol{k}) + \tau\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\left[\frac{\varepsilon_{\boldsymbol{k}} - \mu}{T} + \frac{\partial\mu}{\partial T}\right]\boldsymbol{v}_{\boldsymbol{k}} \cdot (-\boldsymbol{\nabla}T) \,. \end{aligned} \tag{16.3.21}$$

The same result can be obtained using another physical picture if it is noted that the spatial dependence of the distribution function arises from the spatial dependence of the temperature. Electrons of different velocities arrive at point $\boldsymbol{r}$ from different directions – and thus from points of different temperatures. Consider electrons of velocity $\boldsymbol{v}_{\boldsymbol{k}}$ that have flown for an average time τ since their last collision. If $\boldsymbol{v}_{\boldsymbol{k}}$ and $\boldsymbol{\nabla} T$ are parallel, then these electrons come from a point of temperature

$$T - \tau \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{\nabla} T \,. \tag{16.3.22}$$

The same expression gives the effective temperature of electrons when $\boldsymbol{v}_{\boldsymbol{k}}$ and $\boldsymbol{\nabla} T$ are not parallel. Consequently, the distribution function of electrons is such as if their local temperature were not T but $\tau \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{\nabla} T$ less, that is,

$$f(\boldsymbol{k}, T) = f_0(\boldsymbol{k}, T - \tau \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{\nabla} T) \,. \tag{16.3.23}$$

This form of the distribution function is shown in Fig. 16.9. The effective temperatures are different for left- and right-moving electrons. This implies that the sharp boundary of the Fermi sphere is smeared out more on one side than on the other. The figure also shows the distribution function along the axis of decreasing temperature ($\mathrm{d}T/\mathrm{d}x < 0$), as well as the departure from the equilibrium distribution. Since right-moving electrons around $+k_{\mathrm{F}}$ come from warmer places, there are more electrons in the states above the Fermi momentum than there would be in thermal equilibrium at temperature T. By the same token, the occupation of states below the Fermi momentum is lower. The contrary applies to left-moving electrons around $-k_{\mathrm{F}}$: the occupation of states above (below) the Fermi momentum is lower (higher) than in thermal equilibrium.

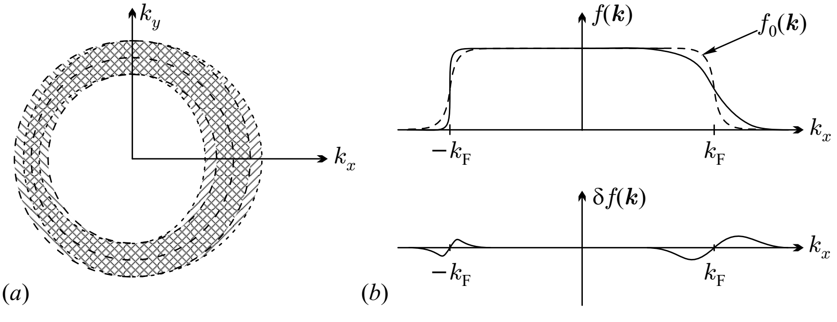


Fig. 16.9. (*a*) Smearing out of the boundary of occupied states at a constant temperature and in the presence of a temperature gradient ($\mathrm{d}T/\mathrm{d}x < 0$). The occupation probability drops from almost one to nearly zero in the hatched region. (*b*) The distribution function along the k_x-axis and its variation upon the application of the temperature gradient

Up to linear order in the temperature gradient,

$$\begin{aligned} f(\boldsymbol{k},T) &= f_0(\boldsymbol{k}) - \tau\frac{\partial f_0}{\partial T}\boldsymbol{v}_{\boldsymbol{k}}\cdot\boldsymbol{\nabla}T \\ &= f_0(\boldsymbol{k}) + \tau\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\left[\frac{\varepsilon_{\boldsymbol{k}}-\mu}{T}+\frac{\partial\mu}{\partial T}\right]\boldsymbol{v}_{\boldsymbol{k}}\cdot(-\boldsymbol{\nabla}T)\,, \end{aligned} \tag{16.3.24}$$

which is in agreement with our earlier result.

If both an electric field and a temperature gradient are present but the variations of the temperature are sufficiently slow then one may make the more general statement that the $\boldsymbol{k}$-space distribution of electrons can be specified by a space-dependent function $f(\boldsymbol{r},\boldsymbol{k})$ whose value is the same as that of the equilibrium distribution function at $\boldsymbol{r}-\boldsymbol{v}_{\boldsymbol{k}}\tau$ for an electron of wave vector $\boldsymbol{k}+e\tau\boldsymbol{E}/\hbar$:

$$f(\boldsymbol{r},\boldsymbol{k}) = f_0\Big(\boldsymbol{r}-\boldsymbol{v}_{\boldsymbol{k}}\tau,\boldsymbol{k}+\frac{e\tau}{\hbar}\boldsymbol{E}\Big)\,, \tag{16.3.25}$$

or

$$f(\boldsymbol{k},T) = f_0\Big(\boldsymbol{k}+\frac{e\tau}{\hbar}\boldsymbol{E},T-\tau\boldsymbol{v}_{\boldsymbol{k}}\cdot\boldsymbol{\nabla}T\Big)\,. \tag{16.3.26}$$

Expanding this expression through linear order in the external perturbations, the following formula is obtained for the nonequilibrium stationary distribution function:

$$f(\boldsymbol{k}) = f_0(\boldsymbol{k}) + \tau\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\boldsymbol{v}_{\boldsymbol{k}}\cdot\left[-e\left(\boldsymbol{E}+\frac{\boldsymbol{\nabla}\mu}{e}\right)+\frac{\varepsilon_{\boldsymbol{k}}-\mu}{T}(-\boldsymbol{\nabla}T)\right]. \tag{16.3.27}$$

Making use of (16.3.20), the previous result is recovered in the $\boldsymbol{E}=0$ case. In what follows, it will be more practical to use this form, since when the electric field is expressed in terms of the scalar potential $\varphi(\boldsymbol{r})$ through $\boldsymbol{E}=-\boldsymbol{\nabla}\varphi$ it is clearer that the electric current is driven by the gradient of the electrochemical potential $\varphi-\mu/e$.

16.3.3 Electric and Heat Currents

An applied electric field induces an electric current, while a temperature gradient causes a heat current to flow in the system. These currents can be simply expressed in terms of the distribution function. The particle current density is

$$\boldsymbol{j}_n = \frac{1}{V}\sum_{\boldsymbol{k},\sigma}\boldsymbol{v}_{\boldsymbol{k}}f(\boldsymbol{k}) = \int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{\boldsymbol{k}}f(\boldsymbol{k})\,. \tag{16.3.28}$$

The electric current density is obtained by multiplying both sides by the electron charge:

$$\boldsymbol{j} = -e\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{\boldsymbol{k}}f(\boldsymbol{k})\,. \tag{16.3.29}$$

The energy current is given by

$$\boldsymbol{j}_E = \int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\varepsilon_{\boldsymbol{k}}\boldsymbol{v}_{\boldsymbol{k}}f(\boldsymbol{k})\,. \tag{16.3.30}$$

However, the heat current cannot be identified directly with the energy current. It follows from the thermodynamic identities $\mathrm{d}Q = T\,\mathrm{d}S$ and $T\,\mathrm{d}S = \mathrm{d}E - \mu\,\mathrm{d}N$ that the heat-current density is

$$\boldsymbol{j}_Q = T\boldsymbol{j}_S = \boldsymbol{j}_E - \mu\boldsymbol{j}_n = \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\,(\varepsilon_{\boldsymbol{k}} - \mu)\,\boldsymbol{v}_{\boldsymbol{k}} f(\boldsymbol{k})\,. \tag{16.3.31}$$

Using (16.3.27) for the distribution function, and making use of the property that no current flows in thermal equilibrium, the electric and heat currents appear as linear responses to external perturbations:

$$\boldsymbol{j} = K_0\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) - K_1\left(-\frac{\boldsymbol{\nabla}T}{T}\right), \tag{16.3.32-a}$$

$$\boldsymbol{j}_Q = -K_1\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) + K_2\left(-\frac{\boldsymbol{\nabla}T}{T}\right), \tag{16.3.32-b}$$

where

$$K_n = e^{2-n}\int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right)\tau(\varepsilon_{\boldsymbol{k}})\frac{1}{3}\boldsymbol{v}_{\boldsymbol{k}}^2\,(\varepsilon_{\boldsymbol{k}} - \mu)^n\,, \tag{16.3.33}$$

and the relaxation time τ may depend on the particle energy. The factor 1/3 appears in front of $\boldsymbol{v}_{\boldsymbol{k}}^2$ because the relationships between currents and external perturbations ($\boldsymbol{E}$, $-\boldsymbol{\nabla}T$) are usually tensorial, and the general expressions for the integrals K_n contain the dyadic product $\boldsymbol{v}_{\boldsymbol{k}} \circ \boldsymbol{v}_{\boldsymbol{k}}$ – however, on account of the isotropic distribution of free electrons in $\boldsymbol{k}$-space, only diagonal terms contribute, and the result is just 1/3 of what would be obtained if the integral were evaluated using $\boldsymbol{v}_{\boldsymbol{k}}^2$. As we shall see in Chapter 24, the appearance of the same integral K_1 in the electric and heat currents is neither accidental nor the result of the approximation: it is the manifestation of the relationships between the transport coefficients in nonequilibrium statistical physics, the *Onsager reciprocal relations.*[26]

It is readily seen that K_0 is just the conductivity:

$$K_0 = \sigma. \tag{16.3.34}$$

Before turning to the study of the roles of other coefficients, let us define the quantity

$$\sigma(\varepsilon) = \frac{2e^2}{3m_{\mathrm{e}}}\rho(\varepsilon)\varepsilon\tau(\varepsilon)\,. \tag{16.3.35}$$

When $\boldsymbol{v}_{\boldsymbol{k}}^2$ is expressed in terms of the electron energy in the defining expression of K_n, and the integration variable $\boldsymbol{k}$ is replaced by the energy, we have

[26] L. ONSAGER, 1931. LARS ONSAGER (1903–1976) was awarded the Nobel Prize in Chemistry in 1968 "for the discovery of the reciprocal relations bearing his name, which are fundamental for the thermodynamics of irreversible processes".

$$K_n = e^{-n} \int \mathrm{d}\varepsilon \left(-\frac{\partial f_0}{\partial \varepsilon} \right) \sigma(\varepsilon)(\varepsilon - \mu)^n \,. \tag{16.3.36}$$

Using the Sommerfeld expansion,

$$K_0 = \sigma(\mu) + \frac{\pi^2}{6}(k_\mathrm{B}T)^2 \left. \frac{\mathrm{d}^2\sigma(\varepsilon)}{\mathrm{d}\varepsilon^2} \right|_{\varepsilon=\mu} , \tag{16.3.37-a}$$

$$K_1 = \frac{\pi^2}{3e}(k_\mathrm{B}T)^2 \left. \frac{\mathrm{d}\sigma(\varepsilon)}{\mathrm{d}\varepsilon} \right|_{\varepsilon=\mu} , \tag{16.3.37-b}$$

$$K_2 = \frac{\pi^2}{3e^2}(k_\mathrm{B}T)^2 \sigma(\mu) \,, \tag{16.3.37-c}$$

where, in line with (16.3.35),

$$\sigma(\mu) = \frac{2e^2}{3m_\mathrm{e}} \rho(\mu)\mu\tau(\mu) \,. \tag{16.3.38}$$

At low temperatures, where corrections of order T^2 can be neglected, the conductivity is given by

$$\sigma_0 = \frac{2e^2}{3m_\mathrm{e}} \rho(\varepsilon_\mathrm{F})\varepsilon_\mathrm{F}\tau = \tfrac{1}{3} e^2 \rho(\varepsilon_\mathrm{F}) v_\mathrm{F}^2 \tau \,, \tag{16.3.39}$$

where τ is the relaxation time of electrons at the Fermi energy. When the density of states is expressed by the electron density using (16.2.55), the well-known Drude formula (16.1.20) is recovered for the conductivity. This form is therefore valid in the Sommerfeld model, too.

Since a temperature gradient induces not only a heat current but, through a cross effect, an electric current as well, pure thermal conduction is obtained when the flow of the electric current is canceled by an applied electric field. This requires

$$K_0 \left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e} \right) = K_1 \left(-\frac{\boldsymbol{\nabla}T}{T} \right) . \tag{16.3.40}$$

Substituting this requirement into the expression for the heat current,

$$\boldsymbol{j}_Q = -\frac{K_1^2}{K_0} \left(-\frac{\boldsymbol{\nabla}T}{T} \right) + K_2 \left(-\frac{\boldsymbol{\nabla}T}{T} \right) . \tag{16.3.41}$$

Since the thermal conductivity λ is defined by the equation

$$\boldsymbol{j}_Q = \lambda \left(-\boldsymbol{\nabla}T \right) , \tag{16.3.42}$$

we have

$$\lambda = \frac{K_2}{T} - \frac{K_1^2}{K_0 T} \,. \tag{16.3.43}$$

The derivative $\mathrm{d}\sigma(\varepsilon)/\,\mathrm{d}\varepsilon$ in K_1 can be determined from (16.3.35). Since close to the Fermi energy the density of states varies little with energy, the derivative is relatively well approximated by $\sigma(\varepsilon_\mathrm{F})/\varepsilon_\mathrm{F}$ in metals, leading to

$$\frac{K_1^2}{K_0K_2} \approx \frac{\pi^2}{3}\left(\frac{k_\mathrm{B}T}{\varepsilon_\mathrm{F}}\right)^2 . \tag{16.3.44}$$

Except for extremely high temperatures, the thermal energy is much smaller than the Fermi energy, thus the second term in (16.3.43) is negligible:

$$\lambda = \frac{K_2}{T}\left\{1 + \mathcal{O}\left(\frac{k_\mathrm{B}T}{\varepsilon_\mathrm{F}}\right)^2\right\} . \tag{16.3.45}$$

By substituting (16.3.37-c) into this expression and using (16.3.39), we finally have

$$\lambda = \frac{\pi^2}{9}k_\mathrm{B}^2 T\rho(\varepsilon_\mathrm{F})v_\mathrm{F}^2\tau . \tag{16.3.46}$$

Comparison with the specific-heat formula (16.2.91) gives

$$\lambda = \tfrac{1}{3}c_\mathrm{el}v_\mathrm{F}^2\tau , \tag{16.3.47}$$

in accordance with the classical expression (16.1.22) of the kinetic theory of gases. However, the temperature dependence does not come from the thermal velocity of electrons now, as in the Drude model, but from the specific heat of electrons.

When the thermal conductivity is expressed in terms of the electrical conductivity, we find

$$\boxed{\lambda = \frac{\pi^2}{3}\left(\frac{k_\mathrm{B}}{e}\right)^2 T\sigma .} \tag{16.3.48}$$

Note that apart from a numerical factor this formula is identical with the Wiedemann–Franz law (16.1.24). The only difference is that the multiplying factor of $(k_\mathrm{B}/e)^2$ in the Lorenz number is now $(\pi^2/3)$ instead of 3/2, i.e.,

$$L = 2.45 \times 10^{-8}\,\mathrm{V}^2\,\mathrm{K}^{-2} . \tag{16.3.49}$$

This value is in good agreement with the experimental data listed in Table 16.4, showing that one of the flaws of the classical model is eliminated by the quantum mechanical treatment.

16.3.4 Thermoelectric Phenomena

It is readily seen from the formulas (16.3.32) of the currents that if there is a temperature gradient between two points of the sample, then an electrostatic potential difference appears between them (and an electric field inside the sample) even when no current is flowing. The *absolute differential thermopower* (also known as thermoelectric power or Seebeck coefficient) S is defined by

$$\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) = S\boldsymbol{\nabla}T , \tag{16.3.50}$$

with the additional requirement that no current should flow. Expressed in terms of the electrochemical potential,

$$S = -\left.\frac{\mathrm{d}(\varphi - \mu/e)}{\mathrm{d}T}\right|_{j=0} . \tag{16.3.51}$$

According to (16.3.32-a),

$$S = -\frac{1}{T}\frac{K_1}{K_0} . \tag{16.3.52}$$

Using the leading terms in (16.3.37) gives the Cutler–Mott formula:[27]

$$S = -\frac{\pi^2 k_{\mathrm{B}}^2 T}{3e} \left.\frac{\mathrm{d}\ln\sigma(\varepsilon)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=\varepsilon_{\mathrm{F}}} . \tag{16.3.53}$$

Substituting the approximation for the derivative of $\sigma(\varepsilon)$ again,

$$S \approx -\frac{\pi^2}{3}\frac{k_{\mathrm{B}}^2 T}{e\varepsilon_{\mathrm{F}}} \tag{16.3.54}$$

is obtained.[28] In the free-electron model the absolute thermoelectric power is negative, and its room-temperature value is a few μV/K. The negative sign is logical. Thermal diffusion gives rise to a flow of electrons from the hot side toward the cold side, resulting in the accumulation of negative charge on the latter, and thus a negative voltage with respect to the hot side, halting the flow of electrons. In reality, a dynamic equilibrium is established. Electrons continue to diffuse toward the cold side (consequently, an electric current flows in the opposite direction). On the other hand, the electric field set up in the sample, which points toward the cold side, drives the electrons toward the hot side. In equilibrium the diffusion and drift currents compensate each other. This is shown in Fig. 16.10(a).

The experimental values of the room-temperature Seebeck coefficient are listed in Table 16.8 for several elements. The data clearly show that the thermoelectric power is positive for many metals, which cannot be interpreted within the Sommerfeld model. Since we shall not deal with the details of thermoelectric phenomena later, it should be noted here that this comes partly from the modifications of the electron states in the periodic potential of the lattice and partly from the energy dependence of the collisions during the diffusion of electrons. These can make the Seebeck coefficient two to three orders of magnitude larger in semiconductors than in metals.

[27] M. Cutler and N. F. Mott, 1969.

[28] Using the relationship $\rho(\varepsilon) \propto \sqrt{\varepsilon}$ for the density of states that is valid for quadratic dispersion relations,

$$S = -\frac{\pi^2}{2}\frac{k_{\mathrm{B}}^2 T}{e\varepsilon_{\mathrm{F}}} .$$

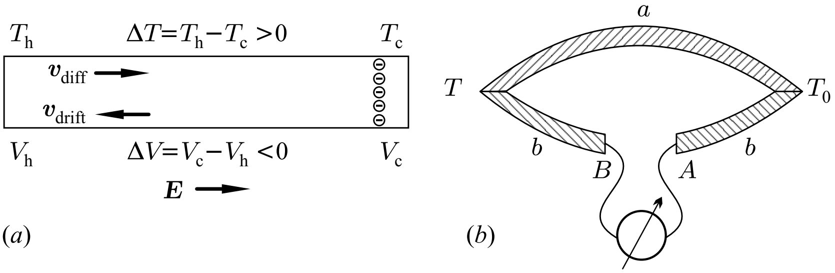


Fig. 16.10. The Seebeck effect. (*a*) The thermoelectric power induced by a temperature gradient. (*b*) The schematic setup of thermoelectric power measurements

Table 16.8. Room-temperature thermoelectric power for several metals

Element	S (μV/K)	Element	S (μV/K)
Li	14	Pd	−9.99
Na	−5	Pt	−5.28
K	−12.5	Cu	1.83
Rb	−8.3	Ag	1.51
V	1.0	Au	1.94
Cr	17.3	Al	−1.8
W	1.07	Pb	−1.05

It is worth noting that phonons also give an indirect contribution to the thermoelectric power: they flow from the hot side to the cold one, and can transfer momentum to the electrons through their interactions. This phenomenon, called *phonon drag*, also leads to the accumulation of electrons on the cold side, and thus gives an additional term in the termoelectric power.

In general, the absolute thermoelectric power itself cannot be measured directly, only the difference of the thermoelectric powers of two metals in contact.[29] The measurement can be performed in the setup shown schematically in Fig. 16.10(*b*), with one point of contact at the reference temperature T_0, while the other at some other temperature T. Because of the unequal absolute differential thermoelectric powers S_a and S_b of the two materials, a potential difference is observed between the two end points of an open system, independently of the temperature T' of the points A and B between which this voltage is measured. If a closed circuit were built of the two materials, a current would be generated. This phenomenon is called the *Seebeck effect*.[30]

The magnitude of the potential difference – also called the thermoelectromotive force – is

[29] Except for the case when a normal metal is in contact with a superconductor, as the thermoelectric power vanishes in superconductors.

[30] T. J. Seebeck, 1821.

$$\begin{aligned} E_{\mathrm{ab}}(T,T_0) &= V(A) - V(B) = -\int\limits_{\mathrm{B}}^{\mathrm{A}} \boldsymbol{E}\cdot \mathrm{d}\boldsymbol{s} \\ &= -\int\limits_{T'}^{T} S_{\mathrm{b}}\mathrm{d}T - \int\limits_{T}^{T_0} S_{\mathrm{a}}\mathrm{d}T - \int\limits_{T_0}^{T'} S_{\mathrm{b}}\mathrm{d}T = \int\limits_{T_0}^{T} (S_{\mathrm{a}} - S_{\mathrm{b}})\mathrm{d}T\,. \end{aligned} \tag{16.3.55}$$

In differential form:

$$S_{\mathrm{ab}} \equiv S_{\mathrm{a}} - S_{\mathrm{b}} = \left.\frac{\partial E_{\mathrm{ab}}(T,T_0)}{\partial T}\right|_{T=T_0}\,. \tag{16.3.56}$$

The coefficient S_{ab} is the relative Seebeck coefficient of material a with respect to material b.

Another phenomenon is observed when electric current is flowing in a system made up of two different metals kept at the same temperature. It follows from (16.3.32) that for $\boldsymbol{\nabla} T = 0$

$$\boldsymbol{j}_Q = -\frac{K_1}{K_0}\boldsymbol{j}\,, \tag{16.3.57}$$

that is, in addition to the electric current, a heat current is also present in each part. The electric current is the same on the two sides of the contact, however K_1 and K_0 depend on material properties, so the magnitude of the heat current is different in the two metals. This is only possible if the system emits or absorbs heat at the contact. This is the *Peltier effect*,[31] while the coefficient

$$\Pi = \left.\frac{\boldsymbol{j}_Q}{\boldsymbol{j}}\right|_{\boldsymbol{\nabla} T=0} = -\frac{K_1}{K_0} \tag{16.3.58}$$

is called the *Peltier coefficient*. Comparison with (16.3.52), the formula for the absolute thermoelectric power S immediately leads to

$$\Pi = ST\,, \tag{16.3.59}$$

which is just the *first Thomson relation* or *Kelvin relation*.[32] As shown in Fig. 16.11(a), the heat emitted or absorbed irreversibly at the contact between two different metals per unit time is

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = (\Pi_{\mathrm{a}} - \Pi_{\mathrm{b}})\,\boldsymbol{j} = (S_{\mathrm{a}} - S_{\mathrm{b}})\,T\boldsymbol{j}\,. \tag{16.3.60}$$

In the setup shown in Fig. 16.11(b) both contacts between the two metals are at the same temperature. When an electric current is circulating, the temperatures of the contacts can be kept equal only if heat is absorbed at one

[31] J. C. A. Peltier, 1834.

[32] W. Thomson (Lord Kelvin), 1854.

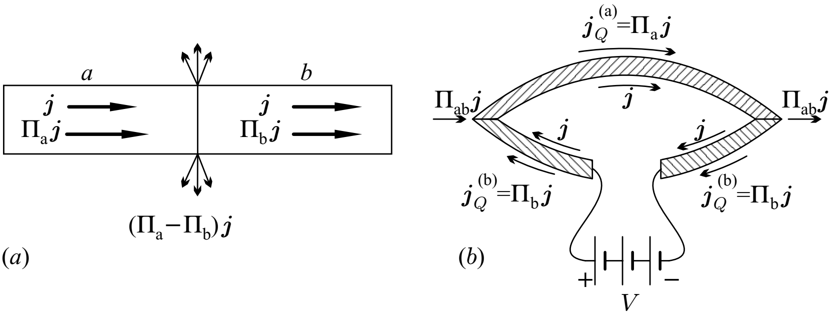


Fig. 16.11. The Peltier effect. (*a*) Emission or absorption of heat at the contact of two current-carrying metals. (*b*) Electric and heat currents in a closed circuit

contact and the same amount is released at the other – that is, heat is pumped from one side to the other. If the same temperature is not maintained by the absorption and release of heat, the voltage induces a temperature difference: one contact cools down while the other warms up. In this sense the Peltier effect can be considered as the inverse of the Seebeck effect.

In a current-carrying conductor heat is reversibly released not only when two different metals are in contact but also when the current flows through a sample made of a single metal but in addition to the potential difference a temperature gradient is also present. This phenomenon is called the *Thomson effect*, and the heat that is released over and above the irreversibly released Joule heat is called the *Thomson heat*. Its magnitude can be evaluated using the continuity equation

$$\frac{\partial n_\mathrm{e}}{\partial t} + \operatorname{div} \boldsymbol{j}_n = 0 \tag{16.3.61}$$

and a similar equation for the energy density $w = E/V$:

$$\frac{\partial w}{\partial t} + \operatorname{div} \boldsymbol{j}_E = \boldsymbol{E} \cdot \boldsymbol{j}\,, \tag{16.3.62}$$

which are the consequences of particle-number and energy conservation, respectively. The term on the right-hand side of the last formula is the change in the energy due to the work of the external field. The heat generated in unit volume ($q = Q/V$) per unit time is then

$$\frac{\partial q}{\partial t} = \boldsymbol{j} \cdot \boldsymbol{E} - \operatorname{div} \boldsymbol{j}_Q\,. \tag{16.3.63}$$

Before transforming this equation, we shall write equations (16.3.32) in an alternative form, using the coefficients introduced above:

$$\boldsymbol{E} = \frac{1}{\sigma}\boldsymbol{j} + S\boldsymbol{\nabla}T\,, \qquad \boldsymbol{j}_Q = \mathit{\Pi}\boldsymbol{j} - \lambda\boldsymbol{\nabla}T\,. \tag{16.3.64}$$

Substituting these formulas into (16.3.63), assuming spatially uniform currents and temperature gradient, and neglecting higher powers of the temperature gradient gives

$$\begin{aligned}\frac{\partial q}{\partial t} &= \frac{1}{\sigma}\boldsymbol{j}^2 + S\boldsymbol{j}\cdot\boldsymbol{\nabla}T - \frac{\partial \Pi}{\partial T}\boldsymbol{j}\cdot\boldsymbol{\nabla}T \\ &= \frac{1}{\sigma}\boldsymbol{j}^2 - T\frac{\partial S}{\partial T}\boldsymbol{j}\cdot\boldsymbol{\nabla}T\,,\end{aligned} \tag{16.3.65}$$

where we have exploited (16.3.59) relating the thermoelectric power to the Peltier coefficient. In the final formula the first term is the Joule heat, while the second is the Thomson heat. By defining the Thomson coefficient μ_{T} from the reversibly released heat as

$$\left(\frac{\partial q}{\partial t}\right)_{\mathrm{rev}} = -\mu_{\mathrm{T}}\boldsymbol{j}\cdot\boldsymbol{\nabla}T\,, \tag{16.3.66}$$

the *second Thomson relation* is obtained:

$$\mu_{\mathrm{T}} = T\frac{\partial S}{\partial T} = T\frac{\partial}{\partial T}\left(\frac{\Pi}{T}\right). \tag{16.3.67}$$

This relation allows us – at least, in principle – to measure indirectly the absolute differential termoelectric power, as

$$S = \int\limits_0^T \frac{\mu_{\mathrm{T}}}{T'}\,\mathrm{d}T'\,. \tag{16.3.68}$$

16.3.5 Galvanomagnetic and Thermomagnetic Phenomena

It was shown in connection with the Hall effect that if a current is flowing in the sample and a perpendicular magnetic field is applied then the Lorentz force on the electrons gives rise to a potential difference that is perpendicular to both the electric and magnetic fields. This phenomenon is easily understood in the classical picture. The quantum mechanical treatment would require the solution of the Schrödinger equation in the presence of a magnetic field. We shall return to this point in Chapters 21 and 22. Below we shall content ourselves with presenting some other effects that arise in magnetic fields.

As has been mentioned, the application of a magnetic field could, in principle, give rise to magnetoresistance – that is, it could lead to a change of the resistance along the electric field direction. However, there is neither transverse nor longitudinal magnetoresistance (in which the electric and magnetic fields are perpendicular and parallel, respectively) in the classical model of free electrons. The situation is similar in the Sommerfeld model as long as only the temperature-independent leading terms are used in the evaluation of the current. However, a more accurate calculation yields

$$\frac{\Delta\varrho}{\varrho_0} = \frac{aB^2}{1+cB^2}\,, \tag{16.3.69}$$

where the coefficients a and c are expressed in terms of the mean free path l as

$$a = \frac{\pi^2}{3}\left(\frac{elk_\mathrm{B}T}{m_\mathrm{e}^2 v_\mathrm{F}^3}\right)^2\,, \qquad c = \left(\frac{el}{m_\mathrm{e}v_\mathrm{F}}\right)^2. \tag{16.3.70}$$

After an initial parabolic increase, the variation of the resistance becomes saturated for strong fields.

The quantum mechanical Sommerfeld model leads to the same Hall coefficient as the classical Drude model. In Hall effect measurements transverse voltages appear because no current is allowed to flow in the direction that is perpendicular both to the magnetic field and the initial current direction. When no transverse voltage is allowed, a secondary electric current starts to flow perpendicular to the primary one. This is the *Corbino effect*.[33]

Since the electric current may be accompanied by a heat current, the magnetic field can also affect the latter. Thus, when electric and magnetic fields and a temperature gradient are all present, the generalization of (16.3.32) gives

$$\begin{aligned}\boldsymbol{j} &= N_{11}\boldsymbol{E} + N_{12}\boldsymbol{\nabla}T + N_{13}\boldsymbol{B}\times\boldsymbol{E} + N_{14}\boldsymbol{B}\times\boldsymbol{\nabla}T\,,\\ \boldsymbol{j}_Q &= N_{21}\boldsymbol{E} + N_{22}\boldsymbol{\nabla}T + N_{23}\boldsymbol{B}\times\boldsymbol{E} + N_{24}\boldsymbol{B}\times\boldsymbol{\nabla}T\end{aligned} \tag{16.3.71}$$

for the electric and heat currents. As the magnetic field itself does not induce a current, only the above terms appear when only terms linear in the magnetic field are allowed. Since in experiments it is easier to control the electric current than the electric field, it is more practical to rewrite the equations in the form

$$\begin{aligned}\boldsymbol{E} &= \frac{1}{\sigma}\boldsymbol{j} + S\boldsymbol{\nabla}T + R_\mathrm{H}\boldsymbol{B}\times\boldsymbol{j} + L\boldsymbol{B}\times\boldsymbol{\nabla}T\,,\\ \boldsymbol{j}_Q &= \Pi\boldsymbol{j} - \lambda\boldsymbol{\nabla}T + M\boldsymbol{B}\times\boldsymbol{j} + N\boldsymbol{B}\times\boldsymbol{\nabla}T\,.\end{aligned} \tag{16.3.72}$$

As indicated in (16.3.64), when no magnetic field is present, the four coefficients correspond to four known quantities: the inverse conductivity, the differential thermoelectric power, the Peltier coefficient, and the thermal conductivity. When a magnetic field is applied, four new coefficients appear. Besides, the former four can also change – moreover, it should also be remembered that the coefficients are no longer necessarily scalars. One of the new coefficients is the previously mentioned Hall coefficient. The others are also related to well-known classical physical phenomena.

Phenomena whose occurrence is due primarily to electric currents are called galvanomagnetic. In connection with the Hall effect, an important feature has not been mentioned yet: while the Hall voltage serves to eliminate

[33] O. M. Corbino, 1911.

transverse electric currents, the sample cannot remain isothermal unless a heat current is flowing in this transverse direction. If there are no heat currents, a transverse temperature gradient appears. This is the *Ettingshausen effect.*[34] The Ettingshausen coefficient is defined by

$$-\boldsymbol{\nabla}T = A_{\mathrm{E}}\boldsymbol{B}\times\boldsymbol{j}\,. \tag{16.3.73}$$

If the electric current is along the x-direction, and the applied magnetic field is along the z-direction, then $\partial T/\partial y$ is finite, and

$$A_{\mathrm{E}} = -\frac{\partial T/\partial y}{B_z j_x}\,. \tag{16.3.74}$$

Making use of general thermodynamic relations, the Ettingshausen coefficient can be related to the Hall coefficient:

$$A_{\mathrm{E}} = \frac{T\mu_{\mathrm{T}}}{\lambda\varrho}R_{\mathrm{H}}\,. \tag{16.3.75}$$

Staying within the framework of the Sommerfeld model, and assuming energy-dependent relaxation times,

$$A_{\mathrm{E}} = -\frac{T}{n_{\mathrm{e}}}\left(\frac{\partial\ln\tau(\varepsilon)}{\partial\varepsilon}\right)_{\varepsilon=\varepsilon_{\mathrm{F}}}\,. \tag{16.3.76}$$

If the primary current is not electric but a heat current induced by the temperature gradient across the sample then we speak of thermomagnetic phenomena. Owing to the magnetic field, a transverse electric field or a transverse secondary temperature gradient appears in this case, too. The first possibility is called the *Nernst effect* or *transverse Nernst–Ettingshausen effect,*[35] while the second is known as the *Righi–Leduc effect.*[36] The corresponding coefficients are defined as

$$\boldsymbol{E} = -A_{\mathrm{N}}\boldsymbol{B}\times\boldsymbol{\nabla}T \qquad \text{and} \qquad \boldsymbol{\nabla}_{\perp}T = A_{\mathrm{RL}}\boldsymbol{B}\times\boldsymbol{\nabla}_{\parallel}T\,. \tag{16.3.77}$$

Assuming once again that the primary current is in the x-direction while the magnetic field is along the z-direction,

$$A_{\mathrm{N}} = -\frac{E_y}{B_z\partial T/\partial x}, \qquad A_{\mathrm{RL}} = \frac{\partial T/\partial y}{B_z\partial T/\partial x}. \tag{16.3.78}$$

[34] A. v. Ettingshausen, 1887.

[35] W. Nernst and A. v. Ettingshausen, 1887. The longitudinal Nernst-Ettingshausen effect is the variation of the thermopower in a transverse magnetic field. Walther Hermann Nernst (1864–1941) was awarded the Nobel Prize in Chemistry in 1920 “in recognition of his work in thermochemistry”.

[36] A. Righi and A. Leduc, 1887.

It should be noted that these coefficients are not independent: they can be expressed in terms of the formerly introduced ones via the *Bridgman relations*:[37]

$$A_{\mathrm{N}} = \frac{\mu_{\mathrm{T}}}{\rho} R_{\mathrm{H}} = \frac{\lambda}{T} A_{\mathrm{E}}\,, \qquad A_{\mathrm{RL}} = R_{\mathrm{H}}\, \sigma\,. \tag{16.3.79}$$

16.4 Scattering of Free Electrons by Impurities

Up to now the system of electrons has been considered as a gas of almost free particles. The word "almost" refers to the fact that the interpretation of conduction phenomena had to be based on the assumption that electrons participate in collision processes from time to time. This was taken into consideration by a phenomenological parameter, the relaxation time. These collision processes can be electron–electron interaction processes as well as scattering on the ions of the lattice. We shall discuss these in more detail later. Below, we shall consider a different kind of scattering events, which is related to a recurrent question in solid-state physics: what happens to a free electron described by a plane-wave wavefunction when it is scattered by an impurity? This question can be addressed in two different ways. The first option is to treat this process as a scattering problem, and solve the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 + V(\boldsymbol{r})\right]\psi(\boldsymbol{r}) = \varepsilon\psi(\boldsymbol{r}) \tag{16.4.1}$$

asymptotically (i.e., far from the impurity) for an electron beam scattered by the potential $V(\boldsymbol{r})$ of the impurity placed at the origin. The other option is to compute the one-particle stationary electron states in the presence of the impurity potential.

16.4.1 Formal Solution of the Schrödinger Equation

Before turning to the study of the scattering problem, we shall briefly outline how to determine the electron states formed around the impurity by means of the formal solution of the Schrödinger equation using Green functions.

Impurities break translational symmetry, and so the wave vector $\boldsymbol{k}$ will no longer be a good quantum number. By writing the energy eigenvalue formally as $\varepsilon = \hbar^2 k^2/2m_{\mathrm{e}}$,

$$\frac{\hbar^2}{2m_{\mathrm{e}}}\left[\boldsymbol{\nabla}^2 + k^2\right]\psi_{\boldsymbol{k}}(\boldsymbol{r}) = V(\boldsymbol{r})\psi_{\boldsymbol{k}}(\boldsymbol{r}) \tag{16.4.2}$$

[37] P. W. Bridgman, 1924. Percy Williams Bridgman (1882–1961) was awarded the Nobel Prize in 1946 "for the invention of an apparatus to produce extremely high pressures, and for the discoveries he made therewith in the field of high pressure".

is obtained after some algebra. The implicit form of the wavefunction can be easily established by the introduction of the free-electron Green function that satisfies the equation

$$\frac{\hbar^2}{2m_\mathrm{e}}\bigl[\boldsymbol{\nabla}^2 + k^2 \mp \mathrm{i}\alpha\bigr]G(\boldsymbol{r},\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}')\,, \tag{16.4.3}$$

where α is an infinitesimal positive number that ensures the appropriate analytic properties of the Green function. Using this defining equation it is readily seen that the function

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} + \int \mathrm{d}\boldsymbol{r}'\, G(\boldsymbol{r},\boldsymbol{r}')V(\boldsymbol{r}')\psi_{\boldsymbol{k}}(\boldsymbol{r}') \tag{16.4.4}$$

is indeed a solution of (16.4.2).

The equation for the Green function can be solved explicitly. Using the formula $\boldsymbol{\nabla}^2(1/r) = -4\pi\delta(\boldsymbol{r})$,

$$G(\boldsymbol{r},\boldsymbol{r}') = -\frac{m_\mathrm{e}}{2\pi\hbar^2}\frac{\mathrm{e}^{\mathrm{i}k|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,, \tag{16.4.5}$$

that is, the wavefunction satisfies the equation

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} - \frac{m_\mathrm{e}}{2\pi\hbar^2}\int \mathrm{d}\boldsymbol{r}'\,\frac{\mathrm{e}^{\mathrm{i}k|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}V(\boldsymbol{r}')\psi_{\boldsymbol{k}}(\boldsymbol{r}')\,. \tag{16.4.6}$$

When the potential is weak, the wavefunction can be calculated from this equation iteratively.

16.4.2 Approach Based on Scattering Theory

When free electrons described by plane waves are scattered elastically by the spherically symmetric potential of an impurity, we shall seek solutions that can be written as the superposition of an incoming plane wave and an outgoing (scattered) spherical wave:

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) \propto \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} + f(\theta)\frac{\mathrm{e}^{\mathrm{i}kr}}{r}\,, \tag{16.4.7}$$

where θ is the angle between the propagation direction $\boldsymbol{k}$ of the incident plane wave and the direction of $\boldsymbol{r}$. Only the the polar angle θ appears, the azimuthal angle φ does not, since the scattered wave exhibits rotational symmetry around the direction of $\boldsymbol{k}$. The coefficient $f(\theta)$ of the outgoing spherical wave is the scattering amplitude; the differential scattering cross section is determined by this term. We shall therefore try to relate $f(\theta)$ to the parameters of the scattering potential.

The spatial part of the wavefunction is determined by the Schrödinger equation (16.4.1). Because of the spherical symmetry of the potential it is practical to use spherical coordinates, and to seek solutions using the method of partial waves. In other words: as the wavefunction in the scattering problem does not depend on the azimuthal angle φ, an expansion in terms of the Legendre polynomials $P_l(\cos\theta)$ associated with the states of angular momentum l is used for the wavefunction:

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}} \sum_{l=0}^{\infty} (2l+1) \mathrm{i}^l R_l(r) P_l(\cos\theta) . \tag{16.4.8}$$

As we shall see, the $\boldsymbol{k}$- and energy dependence appears in $R_l(r)$ in such a way that the true argument turns out to be kr.

In terms of spherical coordinates, the Laplacian in the kinetic energy is

$$\boldsymbol{\nabla}^2 = \frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} + \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial\theta} \sin\theta \frac{\partial}{\partial\theta} + \frac{1}{r^2 \sin^2\theta} \frac{\partial^2}{\partial\varphi^2} . \tag{16.4.9}$$

Rewriting the θ-derivative term using the identity

$$\frac{1}{\sin\theta} \frac{\partial}{\partial\theta} \sin\theta \frac{\partial}{\partial\theta} = \sin^2\theta \frac{\partial^2}{\partial(\cos\theta)^2} - 2\cos\theta \frac{\partial}{\partial(\cos\theta)} , \tag{16.4.10}$$

and exploiting the property that, according to (C.4.15), the Legendre polynomials satisfy the equation

$$\left[(1-\cos^2\theta) \frac{\partial^2}{\partial(\cos\theta)^2} - 2\cos\theta \frac{\partial}{\partial(\cos\theta)} + l(l+1) \right] P_l(\cos\theta) = 0 , \tag{16.4.11}$$

the Schrödinger equation for the radial part $R_l(r)$ takes the form

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}} \frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} + \frac{\hbar^2}{2m_{\mathrm{e}}} \frac{l(l+1)}{r^2} + V(r) \right] R_l(r) = \varepsilon R_l(r) . \tag{16.4.12}$$

Alternatively, this can be written as

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}} \frac{\partial^2}{\partial r^2} + \frac{\hbar^2}{2m_{\mathrm{e}}} \frac{l(l+1)}{r^2} + V(r) \right] (rR_l(r)) = \varepsilon \, (rR_l(r)) , \tag{16.4.13}$$

or, when the energy eigenvalue ε is expressed by a parameter k through $\varepsilon = \hbar^2 k^2 / 2m_{\mathrm{e}}$, as

$$\left[\frac{\partial^2}{\partial r^2} + \frac{2}{r} \frac{\partial}{\partial r} - \frac{l(l+1)}{r^2} - \frac{2m_{\mathrm{e}}}{\hbar^2} V(r) \right] R_l(r) = -k^2 R_l(r) . \tag{16.4.14}$$

A general solution cannot be given, however it is possible to find a solution that is asymptotically valid at large distances from the impurity, where the potential is negligibly small. Note that at distances where the potential vanishes, the radial equation (16.4.14) – when expressed in terms of the variable $kr = z$ – is the same as (C.3.43), the equation for spherical Bessel and

Neumann functions. The asymptotic solution can therefore be written as a linear combination of these functions:

$$R_l(r) = c_l \left[\cos\delta_l \, j_l(kr) - \sin\delta_l \, n_l(kr)\right] . \tag{16.4.15}$$

Using the asymptotic form

$$j_l(kr) \approx \frac{\sin(kr - l\pi/2)}{kr} , \qquad n_l(kr) \approx -\frac{\cos(kr - l\pi/2)}{kr} \tag{16.4.16}$$

of the spherical Bessel and Neumann functions given in Appendix C, we have

$$\psi_{\boldsymbol{k}}(r) \approx \frac{1}{\sqrt{V}} \sum_{l=0}^{\infty} (2l+1)\, \mathrm{i}^l c_l \frac{1}{kr} \sin(kr - l\pi/2 + \delta_l) P_l(\cos\theta) \tag{16.4.17}$$

far from the impurity. Now consider the form (C.4.38) of the incoming plane wave,

$$\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} = \mathrm{e}^{\mathrm{i}kr\cos\theta} = \sum_{l=0}^{\infty} (2l+1) \mathrm{i}^l j_l(kr) P_l(\cos\theta) , \tag{16.4.18}$$

which is valid not only asymptotically but also at $r = 0$ (consequently, it does not contain any second-order spherical Bessel function that is singular at $r = 0$). Taking this expression in the asymptotic region, we have

$$\frac{1}{\sqrt{V}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \approx \frac{1}{\sqrt{V}} \sum_{l=0}^{\infty} (2l+1)\, \mathrm{i}^l \frac{1}{kr} \sin\left(kr - l\pi/2\right) P_l(\cos\theta) . \tag{16.4.19}$$

A comparison of the two formulas shows that scattering by the impurity potential leads to a phase shift δ_l in the lth partial wave.

The coefficient c_l can be determined from the requirement that the wavefunction should indeed be of the form (16.4.7) – that is, it should contain an outgoing spherical wave in addition to the incoming plane wave. Consider now the asymptotic expression for the change of the wavefunction caused by the impurity:

$$\begin{aligned} \psi_{\boldsymbol{k}}(r) - \frac{1}{\sqrt{V}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} = \frac{1}{\sqrt{V}} & \sum_{l=0}^{\infty} (2l+1)\, \mathrm{i}^l P_l(\cos\theta) \frac{1}{kr} \\ & \times \left[c_l \sin\left(kr - l\pi/2 + \delta_l\right) - \sin\left(kr - l\pi/2\right)\right] \end{aligned} \tag{16.4.20}$$

and rewrite the bracketed factor as

$$\frac{c_l}{2\mathrm{i}} \left[\mathrm{e}^{\mathrm{i}(kr - l\pi/2 + \delta_l)} - \mathrm{e}^{-\mathrm{i}(kr - l\pi/2 + \delta_l)}\right] - \frac{1}{2\mathrm{i}} \left[\mathrm{e}^{\mathrm{i}(kr - l\pi/2)} - \mathrm{e}^{-\mathrm{i}(kr - l\pi/2)}\right] . \tag{16.4.21}$$

The terms that contain $\mathrm{e}^{-\mathrm{i}kr}$ – and thus describe an incoming spherical wave – vanish if

$$c_l = \mathrm{e}^{\mathrm{i}\delta_l} . \tag{16.4.22}$$

Since asymptotically

$$\frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \approx \sum_{l=0}^{\infty} \frac{(2l+1)}{2\mathrm{i}kr}\left[\mathrm{e}^{\mathrm{i}kr} - (-1)^l \mathrm{e}^{-\mathrm{i}kr}\right] P_l(\cos\theta) \tag{16.4.23}$$

and

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) \approx \frac{1}{\sqrt{V}} \sum_{l=0}^{\infty} \frac{(2l+1)}{2\mathrm{i}kr}\left[\mathrm{e}^{2\mathrm{i}\delta_l}\mathrm{e}^{\mathrm{i}kr} - (-1)^l \mathrm{e}^{-\mathrm{i}kr}\right] P_l(\cos\theta)\,, \tag{16.4.24}$$

the amplitude of the scattered spherical waves is

$$\begin{aligned} f(\theta) &= \frac{1}{k}\sum_{l=0}^{\infty}(2l+1)(\mathrm{e}^{2\mathrm{i}\delta_l}-1)\frac{1}{2\mathrm{i}}P_l(\cos\theta) \\ &= \frac{1}{k}\sum_{l=0}^{\infty}(2l+1)\mathrm{e}^{\mathrm{i}\delta_l}\sin\delta_l\, P_l(\cos\theta)\,. \end{aligned} \tag{16.4.25}$$

Its square is the differential cross section along the direction θ:

$$\sigma(\theta) = |f(\theta)|^2\,. \tag{16.4.26}$$

Integration over the entire solid angle gives the total cross section of the scattering by the impurity. Exploiting the properties of the Legendre polynomials,

$$\sigma_{\mathrm{tot}} = 2\pi \int_0^{\pi} \sigma(\theta)\sin\theta\,\mathrm{d}\theta = \frac{4\pi}{k^2}\sum_{l=0}^{\infty}(2l+1)\sin^2\delta_l\,. \tag{16.4.27}$$

The influence of the potential thus appears entirely through the phase shifts δ_l.

It follows from the above form of the cross section that there is practically no scattering when the phase shift is an integral multiple of π. As we shall see, for attractive interactions this corresponds to the situation in which electrons are bound to the impurity. Scattering is strongest when $\delta_l = \pi/2, 3\pi/2, \ldots$; in this case we speak of a *resonance* or *virtual bound state.*

Because of their scattering on impurities, electron states have a finite lifetime whose inverse is proportional to the transition probability from a given state to any other state. For spatially disordered impurities an average has to be taken over all possible disordered configurations. Interference terms arising from scattering by different scattering centers then cancel out, resulting in a total transition probability that is proportional to the number of scatterers. As we shall see in Chapter 36 of Volume 3, quantum mechanical interference cannot always be neglected: it can lead to the localization of electron states.

Since the cross section is defined as the ratio of the transition probability and the particle flux, the inverse lifetime is given by

$$\frac{1}{\tau_{\boldsymbol{k}}} = n_{\mathrm{i}} v_{\boldsymbol{k}} \sigma_{\mathrm{tot}}\,, \tag{16.4.28}$$

where n_i is the concentration of impurities. From the expression for the cross section we have

$$\frac{1}{\tau} = \frac{4\pi\hbar n_\mathrm{i}}{m_\mathrm{e}k}\sum_{l=0}^{\infty}(2l+1)\sin^2\delta_l\,. \tag{16.4.29}$$

It will be shown in Chapter 24 that the collision time that appears in the electrical conductivity is different from this lifetime. This is because the current carried by electrons changes only slightly in those scattering processes for which $\theta \approx 0$, while in back-scattering processes of angle $\theta \approx \pi$ the current is significantly reduced. In fact scattering through angle θ should be multiplied by a weight factor $1-\cos\theta$ in the formula for the relaxation time. Thus transport processes are governed by the transport relaxation time defined by

$$\frac{1}{\tau_\mathrm{tr}} = n_\mathrm{i} v_\mathrm{F}\langle\sigma\rangle\,, \tag{16.4.30}$$

where $\langle\sigma\rangle$ is the weighted cross section:

$$\langle\sigma\rangle = 2\pi\int_0^\pi (1-\cos\theta)\sigma(\theta)\sin\theta\,\mathrm{d}\theta = \frac{4\pi}{k^2}\sum_{l=0}^{\infty}(l+1)\sin^2(\delta_l-\delta_{l+1})\,. \tag{16.4.31}$$

Substituting this into the Drude formula, the contribution of impurity scattering to resistivity is

$$\varrho = \frac{m_\mathrm{e}}{n_\mathrm{e}e^2\tau_\mathrm{tr}} = \frac{n_\mathrm{i}m_\mathrm{e}v_\mathrm{F}}{n_\mathrm{e}e^2}\langle\sigma\rangle = \frac{4\pi\hbar n_\mathrm{i}}{n_\mathrm{e}e^2k_\mathrm{F}}\sum_{l=0}^{\infty}(l+1)\sin^2(\delta_l-\delta_{l+1})\,. \tag{16.4.32}$$

This resistivity contribution survives even at very low temperatures where all other scattering processes are frozen in and their contributions vanish. It is therefore called *residual resistivity.*

16.4.3 Friedel Oscillations Around Impurities

It will often prove useful to know how the spatial density of electrons changes around an impurity and how much total charge accumulates there. To calculate these, consider a sphere of radius R centered at the impurity. The contribution of the state of quantum number $\boldsymbol{k}$ to the change in the number of electrons in this sphere is

$$\delta N_{\boldsymbol{k}} = \int_0^{2\pi}\mathrm{d}\varphi\int_0^{\pi}\sin\theta\mathrm{d}\theta\int_0^R\left[\left|\psi_{\boldsymbol{k}}(\boldsymbol{r})\right|^2 - \frac{1}{V}\left|\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\right|^2\right]r^2\,\mathrm{d}r\,. \tag{16.4.33}$$

Writing the wavefunctions in terms of partial waves – see (16.4.8) and (16.4.18) –, integration over the angular variables gives

$$\delta N_{\boldsymbol{k}} = \frac{4\pi}{V}\sum_{l=0}^{\infty}(2l+1)\left[\int_0^R R_l^2(kr)\,r^2\,\mathrm{d}r - \int_0^R j_l^2(kr)\,r^2\,\mathrm{d}r\right]. \tag{16.4.34}$$

To evaluate the integrals, consider (16.4.13), the equation that the radial function R_l has to satisfy. For the sake of conciseness, we shall denote the radial function associated with the wave number k' by R_l'. Multiplying the equation for rR_l by rR_l', and the equation for rR_l' by rR_l, their difference yields

$$rR_l'\frac{\partial^2}{\partial r^2}(rR_l) - rR_l\frac{\partial^2}{\partial r^2}(rR_l') = (k'^2 - k^2)(rR_l)(rR_l')\,. \tag{16.4.35}$$

Integrating both sides from 0 to R,

$$\left[rR_l'\frac{\partial}{\partial r}(rR_l) - rR_l\frac{\partial}{\partial r}(rR_l')\right]_0^R = (k'^2 - k^2)\int_0^R (rR_l)(rR_l')\mathrm{d}r\,. \tag{16.4.36}$$

In the $k' \to k$ limit

$$R_l' = R_l + (k'-k)\frac{\partial R_l}{\partial k}\,, \tag{16.4.37}$$

and so

$$\left[\frac{\partial}{\partial k}(rR_l)\frac{\partial}{\partial r}(rR_l) - rR_l\frac{\partial^2}{\partial r\partial k}(rR_l)\right]_0^R = 2k\int_0^R r^2R_l^2\mathrm{d}r\,. \tag{16.4.38}$$

The left-hand side vanishes at the lower limit. At the upper limit the asymptotic form

$$R_l \approx \frac{\mathrm{e}^{\mathrm{i}\delta_l}}{kr}\sin(kr - l\pi/2 + \delta_l) \tag{16.4.39}$$

of the radial function – implied by (16.4.17) – yields

$$\int_0^R r^2R_l^2\,\mathrm{d}r = \frac{1}{2k^2}\left\{R + \frac{\partial\delta_l}{\partial k} - \frac{1}{2k}\sin\left[2\left(kR + \delta_l - \frac{l\pi}{2}\right)\right]\right\}. \tag{16.4.40}$$

The result for the impurity-free case is obtained from this formula by eliminating the phase shifts. Then, by subtracting one equation from the other, we find

$$\delta N_{\boldsymbol{k}} = \frac{1}{V}\frac{2\pi}{k^2}\sum_{l=0}^{\infty}(2l+1)\left[\frac{\partial\delta_l}{\partial k} - \frac{1}{k}\sin\delta_l\cos(2kR + \delta_l - l\pi)\right]. \tag{16.4.41}$$

Summing the contribution of the occupied states (that is, integrating over the interior of the Fermi sphere and multiplying by a factor 2 for the spin) gives the following formula for the change in the total number of electrons inside a sphere of radius R:

$$\begin{aligned}\delta N &= \frac{V}{4\pi^3}\int\limits_0^{k_\mathrm{F}} \delta N_{\boldsymbol{k}}\, 4\pi k^2\,\mathrm{d}k \\ &= \frac{2}{\pi}\sum_{l=0}^{\infty}(2l+1)\left[\delta_l(k_\mathrm{F}) - \int\limits_0^{k_\mathrm{F}}\frac{\mathrm{d}k}{k}\sin\delta_l\cos\left(2kR+\delta_l-l\pi\right)\right].\end{aligned} \tag{16.4.42}$$

The oscillating terms in this expression can be used to determine the spatial variations of the perturbed charge distribution around the impurity. Denoting the number density at a distance r from the impurity by $n(r)$,

$$\delta N = \int\limits_0^R n(r)4\pi r^2\,\mathrm{d}r\,, \tag{16.4.43}$$

hence

$$\begin{aligned}n(r) &= \frac{1}{4\pi r^2}\frac{\mathrm{d}\delta N(R)}{\mathrm{d}R}\bigg|_{R=r} \\ &= -\frac{1}{2\pi^2 r^3}\sum_l(2l+1)(-1)^l\sin\delta_l\cos(2k_\mathrm{F}r+\delta_l)\,.\end{aligned} \tag{16.4.44}$$

The same result would have been obtained if the variation of the electron density had been determined from the asymptotic form of the perturbed and unperturbed wavefunctions through

$$n(r) = \sum_{\boldsymbol{k},\sigma}\left[\left|\psi_{\boldsymbol{k}}(\boldsymbol{r})\right|^2 - \left|\psi_{\boldsymbol{k}}^{(0)}(\boldsymbol{r})\right|^2\right]. \tag{16.4.45}$$

The perturbation caused by the impurity does not fall off exponentially but much more slowly, as $1/r^3$, and not monotonously but in an oscillatory way. This is called *Friedel oscillation.*[38] The oscillation wavelength is the reciprocal of twice the Fermi wave number, that is, the reciprocal of the diameter of the Fermi sphere. It is important to note that the oscillation is the consequence of the abrupt change in the momentum distribution at the Fermi energy.

At large distances the oscillatory terms drop off. The total accumulated charge around the impurity is then

$$\delta N = \frac{2}{\pi}\sum_{l=0}^{\infty}(2l+1)\delta_l(k_\mathrm{F})\,. \tag{16.4.46}$$

[38] J. Friedel, 1952.

Since electron states are redistributed, electrons are displaced around a charged impurity in such a way that the excess charge should be screened (neutralized) over a relatively short distance – otherwise its contribution to the Coulomb energy would become excessively large –, the phase shifts at the Fermi momentum can be related to the charge Ze of the impurity, which has to be screened:

$$Z = \frac{2}{\pi} \sum_{l=0}^{\infty} (2l+1) \delta_l(k_{\mathrm{F}}) \,. \tag{16.4.47}$$

This formula is known as the *Friedel sum rule.* It can be intuitively understood in the following simple picture. In view of the spherical symmetry of the impurity potential, consider an electron gas with modified boundary conditions: within a sphere of radius R rather than in a rectangular box. By choosing the coefficient c_l in the asymptotic form of the radial wavefunction in such a way that it is normalized in this sphere, we have

$$R_l(r) \approx \frac{1}{\sqrt{2\pi R}} \frac{1}{r} \sin(kr - l\pi/2 + \delta_l) \tag{16.4.48}$$

for large values of r. The allowed values of the wave number k are specified by the requirement that the wavefunction vanish on the surface of the sphere of radius R, that is,

$$kR - \tfrac{1}{2} l\pi + \delta_l(k) = n\pi \,, \qquad n = 0, \pm 1, \pm 2, \dots \,. \tag{16.4.49}$$

Due to the phase shifts, the change of the wave number k with respect to the impurity-free case is

$$\delta k = -\frac{\delta_l(k)}{R} \,. \tag{16.4.50}$$

Since the allowed values of k are separated by regular distances π/R, we find that when the factor 2 arising from spin and the number of states with angular momentum l $(2l+1)$ are taken into account, there are

$$\frac{2}{\pi}(2l+1)\delta_l(k) \tag{16.4.51}$$

allowed electron states in an interval δk. If the impurity possesses an excess charge $\pm Ze$, Z states must be displaced below (or above) the Fermi energy to screen it. This requirement leads to the sum rule (16.4.47).

Using the same argument, the integrated density of states can be determined for an arbitrary energy ε. The change in the number of states below energy ε due to the impurity is

$$\delta N(\varepsilon) = \frac{2}{\pi} \sum_{l=0}^{\infty} (2l+1) \delta_l(\varepsilon) \,. \tag{16.4.52}$$

Its derivative with respect to energy gives the change of the density of states:

$$\delta\rho(\varepsilon) = \frac{2}{\pi}\sum_{l=0}^{\infty}(2l+1)\frac{\mathrm{d}\delta_l(\varepsilon)}{\mathrm{d}\varepsilon}\,. \tag{16.4.53}$$

If the scattering of the lth partial wave features a resonance at energy ε_{r}, then, according to quantum mechanics, the energy dependence of the phase shift is given by

$$\tan\delta_l = \frac{\Gamma}{2(\varepsilon_{\mathrm{r}} - \varepsilon)}\,. \tag{16.4.54}$$

The contribution of the lth partial wave to the change of the density of states is then

$$\delta\rho_l(\varepsilon) = \frac{2}{\pi}(2l+1)\frac{\Gamma/2}{(\varepsilon_{\mathrm{r}} - \varepsilon)^2 + (\Gamma/2)^2}\,, \tag{16.4.55}$$

that is, a Lorentzian peak of half-width Γ appears around ε_{r}.

16.4.4 Bound States Around Impurities

In the foregoing the allowed values of the parameter k were determined from the requirement that the wavefunction vanish on the surface of a sphere of radius R, in other words, that the condition

$$kR - l\pi/2 + \delta_l = n\pi \tag{16.4.56}$$

be met. Because of the phase shift δ_l caused by the impurity, the energies are also shifted relative to those of the ideal electron gas. The perturbed energies are:

$$\varepsilon = \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{n\pi + l\pi/2 - \delta_l}{R}\right)^2\,. \tag{16.4.57}$$

When the impurity potential is weak, δ_l is small, and energies are shifted only slightly. For attractive interactions δ_l is positive and the energy levels are shifted downward, while for repulsive interactions δ_l is negative and they are shifted upward. If δ_l is smaller than π, the nth perturbed level is between the $(n-1)$th and $(n+1)$th unperturbed levels. Since the allowed values of k are separated by regular distances π/R, the energy levels form a quasicontinuum.

A particular situation arises when the attractive potential is sufficiently strong and δ_0 reaches π. The energy of the lowest-lying state (of quantum number $n = 1$ and angular momentum $l = 0$) vanishes – and becomes negative for even stronger potentials. The wave number associated with this level is imaginary, hence the wavefunction decays exponentially with increasing distance from the impurity. Thus, in contrast to the previously discussed electron states that extend over the entire sample, the energy level appearing below the quasicontinuum corresponds to a localized state bound to the impurity.

16.5 Inadequacies of the Free-Electron Model

The inadequacies and failures of the Drude model, based on classical physics, were discussed in Section 16.1.8. The application of quantum mechanics provided a remedy for some of them. The Sommerfeld model gives a much better account of the thermodynamic behavior than the classical model. The electronic specific heat and susceptibility are in order-of-magnitude agreement with experimental data for simple metals, and the theoretical formulas for their temperature dependence are also consistent with measurements. However, the calculated value of the density of states agrees with the measured data only for the simplest metals. When it is estimated from the susceptibility data, it is sometimes several orders of magnitude off the value calculated from the free-electron model. The same is true – though, in part, for different reasons – when the density of states (or the electron mass) is determined from the linear contribution of the specific heat.[39] All this indicates that the free-electron assumption is not justified in most metals.

It is interesting to note that the estimate obtained for the electron mean free path in the Drude model proves incorrect in the Sommerfeld model. In the former, the mean free path of electrons moving at thermal velocities at room temperature is on the order of atomic distances, thus it seemed plausible to assume that the resistivity of the metal is due to the scattering of the electrons by the rigid lattice of ions. In the Sommerfeld model the characteristic velocity of electrons is the Fermi velocity, which is usually an order of magnitude larger than the thermal velocity at room temperature. This characteristic electron velocity is preserved down to low temperatures, and since the resistivity is three or four orders of magnitude smaller in this region, the corresponding mean free path is much larger than the distance between ions. The mean free path is therefore not related to the atomic spacing.

In connection with the electrical and thermal conductivity of metals, a significant improvement is obtained for the value of the Lorenz number provided it is determined from measurements at room temperature. However, the new model cannot eliminate another important shortcoming of the classical treatment: the Wiedemann–Franz law is not satisfied by the resistivity and thermal conductivity data measured at lower temperatures. To illustrate this, the temperature dependence of resistivity and thermal conductivity are shown for some simple metals in Fig. 16.12.

Around room temperature, the resistivity increases linearly with T, while the thermal conductivity is essentially constant, thus the Wiedemann–Franz law is satisfied. As the temperatures is lowered, λ increases rapidly, and after reaching a maximum it vanishes linearly with T at very low temperatures. At such low temperatures the resistivity is practically constant, and the

[39] Note that this linear temperature dependence can be observed only at very low temperatures, leading to particular difficulties in the evaluation of experimental specific heat data. At higher temperatures this is suppressed by nonelectronic – for example, phononic – contributions.

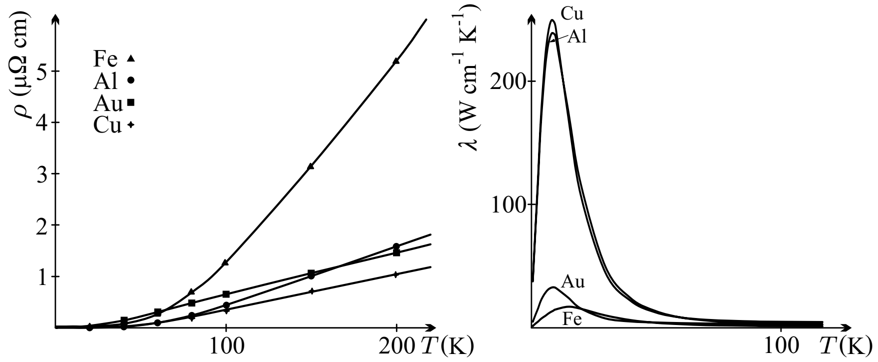


Fig. 16.12. Temperature dependence of resistivity and thermal conductivity for some simple metals

Wiedemann–Franz law is valid again. However, in the intermediate temperature range the description of the scattering processes in terms of a relaxation time – that is equivalent to assuming elastic scattering – is called into question. From the study of lattice vibrations it is known that the energy of the lattice can change in discrete steps that correspond to the creation or annihilation of phonons. It is therefore natural to expect that electrons scattered by the lattice can take part in inelastic collision processes as well. Those processes in which an electron transfers a part of its energy to the lattice while its wave vector changes only slightly contribute to the decay of the heat current, but practically not to the resistivity. Thus our assumptions about elastic scattering and the applicability of the relaxation time will have to be reexamined – and a more precise description of the electron–phonon interactions is called for in the region where they cannot be used.

The Sommerfeld model does not offer an explanation for the observation that the Hall coefficient can be different from its classical value not only in its magnitude but also in its sign. The same applies to the thermoelectric power. Measured data are usually in good agreement with the predictions of the free-electron model for alkali metals but significant deviations are found for various other metals. Moreover, the model itself says nothing about which electrons must or can be considered free and which ones bound to the core. Nonetheless the above findings indicate that the Sommerfeld model can be applicable only to those metals that have a single incomplete shell, the outermost s-shell. It looks as if in solids p- and especially d-electrons were neither almost free nor completely bound to the core.

In connection with electrical resistivity it should also be noted that the resistivity values of order $1-100\,\mathrm{n}\Omega\,\mathrm{m}$ listed in Table 16.2 are typical of good metals. A material is customarily considered a metal if its resistivity at room temperature exceeds $10^6\ (\Omega\,\mathrm{m})^{-1}$. The resistivity of bismuth is on this borderline. There are, however metals with much lower but still finite resistivities,[40] therefore these, too, must contain some "free" electrons. Using the Drude formula one can describe arbitrarily small resistivities by choosing the collision time of charge carriers (electrons) sufficiently small. However, the above picture of electrical conductivity is physically sensible only if the electron mean free path is larger than the electron wavelength, that is, $k_{\mathrm{F}} l > 1$. This leads to the conclusion that there is a minimum metallic conductivity, which would be of order $10^5\,\mathrm{S\,m^{-1}}$ in a three-dimensional sample. When discussing the role of disorder at the end of the third volume we shall see that the situation is much more complex. Here we shall content ourselves with the observation that even in pure materials resistivity is often much lower than the above metallic value – and its temperature dependence can also be very different from that of normal metals. Such materials are called semiconductors. The nature of such materials – just like superconductivity, or the magnetic properties of metals – lies outside the realm of the Sommerfeld model. This clearly indicates the necessity to go beyond the free-electron model in order to understand the true nature of electron states in solids.

Further Reading

1. A. Sommerfeld und H. Bethe, *Elektronentheorie der Metalle*, in Handbuch der Physik, Zweite Auflage, Band XXIV. Zweiter Teil, Verlag von Julius Springer, Berlin (1933); Heidelberger Taschenbücher, Bd. 19, Springer-Verlag, Berlin (1967).
2. A. H. Wilson, *The Theory of Metals*, Second Edition, Cambridge University Press, Cambridge (1958).

[40] The room-temperature resistivity of pure germanium is about $2\,(\Omega\,\mathrm{m})^{-1}$.

17

Electrons in the Periodic Potential of a Crystal

The discussion of the properties of metals in the previous chapter was based on a free-electron model (or rather: a gas of neutral fermionic particles) in an empty box. The classical Drude model and the quantum mechanical Sommerfeld model (based on the Fermi–Dirac statistics) were introduced, and it was shown that a suitable choice of certain parameters leads to a good description of several properties of simple metals. We have to specify which electrons of the atoms remain bound in the ion core and which can be considered free (and hence can participate in conduction in the solid state). The number of conduction electrons is an important parameter even in the Drude model. In much the same way, in the Sommerfeld model various properties of metals are determined through the Fermi energy, by the number of conduction electrons per atom. Since core electrons are ignored, these models obviously cannot account for the electrical properties of ionically or covalently bonded materials, in which electrons are fairly well localized to the ions and covalent bonds. Therefore not even the quantum mechanical model can explain the existence of insulators and semiconductors. However, the conduction electrons are not perfectly free even in metals, since they move through the regular crystalline array of ions, thus their motion is determined by the periodic crystal potential. To resolve these difficulties and contradictions, the behavior of the electrons has to be studied in the presence of the atoms (ions) that make up the crystal lattice.

As a first approximation, we shall consider ions to be fixed at the lattice points, and ignore their vibrations, the phonons. The justification of this approximation and the influence of the motion of ions on the electrons will be discussed in Chapter 23. In the present chapter we shall lump the effects of ions into a local static potential $U_{\text{ion}}(\boldsymbol{r})$ that can be taken as the sum of the individual atomic potentials $v_{\text{a}}(\boldsymbol{r} - \boldsymbol{R}_m)$ of periodically spaced ions.

Taking into account the influence of other electrons is much more difficult. Only by employing the methods of the many-body problem can electron–electron interactions be treated more or less precisely. We shall delve into this complex subject in Volume 3. Below we shall assume that electrons feel the influence of the others only in an average sense, through the averaged potential $U_{\mathrm{el}}(\boldsymbol{r})$ that also possesses the periodicity of the lattice. In this chapter we shall examine how the electronic states can be described in ideal crystals and present some general features of the energy spectrum without any assumptions about the actual form of this periodic potential. At the end of the chapter we shall briefly discuss what happens to the electrons if the structure is not perfectly crystalline. The methods to determine energy eigenvalues for specific periodic potentials will be discussed in the next chapters.

17.1 Band Structure of Electronic States

In this section we shall first introduce the Bloch functions – that is, the Bloch form of electron wavefunctions obtained in the presence of a periodic potential using the Bloch theorem (formulated in full generality in Chapter 6), and then determine some general properties of the electron spectrum.

17.1.1 Bloch States

As mentioned in the introductory part, we shall assume that electrons feel the presence of ions and other electrons only through the spin-independent lattice-periodic potentials $U_{\mathrm{ion}}(\boldsymbol{r})$ and $U_{\mathrm{el}}(\boldsymbol{r})$. Therefore the same one-particle potential,

$$U(\boldsymbol{r}) = U_{\mathrm{ion}}(\boldsymbol{r}) + U_{\mathrm{el}}(\boldsymbol{r}) \tag{17.1.1}$$

acts on each electron. Using this averaged potential, the Hamiltonian of a system of N_{e} electrons is

$$\mathcal{H}_{\mathrm{el}} = -\frac{\hbar^2}{2m_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \frac{\partial^2}{\partial \boldsymbol{r}_i^2} + \sum_{i=1}^{N_{\mathrm{e}}} U(\boldsymbol{r}_i)\,. \tag{17.1.2}$$

This Hamiltonian is the sum of independent one-particle Hamiltonians. Therefore when the solutions of the one-particle Schrödinger equation

$$\mathcal{H}(\boldsymbol{r})\psi_i(\boldsymbol{r}) \equiv \left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r})\right]\psi_i(\boldsymbol{r}) = \varepsilon_i\psi_i(\boldsymbol{r}) \tag{17.1.3}$$

are known, the total wavefunction Ψ of the electron system can be written as the Slater determinant of the wavefunctions of occupied one-particle states, thereby satisfying the requirement that the total wavefunction should be completely antisymmetric. Since the potential is spin-independent, as the spin–orbit interaction is ignored here, only the spatial parts of the wavefunctions are considered. Nevertheless, when writing down the Slater determinant, the requirement of complete antisymmetry applies to the spin variables as well:

$$\Psi = \frac{1}{\sqrt{N_\mathrm{e}!}} \begin{vmatrix} \psi_{i_1,\sigma_1}(\boldsymbol{r}_1, s_1) & \psi_{i_1,\sigma_1}(\boldsymbol{r}_2, s_2) & \dots & \psi_{i_1,\sigma_1}(\boldsymbol{r}_{N_\mathrm{e}}, s_{N_\mathrm{e}}) \\ \psi_{i_2,\sigma_2}(\boldsymbol{r}_1, s_1) & \psi_{i_2,\sigma_2}(\boldsymbol{r}_2, s_2) & \dots & \psi_{i_2,\sigma_2}(\boldsymbol{r}_{N_\mathrm{e}}, s_{N_\mathrm{e}}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{i_{N_\mathrm{e}},\sigma_{N_\mathrm{e}}}(\boldsymbol{r}_1, s_1) & \psi_{i_{N_\mathrm{e}},\sigma_{N_\mathrm{e}}}(\boldsymbol{r}_2, s_2) & \dots & \psi_{i_{N_\mathrm{e}},\sigma_{N_\mathrm{e}}}(\boldsymbol{r}_{N_\mathrm{e}}, s_{N_\mathrm{e}}) \end{vmatrix} . \tag{17.1.4}$$

Because of the periodicity of the crystal structure, the potential $U(\boldsymbol{r})$ satisfies the condition

$$U(\boldsymbol{r} + \boldsymbol{t}_m) = U(\boldsymbol{r}) , \tag{17.1.5}$$

where $\boldsymbol{t}_m$ is an arbitrary translation vector of the crystal lattice. Along with the potential, the full Hamiltonian is also lattice periodic, and satisfies (6.2.1). In accordance with Bloch's theorem, solutions must satisfy the condition

$$\boxed{\psi_{\boldsymbol{k}}(\boldsymbol{r} + \boldsymbol{t}_m) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{t}_m} \psi_{\boldsymbol{k}}(\boldsymbol{r})} \tag{17.1.6}$$

given in (6.2.5), where the vector $\boldsymbol{k}$ can take discrete values allowed by the periodic boundary condition.

The Bloch condition on the wavefunction can be formulated in another way by separating the phase factor $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ off the wavefunction, and by introducing the function $u_{\boldsymbol{k}}(\boldsymbol{r})$ through the definition

$$\boxed{\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{\boldsymbol{k}}(\boldsymbol{r}) .} \tag{17.1.7}$$

It follows directly from (17.1.6) that

$$\psi_{\boldsymbol{k}}(\boldsymbol{r} + \boldsymbol{t}_m) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{t}_m} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{\boldsymbol{k}}(\boldsymbol{r}) . \tag{17.1.8}$$

On the other hand, if the wavefunction given in (17.1.7) is taken at the translated position $\boldsymbol{r} + \boldsymbol{t}_m$, we find

$$\psi_{\boldsymbol{k}}(\boldsymbol{r} + \boldsymbol{t}_m) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{t}_m)} u_{\boldsymbol{k}}(\boldsymbol{r} + \boldsymbol{t}_m) . \tag{17.1.9}$$

Comparison of the two formulas gives

$$\boxed{u_{\boldsymbol{k}}(\boldsymbol{r} + \boldsymbol{t}_m) = u_{\boldsymbol{k}}(\boldsymbol{r}) ,} \tag{17.1.10}$$

that is, the function $u_{\boldsymbol{k}}(\boldsymbol{r})$ obtained by the separation of the phase factor is periodic with the periodicity of the lattice. Bloch's theorem is therefore equivalent to the statement that *the eigenfunctions of a lattice-periodic Hamiltonian can be written in the form (17.1.7), where the $u_{\boldsymbol{k}}(\boldsymbol{r})$ are lattice-periodic functions.* Wavefunctions of this form are called Bloch functions.[1] Such a function $u_{\boldsymbol{k}}(\boldsymbol{r})$ and the real part of the corresponding Bloch function are shown in Fig. 17.1.

[1] Sometimes the lattice-periodic functions $u_{\boldsymbol{k}}(\boldsymbol{r})$ are called Bloch functions in the literature.

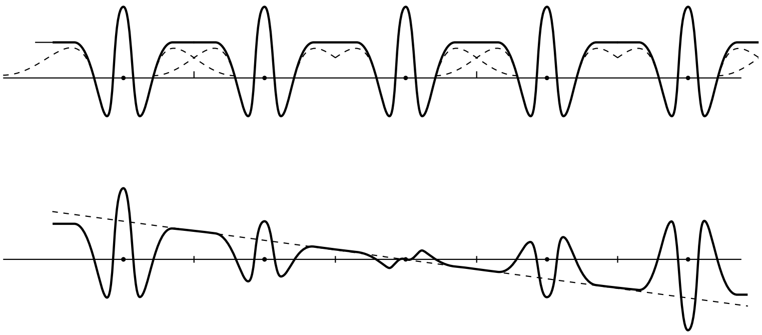


Fig. 17.1. A lattice-periodic function $u(x)$ and the real part of the corresponding Bloch function

In the absence of a periodic potential the wavefunction is a plane wave $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}/\sqrt{V}$. The function $u_{\boldsymbol{k}}(\boldsymbol{r})$ describes the deformation of the wavefunction relative to the plane wave. This deformation is identical for each primitive cell of the crystal. It would, therefore, be more natural to define the Bloch function as

$$\boxed{\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_{\boldsymbol{k}}(\boldsymbol{r})} \tag{17.1.11}$$

instead of (17.1.7). Nonetheless we shall often drop the factor $1/\sqrt{V}$ wherever this does not cause confusion.

The form (17.1.7) of the Bloch function immediately confirms our previous remark that $\hbar\boldsymbol{k}$ is not the momentum of the Bloch state. The momentum operator transforms wavefunctions that satisfy the Bloch theorem into

$$\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}\left(\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_{\boldsymbol{k}}(\boldsymbol{r})\right) = \hbar\boldsymbol{k}\psi_{\boldsymbol{k}}(\boldsymbol{r}) + \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}u_{\boldsymbol{k}}(\boldsymbol{r}) \tag{17.1.12}$$

indicating that $\psi_{\boldsymbol{k}}(\boldsymbol{r})$ is not an eigenstate of the momentum operator.

17.1.2 Energy Levels of Bloch States

Substituting the one-particle wavefunction (17.1.7) into the Schrödinger equation (17.1.3),

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r})\right]\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_{\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_{\boldsymbol{k}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_{\boldsymbol{k}}(\boldsymbol{r})\,. \tag{17.1.13}$$

Differentiating the exponential factor and separating $\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$ on both sides, the function $u_{\boldsymbol{k}}(\boldsymbol{r})$ satisfies the equation

$$\left[\frac{\hbar^2\boldsymbol{k}^2}{2m_{\mathrm{e}}} - \frac{\mathrm{i}\hbar^2}{m_{\mathrm{e}}}\boldsymbol{k}\cdot\boldsymbol{\nabla} - \frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r})\right]u_{\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_{\boldsymbol{k}}u_{\boldsymbol{k}}(\boldsymbol{r})\,. \tag{17.1.14}$$

We shall also use the equivalent form

$$\left[\frac{1}{2m_{\mathrm{e}}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+\hbar\boldsymbol{k}\right)^{2}+U(\boldsymbol{r})\right]u_{\boldsymbol{k}}(\boldsymbol{r})=\varepsilon_{\boldsymbol{k}}u_{\boldsymbol{k}}(\boldsymbol{r})\,. \tag{17.1.15}$$

By introducing the notation

$$\mathcal{H}_{\boldsymbol{k}}=\frac{1}{2m_{\mathrm{e}}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+\hbar\boldsymbol{k}\right)^{2}+U(\boldsymbol{r})\,, \tag{17.1.16}$$

(17.1.15) can be considered as an eigenvalue equation for the function $u_{\boldsymbol{k}}(\boldsymbol{r})$:

$$\mathcal{H}_{\boldsymbol{k}}u_{\boldsymbol{k}}(\boldsymbol{r})=\varepsilon_{\boldsymbol{k}}u_{\boldsymbol{k}}(\boldsymbol{r})\,. \tag{17.1.17}$$

Since $u_{\boldsymbol{k}}(\boldsymbol{r})$ is lattice periodic, it is sufficient to solve this equation for a single primitive cell of the crystal, with periodic boundary conditions, however the solutions have to be found for all possible values of $\boldsymbol{k}$.

The eigenvalue problem has infinitely many solutions for each $\boldsymbol{k}$. We shall label them by a second index, n. The eigenvalue equation to be solved is then

$$\left[\frac{\hbar^{2}\boldsymbol{k}^{2}}{2m_{\mathrm{e}}}-\frac{\mathrm{i}\hbar^{2}}{m_{\mathrm{e}}}\boldsymbol{k}\cdot\boldsymbol{\nabla}-\frac{\hbar^{2}}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^{2}+U(\boldsymbol{r})\right]u_{n\boldsymbol{k}}(\boldsymbol{r})=\varepsilon_{n\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r})\,. \tag{17.1.18}$$

The Bloch functions $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ form a complete orthonormal set:

$$\int\psi^{*}_{n\boldsymbol{k}}(\boldsymbol{r})\psi_{n'\boldsymbol{k}'}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}=\delta_{n,n'}\delta_{\boldsymbol{k},\boldsymbol{k}'}\,, \tag{17.1.19}$$

and

$$\sum_{n\boldsymbol{k}}\psi^{*}_{n\boldsymbol{k}}(\boldsymbol{r})\psi_{n\boldsymbol{k}}(\boldsymbol{r}')=\delta(\boldsymbol{r}-\boldsymbol{r}')\,. \tag{17.1.20}$$

Figure 17.2 shows for each allowed value of $\boldsymbol{k}$ the four lowest energy eigenvalues (obtained for an arbitrarily chosen potential) for chains of the same lattice constant a but different lengths $L=Na$ subject to Born–von Kármán boundary conditions.

In short chains, where the Brillouin zone contains only a few allowed wave numbers k, the location of the energy levels seem to lack any order. When the number of atoms is increased, the allowed k values fill the region $(-\pi/a,\pi/a)$ more densely, and the energy eigenvalues ε_{nk} are arranged in such a way that in the $N\to\infty$ limit they make up continuous curves that are similar to phonon dispersion curves. In finite but sufficiently long chains energy eigenvalues can be labeled in such a manner that for a given n the energies associated with adjacent k values are close – that is, ε_{nk} can be approximated by a continuous function in k-space. If there are several states with close-by energies, one can impose the requirement that the continuous approximation of ε_{nk} should also have a continuous derivative. The energies of the states associated with a particular n are then arranged into bands, which explains why the label n is called the band index. Similarly, band indices can be assigned to the levels of the electronic energy spectrum in three-dimensional crystals, too.

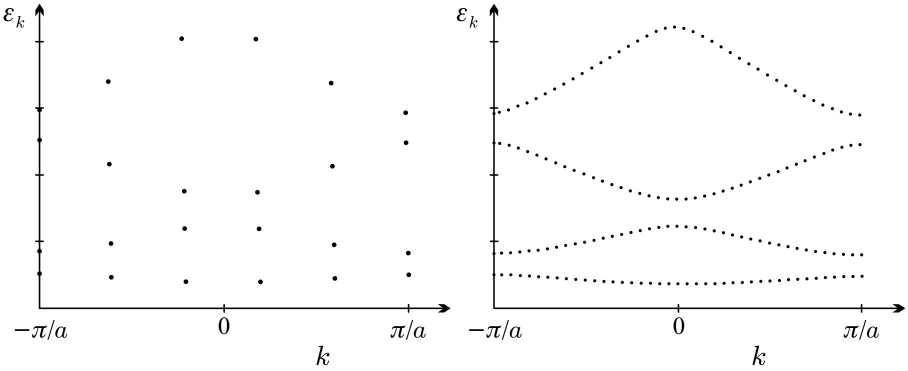


Fig. 17.2. The energy levels of electrons moving in the periodic potential of one-dimensional chains made up of 5 and 40 atoms

17.1.3 Eigenvalue Problem for Equivalent $\boldsymbol{k}$ Vectors

The $\boldsymbol{k}$ vectors that characterize the behavior of the eigenstates of a lattice-periodic system under translations were defined in the cell spanned by the primitive vectors of the reciprocal lattice in Chapter 6. It was then shown that, as far as translational symmetries are concerned, an equivalent description is obtained when the vectors $\boldsymbol{k}$ are replaced by vectors $\boldsymbol{k}' = \boldsymbol{k}+\boldsymbol{G}$ that differ from them in a vector of the reciprocal lattice. Equivalence means that the same results are obtained for measurable quantities such as the energy spectrum or the spatial density of electrons. On the other hand, the wavefunction can receive an extra phase factor. We shall demonstrate these for the states given in terms of Bloch functions.

We shall first show that by replacing $\boldsymbol{k}$ (defined in the primitive cell) by its equivalent $\boldsymbol{k}' = \boldsymbol{k} + \boldsymbol{G}$ (defined in the Brillouin zone), the wavefunction of the state can be written in the Bloch form in terms of the new wave vector. To this end, we shall write the Bloch function

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{17.1.21}$$

as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}'-\boldsymbol{G})\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}} u_{n\boldsymbol{k}'}(\boldsymbol{r}) \,, \tag{17.1.22}$$

where

$$u_{n\boldsymbol{k}'}(\boldsymbol{r}) = u_{n,\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r}) = \mathrm{e}^{-\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r}) \,. \tag{17.1.23}$$

It follows immediately from the lattice periodicity of $u_{n\boldsymbol{k}}(\boldsymbol{r})$ and (5.2.20) that

$$u_{n\boldsymbol{k}'}(\boldsymbol{r} + \boldsymbol{t}_m) = u_{n\boldsymbol{k}'}(\boldsymbol{r}) \,, \tag{17.1.24}$$

that is, this function is also lattice periodic. As the probability of finding the electron at $\boldsymbol{r}$ is $|u_{n\boldsymbol{k}}(\boldsymbol{r})|^2$, and (17.1.23) implies $|u_{n\boldsymbol{k}'}(\boldsymbol{r})|^2 = |u_{n\boldsymbol{k}}(\boldsymbol{r})|^2$, the states associated with $\boldsymbol{k}$ and $\boldsymbol{k} + \boldsymbol{G}$ are equivalent in this sense.

It is worth noting that the Bloch function $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is periodic in reciprocal space and satisfies (17.1.6) in real space, while $u_{n\boldsymbol{k}}(\boldsymbol{r})$ is periodic in real space and satisfies (17.1.23) in reciprocal space.

Let us now examine the eigenvalue problem of the states associated with the equivalent vectors $\boldsymbol{k}+\boldsymbol{G}$. From (17.1.18) we have

$$\begin{aligned}\left[\frac{\hbar^2(\boldsymbol{k}+\boldsymbol{G})^2}{2m_{\mathrm{e}}}-\frac{\mathrm{i}\hbar^2}{m_{\mathrm{e}}}(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{\nabla}-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right]u_{n,\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r})&\\ =\varepsilon_{n,\boldsymbol{k}+\boldsymbol{G}}u_{n,\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r})\,.&\end{aligned}\tag{17.1.25}$$

Expressing $u_{n,\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r})$ in terms of $u_{n\boldsymbol{k}}(\boldsymbol{r})$ through (17.1.23), differentiation of the exponential factor leads to

$$\left[\frac{\hbar^2\boldsymbol{k}^2}{2m_{\mathrm{e}}}-\frac{\mathrm{i}\hbar^2}{m_{\mathrm{e}}}\boldsymbol{k}\cdot\boldsymbol{\nabla}-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right]u_{n\boldsymbol{k}}(\boldsymbol{r})=\varepsilon_{n,\boldsymbol{k}+\boldsymbol{G}}u_{n\boldsymbol{k}}(\boldsymbol{r})\,.\tag{17.1.26}$$

Comparison with (17.1.18) gives

$$\varepsilon_{n,\boldsymbol{k}+\boldsymbol{G}}=\varepsilon_{n\boldsymbol{k}}\,,\tag{17.1.27}$$

that is, equivalent $\boldsymbol{k}$ vectors are associated with the same energy.

17.1.4 Role of the Spin–Orbit Interaction

Up to now electrons were assumed to feel a spin-independent potential, and thus energy eigenstates were independent of the spin orientation: $\varepsilon_{n\boldsymbol{k}\uparrow}=\varepsilon_{n\boldsymbol{k}\downarrow}$. In the absence of a magnetic field energies are therefore doubly degenerate. However, in the field of heavy ions, where relativistic effects cannot be ignored, spin–orbit coupling must also be considered. Using (3.1.30) – or more precisely its spin-dependent part, which gives the most important contribution –, the one-particle states have to be determined from the equation

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})+\frac{\hbar^2}{4\mathrm{i}m_{\mathrm{e}}^2c^2}\boldsymbol{\sigma}\cdot\big((\boldsymbol{\nabla}U(\boldsymbol{r}))\times\boldsymbol{\nabla}\big)\right]\psi_{n\boldsymbol{k}}(\boldsymbol{r})=\varepsilon_{n\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,,\tag{17.1.28}$$

rather than (17.1.13). In this formula $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is a two-component spinor.

Writing the electron wavefunction in its Bloch form, the equation for $u_{n\boldsymbol{k}}(\boldsymbol{r})$ reads

$$\begin{aligned}\bigg\{\frac{1}{2m_{\mathrm{e}}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+\hbar\boldsymbol{k}\right)^2+U(\boldsymbol{r})+\frac{\hbar}{4m_{\mathrm{e}}^2c^2}\boldsymbol{\sigma}\cdot\left[\boldsymbol{\nabla}U(\boldsymbol{r})\times\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}\right]&\\ +\frac{\hbar^2}{4m_{\mathrm{e}}^2c^2}\boldsymbol{k}\cdot\left[\boldsymbol{\sigma}\times\boldsymbol{\nabla}U(\boldsymbol{r})\right]\bigg\}u_{n\boldsymbol{k}}(\boldsymbol{r})=\varepsilon_{n\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r})\,.&\end{aligned}\tag{17.1.29}$$

As we shall see in connection with the band structure of semiconductors in Sec. 20.2, this interaction splits certain – otherwise degenerate – bands. If the crystal does not possess inversion symmetry (which is the case for the sphalerite structure among others), the relation $\varepsilon_{n,\boldsymbol{k},\alpha} = \varepsilon_{n,-\boldsymbol{k},\beta}$, implied by time reversal and Kramers' theorem, continues to hold. This relation establishes that for each state there exists another one with the same energy but opposite wave vector and spin, while for a given $\boldsymbol{k}$ the two spin states are of different energy. This is known as the *Dresselhaus splitting*.[2] On the other hand, when the crystal possesses inversion symmetry, $u_{n,-\boldsymbol{k}}(\boldsymbol{r}) = u_{n,\boldsymbol{k}}(-\boldsymbol{r})$ and $\varepsilon_{n,\boldsymbol{k},\alpha} = \varepsilon_{n,-\boldsymbol{k},\alpha}$. Combined with time reversal, $\varepsilon_{n,\boldsymbol{k},\alpha} = \varepsilon_{n,\boldsymbol{k},\beta}$, which shows that spin degeneracy is preserved.

17.2 Representation of the Band Structure

A complete knowledge of the band structure requires the solution of the eigenvalue problem for each vector $\boldsymbol{k}$ of the primitive cell of the reciprocal lattice (or the Brillouin zone) – that is, the relation between energy and wave vector has to be computed for each band. Since the problem is usually solved numerically, this would require the specification of an excessively large amount of data. Much like for the determination of the phonon spectrum, calculations are usually performed only for some special, high-symmetry directions of the Brillouin zone, and the dispersion curves of Bloch electrons are also displayed only along these directions. For materials that crystallize in fcc structure (whose Brillouin zone is shown in Fig. 7.11), the calculations are usually performed along the lines Δ and Λ – which connect the center Γ of the Brillouin zone with the centers X and L of the square and hexagonal shaped faces –, and perhaps for some other vectors along other characteristic directions. Figure 17.3 shows the band structure of aluminum calculated in this way.

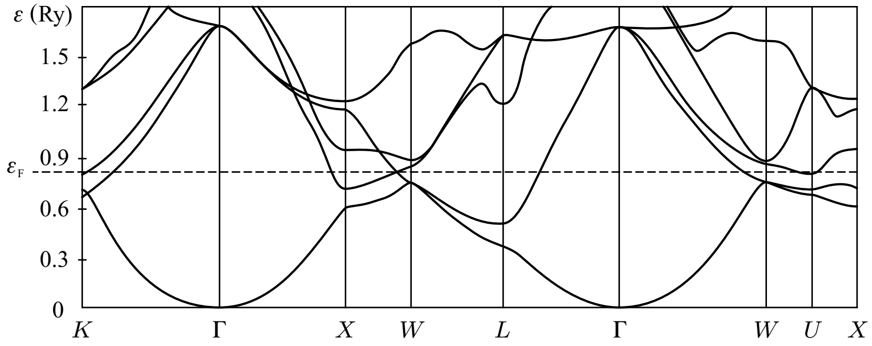


Fig. 17.3. Calculated band structure of aluminum along high-symmetry directions of the Brillouin zone

[2] G. Dresselhaus, 1955.

As the figure shows, if the band index were chosen in such a way that the energies are indexed in ascending order for any $\boldsymbol{k}$ then break points would appear in the $\varepsilon_{n\boldsymbol{k}}$ vs. $\boldsymbol{k}$ plot (with n fixed) wherever two lines intersect. As has been mentioned before, we shall rather require that for a given band index the dispersion relation in $\boldsymbol{k}$ lead to continuous, smooth curves. The dispersion curves of different bands may therefore intersect each other, and the energy regions covered by the bands may overlap.

However, bands can also overlap even when the dispersion curves – or in higher dimensions the $\varepsilon_{n\boldsymbol{k}}$ (hyper)surfaces – do not intersect, nevertheless there are states whose energy is the same although they are associated with different points of the Brillouin zone and belong in different bands.

17.2.1 Reduced-, Repeated-, and Extended-Zone Schemes

As we have seen, the same energy eigenvalue is associated with equivalent $\boldsymbol{k}$ vectors, it is therefore immaterial whether wave vectors defined in the Brillouin zone or in the primitive cell of the reciprocal lattice are used. With wave vectors reduced to the Brillouin zone, the energies of each band can be represented – occasionally separately – in the same Brillouin zone. This is the *reduced-zone scheme.*

Sometimes it is more practical not to restrict wave vectors to the Brillouin zone but use all equivalent vectors $\boldsymbol{k} + \boldsymbol{G}$ as well. By virtue of (17.1.27) the band structure can then be represented by repeating all dispersion curves over the whole $\boldsymbol{k}$-space. This is the *repeated-zone scheme.*

Finally, the infinite number of solutions associated with a given $\boldsymbol{k}$ can also be distributed among the infinite number of vectors $\boldsymbol{k}+\boldsymbol{G}$ in such a way that one solution should belong to each equivalent vector. This can be achieved in two different ways. The first possibility is to draw the Brillouin zones around each lattice point, which gives an unambiguous filling of the entire space. Then, using a predefined – and, to a certain extent, arbitrary – procedure, a vector $\boldsymbol{G}$ of the reciprocal lattice is assigned to each band index, and the states in that band are associated with the wave vectors in the Brillouin zone around that particular $\boldsymbol{G}$.

In the other, more commonly used method for distributing the band states among equivalent $\boldsymbol{k}$ vectors, the notion of higher (second, third, etc.) Brillouin zones is introduced, and the reciprocal space is divided in another way. To this end, we shall generalize Dirichlet's procedure – mentioned in Section 5.1.4 in connection with the construction of the Wigner–Seitz cell, and which is also the method for constructing Brillouin zones in reciprocal space. In this generalization a selected lattice point of the reciprocal lattice is connected with all other lattice points of the reciprocal lattice, and the perpendicular bisecting planes of the segments are drawn. These planes are Bragg planes because any vector $\boldsymbol{k}$ drawn from the selected reciprocal-lattice point to a point of the plane satisfies the Bragg condition (8.1.7). The division of the entire reciprocal space among higher Brillouin zones by means of the Bragg planes is based on the criterion of how many Bragg planes need to be crossed (at least) to reach a particular region from the selected point.

The usual Brillouin zone around the selected point will be the first Brillouin zone. The second Brillouin zone will comprise those regions that can only be reached along a straight line from the selected point by intersecting one Bragg plane. In other words: the second zone comprises those regions that have a common boundary with the first zone.

The third Brillouin zone is made up of regions that have a common boundary with the second zone. To reach such regions from the selected starting point, two Bragg planes need to be crossed. In general: the nth Brillouin zone is reached by crossing $n-1$ Bragg planes. This division of the reciprocal space is shown in Fig. 17.4 for a two-dimensional square lattice.

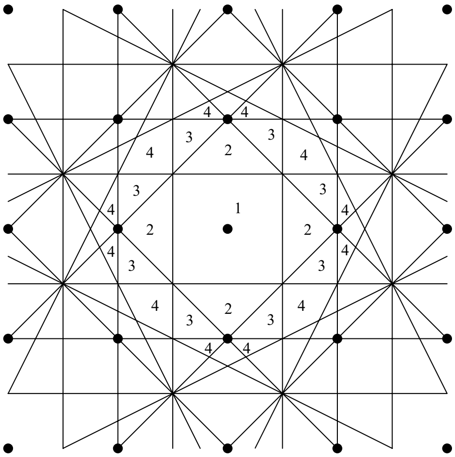


Fig. 17.4. Division of the reciprocal space of a square lattice to first, second, third, etc. Brillouin zones

Higher Brillouin zones may consist of disjoint parts. Nevertheless when these are translated through suitably chosen vectors of the reciprocal lattice, they are found to cover the first Brillouin zone precisely – that is, the total volume of each higher Brillouin zone is the same as that of the first. This is demonstrated in Fig. 17.5 where the parts of the second, third, and fourth Brillouin zones (marked by the corresponding numbers in the previous figure) are moved in such a way that they make up a square.

Figure 17.6 shows the external boundaries of the second, third, and fourth Brillouin zones of a simple cubic lattice, while Fig. 17.7 shows the external boundaries of the first and second Brillouin zones for face- and body-centered cubic lattices. Once again, each higher Brillouin zone has the same total volume as the first, and they can be reduced to the first Brillouin zone by suitable translations.

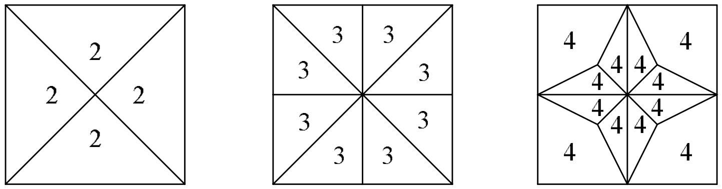


Fig. 17.5. Reduction of the second, third, and fourth Brillouin zones of a square lattice to the first zone

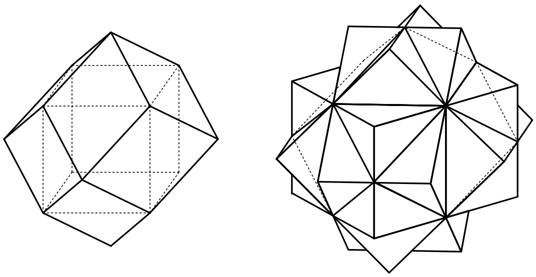


Fig. 17.6. External boundaries of the second, third, and fourth Brillouin zones for a simple cubic lattice

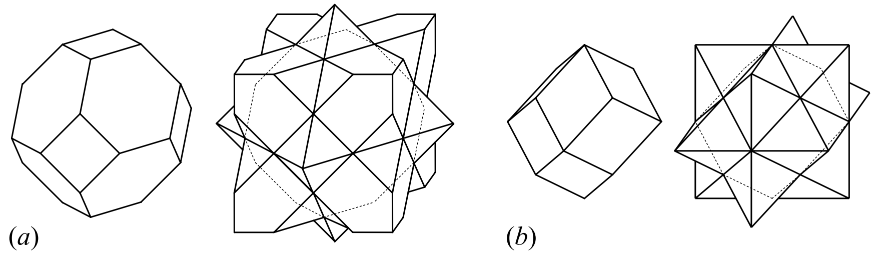


Fig. 17.7. External boundaries of the first and second Brillouin zones for (*a*) face-centered and (*b*) body-centered cubic lattices

The assignment of the states to the zones is then done simply by assigning the states of the first band to the wave vectors in the first Brillouin zone, the states of the second band to the wave vectors in the second Brillouin zone, an so forth. This is how the *extended-zone scheme* is obtained.

In Fig. 17.2 the band structure was plotted as a function of the reduced wave number in the Brillouin zone. The same band structure of the 40-atom chain represented in the repeated- and extended-zone schemes is shown in Fig. 17.8. For the latter the fourth band is outside the displayed region.

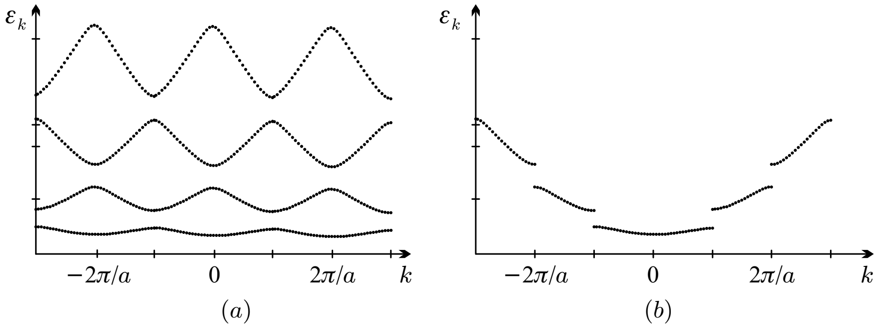


Fig. 17.8. The energy levels shown in Fig. 17.2 in the (*a*) repeated- and (*b*) extended-zone schemes

17.2.2 Constant-Energy Surfaces and the Fermi Surface

The energy of electron states can also be illustrated by specifying the regions of constant energy in $\boldsymbol{k}$-space. In two dimensions lines, while in three dimensions surfaces of constant energy are obtained. The constant-energy surfaces determined for a face-centered cubic lattice with a relatively weak potential are shown in Figure 17.9 for two different energies.

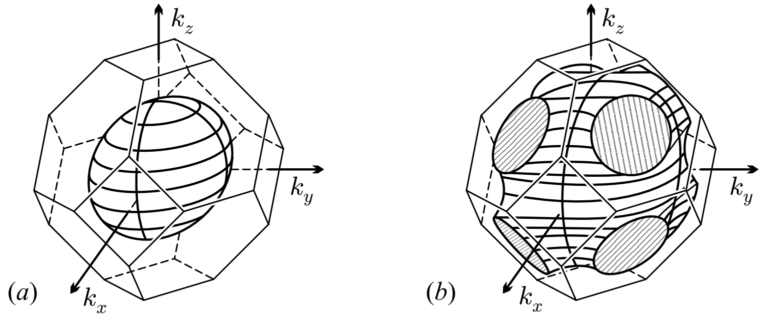


Fig. 17.9. Constant-energy surfaces in the band structure of a face-centered cubic crystal for two values of the energy

Among the constant-energy surfaces particularly important is the surface that contains those $\boldsymbol{k}$ vectors for which the energy is the same as the zero-temperature value of the chemical potential. This energy is called the Fermi energy for Bloch states, too. Since in the ground state all the states whose energy is lower than the chemical potential are occupied while higher-energy states are empty, this surface separates occupied and unoccupied states in

$\boldsymbol{k}$-space. This constant-energy surface is called the *Fermi surface.* As demonstrated for the Sommerfeld model, in the absence of a periodic potential this is just the surface of the Fermi sphere. However, the presence of a periodic potential can drastically distort the spherical shape of the Fermi surface – and, as we shall see, it can even disappear.

For a relatively small number of electrons only the states at the bottom of the lowest-lying band are occupied. The Fermi surface is then a simply connected continuous surface that deviates little from the spherical shape. When the number of electrons is increased, the surface may cease to be simply connected, as illustrated in Fig. 17.9(b). When the number of electrons exceeds twice the number of lattice points, electrons will certainly occupy the states of more than one band. In such cases more than one band can be partially filled. The Fermi surface separating occupied and unoccupied states must then be given for each of these – hence the Fermi surface is made up of several pieces. The reduced- and extended-zone schemes are both used to visualize them.

17.3 Metals, Insulators, Semiconductors

In the previous section it was assumed that there exist Bloch states whose energy is the same as the chemical potential. This is indeed the case when the Fermi energy (the chemical potential) lies inside a band (or several bands). Then the electrons that occupy states with slightly lower energies than the Fermi energy can be easily excited thermally or by an electric field into states with energies in excess of the Fermi energy. The behavior of such materials can be similar to what was presented for the free-electron model. Such materials are metals. As mentioned on page 75, their resistivity is typically on the order of $10\,\mathrm{n}\Omega\,\mathrm{m}$ at room temperature, decreases with decreasing temperature, and in an ideal crystal it would even vanish at $T = 0$. A material is customarily considered metallic if its conductivity exceeds $10^6\,(\Omega\,\mathrm{m})^{-1}$ (i.e., its resistivity $\varrho < 1\mu\Omega\,\mathrm{m}$).

However, the chemical potential may just as well be not inside a band but between two nonoverlapping bands, in which case there is no Bloch state whose energy is the same as the chemical potential. One cannot even speak of a Fermi surface then. This situation arises only when there are forbidden regions of finite width called *energy gaps* or *band gaps* such that the energies inside the gap do not appear as the energy of any band state. The most relevant band gap is the one that separates the bands that are filled completely in the ground state from those that are completely empty. By way of example, the band structure of diamond is shown in Fig. 17.10(a).

This can occur when the number of electrons per atom is even, and they completely fill one or more bands in the ground state, and the next, empty band is separated by a finite energy gap from the occupied ones. In such

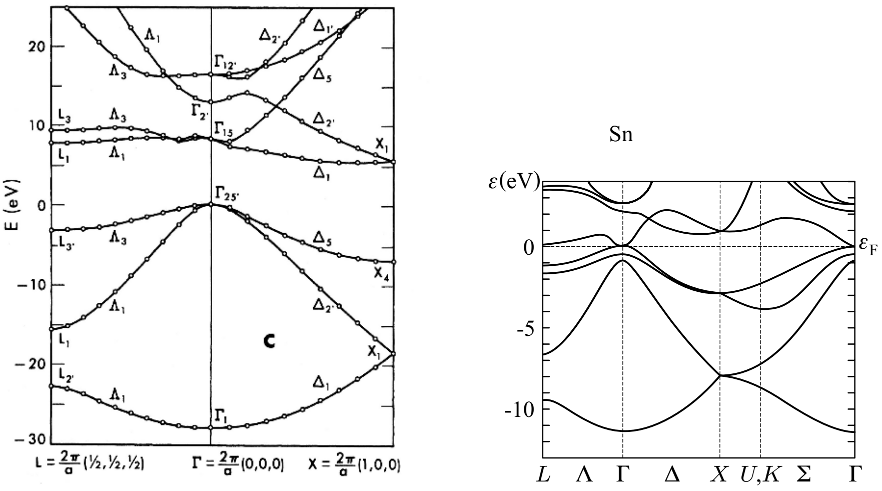


Fig. 17.10. Calculated band structures in high-symmetry directions of the Brillouin zone for (*a*) diamond [Reprinted with permission from W. Saslow, T. K. Bergstresser, and M. L. Cohen, *Phys. Rev. Lett.* **16**, 354 (1966). ©1966 by the American Physical Society] and (*b*) gray tin [Reprinted with permission from J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**, 556 (1976). ©1976 by the American Physical Society]

materials electric current starts to flow only above a certain (rather large) threshold voltage. They are called *insulators*. At $T = 0$ the resistivity of insulators is infinitely large.

This implies that if the band structure is known, a sharp line can be drawn between metals and insulators: the chemical potential in the ground state is inside a band in the former and inside a gap in the latter. However, when physical properties measured at finite temperatures are considered, continuous variations are found. At finite temperatures those states whose energy is larger than the chemical potential can be partially filled by thermally excited electrons, and lower-energy states can be partially empty. Therefore, strictly speaking, the conductivity does not vanish even when the energy gap exceeds the thermal energy $k_{\mathrm{B}}T$. A material is considered to be electrically insulating if its conductivity is less than $10^{-8}\,(\Omega\,\mathrm{m})^{-1}$. In general, those materials that are insulators at low temperatures remain insulators at room temperature, too. However, there is a class of materials whose conductivity vs. temperature plot has a jump at a certain temperature or changes several orders of magnitude in a narrow temperature range. This phenomenon is called the *metal–insulator transition*.

Among the materials whose band structure has a finite gap around the chemical potential particular behavior is observed in those for which the energy gap is larger than the thermal energy at room temperature but nonetheless sufficiently small, consequently electrons excited thermally across the gap carry an electric current that can be easily measured and has substantial effects. Such materials are called *semiconductors.*

Compared to the usual values in metals, smaller conductivities are obtained when the energy gap is extremely narrow – or even nonexistent – around the Fermi energy but the electronic density of states is very low there. Materials featuring such a band structure are called *semimetals.* Examples include elements of the nitrogen group (group 15 [VA] of the periodic table): arsenic, antimony, and bismuth, as well as graphite, a modification of carbon (group 14 [IVA]). As listed in Table 16.2, the conductivity of bismuth is close to the lower limit set for metals. Gray tin is the example for vanishing gap at the Fermi energy, as shown in Fig. 17.10(*b*): the two bands touch each other at the Fermi energy. Nevertheless, on account of the low density of states, thermally excited electrons do not reach the states directly above the Fermi level but those in a somewhat higher-lying local minimum with a high density of states, therefore, from an electric point of view, gray tin is a semiconductor.

In materials that crystallize in a structure with a monatomic basis each band contains the same number of $\boldsymbol{k}$ states as there are lattice points. Among them those materials that have an odd number of electrons per atom behave electrically as metals since the topmost band that contains electrons in the ground state cannot be completely filled on account of spin degeneracy. Indeed, the alkali metals in the first column of the periodic table are all good metallic conductors. The single electron on the outermost s-shell in the atomic state finds itself in a band that is only half filled. The noble metals of group 11 (IB) and the elements of the boron group (group 13 [IIIA]) have more complex band structures, nonetheless they are metals – with the sole exception of boron. Boron itself crystallizes in a rhombohedral structure with a polyatomic basis, the chemical potential is inside a 1.5 eV wide gap, therefore it is a semiconductor.

Insulators and semiconductors are expected to be found among the elements of even-numbered columns of the periodic table, to the left and right of transition metals. However, elements of group 2 (IIA), alkaline-earth metals, are all metals, since the overlap of bands leads to the formation of two partially filled bands. At the other extremity of the periodic table nonmetallic elements are found. The halogens in group 17 (VIIA) and the noble gases in group 18 (VIIIA) are not even crystalline at room temperature.[3] Semiconductors should be looked for in groups 14 (IVA) and 16 (VIA).

Instead of their conductivity, materials can also be characterized by the number of charge carriers. In metals the room-temperature concentration of conduction electrons usually exceeds $10^{22}/\mathrm{cm}^3$, while it is between 10^{17} and $10^{21}/\mathrm{cm}^3$ in semimetals. It is even lower in semiconductors, e.g., $10^{13}/\mathrm{cm}^3$ in germanium and $10^{10}/\mathrm{cm}^3$ in silicon at room temperature.

[3] In spite of their odd number of electrons, halogens are not metallic even at low temperatures as their centered monoclinic or orthorhombic lattices are decorated with polyatomic bases.

This classification of crystalline solids is based on the assumption that all interactions can be lumped into a periodic potential, and then the state of the electron system can be given as the superposition of one-particle states. However, this assumption is not always valid. Those insulators whose properties can be interpreted in this simple band picture are called band insulators or *Bloch–Wilson insulators.* Nevertheless under certain circumstances the interaction of electrons with lattice vibrations or with each other can also render the material insulator. We speak of *Peierls insulators* in the first case and *Mott insulators* in the second. Finally, disorder can also make the conductivity vanish. Such materials are called *Anderson insulators.* The study of their physical properties, which are essentially determined by the interactions, will be the subject of later chapters.

17.4 Bloch Electrons as Quasiparticles

Whether the thermodynamic properties are examined or the response to applied electromagnetic fields, real metals often show very similar qualitative behavior to a free electron gas. This is not surprising in view of the band structure constructed from one-particle states, however the question remains: how come electron–electron interactions can be lumped into a one-particle potential? This question will be revisited in Volume 3. Below we shall demonstrate that the system of electrons interacting with the periodic potential of the crystal can be described as a gas of fictitious noninteracting particles obeying the Fermi–Dirac statistics. These fermionic *quasiparticles* are called Bloch electrons. Quasiparticles are defined only inside solids, and their quantum mechanical state cannot be identified with any state of a single real electron, but they offer a simple physical interpretation of the behavior of the electron system. Using this picture we shall show that the thermodynamic properties of electrons moving in a periodic potential can be easily calculated using essentially a few parameters for simple metals.

17.4.1 Creation and Annihilation Operators of Bloch States

It is often more convenient to treat the electron system in the occupation-number formalism (second quantization). Following the prescriptions of Appendix H, we shall introduce the creation and annihilation operators of Bloch states, $c^{\dagger}_{n\boldsymbol{k}\sigma}$ and $c_{n\boldsymbol{k}\sigma}$, where σ is the spin quantum number. These operators obey the fermionic anticommutation relations,

$$\left[c_{n\boldsymbol{k}\sigma}, c^{\dagger}_{n'\boldsymbol{k}'\sigma'}\right]_{+} = \delta_{nn'}\delta_{\boldsymbol{k}\boldsymbol{k}'}\delta_{\sigma\sigma'} \,. \tag{17.4.1}$$

Using these operators the state of N_{e} electrons in which individual Bloch electrons occupy one-particle states associated with the wavefunctions $\psi_{n_1\boldsymbol{k}_1\sigma_1}$, $\psi_{n_2\boldsymbol{k}_2\sigma_2}, \dots$ (and which could also be written as a Slater determinant) can be expressed as

$$\Psi = c^{\dagger}_{n_1 \boldsymbol{k}_1 \sigma_1} c^{\dagger}_{n_2 \boldsymbol{k}_2 \sigma_2} \cdots c^{\dagger}_{n_{N_{\mathrm{e}}} \boldsymbol{k}_{N_{\mathrm{e}}} \sigma_{N_{\mathrm{e}}}} |0\rangle \,, \tag{17.4.2}$$

where $|0\rangle$ is the vacuum of Bloch electrons.

If the Hamiltonian of the system can be written as the sum of one-particle Hamiltonians over the individual particles – as was assumed in (17.1.2) – then, according to (H.2.9) and (H.2.10), it can be rewritten in the occupation-number representation as the following bilinear combination of creation and annihilation operators:

$$\mathcal{H} = \sum_{\substack{nn' \\ \boldsymbol{k}\boldsymbol{k}'\sigma\sigma'}} c^{\dagger}_{n\boldsymbol{k}\sigma} H(n, n', \boldsymbol{k}, \boldsymbol{k}', \sigma, \sigma') c_{n'\boldsymbol{k}'\sigma'} \,, \tag{17.4.3}$$

where $H(n, n', \boldsymbol{k}, \boldsymbol{k}', \sigma, \sigma')$ is the matrix element of the one-particle Hamiltonian $\mathcal{H}(\boldsymbol{r})$ between one-particle states:

$$H(n, n', \boldsymbol{k}, \boldsymbol{k}', \sigma, \sigma') = \int \psi^{*}_{n\boldsymbol{k}\sigma}(\boldsymbol{r}) \mathcal{H}(\boldsymbol{r}) \psi_{n'\boldsymbol{k}'\sigma'}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} \,. \tag{17.4.4}$$

Since according to (17.1.3) the orthonormalized Bloch functions $\psi_{n\boldsymbol{k}\sigma}(\boldsymbol{r})$ are the eigenfunctions of the one-particle Hamiltonian with energy eigenvalues $\varepsilon_{n\boldsymbol{k}}$,

$$\boxed{\mathcal{H} = \sum_{n,\boldsymbol{k},\sigma} \varepsilon_{n\boldsymbol{k}} c^{\dagger}_{n\boldsymbol{k}\sigma} c_{n\boldsymbol{k}\sigma} \,.} \tag{17.4.5}$$

The Hamiltonian of the electron system is that of a noninteracting gas of fermions (Bloch electrons) of energy $\varepsilon_{n\boldsymbol{k}}$. Compared to the gas of free electrons, the energy $\hbar^2 k^2 / 2m_{\mathrm{e}}$ containing the electron mass is replaced by $\varepsilon_{n\boldsymbol{k}}$ obtained for Bloch electrons. In the ground state Bloch electrons fill the Fermi sea completely – that is, all states whose energy is smaller than the Fermi energy are occupied. At finite temperatures states above the Fermi energy may also be occupied. According to the Fermi–Dirac distribution function, the occupation probability – i.e., the mean number of electrons with wave vector $\boldsymbol{k}$ and spin σ in the nth band – is given by

$$\langle n_{n\boldsymbol{k}\sigma} \rangle = \langle c^{\dagger}_{n\boldsymbol{k}\sigma} c_{n\boldsymbol{k}\sigma} \rangle = \frac{1}{\mathrm{e}^{(\varepsilon_{n\boldsymbol{k}} - \mu)/k_{\mathrm{B}}T} + 1} \,. \tag{17.4.6}$$

The variations of the momentum distribution function at finite temperature with respect to the ground-state distribution is interpreted as being due to the thermal excitation of fermionic quasiparticles.

17.4.2 Effective Mass of Bloch Electrons

At the bottom of the bands, close to the minimum, the dispersion relation of Bloch electrons can be approximated by a quadratic expression of the wave vectors. In a cubic crystal, when the minimum is at a high-symmetry point $\boldsymbol{k}_0$, the energy can be approximated by

$$\varepsilon_{\boldsymbol{k}} \approx \varepsilon_{\boldsymbol{k}_0} + A(\boldsymbol{k} - \boldsymbol{k}_0)^2 \,. \tag{17.4.7}$$

By writing the coefficient A as

$$A = \frac{\hbar^2}{2m^*} \,, \tag{17.4.8}$$

the dispersion relation is similar to that of free electrons but the electron mass m_e is replaced by another value, m^*, which depends on the periodic potential and can also be defined as

$$\boxed{\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}^2}} \,. \tag{17.4.9}$$

By analogy, this parameter is called the *effective mass* of the Bloch electron.

Naturally, the effective mass determined from (17.4.9) is different from the free electron mass, moreover, it takes different values in each band – and if there are several minima in the band, it also depends on the specific point of the Brillouin zone at which the second derivative is taken. The notion of effective mass is particularly useful when the electrons that determine the thermal properties of the crystal are in that region of a band where the quadratic approximation of the energy is justified. A barely filled metallic band or the conduction and valence bands in semiconductors are good examples. In these cases the system of Bloch electrons behaves just like a system of free electrons, except that the electron mass is replaced by an effective mass.

In noncubic crystals the dispersion relation is not spherically symmetric. If close to the minimum the local symmetry of the Brillouin zone is orthorhombic, the series expansion of the energy is

$$\varepsilon_{\boldsymbol{k}} \approx \varepsilon_{\boldsymbol{k}_0} + A_x(k_x - k_{0x})^2 + A_y(k_y - k_{0y})^2 + A_z(k_z - k_{0z})^2 \,. \tag{17.4.10}$$

Rewriting this expression as

$$\varepsilon_{\boldsymbol{k}} \approx \varepsilon_{\boldsymbol{k}_0} + \frac{\hbar^2(k_x - k_{0x})^2}{2m_x^*} + \frac{\hbar^2(k_y - k_{0y})^2}{2m_y^*} + \frac{\hbar^2(k_z - k_{0z})^2}{2m_z^*} \,, \tag{17.4.11}$$

the behavior of the electrons is characterized by the triplet m_x^*, m_y^*, m_z^*. Even more generally, when it is sufficient to keep quadratic terms in the series expansion around the minimum, an inverse effective-mass tensor can be introduced instead of the scalar effective mass using the definition

$$\boxed{\left(\frac{1}{\mathsf{M}^*}\right)_{\alpha\beta} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\boldsymbol{k}}}{\partial k_\alpha \partial k_\beta} \,, \qquad \alpha, \beta = x, y, z \,.} \tag{17.4.12}$$

The energy of the band states then reads

$$\varepsilon_{\boldsymbol{k}} \approx \varepsilon_{\boldsymbol{k}_0} + \frac{\hbar^2}{2}\sum_{\alpha\beta}\left(\frac{1}{\mathsf{M}^*}\right)_{\alpha\beta}(k_\alpha - k_{0\alpha})(k_\beta - k_{0\beta}) \,. \tag{17.4.13}$$

The effective-mass tensor M^* is the inverse of the above-defined $1/\mathsf{M}^*$. It takes a particularly simple form when the inverse mass tensor is diagonal: if

$$\frac{1}{\mathsf{M}^*} = \begin{pmatrix} \dfrac{1}{m_1^*} & 0 & 0 \\ 0 & \dfrac{1}{m_2^*} & 0 \\ 0 & 0 & \dfrac{1}{m_3^*} \end{pmatrix} \tag{17.4.14}$$

then the effective-mass tensor is also diagonal and

$$\mathsf{M}^* = \begin{pmatrix} m_1^* & 0 & 0 \\ 0 & m_2^* & 0 \\ 0 & 0 & m_3^* \end{pmatrix}. \tag{17.4.15}$$

17.4.3 Bloch Electrons and Holes

The description in terms of the effective mass is usually satisfactory at low band filling. We shall often encounter the opposite case where the band is almost completely filled and only a small number of states close to the top of the band remain unoccupied. This situation can be regarded as if some electrons in the vicinity of the maximum were removed from the completely filled band – that is, holes were generated.

When a band is completely filled with electrons the sum of the wave vectors $\boldsymbol{k}$ of occupied states (i.e., the sum over the entire Brillouin zone) is zero:

$$\sum_{\boldsymbol{k}\in\mathrm{BZ}} \boldsymbol{k} = 0\,. \tag{17.4.16}$$

Now let us remove an electron of wave vector $\boldsymbol{k}_{\mathrm{e}}$ from this completely filled band. The wave vector of this state,

$$\boldsymbol{k}_{\mathrm{h}} = \sum_{\boldsymbol{k}\in\mathrm{BZ}} \boldsymbol{k} - \boldsymbol{k}_{\mathrm{e}} = -\boldsymbol{k}_{\mathrm{e}}\,, \tag{17.4.17}$$

is just the negative of the electron's wave vector. When a hole is created, the change in the energy of the system is the negative of the energy of the removed particle, therefore the energy $\varepsilon_{\mathrm{h}}(\boldsymbol{k}_{\mathrm{h}})$ of the hole satisfies

$$\varepsilon_{\mathrm{h}}(\boldsymbol{k}_{\mathrm{h}}) = -\varepsilon(\boldsymbol{k}_{\mathrm{e}})\,. \tag{17.4.18}$$

Provided that close to the top of the band the dispersion relation is isotropic, the leading correction of the expansion gives

$$\varepsilon_{\boldsymbol{k}} \approx \varepsilon_{\boldsymbol{k}_0} - A(\boldsymbol{k}-\boldsymbol{k}_0)^2\,, \qquad A > 0\,. \tag{17.4.19}$$

According to the defining equation (17.4.9), the effective mass is negative for electrons with energies close to the maximum:

$$\frac{1}{m^*} = -\frac{2A}{\hbar^2}\,. \tag{17.4.20}$$

However, when hole energies are considered, the sign of the quadratic term is reversed, and the effective mass of the hole defined by

$$\frac{1}{m_{\mathrm{h}}^*} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\mathrm{h}}(\boldsymbol{k}_{\mathrm{h}})}{\partial \boldsymbol{k}_{\mathrm{h}}^2} \tag{17.4.21}$$

is positive:

$$m_{\mathrm{h}}^* = -m^*\,. \tag{17.4.22}$$

For general dispersion curves the inverse effective-mass tensor

$$\left(\frac{1}{M_{\mathrm{h}}^*}\right)_{\alpha\beta} = -\frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\boldsymbol{k}}}{\partial k_\alpha \partial k_\beta} \tag{17.4.23}$$

has to be used once again. If a band is almost completely filled, treating the small number of unoccupied electron states as positive-effective-mass holes proves to be convenient especially in the description of the dynamics of electrons and the discussion of semiconductors.

17.4.4 Density of States for Bloch Electrons

Just like in the free-electron case, we shall often need the sum of some quantity $g(\boldsymbol{k})$ – for example the energy – over occupied $\boldsymbol{k}$-space states in order to determine certain properties of the electron system. For large samples containing a great number of primitive cells the allowed values of $\boldsymbol{k}$ fill the Brillouin zone densely, therefore the sum can be replaced by an integral. Considering the contribution of electrons in each band individually,

$$\boxed{\sum_{\boldsymbol{k}} g(\boldsymbol{k}) f_0(\varepsilon_{n\boldsymbol{k}}) \to \frac{V}{(2\pi)^3}\int g(\boldsymbol{k}) f_0(\varepsilon_{n\boldsymbol{k}})\,\mathrm{d}\boldsymbol{k}\,.} \tag{17.4.24}$$

In a lot of cases quantities that depend on the energy alone need to be summed, therefore we can once again introduce the density of states by stipulating that $\rho_n(\varepsilon)\,\mathrm{d}\varepsilon$ is the number of states in the nth band with energies between ε and $\varepsilon + \mathrm{d}\varepsilon$. The $\boldsymbol{k}$-sum or $\boldsymbol{k}$-integral can then be rewritten as an energy integral of the density of states:

$$\sum_{\boldsymbol{k}\sigma} g(\varepsilon_{n\boldsymbol{k}}) f_0(\varepsilon_{n\boldsymbol{k}}) = V\int g(\varepsilon) f_0(\varepsilon)\rho_n(\varepsilon)\,\mathrm{d}\varepsilon\,. \tag{17.4.25}$$

Following the steps in Chapter 12, where the phonon density of states was derived, we get

$$\boxed{\rho_n(\varepsilon) = \frac{2}{(2\pi)^3} \int\limits_{S(\varepsilon)} \frac{\mathrm{d}S}{|\boldsymbol{\nabla}_{\boldsymbol{k}} \varepsilon_{n\boldsymbol{k}}|},} \tag{17.4.26}$$

where the integral has to be evaluated for the constant-energy surface $S(\varepsilon)$. The factor two comes from the spin quantum number. The total electronic density of states is a sum over bands:

$$\rho(\varepsilon) = \sum_n \rho_n(\varepsilon) . \tag{17.4.27}$$

If the energy of the electrons depends on spin, so does the density of states:

$$\rho_{n\sigma}(\varepsilon) = \frac{1}{(2\pi)^3} \int\limits_{S(\varepsilon)} \frac{\mathrm{d}S}{|\boldsymbol{\nabla}_{\boldsymbol{k}} \varepsilon_{n\boldsymbol{k}\sigma}|} , \tag{17.4.28}$$

and the total density of states is then given by

$$\rho(\varepsilon) = \sum_{n\sigma} \rho_{n\sigma}(\varepsilon) . \tag{17.4.29}$$

Just like for phonons, the Van Hove singularities discussed in Chapter 12 also appear in the electronic density of states. There is, nevertheless, an essential difference: for electrons the dispersion relation is quadratic at the bottom of the band, therefore a square-root singularity (P_0-type critical point) appears at the bottom of each band. It is quite natural that each band has a minimum and a maximum, i.e., the density of states features a P_0- and a P_3-type point. It can also be shown that each band has at least three P_1- and three P_2-type critical points as well. As we shall see in the tight-binding approximation, the simple cubic crystal provides an example for the minimum number of critical points if the dispersion relation is approximated by

$$\varepsilon_{\boldsymbol{k}} = \alpha \left[3 - \cos k_x a - \cos k_y a - \cos k_z a \right] . \tag{17.4.30}$$

The minimum is at the center Γ of the Brillouin zone, and the maximum is at the corner points R. P_1-type saddle points are found at the face centers X, since within the face the dispersion relation has its minimum at the center, while in the perpendicular direction it has its maximum there. On the other hand, P_2-type saddle points are found at the edge centers, since the dispersion relation along the edge has its minimum there, while it has its maximum in the same point along the lines joining the centers of the two adjacent faces.

If in the general case the dispersion relation transformed to the principal axes can be approximated by

$$\varepsilon_{\boldsymbol{k}} = \varepsilon_{\min} + \frac{\hbar^2 k_1^2}{2m_1^*} + \frac{\hbar^2 k_2^2}{2m_2^*} + \frac{\hbar^2 k_3^2}{2m_3^*} \tag{17.4.31}$$

in the vicinity of the minimum, then the density of states is

$$\rho(\varepsilon) = \frac{\sqrt{2}}{\pi^2\hbar^3}\left(m_1^* m_2^* m_3^*\right)^{1/2}\sqrt{\varepsilon - \varepsilon_{\text{min}}}\,. \tag{17.4.32}$$

If the dispersion relation can be written as

$$\varepsilon_{\boldsymbol{k}} = \varepsilon_{\text{max}} - \frac{\hbar^2 k_1^2}{2m_1^*} - \frac{\hbar^2 k_2^2}{2m_2^*} - \frac{\hbar^2 k_3^2}{2m_3^*} \tag{17.4.33}$$

in the vicinity of the maximum, then, by the same token, the density of states reads

$$\rho(\varepsilon) = \frac{\sqrt{2}}{\pi^2\hbar^3}\left(m_1^* m_2^* m_3^*\right)^{1/2}\sqrt{\varepsilon_{\text{max}} - \varepsilon}\,. \tag{17.4.34}$$

Around a P_1-type saddle point, where the dispersion relation is

$$\varepsilon_{\boldsymbol{k}} = \varepsilon_{\text{c}} + \frac{\hbar^2 k_1^2}{2m_1^*} + \frac{\hbar^2 k_2^2}{2m_2^*} - \frac{\hbar^2 k_3^2}{2m_3^*}\,, \tag{17.4.35}$$

the density of states has a square-root-type energy dependence in the region $\varepsilon < \varepsilon_{\text{c}}$:

$$\rho(\varepsilon) = C - \frac{\sqrt{2}}{\pi^2\hbar^3}\left(m_1^* m_2^* m_3^*\right)^{1/2}\sqrt{\varepsilon_{\text{c}} - \varepsilon}\,. \tag{17.4.36}$$

For P_2-type saddle points a similar formula is obtained for energies above ε_{c}:

$$\rho(\varepsilon) = C - \frac{\sqrt{2}}{\pi^2\hbar^3}\left(m_1^* m_2^* m_3^*\right)^{1/2}\sqrt{\varepsilon - \varepsilon_{\text{c}}}\,. \tag{17.4.37}$$

These formulas are straightforward to derive from the expressions for the Van Hove singularities for phonons established in Chapter 12: one has to substitute $\alpha_1 = \hbar^2/2m_1^*$, $\alpha_2 = \hbar^2/2m_2^*$, $\alpha_3 = \hbar^2/2m_3^*$, and include an extra factor of two arising from the electron spin.

Just like for phonons, the density of states features an inverse-square-root singularity at the bottom and top of the band in one-dimensional electron systems. In two-dimensional systems a logarithmic singularity appears at the energy associated with the saddle point:

$$\rho(\varepsilon) = C - \frac{1}{\pi^2\hbar^2}\left(m_1^* m_2^*\right)^{1/2}\ln\left|1 - \frac{\varepsilon}{\varepsilon_{\text{c}}}\right| + \mathcal{O}(\varepsilon - \varepsilon_{\text{c}})\,, \tag{17.4.38}$$

but at the energy of the bottom of the band the density of states jumps from zero to a finite value, and at the energy of the top of the band it drops from a finite value to zero.

17.4.5 Specific Heat and Susceptibility of Bloch Electrons

Bloch electrons also obey the Fermi–Dirac statistics at finite temperatures. By repeating the steps of the procedure for free electrons, the thermal energy and specific heat of electrons moving in a periodic potential can be determined.

Once again, we start with (16.2.88), however $\varepsilon_{\boldsymbol{k}}$ is the energy of Bloch electrons and $\rho(\varepsilon)$ the density of states of Bloch electrons this time. If the variation of the density of states is so slow in a region of width $k_{\mathrm{B}}T$ around the Fermi energy that it is justified to keep only the leading order term in the Sommerfeld expansion, then the specific heat takes the same form as (16.2.91):

$$\boxed{c_{\mathrm{el}} = \frac{\pi^2}{3} k_{\mathrm{B}}^2 T \rho(\varepsilon_{\mathrm{F}})\,,} \tag{17.4.39}$$

where $\rho(\varepsilon_{\mathrm{F}})$ is the Bloch electron density of states at the Fermi energy. This can be understood relatively simply. Because of the Fermi–Dirac statistics, the thermal properties are determined by electrons whose energies are within a few times $k_{\mathrm{B}}T$ of the Fermi energy. As $k_{\mathrm{B}}T \ll \varepsilon_{\mathrm{F}}$ at the usual temperatures, the specific heat is independent of the density of states far from the Fermi energy.

When the dispersion relation of Bloch electrons can be approximated by a quadratic function and the effective mass is a scalar, the coefficient of the linear term in the specific heat can be expressed by the effective mass just like in (16.2.92):

$$\gamma = \frac{k_{\mathrm{B}}^2 m^*}{3\hbar^2} \left(3\pi^2 n_{\mathrm{e}}\right)^{1/3}. \tag{17.4.40}$$

In noncubic systems, where the dispersion relation must be given in terms of an effective-mass tensor rather than a single scalar parameter, the density of states has to be determined using (17.4.26) as an integral over the Fermi surface. It can be shown that the density of states can be recast in the same form as for free electrons but the electron mass is replaced by

$$m_{\mathrm{ds}}^* = \left[\det(m_{ij}^*)\right]^{1/3}, \tag{17.4.41}$$

which corresponds to the combination

$$m_{\mathrm{ds}}^* = \left(m_x^* m_y^* m_z^*\right)^{1/3} \tag{17.4.42}$$

in the orthorhombic case. This quantity is called the *density-of-states mass* in order to distinguish it from other combinations encountered in other physical quantities. Since this effective mass appears in the specific heat, the term *thermal mass* is also used.

This explains in part why the Sommerfeld coefficient (the proportionality factor of the temperature in the specific heat) of metals differs from the free-electron value. Nevertheless, the quantitative understanding of the so-called heavy-fermionic behavior, the extremely large effective mass, requires a more precise account of electron–electron interactions.

To determine the Pauli susceptibility arising from the spins of Bloch electrons, the calculation performed in the free-electron model is repeated. If the density of states varies little over a region of width $\mu_{\mathrm{B}}B$ around the Fermi energy, a similar formula is found:

$$\boxed{\chi_{\mathrm{P}} = \tfrac{1}{4}\mu_0 (g_{\mathrm{e}}\mu_{\mathrm{B}})^2 \rho(\varepsilon_{\mathrm{F}})\,,} \tag{17.4.43}$$

with the single difference that the density of states at the Fermi energy is now that of the Bloch electrons. Therefore the periodic potential modifies the Pauli susceptibility, too, through the effective mass of Bloch electrons. As the same effective mass appears in the specific heat and the susceptibility, the Wilson ratio for noninteracting Bloch electrons is the same as for free electrons.

17.5 Wannier States

The one-particle wavefunction of electrons was written in the Bloch form in the previous sections. However, this is not the only possibility: in certain cases another representation is more practical.

17.5.1 Wannier Functions

We shall make use of the previously established relationship

$$\psi_{n,\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}} u_{n,\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{n\boldsymbol{k}}(\boldsymbol{r}) = \psi_{n\boldsymbol{k}}(\boldsymbol{r})\,, \tag{17.5.1}$$

which states that when $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is considered as a function of $\boldsymbol{k}$ at a fixed $\boldsymbol{r}$, it is periodic in the reciprocal lattice. It can therefore be expanded into a Fourier series; the vectors that appear in this representation are the translation vectors of the reciprocal of the reciprocal lattice – that is, the lattice vectors of the original direct lattice. A convenient choice of normalization is

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{R}_j} \phi_n(\boldsymbol{r}, \boldsymbol{R}_j) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\,. \tag{17.5.2}$$

Making use of (C.1.47) leads to

$$\phi_n(\boldsymbol{r}, \boldsymbol{R}_j) = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{k}\in\mathrm{BZ}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} \psi_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{17.5.3}$$

for the Fourier coefficients. In what follows, even when it is not explicitly indicated, summation over $\boldsymbol{k}$ refers to summation over the wave vectors in the Brillouin zone.

We shall first demonstrate that the above-defined $\phi_n(\boldsymbol{r}, \boldsymbol{R}_j)$ is indeed a function of $\boldsymbol{r} - \boldsymbol{R}_j$ alone. To this end we shall translate both $\boldsymbol{R}_j$ and $\boldsymbol{r}$ by $\boldsymbol{t}_m$, and make use of the translational properties of $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$:

$$\begin{aligned}\phi_n(\boldsymbol{r}+\boldsymbol{t}_m,\boldsymbol{R}_j+\boldsymbol{t}_m) &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_j+\boldsymbol{t}_m)}\psi_{n\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{t}_m)\\ &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_j+\boldsymbol{t}_m)}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{t}_m}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\\ &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\\ &= \phi_n(\boldsymbol{r},\boldsymbol{R}_j)\,. \end{aligned}\tag{17.5.4}$$

By choosing $\boldsymbol{t}_m = -\boldsymbol{R}_j$,

$$\phi_n(\boldsymbol{r},\boldsymbol{R}_j) = \phi_n(\boldsymbol{r}-\boldsymbol{R}_j,0)\,,\tag{17.5.5}$$

which shows that the genuine variable is the difference of $\boldsymbol{r}$ and $\boldsymbol{R}_j$. The functions $\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)$ are called *Wannier functions.*[4] According to (17.5.3), they can be related to the Bloch function by

$$\begin{aligned}\phi_n(\boldsymbol{r}-\boldsymbol{R}_j) &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\\ &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{R}_j)}u_{n\boldsymbol{k}}(\boldsymbol{r})\,,\end{aligned}\tag{17.5.6}$$

while the inverse relationship is

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_j} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\,,\tag{17.5.7-a}$$

$$u_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_j} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{R}_j)}\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\,.\tag{17.5.7-b}$$

It is readily seen that the functions $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ expressed in terms of the Wannier functions satisfy the conditions (6.2.5) and (17.1.6) imposed on the Bloch functions. Taking the function at the position translated through $\boldsymbol{t}_m$,

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{t}_m) = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_j} \phi_n(\boldsymbol{r}+\boldsymbol{t}_m-\boldsymbol{R}_j)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\,.\tag{17.5.8}$$

Indexing the sum by the translated lattice point $\boldsymbol{R}_l = \boldsymbol{R}_j - \boldsymbol{t}_m$ instead of the original $\boldsymbol{R}_j$, the sum remains unaltered on account of the periodic boundary conditions, thus

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{t}_m) = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_l} \phi_n(\boldsymbol{r}-\boldsymbol{R}_l)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_l+\boldsymbol{t}_m)} = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{t}_m}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,.\tag{17.5.9}$$

[4] G. H. Wannier, 1937.

Whether the Bloch functions or the Wannier representation is used, we have a complete orthogonal set of wavefunctions, since the orthogonality of the Bloch functions implies the orthogonality of the Wannier functions associated with different bands and lattice points:

$$\begin{aligned}\int \phi_n^*(\boldsymbol{r}-\boldsymbol{R}_j)\phi_{n'}(\boldsymbol{r}-\boldsymbol{R}_{j'})\,\mathrm{d}\boldsymbol{r} &= \frac{1}{N}\sum_{\boldsymbol{k},\boldsymbol{k}'}\int \mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{R}_j-\boldsymbol{k}'\cdot\boldsymbol{R}_{j'})}\psi_{n\boldsymbol{k}}^*(\boldsymbol{r})\psi_{n'\boldsymbol{k}'}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\\ &= \frac{1}{N}\sum_{\boldsymbol{k}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_j-\boldsymbol{R}_{j'})}\delta_{n,n'}\\ &= \delta_{\boldsymbol{R}_j,\boldsymbol{R}_{j'}}\delta_{n,n'}. \end{aligned} \qquad (17.5.10)$$

The completeness of the Wannier functions can be demonstrated along the same lines:

$$\sum_{n,\boldsymbol{R}_j}\phi_n^*(\boldsymbol{r}-\boldsymbol{R}_j)\phi_n(\boldsymbol{r}'-\boldsymbol{R}_j)=\delta(\boldsymbol{r}-\boldsymbol{r}')\,. \qquad (17.5.11)$$

Since the phase of Wannier functions can be chosen arbitrarily, it is possible to construct Wannier functions $\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)$ that take large values only around the lattice point $\boldsymbol{R}_j$, and drop off exponentially with distance in other cells. To demonstrate this, consider a simple example. For free electrons the function $u_{n\boldsymbol{k}}(\boldsymbol{r})$ appearing in the Bloch function is independent of $\boldsymbol{k}$: $\mathrm{e}^{\mathrm{i}\boldsymbol{G}_n\cdot\boldsymbol{r}}$. Assuming a slightly more general but still $\boldsymbol{k}$-independent function $u_n(\boldsymbol{r})$, the Wannier functions that correspond to the Bloch functions

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r})=\frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_n(\boldsymbol{r}) \qquad (17.5.12)$$

can be determined exactly:

$$\begin{aligned}\phi_n(\boldsymbol{r}-\boldsymbol{R}_j) &= \frac{1}{\sqrt{NV}}\sum_{\boldsymbol{k}}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_n(\boldsymbol{r}) \qquad (17.5.13)\\ &= \frac{1}{N\sqrt{v}}\sum_{\boldsymbol{k}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{R}_j)}u_n(\boldsymbol{r}) = \varPhi(\boldsymbol{r}-\boldsymbol{R}_j)u_n(\boldsymbol{r})\,,\end{aligned}$$

where $\varPhi$ depends on the lattice constants a, b, c as

$$\varPhi(\boldsymbol{r})=\frac{1}{\sqrt{v}}\frac{\sin\pi x/a}{\pi x/a}\,\frac{\sin\pi y/b}{\pi y/b}\,\frac{\sin\pi z/c}{\pi z/c}\,. \qquad (17.5.14)$$

It is readily seen that the Wannier function drops off with increasing distance, while oscillating sinusoidally with a period of twice the lattice constant. The Wannier functions are usually well localized in space for more realistic functions $u_{n\boldsymbol{k}}(\boldsymbol{r})$ as well.

The Wannier states are not eigenstates of the one-particle Hamiltonian; $\mathcal{H}$ couples Wannier functions associated with different lattice points but with the same band. Using (17.1.3),

$$\begin{aligned}\mathcal{H}\,\phi_n(\boldsymbol{r}-\boldsymbol{R}_j) &= \mathcal{H}\,\frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) \\ &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\varepsilon_{n\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) \qquad (17.5.15)\\ &= \frac{1}{N}\sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\varepsilon_{n\boldsymbol{k}}\sum_{\boldsymbol{R}_l}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l}\phi_n(\boldsymbol{r}-\boldsymbol{R}_l)\,.\end{aligned}$$

This can be rewritten as

$$\mathcal{H}\,\phi_n(\boldsymbol{r}-\boldsymbol{R}_j) = \sum_{\boldsymbol{R}_l} t_{n,jl}\phi_n(\boldsymbol{r}-\boldsymbol{R}_l) \qquad (17.5.16)$$

with

$$t_{n,jl} = \frac{1}{N}\sum_{\boldsymbol{k}} \varepsilon_{n\boldsymbol{k}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_l-\boldsymbol{R}_j)}\,. \qquad (17.5.17)$$

The orthonormality of the Wannier functions implies

$$t_{n,jl} = \int \phi_n^*(\boldsymbol{r}-\boldsymbol{R}_l)\mathcal{H}(\boldsymbol{r})\,\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\,\mathrm{d}\boldsymbol{r}\,. \qquad (17.5.18)$$

Since the one-particle Hamiltonian is local, this coefficient vanishes unless the Wannier functions of the two lattice points overlap. Consequently it is often sufficient to consider only nearest neighbors and choose the Wannier function $\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)$ as the wavefunction of a core electron of the atom at $\boldsymbol{R}_j$, even though the orthogonality is lost with this choice.

17.5.2 Creation and Annihilation Operators of Wannier States

The creation and annihilation operators ($c_{nj\sigma}^\dagger$ and $c_{nj\sigma}$) can be introduced for Wannier states as well; in this case they change the occupation of the Wannier state at $\boldsymbol{R}_j$ in the nth band by adding or removing an electron. In terms of these operators the Hamiltonian reads

$$\boxed{\mathcal{H} = \sum_{n,\sigma}\sum_{j,l} t_{n,jl}c_{nl\sigma}^\dagger c_{nj\sigma}\,.} \qquad (17.5.19)$$

In this representation the Hamiltonian is nondiagonal. It describes the hopping of an electron in the Wannier state centered on the jth atom to a state centered on the lth atom. The probability of transition is the absolute square of the *hopping matrix element* $t_{n,jl}$.

Using the creation and annihilation operators of Bloch electrons means working in reciprocal space, while using those of the Wannier states means working in real space. The two representations are related by the discrete Fourier transforms

$$c^\dagger_{nj\sigma} = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{k}} c^\dagger_{n\boldsymbol{k}\sigma} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} , \quad c_{nj\sigma} = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{k}} c_{n\boldsymbol{k}\sigma} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} \tag{17.5.20}$$

and

$$c^\dagger_{n\boldsymbol{k}\sigma} = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{R}_j} c^\dagger_{nj\sigma} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} , \quad c_{n\boldsymbol{k}\sigma} = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{R}_j} c_{nj\sigma} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} , \tag{17.5.21}$$

in analogy with the mutual relations between Bloch and Wannier functions. These expressions can also be viewed as unitary transformations that diagonalize the Hamiltonian (17.5.19). This transformation also establishes the relationship between the energy of Bloch electrons and the hopping matrix elements (17.5.17):

$$\varepsilon_{n\boldsymbol{k}} = \sum_{\boldsymbol{R}_l} t_{n,jl} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_l - \boldsymbol{R}_j)} . \tag{17.5.22}$$

17.6 Electron States Around Impurities

Before turning to the subject matter of the next chapters, the technical details of computing the band structure in an ideal crystal, it should be noted that only in ideal crystals do the Bloch and Wannier states exist in the form presented in the foregoing. In real materials impurities and defects are always present, and the electronic spectrum is therefore modified with respect to the ideal case. Adapting the procedure used in the study of localized lattice vibrations, it can be demonstrated that the energies of the N electron states making up the quasicontinuous band are modified in such a way that $N-1$ states remain inside the original band but one state can move outside. The difference with localized lattice vibrations is that the bound state can appear below or above the band (depending on whether the potential is attractive or repulsive). The spectrum of free electrons showed a similar pattern in the vicinity of an impurity, however, the free-electron spectrum being unbounded from above, bound states appeared only for attractive potentials, at negative energies. Below, we shall generalize the free-electron results to electrons moving in a periodic potential.

Suppose that an impurity atom at lattice site $\boldsymbol{R}_0$ gives rise to an additional short-range potential $V(\boldsymbol{r}-\boldsymbol{R}_0)$ relative to the periodic potential $U(\boldsymbol{r})$ of the ideal crystal. Since the impurity breaks discrete translational symmetry, the states can no longer be characterized by a wave vector $\boldsymbol{k}$. Using an index μ as quantum number, the Schrödinger equation that determines the electron states reads

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}} \boldsymbol{\nabla}^2 + U(\boldsymbol{r}) + V(\boldsymbol{r}-\boldsymbol{R}_0) \right] \psi_\mu(\boldsymbol{r}) = \varepsilon_\mu \psi_\mu(\boldsymbol{r}) . \tag{17.6.1}$$

Suppose, furthermore, that the electron states of the ideal crystal – that is, the eigenfunctions and eigenvalues of the Schrödinger equation

$$\mathcal{H}_0\psi_{n\boldsymbol{k}}(\boldsymbol{r}) \equiv \left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r})\right]\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_{n\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{17.6.2}$$

are known.

Equation (17.6.1) can be solved much in the same manner as (16.4.2), which describes the interaction of a free electron with an impurity. We find

$$\psi_\mu(\boldsymbol{r}) = \psi_{n\boldsymbol{k}}(\boldsymbol{r}) + \int G_{\varepsilon_\mu}(\boldsymbol{r},\boldsymbol{r}')V(\boldsymbol{r}')\psi_\mu(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'\,, \tag{17.6.3}$$

where now the Green function $G_\varepsilon(\boldsymbol{r},\boldsymbol{r}')$ is the solution of

$$(\varepsilon-\mathcal{H}_0)G_\varepsilon(\boldsymbol{r},\boldsymbol{r}') \equiv \left[\varepsilon + \frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 - U(\boldsymbol{r})\right]G_\varepsilon(\boldsymbol{r},\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}')\,. \tag{17.6.4}$$

Expanding the Green function in terms of Bloch states as

$$G_\varepsilon(\boldsymbol{r}-\boldsymbol{r}') = \sum_{n\boldsymbol{k}} a_{n\boldsymbol{k}}(\boldsymbol{r}')\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,, \tag{17.6.5}$$

and substituting this form into the equation for the Green function, the eigenvalue equation

$$\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r}) - \varepsilon_{n\boldsymbol{k}}\right]\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = 0 \tag{17.6.6}$$

of the Bloch states and the orthogonality of Bloch functions imply that

$$a_{n\boldsymbol{k}}(\boldsymbol{r}') = \frac{\psi^*_{n\boldsymbol{k}}(\boldsymbol{r}')}{\varepsilon-\varepsilon_{n\boldsymbol{k}}}\,, \tag{17.6.7}$$

or

$$G_\varepsilon(\boldsymbol{r}-\boldsymbol{r}') = \sum_{n\boldsymbol{k}} \frac{\psi^*_{n\boldsymbol{k}}(\boldsymbol{r}')\psi_{n\boldsymbol{k}}(\boldsymbol{r})}{\varepsilon-\varepsilon_{n\boldsymbol{k}}}\,. \tag{17.6.8}$$

The wavefunction can then be written as

$$\psi_\mu(\boldsymbol{r}) = \psi_{n\boldsymbol{k}}(\boldsymbol{r}) + \sum_{n'\boldsymbol{k}'}\int \frac{\psi_{n'\boldsymbol{k}'}(\boldsymbol{r})\psi^*_{n'\boldsymbol{k}'}(\boldsymbol{r}')}{\varepsilon-\varepsilon_{n'\boldsymbol{k}'}+\mathrm{i}\alpha}V(\boldsymbol{r}')\psi_\mu(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'\,, \tag{17.6.9}$$

where α is an infinitesimally small positive quantity that ensures the required analytic properties.

If the influence of the potential is limited to the vicinity of the impurity, as has been assumed, then it is more convenient to use Wannier functions than Bloch functions to calculate the matrix element. If the impurity is located at $\boldsymbol{R}_0$ then only the Wannier functions associated with the same lattice point have nonvanishing matrix elements, and among them the matrix element of the transition into the same band is the most important. Keeping only this one, we have

$$\int \phi_n^*(\boldsymbol{r}-\boldsymbol{R}_j)V(\boldsymbol{r}-\boldsymbol{R}_0)\phi_{n'}(\boldsymbol{r}-\boldsymbol{R}_{j'})\,\mathrm{d}\boldsymbol{r} = V_{nn}\delta_{nn'}\delta(\boldsymbol{R}_j-\boldsymbol{R}_0)\delta(\boldsymbol{R}_{j'}-\boldsymbol{R}_0)\,, \tag{17.6.10}$$

which is tantamount to stipulating that scattering by the impurity does not couple states in different bands. The state ψ_μ can then be expanded in Wannier states associated with a single band:

$$\psi_\mu(\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_j c(\boldsymbol{R}_j)\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\,. \tag{17.6.11}$$

Substituting this into (17.6.9), and using an expansion in Wannier functions rather than Bloch functions,

$$c(\boldsymbol{R}_j) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} + \frac{1}{N}\sum_{\boldsymbol{k}'}\frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot(\boldsymbol{R}_j-\boldsymbol{R}_0)}}{\varepsilon-\varepsilon_{n\boldsymbol{k}'}+\mathrm{i}\alpha}V_{nn}c(\boldsymbol{R}_0)\,. \tag{17.6.12}$$

The solution for $\boldsymbol{R}_j = \boldsymbol{R}_0$ is

$$c(\boldsymbol{R}_0) = \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_0}}{1 - V_{nn}\dfrac{1}{N}\displaystyle\sum_{\boldsymbol{k}'}\dfrac{1}{\varepsilon-\varepsilon_{n\boldsymbol{k}'}+\mathrm{i}\alpha}}\,. \tag{17.6.13}$$

The sum over $\boldsymbol{k}'$ can be replaced by the energy integral

$$\begin{aligned}\frac{1}{N}\sum_{\boldsymbol{k}'}\frac{1}{\varepsilon-\varepsilon_{n\boldsymbol{k}'}+\mathrm{i}\alpha} &= \frac{V}{N}\int\frac{\rho_n(\varepsilon')}{\varepsilon-\varepsilon'+\mathrm{i}\alpha}\mathrm{d}\varepsilon' \\ &= \frac{V}{N}\left[\mathrm{P}\int\frac{\rho_n(\varepsilon')}{\varepsilon-\varepsilon'}\mathrm{d}\varepsilon' - \mathrm{i}\pi\rho_n(\varepsilon)\right].\end{aligned} \tag{17.6.14}$$

Using the notation

$$F_n(\varepsilon) = \mathrm{P}\int\frac{\rho_n(\varepsilon')}{\varepsilon-\varepsilon'}\,\mathrm{d}\varepsilon' \tag{17.6.15}$$

for the principal value,

$$c(\boldsymbol{R}_0) = \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_0}}{1 - V_{nn}(V/N)F_n(\varepsilon) + \mathrm{i}\pi V_{nn}(V/N)\rho_n(\varepsilon)}\,. \tag{17.6.16}$$

The amplitude $c(\boldsymbol{R}_0)$ of the wavefunction is large at the impurity for energy values ε_0 that satisfy

$$1 - V_{nn}(V/N)F_n(\varepsilon_0) = 0\,. \tag{17.6.17}$$

When this condition is satisfied inside a band, the amplitude remains finite on account of the imaginary part but the $|c(\boldsymbol{R}_0)|^2$ vs. ε function has a resonance-like maximum. This corresponds to a virtual bound state, as a Lorentzian peak

$$\rho(\varepsilon) = \rho_n(\varepsilon) + \frac{2}{\pi} \frac{\mathit{\Gamma}/2}{(\varepsilon - \varepsilon_0)^2 + (\mathit{\Gamma}/2)^2} \tag{17.6.18}$$

appears in the density of states, where

$$\mathit{\Gamma}/2 = \frac{\pi \rho_n(\varepsilon_0)}{F'(\varepsilon_0)} \,. \tag{17.6.19}$$

The corresponding density of states is plotted in Fig. 17.11.

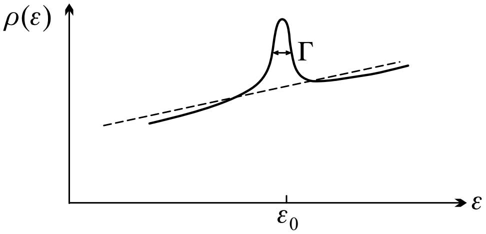


Fig. 17.11. The appearance of virtual bound states in the density of states

On the other hand, when the condition (17.6.17) is satisfied outside the band, and thus the imaginary part vanishes, the amplitude of the Wannier function at the impurity becomes infinitely large, indicating a real bound state localized around the impurity. For repulsive potentials ($V_{nn} > 0$) the bound state can appear above the band, at energy $\varepsilon_0 = \max \varepsilon_{n\boldsymbol{k}} + \mathit{\Delta}$, while for attractive potentials ($V_{nn} < 0$) below the band, at energy $\varepsilon_0 = \min \varepsilon_{n\boldsymbol{k}} - \mathit{\Delta}$. The binding energy $\mathit{\Delta}$ can be determined from the equations

$$1 = \frac{V_{nn}}{N} \sum_{\boldsymbol{k}} \frac{1}{\mathit{\Delta} + \max \varepsilon_{n\boldsymbol{k}} - \varepsilon_{n\boldsymbol{k}}} \tag{17.6.20}$$

and

$$1 = \frac{|V_{nn}|}{N} \sum_{\boldsymbol{k}} \frac{1}{\mathit{\Delta} + \varepsilon_{n\boldsymbol{k}} - \min \varepsilon_{n\boldsymbol{k}}} \,. \tag{17.6.21}$$

As the dispersion relation is quadratic in the vicinity of the top and bottom of the band, the sum on the right-hand side diverges at $\mathit{\Delta} = 0$ in one- and two-dimensional systems, thus there exists a bound state with finite binding energy no matter how weak the impurity potential is. In three-dimensional systems bound states exist only for potentials V_{nn} satisfying either

$$V_{nn} > \left[\frac{1}{N} \sum_{\boldsymbol{k}} \frac{1}{\max \varepsilon_{n\boldsymbol{k}} - \varepsilon_{n\boldsymbol{k}}} \right]^{-1} \tag{17.6.22}$$

or

$$|V_{nn}| > \left[\frac{1}{N} \sum_{\boldsymbol{k}} \frac{1}{\varepsilon_{n\boldsymbol{k}} - \min \varepsilon_{n\boldsymbol{k}}} \right]^{-1} \,. \tag{17.6.23}$$

Further Reading

1. S. L. Altmann, *Band Theory of Solids, An Introduction from the Point of View of Symmetry*, Clarendon Press, Oxford (1991).
2. H. Jones, *The Theory of Brillouin Zones and Electronic States in Crystals*, 2nd revised edition, North-Holland Publishing Co., Amsterdam (1975).
3. S. Raimes, *The Wave Mechanics of Electrons in Metals*, North-Holland Publishing Company, Amsterdam (1961).

18

Simple Models of the Band Structure

The knowledge of the band structure, the energies of one-particle states, and the probability of their occupation are of fundamental importance when a quantitative explanation is sought for those properties of crystalline materials that are determined by electrons. Therefore the theoretical (numerical) calculation and experimental determination of the band structure is a very important chapter of solid-state physics. Before giving a concise summary of the most commonly employed methods in the next chapter, we present two simple models below that give a qualitatively correct description of the band structure of simple metals.

To form an intuitive picture of the electron states in crystalline materials, we shall treat the problem in two opposite limits: either the electrons are considered nearly free and atomic potentials weak, or electrons are assumed to be bound to atoms, and their hopping to adjacent atoms is treated as a perturbation.

18.1 Nearly-Free-Electron Approximation

In view of the success of the Sommerfeld model, it is straightforward to assume that the periodic potential is a weak perturbation for the conduction electrons in metals, and its effects can be determined in perturbation theory. In zeroth order, called the *empty-lattice approximation*, the periodic potential of the crystal is treated as negligibly weak. This can serve as a fairly good starting point for understanding the band structure, since the states determined in this approximation evolve smoothly into the Bloch states when the potential is switched on. Moreover, the dispersion relation differs appreciably from that obtained in the empty-lattice approximation only at the boundaries and center of the Brillouin zone.

18.1.1 Band Structure in the Empty Lattice

The problem of electrons in an empty box is in fact the Sommerfeld model. It was shown that the wave vectors $\boldsymbol{k}$ are quantized, and the allowed values are given by (16.2.16), where the n_α can take arbitrary integer values. There is no upper bound on the magnitude of the vectors $\boldsymbol{k}$. Apart from the spin, this vector is the only quantum number – that is, two states are possible for each value of $\boldsymbol{k}$.

Even though the lattice-periodic potential due to the ions and other electrons vanishes in the empty-lattice approximation as well, the latter nevertheless differs from the problem of electrons in an empty box since the full translational symmetry is assumed to be broken, and the system to be invariant under translations through lattice vectors. Formally, a one-electron Schrödinger equation has to be solved with the strength of the potential approaching zero.

In the presence of a finite periodic potential the allowed wave vectors $\boldsymbol{k}$ are defined in the Brillouin zone or in an equivalent volume of $\boldsymbol{k}$-space, however there exist an infinite number of solutions for each $\boldsymbol{k}$ as the band index can take infinitely many values. Consequently, we shall characterize electron states in an empty lattice by the $\boldsymbol{k}$ vectors defined in the Brillouin zone of the reciprocal lattice and a band index. The task is to establish a relationship between the plane-wave states of free electrons and the states in an empty lattice.

Using the Bloch form for the electron wavefunction, the lattice-periodic part $u_{n\boldsymbol{k}}(\boldsymbol{r})$ can be represented by a Fourier series that contains only the vectors of the reciprocal lattice. The same applies to the periodic potential. Instead of (C.1.36), it is more convenient to eliminate the volume factor and write the Fourier series of lattice-periodic functions as

$$u_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}_j} c_{n\boldsymbol{k}}(\boldsymbol{G}_j) \mathrm{e}^{\mathrm{i}\boldsymbol{G}_j\cdot\boldsymbol{r}}\,, \qquad U(\boldsymbol{r}) = \sum_{\boldsymbol{G}_j} U(\boldsymbol{G}_j) \mathrm{e}^{\mathrm{i}\boldsymbol{G}_j\cdot\boldsymbol{r}}\,. \tag{18.1.1}$$

By substituting these formulas into (17.1.18), multiplying it by $\mathrm{e}^{-\mathrm{i}\boldsymbol{G}_i\cdot\boldsymbol{r}}$ from the left, and integrating over the volume v of the primitive cell, the following equation is obtained for the Fourier coefficients:

$$\left[\frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2 - \varepsilon_{n\boldsymbol{k}}\right] c_{n\boldsymbol{k}}(\boldsymbol{G}_i) + \sum_{\boldsymbol{G}_j} U(\boldsymbol{G}_i - \boldsymbol{G}_j) c_{n\boldsymbol{k}}(\boldsymbol{G}_j) = 0\,. \tag{18.1.2}$$

The solution in the $U \to 0$ limit is

$$c_{n\boldsymbol{k}}(\boldsymbol{G}_i) = 0 \qquad \text{or} \qquad \varepsilon_{n\boldsymbol{k}} = \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} \equiv \frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2\,. \tag{18.1.3}$$

The solution in the empty lattice for the band of index n is therefore particularly simple: apart from a single $\boldsymbol{G}_i$, all reciprocal-lattice vectors have vanishing coefficients in the expansion (18.1.1).[1] The normalized Bloch function of this state is then

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \mathrm{e}^{\mathrm{i}\boldsymbol{G}_i\cdot\boldsymbol{r}} . \tag{18.1.4}$$

It is readily seen that the nonvanishing coefficient is unity. This procedure establishes a one-to-one correspondence between the band indices n and the reciprocal-lattice vectors $\boldsymbol{G}_i$, moreover, it guarantees that the dispersion relations $\varepsilon_{n\boldsymbol{k}}$ have continuous derivatives in $\boldsymbol{k}$-space. Nevertheless, this choice has a serious drawback: as we shall see, the states of different bands are mixed by the periodic potential.

As the simplest example, consider a one-dimensional system of length L. Imposing periodic boundary conditions, the wave number can take the discrete values $k = m\,2\pi/L$, where m is an arbitrary integer. For free electrons the energy is quadratic in the wave number:

$$\varepsilon_{\boldsymbol{k}}^{(0)} = \frac{\hbar^2}{2m_\mathrm{e}} k^2 . \tag{18.1.5}$$

If the system of length L is in fact a chain of lattice constant a containing N atoms then the allowed wave numbers are, once again, $k = m\,2\pi/L = m\,2\pi/(Na)$, but – provided only the wave numbers defined in the first Brillouin zone $-\pi/a < k \le \pi/a$ are considered (i.e., the equivalent wave numbers in higher Brillouin zones are reduced to the first Brillouin zone) – m can take values only in the interval $-N/2 < m \le N/2$. By taking the vectors $G = n\,2\pi/a$ of the reciprocal lattice of the linear chain for each integer value of n, the energy eigenvalues are given by

$$\varepsilon_{nk} = \frac{\hbar^2}{2m_\mathrm{e}}\left(k + n\frac{2\pi}{a}\right)^2 = \frac{\hbar^2}{2m_\mathrm{e}}\left(\frac{2\pi}{Na}\right)^2 (m+nN)^2 \tag{18.1.6}$$

in the empty-lattice approximation according to (18.1.3). These energies are shown in Fig. 18.1(a).

At first sight, the spectrum obtained for electrons moving in an empty lattice looks very different from the well-known parabolic spectrum of free electrons – even though they describe the same system of electrons. The proper relation between the results of the two approximations is established when the repeated- or extended-zone scheme is used instead of restricting the k values to the first Brillouin zone. The energies associated with the wave numbers in the interval $-\pi/a < k \le \pi/a$ are also associated with all equivalent wave numbers $k + G$ in the repeated-zone scheme, as shown in Fig. 18.1(b). The dispersion curves are then continuous across the boundaries of the Brillouin

[1] Since the wavefunction is constructed from a single plane wave, the method is also known as the single-OPW approximation, where OPW stands for orthogonalized plane wave.

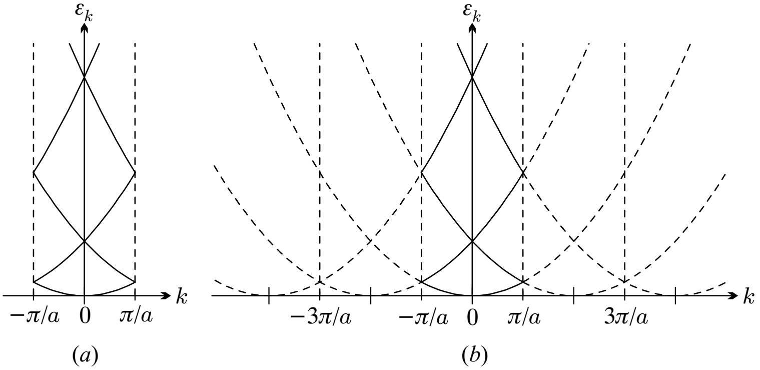


Fig. 18.1. Energy vs. wave number plot for a one-dimensional electron system in the empty-lattice approximation, represented in the (a) reduced-zone and (b) repeated-zone schemes

zone, and the characteristic parabolas of the free-electron model are recovered – in a repeated pattern.

The parabolic dispersion curve of free electrons is recovered even more directly in the extended-zone scheme. The energies given by (18.1.6) are then represented by plotting the state of index n not in the interval $-\pi/a < k \leq \pi/a$ but in an equivalent interval $(2n-1)\pi/a < k \leq (2n+1)\pi/a$. The dispersion relation obtained in this way is shown in Fig. (18.2).

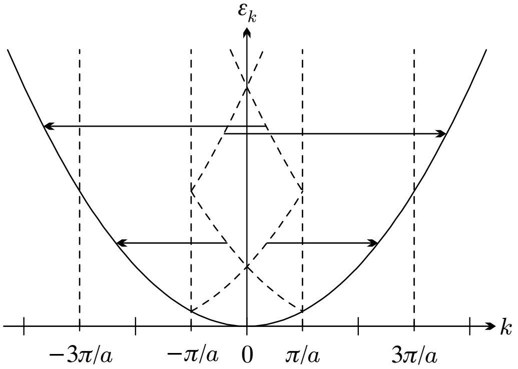


Fig. 18.2. Energy vs. wave number plot for a one-dimensional electron system in the empty-lattice approximation, represented in the extended-zone scheme

Conversely, one may say that the energy eigenvalues of the electrons moving through the empty lattice can be obtained for the k values in the reduced-zone scheme from the free-electron dispersion curve by finding the equivalent k values inside the first Brillouin zone for each wave number outside of it, and then shifting the energy eigenvalue to this k value. This procedure is called *zone folding.*

After the one-dimensional chain consider a somewhat more complicated case, a simple cubic crystal with a monatomic basis. The bands made up of the first few energy eigenvalues are shown for two high-symmetry directions in Fig. 18.3. The numbers appearing next to the band indices A, B, C, ... indicate the degeneracy of the bands.

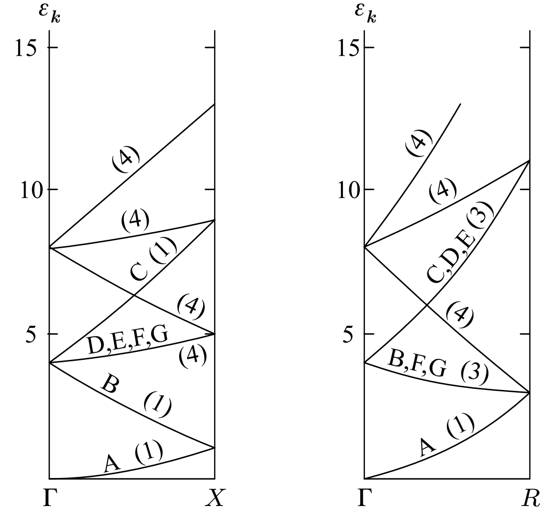


Fig. 18.3. Energy bands for an empty simple cubic lattice in two special directions of the Brillouin zone. Energy is given in units of $(\hbar^2/2m_\mathrm{e})(\pi/a)^2$

One of the directions is along line $\varDelta$ connecting the center $\varGamma = (0,0,0)$ of the Brillouin zone and $X = (\pi/a)(0,0,1)$ (see Fig. 7.2). The lowest-lying band (A) belongs to the vector $\boldsymbol{G} = 0$ of the reciprocal lattice. The energy of the state associated with the point $\varDelta = (\pi/a)(0,0,\xi)$ is

$$\varepsilon_\mathrm{A}(0,0,\xi) = \frac{\hbar^2}{2m_\mathrm{e}}\left(\frac{\pi}{a}\right)^2 \xi^2 . \tag{18.1.7}$$

The next band (B) corresponds to the reciprocal-lattice vector $\boldsymbol{G} = (2\pi/a)$ $(0,0,\bar{1})$; its energy is

$$\varepsilon_\mathrm{B}(0,0,\xi) = \frac{\hbar^2}{2m_\mathrm{e}}\left(\frac{\pi}{a}\right)^2 (\xi-2)^2 . \tag{18.1.8}$$

The vector $\boldsymbol{G} = (2\pi/a)(0,0,1)$ is associated with band C of energy

$$\varepsilon_\mathrm{C}(0,0,\xi) = \frac{\hbar^2}{2m_\mathrm{e}}\left(\frac{\pi}{a}\right)^2 (\xi+2)^2 , \tag{18.1.9}$$

while the bands D, E, F, and G, which belong to points $\boldsymbol{G} = (2\pi/a)(1,0,0)$, $(2\pi/a)(0,1,0)$, $(2\pi/a)(\bar{1},0,0)$, and $(2\pi/a)(0,\bar{1},0)$, are degenerate along the line $\varGamma X$. Their energy is

$$\varepsilon_{\mathrm{D,E,F,G}}(0,0,\xi) = \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{\pi}{a}\right)^2(\xi^2+4)\,. \tag{18.1.10}$$

The other direction is along the line $\varLambda$ joining the center $\varGamma$ and $R = (\pi/a)(1,1,1)$. The bands correspond to the same vectors of the reciprocal lattice as above, therefore the same indices are used. The energy values at $\varLambda = (\pi/a)(\xi,\xi,\xi)$ are

$$\begin{aligned}
\varepsilon_{\mathrm{A}}(\xi,\xi,\xi) &= \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{\pi}{a}\right)^2 3\xi^2\,,\\
\varepsilon_{\mathrm{B,F,G}}(\xi,\xi,\xi) &= \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{\pi}{a}\right)^2\left[(\xi-2)^2+2\xi^2\right],\\
\varepsilon_{\mathrm{C,D,E}}(\xi,\xi,\xi) &= \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{\pi}{a}\right)^2\left[(\xi+2)^2+2\xi^2\right].
\end{aligned} \tag{18.1.11}$$

A similar procedure is followed for other lattice types, too. For a body-centered cubic crystal the Brillouin zone is shown in Fig. 7.7. Its special points are the center $\varGamma = (2\pi/a)(0,0,0)$, the vertices H and P – e.g., $H = (2\pi/a)(1,0,0)$ and $P = (2\pi/a)(\frac{1}{2},\frac{1}{2},\frac{1}{2})$ –, and the face centers – e.g., $N = (2\pi/a)(\frac{1}{2},\frac{1}{2},0)$. Figure 18.4 shows the energy spectrum for the wave vectors along the lines joining them calculated in the empty-lattice approximation.

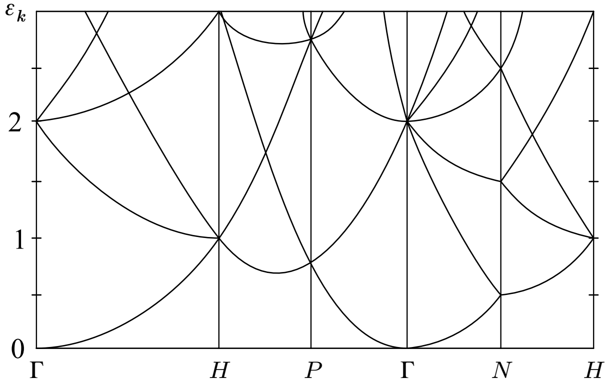


Fig. 18.4. Energy bands for an empty body-centered cubic lattice. Energy is given in units of $(\hbar^2/2m_{\mathrm{e}})(2\pi/a)^2$

The energy values obtained in the empty-lattice approximation for a face-centered cubic lattice will be presented in Chapter 20 on the band structure of semiconductors.

18.1.2 Fermi Surface in the Empty Lattice

As a consequence of the Fermi–Dirac statistics, the bands are filled up to the Fermi energy in the ground state. Since the physical properties of metals are determined by electrons occupying states near the Fermi energy, the knowledge of these states – especially the density of states at the Fermi energy and the characteristic features of the Fermi surface – are particularly important for understanding these properties. A good starting point for this is the Fermi surface in an empty lattice, which can be obtained via the *Harrison construction.*[2]

As demonstrated in the previous subsection, the dispersion relation obtained in the extended-zone scheme in the empty-lattice approximation is identical to the quadratic dispersion relation of free electrons. The natural labeling of the bands also comes from the free-electron model: the nth band is the part of the free-electron spectrum that falls into the nth Brillouin zone. The only difference in the reduced-zone scheme is that the states associated with wave vectors $\boldsymbol{k}$ that lie outside the first Brillouin zone are reduced to equivalent $\boldsymbol{k}$ vectors inside the first zone. Thus the Fermi surface in the empty-lattice approximation can be constructed by drawing a Fermi sphere of radius k_{F} in the reciprocal lattice such that the number of allowed electron states within the Fermi sphere should be equal to the number of conduction electrons – and then, if necessary, reducing the parts of this sphere lying outside the first Brillouin zone to the interior of the zone.

Denoting the number of conduction electrons per atom by z, the Fermi momentum k_{F} of the gas of free electrons can be calculated from

$$\frac{N_{\mathrm{e}}}{V} = \frac{zN}{V} = \frac{k_{\mathrm{F}}^3}{3\pi^2}\,, \tag{18.1.12}$$

which is a consequence of (16.2.24). For metals that crystallize in a simple cubic structure with a monatomic basis, $V = Na^3$, so

$$k_{\mathrm{F}} = \left(\frac{\pi}{a}\right)\left(\frac{3z}{\pi}\right)^{1/3}, \tag{18.1.13}$$

and the Fermi energy is

$$\varepsilon_{\mathrm{F}} = \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{\pi}{a}\right)^2\left(\frac{3z}{\pi}\right)^{2/3}. \tag{18.1.14}$$

For one conduction electron per atom,

$$k_{\mathrm{F}} = 0.985\left(\frac{\pi}{a}\right), \qquad \varepsilon_{\mathrm{F}} = 0.970\frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{\pi}{a}\right)^2. \tag{18.1.15}$$

[2] W. A. Harrison, 1960.

Since the diameter $2k_{\mathrm{F}}$ of the Fermi sphere is smaller than the edge of the cubic Brillouin zone, the entire Fermi sphere is inside the first Brillouin zone, and all occupied states are in the lowest-lying band.

If the gas of free electrons contains two electrons per lattice point,

$$k_{\mathrm{F}} = 1.241 \left(\frac{\pi}{a}\right), \qquad \varepsilon_{\mathrm{F}} = 1.539 \frac{\hbar^2}{2m_{\mathrm{e}}} \left(\frac{\pi}{a}\right)^2 . \tag{18.1.16}$$

In this case k_{F} is larger than the distance ΓX but smaller than the distance ΓR. Thus, when the lowest-energy states are filled gradually by electrons in the ground state (as required by the Fermi–Dirac statistics), the lowest-lying states in the second band become occupied before the highest-lying states in the first band. (The latter are states whose wave vectors are close to the corner point R of the Brillouin zone.)

It was mentioned in the previous chapter that the prerequisite for metallic behavior is that the Fermi energy should fall inside a band. This is the case for two partially filled bands of the previous example. Moreover, no matter up to what energy the levels are filled, the system is always metallic in the empty-lattice approximation as there are no gaps (forbidden energies) in the band structure.

The radius of the Fermi sphere can be calculated similarly for crystals displaying other symmetries. In metals with a body-centered cubic crystal structure, where the Bravais cell of side a contains two atoms, the Fermi wave number is

$$k_{\mathrm{F}} = \left(\frac{2\pi}{a}\right) \left(\frac{3z}{4\pi}\right)^{1/3} , \tag{18.1.17}$$

while in metals with a face-centered cubic crystal structure, where the Bravais cell of side a contains four atoms,

$$k_{\mathrm{F}} = \left(\frac{2\pi}{a}\right) \left(\frac{3z}{2\pi}\right)^{1/3} , \tag{18.1.18}$$

and in a hexagonal close-packed structure

$$k_{\mathrm{F}} = \left(\frac{\pi}{a}\right) \left(\frac{2\sqrt{3}z}{\pi c/a}\right)^{1/3} . \tag{18.1.19}$$

The surface of this sphere is the Fermi sphere in the empty-lattice approximation in the extended-zone scheme. When working in the reduced-zone scheme, the Fermi surfaces of individual bands are obtained by reducing the Fermi sphere to the first Brillouin zone. For ease of illustration, we shall first consider a two-dimensional square lattice of lattice constant a, with z electrons per atom (and so a total of $N_{\mathrm{e}} = zN$ electrons). The radius of the "Fermi sphere" is then determined from the formula

$$zN = 2k_{\mathrm{F}}^2 \pi \frac{V}{(2\pi)^2} = 2k_{\mathrm{F}}^2 \pi \frac{Na^2}{(2\pi)^2} , \tag{18.1.20}$$

which gives

$$k_{\mathrm{F}} = \left(\frac{2z}{\pi}\right)^{1/2} \frac{\pi}{a} . \tag{18.1.21}$$

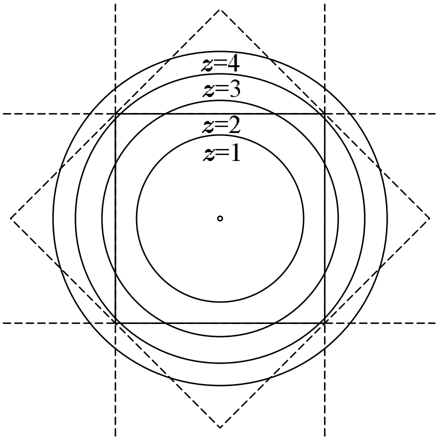


Fig. 18.5. Fermi sphere of two-dimensional systems with one, two, three, and four electrons per primitive cell in the empty-lattice approximation, represented in the extended-zone scheme

As illustrated in Fig. 18.5, the "Fermi sphere" reaches beyond the boundaries of the Brillouin zone – a square of side $2\pi/a$ – if $z \geq 2$. When the extended-zone scheme is used, the Fermi surface is in the first and second Brillouin zones for $z = 2$ and $z = 3$, while for $z = 4$ the states of the first Brillouin zone are all occupied, and the Fermi surface is in the second, third, and fourth Brillouin zones.

When the pieces lying outside the first Brillouin zone in the extended-zone scheme are reduced to the first Brillouin zone, the Fermi surface in the $z = 2$ case – shown in Fig. 18.6 – is seen to be composed of disjoint parts in the first and second bands.

Another representation, which will prove particularly useful later, is obtained when the Fermi sphere is drawn around each point of the reciprocal lattice in the repeated-zone scheme. As shown in Fig. 18.7, certain regions around the corners M of the first Brillouin zone are not covered, while certain closed regions around the edge centers X are covered twice.

It is immediately seen that if the wave vectors are not reduced to the Brillouin zone centered at Γ but to an equivalent region centered at a vertex or edge center of the first Brillouin zone, then the pieces in the first and second bands that are disjoint in the customary representation make up a closed Fermi surface. This is shown in Fig. 18.8.

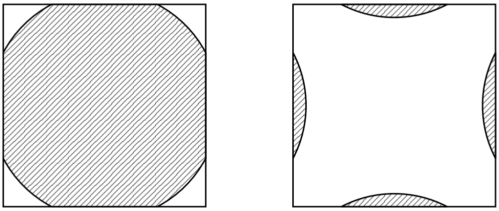


Fig. 18.6. Fermi surface in two partially filled bands of a two-dimensional system with two electrons per primitive cell in the empty-lattice approximation, represented in the reduced-zone scheme

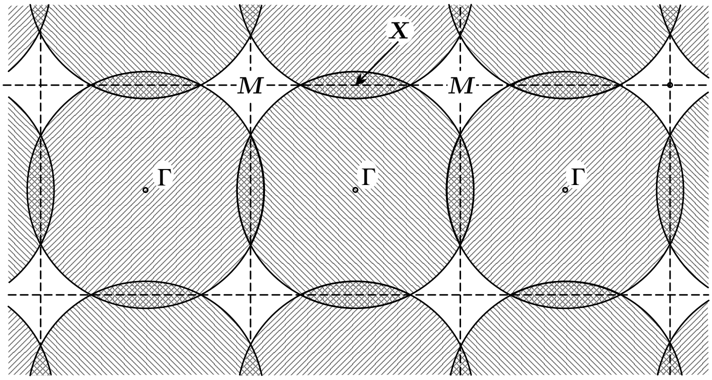


Fig. 18.7. Fermi spheres in the empty-lattice approximation for a two-dimensional square lattice with two electrons per primitive cell, represented in the repeated-zone scheme

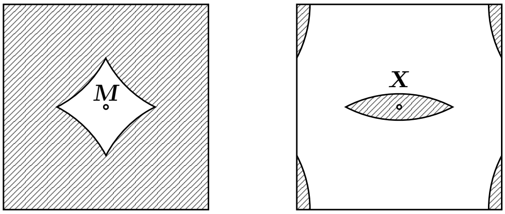


Fig. 18.8. Closed Fermi surface of the system shown in the two previous figures when wave vectors are reduced to regions centered at M or X rather than Γ

The Fermi surface in the first band – that is, the surface made up of the pieces that belonged initially in the first Brillouin zone – surrounds empty states. It is therefore said to be a hole-type Fermi surface. The states of the second band make up an electron-like Fermi surface. These closed Fermi surfaces are not at all spherical.

A similar procedure is followed when each atom has four electrons. The overlapping Fermi spheres of the repeated-zone scheme are shown in the upper part of Fig. 18.9. The regions covered twice, three, or four times mark the parts of the Fermi surface that belong in the second, third, and fourth bands. These are shown in the lower part, represented in the reduced-zone scheme. Since the states of the first band are all filled, there is no Fermi surface there. The Fermi surface in the third and fourth bands are closed provided the equivalent vectors $\boldsymbol{k}$ are reduced to the region centered at M.

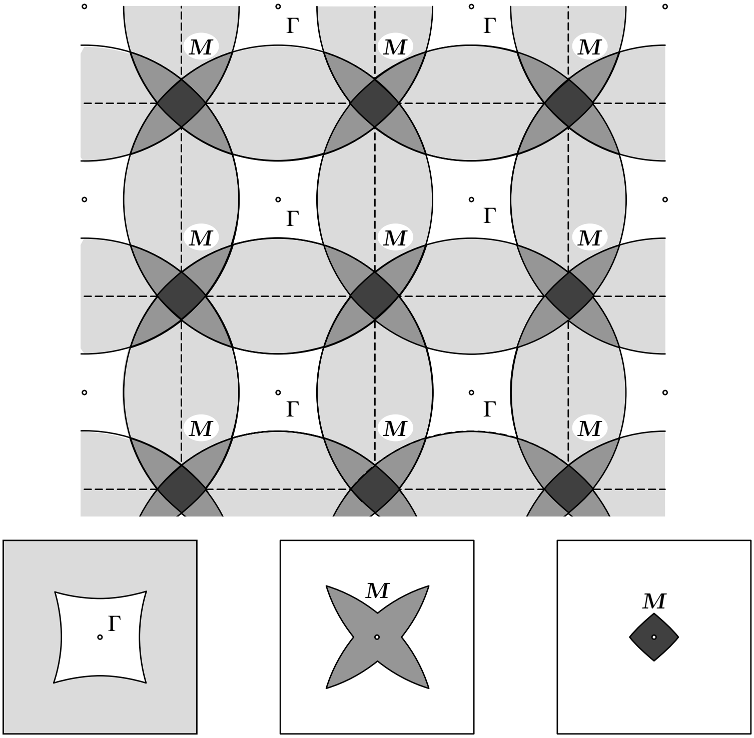


Fig. 18.9. Fermi surfaces of a two-dimensional system with four electrons per primitive cell in the empty-lattice approximation, represented in the repeated- and reduced-zone schemes. The pieces of the Fermi surface in the third and fourth bands are now reduced to the region around point M rather than Γ

By employing the same procedure, the Fermi surfaces can be easily constructed for two highly important three-dimensional cubic crystal structures: body- and face-centered cubic lattices. Figures 18.10 and 18.11 show the parts of the Fermi surface that belong to each band in the reduced-zone scheme for mono-, di-, tri-, and tetravalent metals. For the sake of better illustration, the surfaces are sometimes shown not around the zone center Γ but other high-symmetry points of the Brillouin zone. Convex portions of the Fermi surface surround electron states, while concave portions hole states.

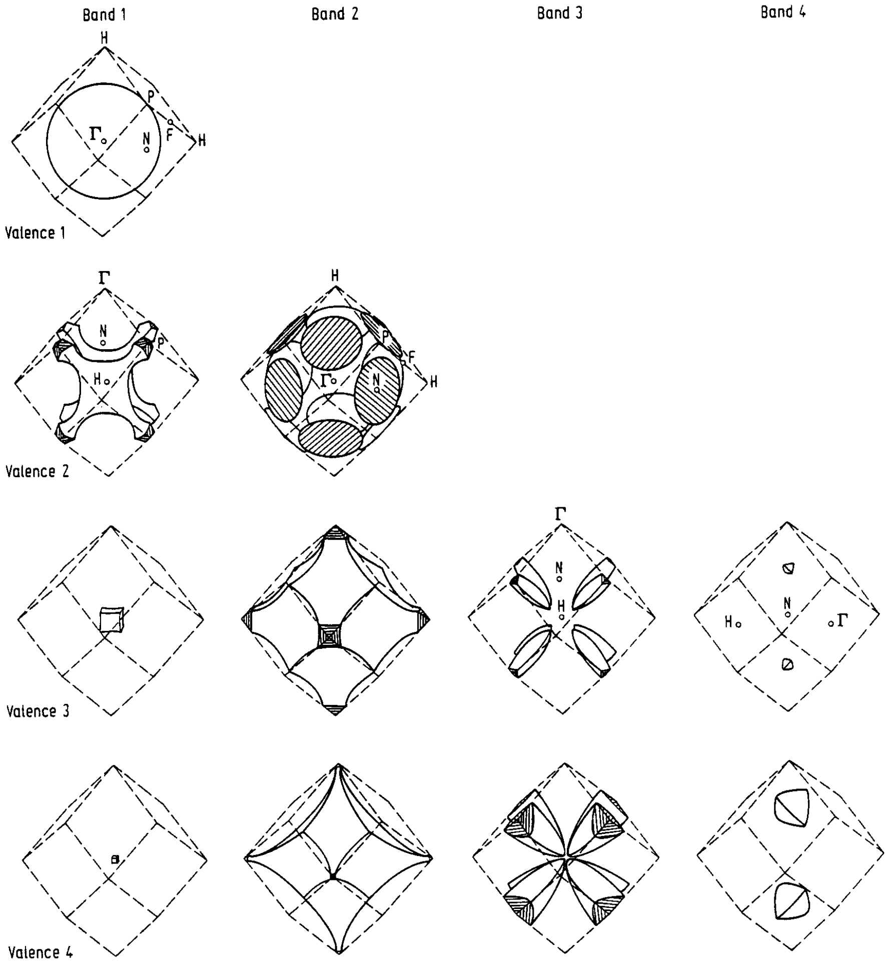


Fig. 18.10. Fermi surfaces of mono-, di-, tri-, and tetravalent metals in an empty body-centered cubic lattice. In certain cases the reduced zone is centered at point H or N rather than Γ [Reprinted with permission from W. A. Harrison, *Phys. Rev.* **118**, 1190 (1960). ©1960 by the American Physical Society]

18.1.3 Effects of a Weak Periodic Potential

The finite periodic potential $U(\boldsymbol{r})$ due to the ions and other electrons modifies the energy spectrum obtained in the empty-lattice approximation. If the potential is weak, a complete solution of the system of equations (18.1.2) is not necessary: the effects of the potential can be taken into account using perturbation theory, keeping only the lowest-order nonvanishing corrections. This is the *nearly-free-electron model.* Rough as it may seem, this approximation provides qualitatively correct information about the bands and Fermi surfaces in metals with s- and p-electrons. An important feature of the band

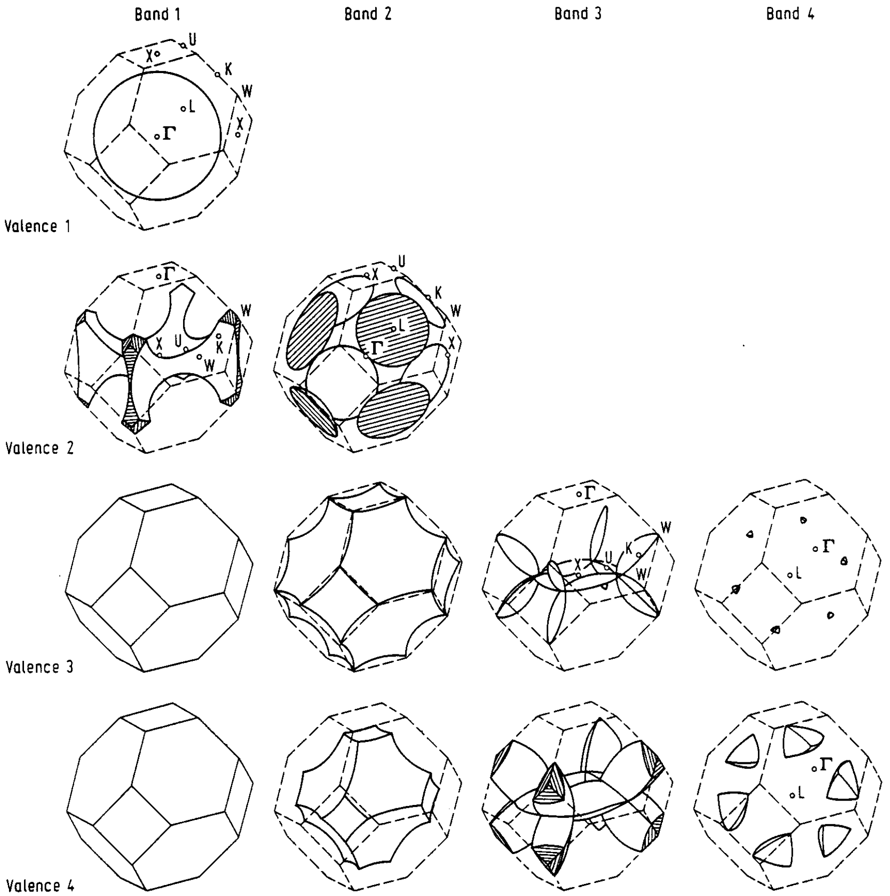


Fig. 18.11. Fermi surfaces of mono-, di-, tri-, and tetravalent metals in an empty face-centered cubic lattice. In certain cases the reduced zone is centered at point X or L rather than Γ [W. A. Harrison, *ibid.*]

structure can be observed even in this approximation: energy bands may be separated by gaps. Gaps arise because the degeneracies present in the empty-lattice approximation at the center and boundary of the Brillouin zone are lifted by the periodic potential, and thus certain states are shifted upward, while others downward.

We shall determine the energy shift due to the periodic potential for a state for which the unperturbed energy in the empty lattice, $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i}$, is relatively far from the energies $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}$ of states with the same $\boldsymbol{k}$ in all other bands, that is, for all $\boldsymbol{G}_j \neq \boldsymbol{G}_i$:

$$|\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}| \gg U\,, \qquad (18.1.22)$$

where U is some kind of average of the periodic potential of the lattice. The unperturbed wavefunction (18.1.4) of the selected state contains a single vector $\boldsymbol{G}_i$, that is, the coefficient $c_{n\boldsymbol{k}}(\boldsymbol{G}_i)$ of this vector is unity in the expansion (18.1.1), while $c_{n\boldsymbol{k}}(\boldsymbol{G}_j) = 0$ for all other $\boldsymbol{G}_j$. Since the potential is assumed to be weak on the scale of the kinetic energy, $c_{n\boldsymbol{k}}(\boldsymbol{G}_i)$ is expected to be of order unity even in the presence of a perturbing potential, while all other coefficients are expected to be proportional to U or its higher powers. Therefore the equations for $c_{n\boldsymbol{k}}(\boldsymbol{G}_i)$ and $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ will be treated separately.

For simplicity, we shall assume that the component for $\boldsymbol{G}_i = 0$ vanishes in the Fourier series of the potential – that is, the integral of the potential over the entire volume is zero. (If this were not the case, the Fourier component associated with the zero vector of the reciprocal lattice could be eliminated by adding a constant to the potential. Such a constant would shift all energy values by the same amount, and thus would not modify the dispersion relation.) Ignoring the term $\boldsymbol{G}_j = \boldsymbol{G}_i$ in the second term of (18.1.2),

$$\left[\frac{\hbar^2}{2m_{\mathrm{e}}}(\boldsymbol{k}+\boldsymbol{G}_i)^2 - \varepsilon_{n\boldsymbol{k}}\right] c_{n\boldsymbol{k}}(\boldsymbol{G}_i) + \sum_{\boldsymbol{G}_j \neq \boldsymbol{G}_i} U(\boldsymbol{G}_i - \boldsymbol{G}_j) c_{n\boldsymbol{k}}(\boldsymbol{G}_j) = 0 \quad (18.1.23)$$

is obtained. The second term contains only the coefficients $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ that are assumed to be small, multiplied by the weak potential – therefore this term is second- or higher-order in the potential. This implies that it is sufficient to determine the small coefficients $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ up to linear order in the potential. If the equations for $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ are derived from (18.1.2), and the term containing $\boldsymbol{G}_i$ is separated from the sum over the other reciprocal-lattice vectors,

$$\begin{aligned}\left[\frac{\hbar^2}{2m_{\mathrm{e}}}(\boldsymbol{k}+\boldsymbol{G}_j)^2 - \varepsilon_{n\boldsymbol{k}}\right] c_{n\boldsymbol{k}}(\boldsymbol{G}_j) &+ U(\boldsymbol{G}_j - \boldsymbol{G}_i) c_{n\boldsymbol{k}}(\boldsymbol{G}_i) \\ &+ \sum_{\boldsymbol{G}_k \neq \boldsymbol{G}_i} U(\boldsymbol{G}_j - \boldsymbol{G}_k) c_{n\boldsymbol{k}}(\boldsymbol{G}_k) = 0\end{aligned} \quad (18.1.24)$$

is obtained. The third term on the left-hand side is a second-order correction and can be neglected. This leaves us with

$$c_{n\boldsymbol{k}}(\boldsymbol{G}_j) = \frac{U(\boldsymbol{G}_j - \boldsymbol{G}_i)}{\varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}} c_{n\boldsymbol{k}}(\boldsymbol{G}_i)\,. \quad (18.1.25)$$

Substituting this into (18.1.23),

$$\left[\varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i}\right] c_{n\boldsymbol{k}}(\boldsymbol{G}_i) = \sum_{\boldsymbol{G}_j \neq \boldsymbol{G}_i} \frac{U(\boldsymbol{G}_i - \boldsymbol{G}_j) U(\boldsymbol{G}_j - \boldsymbol{G}_i)}{\varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}} c_{n\boldsymbol{k}}(\boldsymbol{G}_i)\,, \quad (18.1.26)$$

and so

$$\varepsilon_{n\boldsymbol{k}} = \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} + \sum_{\boldsymbol{G}_j \neq \boldsymbol{G}_i} \frac{|U(\boldsymbol{G}_i - \boldsymbol{G}_j)|^2}{\varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}}\,. \quad (18.1.27)$$

The self-consistent solution of this equation is the perturbed energy value.

Since the neglected terms are of order U^3, the same accuracy is kept if $\varepsilon_{n\boldsymbol{k}}$ is replaced by the unperturbed $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i}$ in the energy denominator on the right-hand side:

$$\varepsilon_{n\boldsymbol{k}} = \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} + \sum_{\boldsymbol{G}_j \neq \boldsymbol{G}_i} \frac{|U(\boldsymbol{G}_i - \boldsymbol{G}_j)|^2}{\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}} + \mathcal{O}(U^3)\,. \tag{18.1.28}$$

This formula shows that energy levels repel each other. In those terms where the unperturbed energy $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}$ is larger than $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i}$ the correction is negative. The selected level is thus shifted downward by the levels above it. Similarly, the mixing with levels of lower energy gives rise to an upward shift.

This correction is usually small provided the potential is weak. When the unperturbed energies determined in the empty-lattice approximation are not too close together, the energy levels are hardly shifted. The situation is radically different when the energy of the selected state is close to that of another state with the same (or an equivalent) vector $\boldsymbol{k}$. The energy denominator in (18.1.25) is then small, and therefore the coefficient $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ will also be large. The situation is particularly interesting when there is a wave vector $\boldsymbol{k}$ for which the unperturbed energy levels associated with the vectors $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$ of the reciprocal lattice are equal:

$$\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} = \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}\,. \tag{18.1.29}$$

Barring accidental situations, this occurs at the center or boundary of the Brillouin zone. Note that the previous formula is satisfied for a quadratic dispersion relation only if

$$|\boldsymbol{k} + \boldsymbol{G}_i| = |\boldsymbol{k} + \boldsymbol{G}_j|\,. \tag{18.1.30}$$

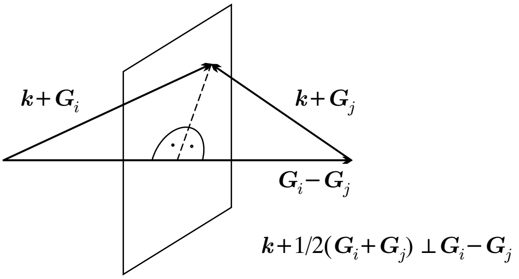

$\boldsymbol{k}+\boldsymbol{G}_i$
$\boldsymbol{k}+\boldsymbol{G}_j$
$\boldsymbol{G}_i-\boldsymbol{G}_j$
$\boldsymbol{k}+1/2(\boldsymbol{G}_i+\boldsymbol{G}_j) \perp \boldsymbol{G}_i-\boldsymbol{G}_j$

Fig. 18.12. Two vectors satisfying the condition $|\boldsymbol{k}+\boldsymbol{G}_i| = |\boldsymbol{k}+\boldsymbol{G}_j|$. Their starting points are the tail and tip of the reciprocal-lattice vector $\boldsymbol{G}_i - \boldsymbol{G}_j$, while their common end point is a point of the perpendicular bisecting plane (Bragg plane) of $\boldsymbol{G}_i - \boldsymbol{G}_j$

If $|\boldsymbol{G}_i| = |\boldsymbol{G}_j|$ then the energies are equal at the zone center. Otherwise the equality is satisfied at the boundary of a (possibly higher) Brillouin zone, since the previous condition implies

$$\boldsymbol{k} \cdot (\boldsymbol{G}_i - \boldsymbol{G}_j) + \boldsymbol{G}_i^2 - \boldsymbol{G}_j^2 = 0\,, \tag{18.1.31}$$

which is equivalent to the condition that $\boldsymbol{k} + \frac{1}{2}(\boldsymbol{G}_i + \boldsymbol{G}_j)$ should be perpendicular to $\boldsymbol{G}_i - \boldsymbol{G}_j$. As shown in Fig. 18.12, the vector $\boldsymbol{k} + \boldsymbol{G}_i$ is then in the perpendicular bisecting plane of $\boldsymbol{G}_i - \boldsymbol{G}_j$, and this plane is, by definition, a Bragg plane, a zone boundary.

The unperturbed energies that belong to $\boldsymbol{k} + \boldsymbol{G}_i$ and $\boldsymbol{k} + \boldsymbol{G}_j$ are equal but this degeneracy may be lifted by the periodic potential. The calculation of the splitting requires the application of degenerate perturbation theory. Owing to the mixing of the degenerate states, the coefficients $c_{n\boldsymbol{k}}(\boldsymbol{G}_i)$ and $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ associated with the vectors $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$ can be large. By neglecting the terms associated with other vectors of the reciprocal lattice in the expansion (18.1.1), since they are proportional to $c_{n\boldsymbol{k}}(\boldsymbol{G}_k)$ and vanish in the $U \to 0$ limit, the following equations are obtained for the energies:

$$\begin{aligned} \left[\varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i}\right] c_{n\boldsymbol{k}}(\boldsymbol{G}_i) &= U(\boldsymbol{G}_i - \boldsymbol{G}_j) c_{n\boldsymbol{k}}(\boldsymbol{G}_j)\,, \\ \left[\varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}\right] c_{n\boldsymbol{k}}(\boldsymbol{G}_j) &= U(\boldsymbol{G}_j - \boldsymbol{G}_i) c_{n\boldsymbol{k}}(\boldsymbol{G}_i)\,. \end{aligned} \tag{18.1.32}$$

This homogeneous system of equations has a nontrivial solution for $c_{n\boldsymbol{k}}$ if the determinant vanishes:

$$\begin{vmatrix} \varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} & -U(\boldsymbol{G}_i - \boldsymbol{G}_j) \\ -U^*(\boldsymbol{G}_i - \boldsymbol{G}_j) & \varepsilon_{n\boldsymbol{k}} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j} \end{vmatrix} = 0\,. \tag{18.1.33}$$

The perturbed energies are then

$$\varepsilon_{n\boldsymbol{k}} = \tfrac{1}{2}\left(\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} + \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}\right) \pm \left\{\tfrac{1}{4}\left(\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} - \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}\right)^2 + |U(\boldsymbol{G}_i - \boldsymbol{G}_j)|^2\right\}^{1/2}\,. \tag{18.1.34}$$

The new energy eigenvalues are shown schematically in Fig. 18.13 along a particular direction of the Brillouin zone. The degeneracy is lifted at the boundary (or center) of the Brillouin zone by the periodic potential and a gap appears between the two bands. The splitting is proportional to $|U(\boldsymbol{G}_i - \boldsymbol{G}_j)|$.

It follows from the expression for the unperturbed energies that at the zone boundary

$$\boldsymbol{\nabla}_{\boldsymbol{k}} \varepsilon_{n\boldsymbol{k}} = \frac{\hbar^2}{m_{\mathrm{e}}}\left[\boldsymbol{k} + \tfrac{1}{2}(\boldsymbol{G}_i + \boldsymbol{G}_j)\right]\,, \tag{18.1.35}$$

i.e., the $\boldsymbol{k}$-space gradient of $\varepsilon_{n\boldsymbol{k}}$ is along the plane of the zone boundary. Thus the group velocity, which is proportional to this gradient, has a vanishing component perpendicular to the zone boundary. Since this gradient is perpendicular to the constant-energy surface, the dispersion curves, when considered

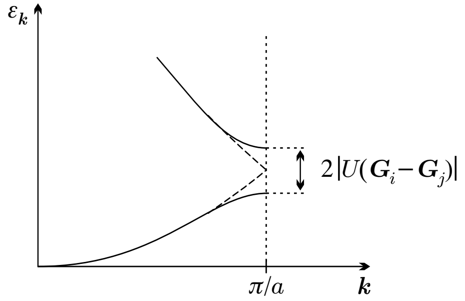


Fig. 18.13. Upward and downward shift of the free-electron levels at the zone boundary due to a weak periodic potential

along lines perpendicular to a Bragg plane, reach this plane with a vanishing slope. This also implies that the constant-energy surfaces are perpendicular to the zone boundary.

When the foregoing results are applied to a one-dimensional lattice, the periodic potential is seen to lift all the degeneracies at the center and boundaries of the Brillouin zone that were obtained in the previous subsection. Figure 18.14 shows the resulting band pattern in the first Brillouin zone, and – by suitable folding of the bands – in the extended- and repeated-zone scheme. In contrast to the energy spectra in Figs. 18.1 and 18.2, where all energy values were allowed, distinctly separate bands (and gaps between them) are observed now.

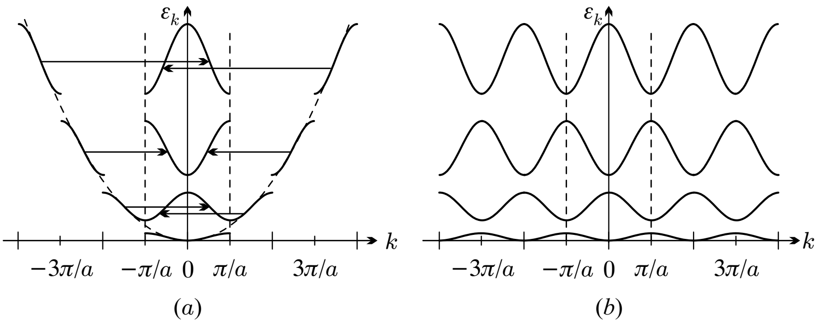


Fig. 18.14. Band structure of a one-dimensional system in the nearly-free-electron approximation (*a*) in the reduced- and extended-zone schemes, and (*b*) in the repeated-zone scheme

Based on these findings one may say that in much of the Brillouin zone, apart from the center and boundaries, the energy of electrons in the nearly-free-electron model is fairly close to the energy of free electrons in an empty lattice. Significant differences are observed only at the center and boundaries of the zone as degeneracies are – at least, partially – lifted there. This leads to the vanishing of the group velocity at the zone center and of its perpendicular component at the zone boundary, provided the zone boundary is separated from its mirror image by a reciprocal-lattice vector.

The band structure of the empty lattice may exhibit not only double but also higher degeneracies. This can occur, for example, at the edges or vertices of the Brillouin zone. In these geometries degenerate perturbation theory has to be applied to four- and eightfold degenerate states. The degenerate levels usually split, even though this may depend on the particular choice of the potential. Below we shall present a method that allows us to determine the character of the splitting from the crystal symmetries, without knowing the exact form of the potential. One can thus specify in full generality which degeneracies are accidental, which ones can be lifted by an applied potential, and which ones are preserved even in the presence of the lattice potential.

18.1.4 Lifting of Accidental Degeneracies

Suppose that a solution $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ of the Schrödinger equation and the corresponding energy eigenvalue $\varepsilon_{n\boldsymbol{k}}$ are known. As demonstrated in Chapter 6, the energy is the same for all vectors $\boldsymbol{k}'$ that are related to $\boldsymbol{k}$ by a symmetry transformation. Below we shall examine under what circumstances do two or more states with the same wave vector but different wavefunctions have the same energy – that is, when there is a degeneracy imposed by the symmetry. If there is not, then any degeneracy obtained in the band-structure calculation is accidental.

Consider the symmetry group of the Hamiltonian. For each element P of the group the corresponding symmetry operation commutes with the Hamiltonian:

$$P\mathcal{H} = \mathcal{H}P\,. \tag{18.1.36}$$

Recall from Chapter 6 that if $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is an eigenfunction of the Hamiltonian then any function

$$P\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \psi_{n\boldsymbol{k}}(P^{-1}\boldsymbol{r}) \tag{18.1.37}$$

obtained by a symmetry transformation is also an eigenfunction with the same energy. It was also mentioned that these functions belong to different wave vectors $\boldsymbol{k}$ as

$$\psi_{n\boldsymbol{k}}(P^{-1}\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(P^{-1}\boldsymbol{r})}u_{n\boldsymbol{k}}(P^{-1}\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}(P\boldsymbol{k})\cdot\boldsymbol{r}}u'_{n\boldsymbol{k}}(\boldsymbol{r})\,. \tag{18.1.38}$$

Now select those symmetry elements T that take $\boldsymbol{k}$ into itself or into an equivalent vector ($T\boldsymbol{k} \equiv \boldsymbol{k}$). These symmetry elements constitute the *little group* of $\boldsymbol{k}$. A representation of this little group can be constructed on the basis of the functions $T\psi_{n\boldsymbol{k}}(\boldsymbol{r})$. According to Wigner's theorem, unless a special potential is chosen, accidental degeneracies can be ignored, and the representation must be irreducible. The bands can then be characterized by the irreducible representation of the little group of $\boldsymbol{k}$ according to which the wavefunction transforms. For this reason, the band structure is often specified by indexing the bands by letters that refer to the irreducible representations of the states. The degeneracy of a band is determined by the dimension of the irreducible representation.

Whenever the energies are the same for some wavefunctions but the representation of the little group on these functions is reducible, the degeneracy is accidental: it is an artefact of the approximation, e.g., the consequence of a particular, oversimplified choice of the potential. The degeneracy of the energy levels is removed by choosing a more general potential that nevertheless respects the symmetries of the system. Consequently, whether the degeneracy at a particular wave vector $\boldsymbol{k}$ is dictated by underlying symmetries can be determined by representing the little group of $\boldsymbol{k}$ on the space of the degenerate wavefunctions and reducing the representation. The states whose wavefunctions belong to the same irreducible representation are necessarily of the same energy. These degeneracies are thus required by symmetry. On the other hand, the energy of the wavefunctions that belong to different irreducible representations can be the same only accidentally.

By way of example, let us consider once again a simple cubic crystal – whose empty-lattice band structure was determined in the previous subsection. Figure 18.3 shows the energies of the electrons along two directions of the Brillouin zone. In the $\varGamma X$ direction the states D, E, F, and G form a fourfold degenerate band, while along the direction $\varGamma R$ the bands B, F, G and C, D, E, respectively, are degenerate. Further degeneracies appear at the center and boundaries of the Brillouin zone in the empty-lattice approximation. The question is: What happens to these degeneracies when the periodic potential is turned on? Are they removed, or preserved as a result of an underlying symmetry? We shall use group-theoretical methods to answer this question.

Let us first examine what happens at the center of the Brillouin zone, in point $\varGamma$. The state that belongs to the lowest band, A, is not degenerate, while the bands B, C, D, E, F, and G are all of the same energy in $\varGamma$. Is this degeneracy symmetry-related?

In the empty-lattice approximation the wavefunctions that correspond to the six $k = 0$ states are readily expressed in terms of the reciprocal-lattice vectors associated with the branches (see Sec. 18.1.1):

$$\begin{aligned} &\phi_{\mathrm{B}} = \mathrm{e}^{-2\pi \mathrm{i} z/a}\,, &&\phi_{\mathrm{C}} = \mathrm{e}^{2\pi \mathrm{i} z/a}\,, &&\phi_{\mathrm{D}} = \mathrm{e}^{2\pi \mathrm{i} x/a}\,,\\ &\phi_{\mathrm{E}} = \mathrm{e}^{-2\pi \mathrm{i} x/a}\,, &&\phi_{\mathrm{F}} = \mathrm{e}^{2\pi \mathrm{i} y/a}\,, &&\phi_{\mathrm{G}} = \mathrm{e}^{-2\pi \mathrm{i} y/a}\,. \end{aligned} \tag{18.1.39}$$

Next, the elements of the little group of the point $\varGamma = (0, 0, 0)$ are represented using these functions. The little group contains 48 elements, as all symmetry operations of the cube – listed in Tables 5.1 and 5.4 – take the vector $\boldsymbol{k} = 0$ into itself. As discussed in Chapter 6, the 48 group elements constitute 10 classes, and so there are 10 irreducible representations: 4 one-dimensional, 2 two-dimensional, and 4 three-dimensional. Their characters are given in Table D.1 of Volume 1.

As there are no six-dimensional irreducible representations, the little group obviously transforms reducibly on the space of the functions given in (18.1.39). To reduce it, consider the representation matrix for one typical element of each class:

$$\begin{gathered}
D(E)=\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}, \qquad D(C_{4z})=\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&0&0&0&1\\0&0&0&0&1&0\\0&0&1&0&0&0\\0&0&0&1&0&0\end{pmatrix},\\
D(C_{2z})=\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&0&1&0&0\\0&0&1&0&0&0\\0&0&0&0&0&1\\0&0&0&0&1&0\end{pmatrix}, \qquad D(C_{2a})=\begin{pmatrix}0&1&0&0&0&0\\1&0&0&0&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\\0&0&1&0&0&0\\0&0&0&1&0&0\end{pmatrix},\\
D(C_{3a})=\begin{pmatrix}0&0&0&0&0&1\\0&0&0&0&1&0\\0&1&0&0&0&0\\1&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\end{pmatrix}, \qquad D(I)=\begin{pmatrix}0&1&0&0&0&0\\1&0&0&0&0&0\\0&0&0&1&0&0\\0&0&1&0&0&0\\0&0&0&0&0&1\\0&0&0&0&1&0\end{pmatrix}.
\end{gathered} \tag{18.1.40}$$

The characters can be read off immediately, and are listed in Table 18.1.

Table 18.1. Characters of the representation of the little group of point Γ on the six degenerate functions

	E	$3C_{2m}$	$6C_{4m}$	$6C_{2p}$	$8C_{3j}$	I	$3IC_{2m}$	$6IC_{4m}$	$6IC_{2p}$	$8IC_{3j}$
χ	6	2	2	0	0	0	4	0	2	0

When the reduction is performed using (D.1.30), the representation is decomposed into the sum of a one-dimensional (Γ_1), a two-dimensional (Γ_{12}), and a three-dimensional (Γ_{15}) representation. The wavefunctions serving as basis can also be determined using (D.1.43). The basis function for the one-dimensional representation is

$$\psi_{\Gamma_1}=\cos(2\pi x/a)+\cos(2\pi y/a)+\cos(2\pi z/a)\,, \tag{18.1.41}$$

those of the two-dimensional are

$$\begin{aligned}\psi^{(1)}_{\Gamma_{12}} &= \cos(2\pi x/a) - \cos(2\pi y/a)\,,\\ \psi^{(2)}_{\Gamma_{12}} &= \cos(2\pi z/a) - \tfrac{1}{2}\big[\cos(2\pi x/a) + \cos(2\pi y/a)\big]\,,\end{aligned} \qquad (18.1.42)$$

and those of the three-dimensional are

$$\psi^{(1)}_{\Gamma_{15}} = \sin(2\pi x/a)\,, \quad \psi^{(2)}_{\Gamma_{15}} = \sin(2\pi y/a)\,, \quad \psi^{(3)}_{\Gamma_{15}} = \sin(2\pi z/a)\,. \qquad (18.1.43)$$

Note that ψ_{Γ_1} possesses s-type symmetry, the three $\psi_{\Gamma_{15}}$ p-type symmetry, while the two $\psi_{\Gamma_{12}}$ the symmetries of the functions $d_{x^2-y^2}$ and d_{z^2}. These symmetries can be used to label the bands.

Since states characterized by functions that belong to the same irreducible representation are of the same energy, no matter how the periodic potential is chosen (as long as it respects the symmetry), a doubly and a triply degenerate energy level persist even after it is turned on. On the other hand, by choosing an arbitrary potential of cubic symmetry, the matrix element of the potential between wavefunctions that belong to different irreducible representations will vanish. These states are not mixed by the potential. Therefore, accidental symmetries aside, the energies of the three irreducible representations are different in general.

Let us now examine what happens when the points $\boldsymbol{k} = (\pi/a)(0,0,\xi)$ of line Δ connecting Γ and X are considered. The elements of the little group for line Δ are E, C_{4z}, C_{2z}, C^3_{4z}, $IC_{2x} = \sigma_x$, $IC_{2y} = \sigma_y$, $IC_{2a} = \sigma_a$, and $IC_{2b} = \sigma_b$. Inversion itself is not a symmetry element any more, only in combination with rotations around axes lying in the xy-plane. The eight elements of the group can be divided into five classes, thus the little group has a two-dimensional and four one-dimensional irreducible representations. The character table of the irreducible representations is given in Table 18.2. Each class is represented by a typical element, and the number of elements is also given.

Table 18.2. Character table for the irreducible representations of the little group of line Δ

	E	C_{2z}	$2C_{4z}$	$2IC_{2x}$	$2IC_{2a}$
Δ_1	1	1	1	1	1
Δ_2	1	1	-1	1	-1
Δ_2'	1	1	-1	-1	1
Δ_1'	1	1	1	-1	-1
Δ_5	2	-2	0	0	0

Now consider the wavefunctions obtained in the empty-lattice approximation. Since branch A belongs to vector $\boldsymbol{G} = 0$, the wavefunction of the state associated with the vector $\boldsymbol{k} = (\pi/a)(0,0,\xi)$ is

$$\psi_{\mathrm{A}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} = \mathrm{e}^{\mathrm{i}\pi\xi z/a}\,. \qquad (18.1.44)$$

This wavefunction is invariant under the action of the symmetry elements of the little group of line $\varDelta$ (listed above), that is, the wavefunction transforms according to the irreducible representation $\varDelta_1$. This representation is one-dimensional, therefore, inside the Brillouin zone a nondegenerate band is formed by these states.

The branch marked B belongs to the vector $\boldsymbol{G} = (2\pi/a)(0,0,\bar{1})$, and its wavefunction is

$$\psi_{\mathrm{B}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\pi(\xi-2)z/a}\,. \tag{18.1.45}$$

A similar expression is found for the wavefunction of branch C:

$$\psi_{\mathrm{C}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\pi(\xi+2)z/a}\,. \tag{18.1.46}$$

The symmetry elements of the little group of point $\varDelta$ leave these wavefunctions invariant as well – so they, too, transform according to the one-dimensional representation $\varDelta_1$. This gives two more nondegenerate bands inside the Brillouin zone.

In the empty-lattice approximation the bands D, E, F, and G are degenerate along the line $\varDelta$ $(\varGamma X)$. The wavefunctions that correspond to the four bands are

$$\begin{aligned} \psi_{\mathrm{D}}(\boldsymbol{r}) &= \mathrm{e}^{\mathrm{i}\pi\xi z/a}\,\mathrm{e}^{2\pi\mathrm{i}x/a}\,, & \psi_{\mathrm{E}}(\boldsymbol{r}) &= \mathrm{e}^{\mathrm{i}\pi\xi z/a}\,\mathrm{e}^{-2\pi\mathrm{i}x/a}\,,\\ \psi_{\mathrm{F}}(\boldsymbol{r}) &= \mathrm{e}^{\mathrm{i}\pi\xi z/a}\,\mathrm{e}^{2\pi\mathrm{i}y/a}\,, & \psi_{\mathrm{G}}(\boldsymbol{r}) &= \mathrm{e}^{\mathrm{i}\pi\xi z/a}\,\mathrm{e}^{-2\pi\mathrm{i}y/a}\,. \end{aligned} \tag{18.1.47}$$

By choosing these four functions as the basis of the representation, a reducible representation of the little group is obtained. This representation is the direct sum of the irreducible representations $\varDelta_1$, $\varDelta_2$, and $\varDelta_5$, so the fourfold degenerate level splits into two nondegenerate and a doubly degenerate level. By determining the linear combinations that can be chosen as basis functions of the irreducible representations, the function

$$\psi_{\varDelta_1} = \mathrm{e}^{\mathrm{i}\pi\xi z/a}\big[\cos(2\pi x/a) + \cos(2\pi y/a)\big] \tag{18.1.48}$$

is seen to transform according to the representation $\varDelta_1$, the function

$$\psi_{\varDelta_2} = \mathrm{e}^{\mathrm{i}\pi\xi z/a}\big[\cos(2\pi x/a) - \cos(2\pi y/a)\big] \tag{18.1.49}$$

according to $\varDelta_2$, while the linear combinations for the two-dimensional representation $\varDelta_5$ are

$$\psi^{(1)}_{\varDelta_5} = \mathrm{e}^{\mathrm{i}\pi\xi z/a}\sin(2\pi x/a) \qquad \text{and} \qquad \psi^{(2)}_{\varDelta_5} = \mathrm{e}^{\mathrm{i}\pi\xi z/a}\sin(2\pi y/a)\,. \tag{18.1.50}$$

In our previous calculations we saw that the sixfold degenerate level in point $\varGamma = (0,0,0)$ arising from the states B, C, D, E, F, and G in the empty-lattice approximation splits into levels that transform according to the one-dimensional $\varGamma_1$, the two-dimensional $\varGamma_{12}$, and the three-dimensional $\varGamma_{15}$ representations when the periodic potential is turned on. Close to $\varGamma$ and along the line $\varDelta$, the one-dimensional representation $\varDelta_1$ appears three times, and both the one-dimensional $\varDelta_2$ and the three-dimensional $\varDelta_5$ once. How does this splitting occur when $\boldsymbol{k}$ moves from $\varGamma$ toward X? There are two possibilities. Either the degeneracy of the level that is doubly degenerate in $\varGamma$ is conserved and the triply degenerate level splits into three nondegenerate ones, or the doubly degenerate level splits into two nondegenerate ones and the triply degenerate level into a nondegenerate and a doubly degenerate one. We shall now show that the answer can be deduced from the compatibility conditions for the irreducible representations, regardless the specific form of the potential.

The little group of line $\varDelta$ is a subgroup of the little group of point $\varGamma$. The irreducible representations of point $\varGamma$ can be reducible on this subgroup, and so they may be reduced to the irreducible representations of $\varDelta$. The irreducible representations that occur in this reduction procedure are said to be compatible with the irreducible representation of point $\varGamma$.

A comparison of the character tables reveals that $\varGamma_1$ is compatible with $\varDelta_1$. By breaking the cubic symmetry, the function

$$\psi_{\varDelta_1} = a_1(\xi)\cos(2\pi z/a) + a_2(\xi)\big[\cos(2\pi x/a) + \cos(2\pi y/a)\big]\,, \qquad (18.1.51)$$

which transforms according to the representation $\varDelta_1$, is obtained from function (18.1.41), which transforms according to the representation $\varGamma_1$. In $\varGamma$, $a_1(0) = a_2(0)$, while in other points of line $\varDelta$ these coefficients take different values, depending on the potential and the wave number. Naturally, the coefficients change continuously with the wave number.

The representation $\varGamma_{12}$ is compatible with $\varDelta_1$ and $\varDelta_2$, since $\chi_{\varGamma_{12}}(P) = \chi_{\varDelta_1}(P) + \chi_{\varDelta_2}(P)$ holds for the character of all symmetry elements P of the line $\varDelta$. Therefore functions (18.1.42) that transform according to $\varGamma_{12}$ in point $\varGamma$ should continuously evolve into functions that transform according to $\varDelta_1$ and $\varDelta_2$ along line $\varGamma X$. It is readily seen that one of the basis functions,

$$\cos(2\pi x/a) - \cos(2\pi y/a)\,, \qquad (18.1.52)$$

transforms according to $\varDelta_2$. Starting with the other basis function of $\varGamma_{12}$ in $\varGamma$, it then evolves into the function

$$a_2(\xi)\cos(2\pi z/a) - \tfrac{1}{2}a_1(\xi)\big[\cos(2\pi x/a) + \cos(2\pi y/a)\big]\,, \qquad (18.1.53)$$

which transforms according to $\varDelta_1$, and is orthogonal to the function in (18.1.51) that also transforms according to $\varDelta_1$.

For the symmetry elements of the line $\varDelta$, the representation $\varGamma_{15}$ is decomposed into the irreducible representations $\varDelta_1$ and $\varDelta_5$. Of the three basis functions in (18.1.43), $\sin(2\pi z/a)$ is separated from the others. The level that is triply degenerate in point $\varGamma$ splits into a nondegenerate and a doubly degenerate level along line $\varDelta$. These compatibility relations are summarized in Table 18.3.

Table 18.3. Compatibility of the irreducible representations of lines $\varDelta$ and $\varLambda$ and the irreducible representations of point $\varGamma$

$\varGamma_1$	$\varGamma_2$	$\varGamma_{12}$	$\varGamma_{15}$	$\varGamma_{25}$
$\varDelta_1$	$\varDelta_2$	$\varDelta_1\ \varDelta_2$	$\varDelta_1\ \varDelta_5$	$\varDelta_2\ \varDelta_5$
$\varLambda_1$	$\varLambda_2$	$\varLambda_3$	$\varLambda_1\ \varLambda_3$	$\varLambda_2\ \varLambda_3$

This means that the symmetry-related degeneracy in point $\varGamma$ is lifted in nearby points as described by the second option above. Just like Fig. 18.3, the left panel of Fig. 18.15 also shows the band structure obtained in the empty-lattice approximation along the line joining $\varGamma$ and X, however bands are now indexed according to the irreducible representations of the states. The right-hand side shows the band structure when band splitting is also taken into account. The diagram is schematic: the character of the splitting can be determined by symmetry considerations alone but the amount of splitting and the order of levels cannot.

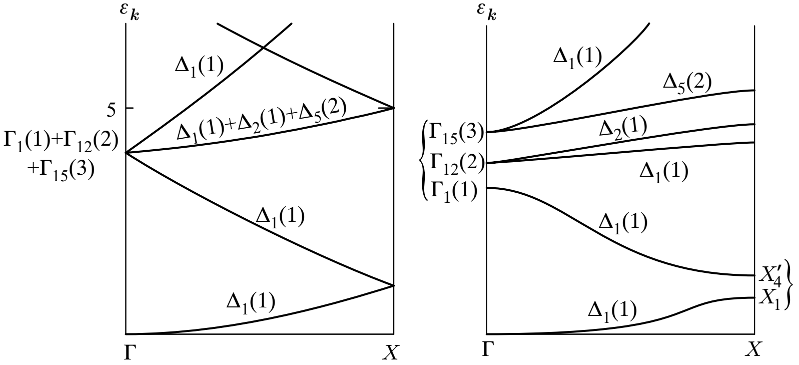


Fig. 18.15. Band structure in the empty-lattice approximation and after the lifting of degeneracies along the line $\varGamma X$ in a simple cubic crystal. Bands are indexed by the irreducible representations of the states. The numbers in the brackets show the dimension of the representation, i.e., the degree of degeneracy for the level

We shall examine what exactly happens to these bands in point X only for the two lowest-energy bands, denoted by A and B in the empty-lattice approximation. In the absence of the periodic potential they are degenerate in point X, however this degeneracy is expected to be removed when the periodic potential is turned on. We shall prove that this splitting must indeed occur for symmetry reasons. For this, we have to demonstrate that the functions $\psi_A(\boldsymbol{r})$ and $\psi_B(\boldsymbol{r})$ are not the basis functions of a two-dimensional irreducible representation in point X – i.e., the representation of the little group of the symmetries of point X is reducible, and can be decomposed into two one-dimensional irreducible representations. As the wavefunctions are not related by any symmetry, the corresponding energies need not be equal.

The symmetry is higher in point X at the zone boundary than along the line Δ inside the zone because of the new symmetry operations, e.g., inversion, that transform point $X = (\pi/a)(0,0,1)$ into the equivalent point $X' = (\pi/a)(0,0,\bar{1})$. The symmetry elements making up the little group of X are thus: the identity element E, the fourfold rotation C_{4z} around [001], the twofold rotations C_{2x} and C_{2y} around [100] and [010], the twofold rotations C_{2a} and C_{2b} around [110] and [$\bar{1}$10], the inversion I, and the combination of inversion and the rotations. The 16 symmetry elements are divided into 10 classes – therefore the little group has 10 irreducible representations: 8 one-dimensional and 2 two-dimensional. The character table of the representations is given in Table 18.4.

Table 18.4. Character table of the irreducible representations for the little group of points X

	E	C_{2z}	$2C_{4z}$	$2C_{2x}$	$2C_{2a}$	I	IC_{2z}	$2IC_{4z}$	$2IC_{2x}$	$2IC_{2a}$
X_1	1	1	1	1	1	1	1	1	1	1
X_2	1	1	−1	1	−1	1	1	−1	1	−1
X_3	1	1	−1	−1	1	1	1	−1	−1	1
X_4	1	1	1	−1	−1	1	1	1	−1	−1
X_5	2	−2	0	0	0	2	−2	0	0	0
X'_1	1	1	1	1	1	−1	−1	−1	−1	−1
X'_2	1	1	−1	1	−1	−1	−1	1	−1	1
X'_3	1	1	−1	−1	1	−1	−1	1	1	−1
X'_4	1	1	1	−1	−1	−1	−1	−1	1	1
X'_5	2	−2	0	0	0	−2	2	0	0	0

Consider the representation of this group on the space of the unperturbed functions ψ_{A} and ψ_{B}. The wavefunctions at point $X = (\pi/a)(0,0,1)$ of the Brillouin zone are

$$\psi_{\mathrm{A}}(\boldsymbol{r}) = \mathrm{e}^{\pi \mathrm{i} z/a}\,, \qquad \psi_{\mathrm{B}}(\boldsymbol{r}) = \mathrm{e}^{-\pi \mathrm{i} z/a}\,. \tag{18.1.54}$$

It is straightforward to calculate the matrices of the representation by applying the symmetry operations of the little group on these wavefunctions. Since we are interested only in the characters of the representation, it is sufficient to write them down for a single element in each of the 10 classes:

$$
\begin{aligned}
D(E) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad D(C_{2z}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad D(C_{4z}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\
D(C_{2x}) &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad D(C_{2a}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\
D(I) &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad D(IC_{2z}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad D(IC_{4z}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\
D(IC_{2x}) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad D(IC_{2a}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\end{aligned}
$$

The characters can be immediately read off from the matrices; the results are listed in Table 18.5.

Table 18.5. Characters of the representation of the little group of point X on the functions ψ_A and ψ_B

	E	C_{2z}	$2C_{4z}$	$2C_{2x}$	$2C_{2a}$	I	IC_{2z}	$2IC_{4z}$	$2IC_{2x}$	$2IC_{2a}$
χ	2	2	2	0	0	0	0	0	2	2

Comparison of Tables 18.4 and 18.5 shows that these characters are just the sums of the corresponding characters of the irreducible representations X_1 and X_4'. According to the addition theorem for characters, the above representation is the direct sum of these two irreducible representations. The wavefunctions (the basis functions of the irreducible representations) can also be obtained easily. It is straightforward to show that the linear combinations $\psi_{X_1} = \cos(\pi z/a)$ and $\psi_{X_4'} = \sin(\pi z/a)$ transform according to X_1 and X_4', respectively.

Since the two wavefunctions belong to different irreducible representations, the matrix element of the lattice potential $U(\boldsymbol{r})$ vanishes between them:

$$
\int \psi_{X_1}^*(\boldsymbol{r}) U(\boldsymbol{r}) \psi_{X_4'}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} = 0 . \tag{18.1.55}
$$

Using this basis, it is not necessary to make recourse to degenerate perturbation theory. In the first order of perturbation theory the energy correction can be calculated separately for the two wavefunctions from

$$
\varepsilon_i^{(1)} = \frac{\int \psi_i^*(\boldsymbol{r}) U(\boldsymbol{r}) \psi_i(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r}}{\int \psi_i^*(\boldsymbol{r}) \psi_i(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r}} , \qquad i = X_1, X_4' . \tag{18.1.56}
$$

The levels are indeed split in point X.

In our earlier considerations we used the functions

$$
\psi_{\mathrm{A}}(\boldsymbol{r}) = \mathrm{e}^{\pi \mathrm{i} \xi z/a} , \qquad \psi_{\mathrm{B}}(\boldsymbol{r}) = \mathrm{e}^{\pi \mathrm{i}(\xi - 2) z/a} , \tag{18.1.57}
$$

given in (18.1.44) and (18.1.45) for the states of bands A and B in general points of line Δ. However, for $\xi \to 1$ they do not tend continuously to the functions obtained in point X,

$$\psi_{X_1} = \cos(\pi z/a)\,, \qquad \psi_{X_4'} = \sin(\pi z/a)\,, \tag{18.1.58}$$

even though the wavefunctions – just like the energy – must change continuously in a band. This contradiction can be resolved by noting that the functions ψ_{A} and ψ_{B} transform according to the same representation, consequently one may just as well take any linear combination. A more appropriate form of the wavefunctions satisfying the orthogonality condition is

$$\begin{aligned}\psi_{\mathrm{A}}(\boldsymbol{r}) &= a_1(\xi)\mathrm{e}^{\mathrm{i}\pi\xi z/a} + a_2(\xi)\mathrm{e}^{\mathrm{i}\pi(\xi-2)z/a}\\ &= \mathrm{e}^{\mathrm{i}\pi\xi z/a}\left[a_1(\xi) + a_2(\xi)\mathrm{e}^{-2\pi\mathrm{i}z/a}\right],\end{aligned} \tag{18.1.59-a}$$

$$\begin{aligned}\psi_{\mathrm{B}}(\boldsymbol{r}) &= -a_2(\xi)\mathrm{e}^{\mathrm{i}\pi\xi z/a} + a_1(\xi)\mathrm{e}^{\mathrm{i}\pi(\xi-2)z/a}\\ &= \mathrm{e}^{\mathrm{i}\pi\xi z/a}\left[-a_2(\xi) + a_1(\xi)\mathrm{e}^{-2\pi\mathrm{i}z/a}\right].\end{aligned} \tag{18.1.59-b}$$

The coefficients a_1 and a_2 depend on the strength of the interaction and the location of the point Δ (ξ). In the empty-lattice approximation $a_1 = 1$ and $a_2 = 0$, and the wavefunctions given in (18.1.44) and (18.1.45) are recovered. These values are modified by the lattice potential. Far from point X, where the difference of the energies is large, there is hardly any mixing: $a_1 \approx 1$ and $a_2 \approx 0$. As X is approached, a_1 decreases and a_2 increases, and the two become equal in X ($\xi = 1$). The wavefunctions (18.1.58) are then recovered.

Using a similar procedure, it is straightforward to calculate the degree of the symmetry-related (nonaccidental) degeneracy of the band states and the lifting of the degeneracy with respect to the empty-lattice approximation shown in Fig. 18.3 for points $\Lambda = (\pi/a)(\xi,\xi,\xi)$ along the line joining Γ and R. In addition to the identity element E, the little group of the line Λ contains rotations through 120° and 240° around the direction [111], and reflections in the planes of normal $[1\bar{1}0]$, $[10\bar{1}]$, and $[01\bar{1}]$. The character table of the irreducible representations of the little group is given in Table 18.6.

Table 18.6. Character table of the irreducible representations for the little group of line Λ

	E	$2C_3$	$3IC_{2p}$
Λ_1	1	1	1
Λ_2	1	1	-1
Λ_3	2	-1	0

It is readily seen that the representation Γ_1 is compatible with Λ_1, Γ_{12} with Λ_3, and Γ_{15} with Λ_1 and Λ_3. This means that the nondegenerate state Γ_1 goes over continuously into a state Λ_1 as $\boldsymbol{k}$ moves from the zone center toward R. Similarly, the doubly degenerate state Γ_{12} goes over continuously into the state Λ_3 and remains doubly degenerate. However, the triply degenerate state Γ_{15} is split into a nondegenerate level Λ_1 and a doubly degenerate level Λ_3 along the line Λ. Just like Fig. 18.15, Fig. 18.16 also shows the band structure presented in Fig. 18.3, however bands are now indexed by the irreducible representations of the states, while the right-hand side shows the band structure obtained by taking splitting into account.

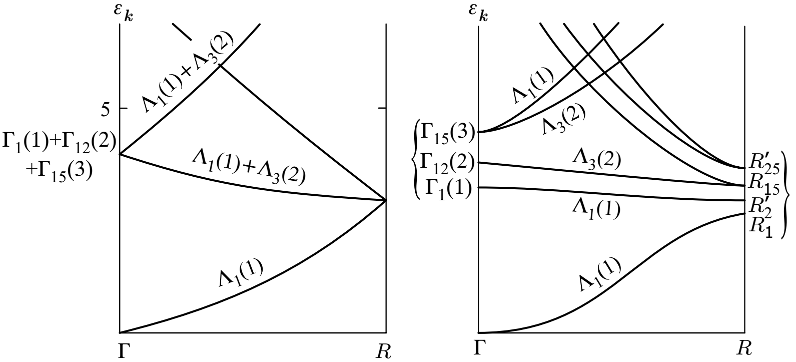


Fig. 18.16. Band structure in the empty-lattice approximation and after the lifting of degeneracies along the line ΓR in a simple cubic crystal. Bands are indexed by the irreducible representations of the states

The figure also indicates what may happen to the split levels in point R. It can be demonstrated that the lowest-lying, eightfold degenerate level obtained in the empty-lattice approximation splits into two nondegenerate and two triply degenerate levels in R when the periodic potential is turned on. Then each triply degenerate level splits further into a nondegenerate and a doubly degenerate level along the line ΓR. Using group-theoretical considerations alone, it is impossible to determine which state of point R will correspond to a particular band starting from Γ.

18.1.5 Fermi Surface for Nearly Free Electrons

Even though it contains less information, one often visualizes the band structure in higher dimensions by displaying the constant-energy surfaces rather than the dispersion relation. For free electrons these surfaces are spherical. Represented in the extended-zone scheme, these surfaces are also spherical in the nearly-free-electron model, as long as they are far from the zone boundaries. When the boundaries are approached, the surfaces become distorted so that they can reach the boundary perpendicularly, as the $\boldsymbol{k}$-space gradient of

$\varepsilon_{n\boldsymbol{k}}$ must lie in a Bragg plane, in the zone boundary. Constant-energy surfaces in an empty lattice and in the presence of a weak potential are shown in Fig. 18.17 for a two-dimensional square lattice.

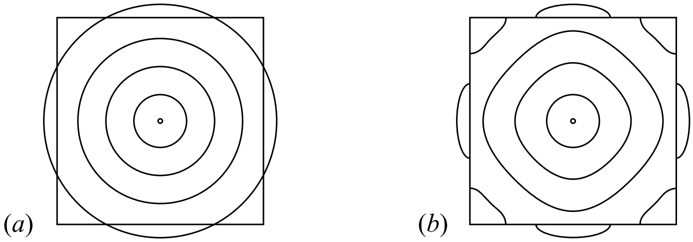


Fig. 18.17. Constant-energy surfaces in a square lattice (a) in the empty-lattice approximation and (b) in the nearly-free-electron approximation

Among the constant-energy surfaces the one that corresponds to the Fermi energy – the Fermi surface – is particularly important. For a given band structure the energy of the highest occupied levels in the ground state – and so the shape of the Fermi surface – depends on the number of electrons and the filling of the band. Figure 18.18 shows the Fermi surface in the extended-zone scheme for a two-dimensional square lattice at three different electron numbers. Comparison with Fig. 18.5 shows the distortion of the Fermi surface due to the periodic potential.

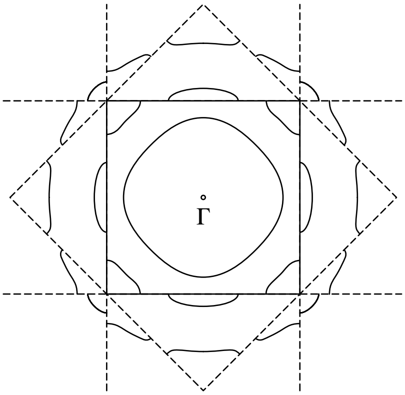


Fig. 18.18. Fermi surface distorted by a weak periodic potential in the extended-zone scheme for a two-dimensional system with one, two, and four electrons per primitive cell

Represented in the reduced-zone scheme, the shape of the Fermi surface is shown in Fig. 18.19. In some cases the occupied states are drawn around points M or X rather than Γ.

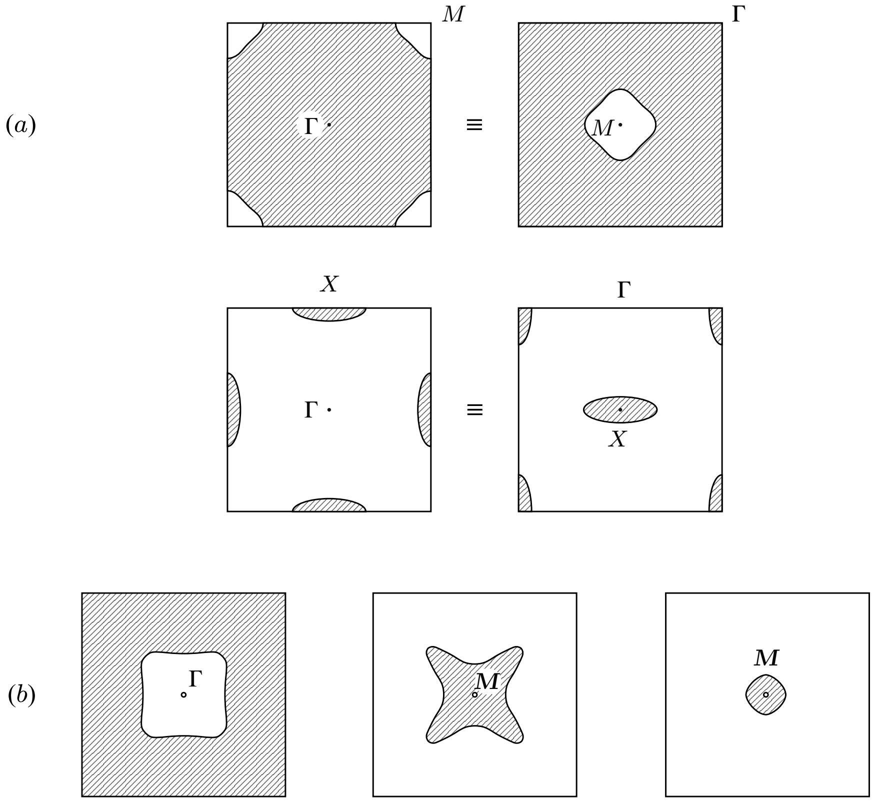


Fig. 18.19. Fermi surface in a square lattice, in the presence of a weak periodic potential, represented in the reduced-zone scheme for (a) two and (b) four electrons per primitive cell

Compared to the Fermi surfaces obtained in the empty-lattice approximation and shown in Figs. 18.6, 18.8, and 18.9, it is readily seen that the periodic potential and the resulting distortion of the Fermi surface round off the sharp corners of the Fermi surface obtained in the empty-lattice approximation. Owing to this distortion more states may be accommodated below the Fermi energy in the second and third Brillouin zones than in an empty lattice, and thus the number of occupied states may be reduced, or may even vanish in the fourth Brillouin zone. The total volume enclosed by the Fermi surface – which is the sum of the parts in the various zones – is, nevertheless, independent of the strength of the periodic potential, in agreement with Luttinger's theorem,[3] which stipulates that the interaction does not change the volume of the $\boldsymbol{k}$-space region enclosed by the Fermi surface.

The situation is similar in three dimensions. The constant-energy surfaces are almost spherical for energies close to the bottom of the lowest-lying band, and the states are essentially free-electron-like. For higher energies, where the constant-energy surface gets close to the zone boundaries, its spherical shape becomes distorted. Figure 18.20 shows this in a section of the Brillouin zone of a face-centered cubic crystal for two energy values.

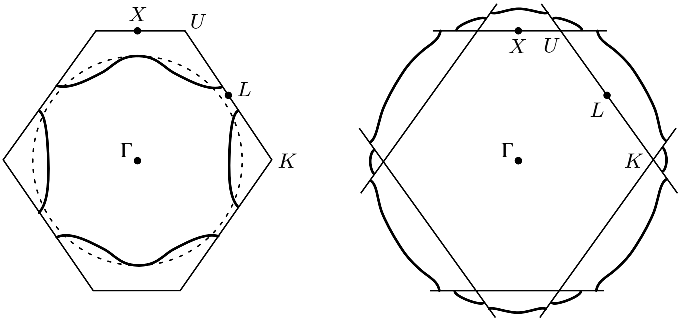


Fig. 18.20. The distortion of the constant-energy surfaces in a face-centered cubic lattice in the nearly-free-electron approximation at two different energies and occupation values

Similar methods can be used for three-dimensional systems to derive the realistic Fermi surface, distorted by the periodic potential, from the Fermi surface obtained in an empty three-dimensional lattice using the Harrison construction. For symmetry reasons, the Fermi surface must be perpendicular to those boundaries of the Brillouin zone that are separated from their mirror images by a reciprocal-lattice vector.

18.2 Tight-Binding Approximation

In the previous section the band structure of electrons moving in a crystal was presented using free electrons as the starting point. A completely different – and, in a sense, exactly opposite – approach to calculate the band structure is obtained when the electrons are supposed to be bound to atoms, their wavefunction to be that of an atomic eigenstate, and the effects of other ions are treated as a perturbation.

[3] J. M. Luttinger, 1960.

18.2.1 Broadening of Atomic Levels into Bands

To understand the formation of the band structure imagine that the crystal is constructed by placing the atoms in a crystalline order, however, with a very large initial lattice constant, and then the lattice constant is gradually reduced. If the atom at $\boldsymbol{R}_j$ acts on the surrounding electrons via a potential $v_\mathrm{a}(\boldsymbol{r}-\boldsymbol{R}_j)$, then the atomic wavefunctions are the solutions of the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 + v_\mathrm{a}(\boldsymbol{r}-\boldsymbol{R}_j)\right] w_\alpha(\boldsymbol{r}-\boldsymbol{R}_j) = \varepsilon_\alpha w_\alpha(\boldsymbol{r}-\boldsymbol{R}_j)\,. \tag{18.2.1}$$

Because of the spherical symmetry of the atomic potential, states are characterized, as usual, by the quantum numbers n, l, and m_l, which we shall collectively denote by α. As long as the atoms are far from each other, the same atomic energy levels ε_α appear for each atom. For N atoms the levels are at least N-fold degenerate. Electrons occupy these highly degenerate atomic states.

When the lattice constant is reduced, the overlap between atomic wavefunctions gradually increases, energies become shifted from the atomic values, and the multiple degeneracy is lifted. As shown in Fig. 18.21 schematically, the energies form broader and broader bands as the lattice constant decreases.

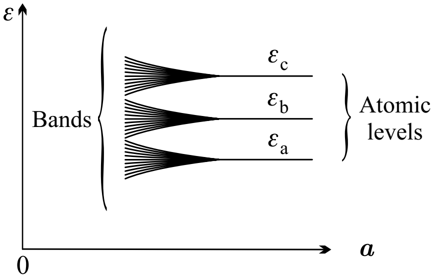


Fig. 18.21. The formation of energy bands from atomic levels as the lattice constant is reduced

If the overlap is not strong even when the real lattice constant is reached, and the mixing of atomic states of unequal energies can be ignored, then, following BLOCH's proposal,[4] the wavefunctions associated with the band states can be written as linear combinations of the atomic wavefunctions:

$$\psi_\alpha(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} w_\alpha(\boldsymbol{r}-\boldsymbol{R}_j) \tag{18.2.2}$$

This form satisfies the relation (17.1.6) for Bloch functions as

[4] F. BLOCH, 1928.

$$\begin{aligned}\psi_\alpha(\boldsymbol{k},\boldsymbol{r}+\boldsymbol{R}_n) &= \frac{1}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} w_\alpha(\boldsymbol{r}+\boldsymbol{R}_n-\boldsymbol{R}_j)\\ &= \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_n}\frac{1}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_j-\boldsymbol{R}_n)} w_\alpha(\boldsymbol{r}-(\boldsymbol{R}_j-\boldsymbol{R}_n))\\ &= \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_n}\psi_\alpha(\boldsymbol{k},\boldsymbol{r})\,. \end{aligned} \tag{18.2.3}$$

This representation of the wavefunction resembles the relation between Bloch and Wannier functions. However, as the atomic functions of neighboring atoms are not orthogonal, in contrast to the Wannier functions, these are not genuine Bloch functions, since they do not satisfy the orthogonality condition (17.1.18).

The correct normalization of the wavefunction is ensured by a coefficient $c \neq 1$, that is,

$$\psi_\alpha(\boldsymbol{k},\boldsymbol{r}) = \frac{c}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} w_\alpha(\boldsymbol{r}-\boldsymbol{R}_j)\,. \tag{18.2.4}$$

In what follows, we shall assume that the atomic wavefunctions drop off sufficiently rapidly so that there is no significant overlap between them even when the atoms are nearest neighbors. The normalization factor can then be approximated by unity.

The method based on the wavefunctions built up of atomic states in this way is called the *tight-binding approximation.* According to the foregoing, this approach can also be considered as an approximation of the orthogonal Wannier functions by the atomic wavefunctions.

There is no unknown parameter in the chosen wavefunction, which is generally not an exact eigenstate either. The energy of the state is therefore given by

$$\varepsilon_\alpha(\boldsymbol{k}) = \frac{\displaystyle\int \psi^*_\alpha(\boldsymbol{k},\boldsymbol{r})\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right]\psi_\alpha(\boldsymbol{k},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}}{\displaystyle\int \psi^*_\alpha(\boldsymbol{k},\boldsymbol{r})\psi_\alpha(\boldsymbol{k},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}}\,. \tag{18.2.5}$$

Inserting (18.2.2) into this expression, and assuming that overlaps can be neglected for normalization purposes,

$$\varepsilon_\alpha(\boldsymbol{k}) \approx \frac{1}{N}\sum_{j,j'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_j-\boldsymbol{R}_{j'})}\int w^*_\alpha(\boldsymbol{r}-\boldsymbol{R}_{j'})\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right] w_\alpha(\boldsymbol{r}-\boldsymbol{R}_j)\,\mathrm{d}\boldsymbol{r}\,. \tag{18.2.6}$$

One sum in the double sum over lattice points can be calculated, since the integral

$$\int w^*_\alpha(\boldsymbol{r}-\boldsymbol{R}_{j'})\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right] w_\alpha(\boldsymbol{r}-\boldsymbol{R}_j)\,\mathrm{d}\boldsymbol{r} \tag{18.2.7}$$

depends only on the difference $\boldsymbol{R}_l = \boldsymbol{R}_j - \boldsymbol{R}_{j'}$ on account of the periodicity of $U(\boldsymbol{r})$. Then

$$\varepsilon_\alpha(\boldsymbol{k}) \approx \sum_l \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l} \int w_\alpha^*(\boldsymbol{r}+\boldsymbol{R}_l) \left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r})\right] w_\alpha(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{18.2.8}$$

Because of the weak overlap between atomic wavefunctions the largest contribution obviously comes from the term $\boldsymbol{R}_l = 0$. If all other contributions were zero, electrons would be bound to atoms, and would not propagate in the lattice. Energies would then be independent of $\boldsymbol{k}$. The weak, nonetheless finite overlap, which is irrelevant for normalization, becomes essential when the dispersion relation is determined, since it allows electrons to hop from one atom to its neighbor.

The energy formula can be further simplified by making use of the Schrödinger equation (18.2.1) of the atomic problem:

$$\varepsilon_\alpha(\boldsymbol{k}) \approx \varepsilon_\alpha + \sum_l \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l} \int w_\alpha^*(\boldsymbol{r}+\boldsymbol{R}_l)\,[U(\boldsymbol{r}) - v_\mathrm{a}(\boldsymbol{r})]\,w_\alpha(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{18.2.9}$$

The first term on the right-hand side is the energy of the electron in the atomic state. It looks as if the absence of overlap of atomic wavefunctions had to be exploited once again – however the same term appears when the assumption $c \approx 1$ is not made. In the second term the $\boldsymbol{R}_l = 0$ term of the sum over the lattice points is expected to give a negative contribution, as it is plausible to assume that the potential around an atom in a crystal is lower than around a free atom ($U(\boldsymbol{r}) - v_\mathrm{a}(\boldsymbol{r}) < 0$). This term,

$$\Delta\varepsilon_\alpha = \int w_\alpha^*(\boldsymbol{r})\,[U(\boldsymbol{r}) - v_\mathrm{a}(\boldsymbol{r})]\,w_\alpha(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,, \tag{18.2.10}$$

shifts the energy of each state in the band by the same amount. This shift will therefore be ignored below. The broadening of atomic energy levels into bands is described by the terms $\boldsymbol{R}_l \neq 0$ in the second term of (18.2.9); they are also proportional to the difference between the real potential and the atomic potential. Introducing the notation

$$\gamma_\alpha(\boldsymbol{R}_l) = -\int w_\alpha^*(\boldsymbol{r}+\boldsymbol{R}_l)\,[U(\boldsymbol{r}) - v_\mathrm{a}(\boldsymbol{r})]\,w_\alpha(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,, \tag{18.2.11}$$

the energies in band α are

$$\varepsilon_\alpha(\boldsymbol{k}) \approx \varepsilon_\alpha - {\sum_l}' \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l}\gamma_\alpha(\boldsymbol{R}_l)\,, \tag{18.2.12}$$

where the sum is over the lattice points $\boldsymbol{R}_l \neq 0$. Since the overlap between distant neighbors can be exponentially small, the dominant contribution usually comes from nearest neighbors. The energy is then

$$\varepsilon_\alpha(\boldsymbol{k}) \approx \varepsilon_\alpha - {\sum_l}' \gamma_\alpha(\boldsymbol{\delta}_l)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\delta}_l}\,, \tag{18.2.13}$$

where the sum is over nearest neighbors only.

This method is expected to work well for those states that overlap only slightly even in the crystal – that is, for inner shells. Nevertheless it sometimes provides an adequate description of the states of conduction electrons as well. It does not only propose a simple picture of the formation of bands but also models properly some of their characteristics. Since according to our assumptions there is no mixing between atomic states of different energies, depending on which atomic states are used as the starting point, we speak of s-, p-, and d-bands.

It should be noted that the simple form derived above is valid only when the primitive cell contains a single atom and only one electron state per atom is considered. For crystals with polyatomic bases or for p- and d-bands the linear combinations of degenerate states must be chosen as atomic states, and the energy eigenvalues should be determined by diagonalizing a matrix of the corresponding size. Below we shall examine a few examples.

18.2.2 Band of s-Electrons

Let us first examine the case where the tight-binding band is formed from the atomic s-states (azimuthal quantum number: $l = 0$) with the same principal quantum number, n. Owing to the spherical symmetry of the atomic wavefunction, $\gamma_{n,0}(\boldsymbol{R}_l)$ exhibits the same symmetry as the potential $U(\boldsymbol{r})$. For a cubic crystal – where nearest neighbors are all located at equal distances, at such positions that they can be transformed into each other by symmetry transformations – $\gamma_{n,0}(\boldsymbol{\delta}_l)$ can be replaced by a constant. For a simple cubic lattice, where nearest neighbors are located at $a(\pm 1, 0, 0)$, $a(0, \pm 1, 0)$, and $a(0, 0, \pm 1)$,

$$\varepsilon_{n,0}(\boldsymbol{k}) = \varepsilon_{n,0} - 2\gamma(1,0,0)[\cos k_x a + \cos k_y a + \cos k_z a] . \tag{18.2.14}$$

In a face-centered cubic crystal, where there are 12 nearest neighbors, located at $(a/2)(\pm 1, \pm 1, 0)$, $(a/2)(\pm 1, 0, \pm 1)$, and $(a/2)(0, \pm 1, \pm 1)$, the formula

$$\begin{aligned}
\mathrm{e}^{\mathrm{i}a(k_x+k_y)/2} + \mathrm{e}^{\mathrm{i}a(k_x-k_y)/2} + \mathrm{e}^{\mathrm{i}a(-k_x+k_y)/2} + \mathrm{e}^{\mathrm{i}a(-k_x-k_y)/2}& \\
&= \left[\mathrm{e}^{\mathrm{i}ak_x/2} + \mathrm{e}^{-\mathrm{i}ak_x/2}\right]\left[\mathrm{e}^{\mathrm{i}ak_y/2} + \mathrm{e}^{-\mathrm{i}ak_y/2}\right] \\
&= 4\cos(k_x a/2)\cos(k_y a/2)
\end{aligned} \tag{18.2.15}$$

leads to

$$\begin{aligned}
\varepsilon_{n,0}(\boldsymbol{k}) = \varepsilon_{n,0} - 4\gamma(\tfrac{1}{2}\tfrac{1}{2}0)\,&[\cos(k_x a/2)\cos(k_y a/2) \\
&+ \cos(k_y a/2)\cos(k_z a/2) + \cos(k_z a/2)\cos(k_x a/2)] .
\end{aligned} \tag{18.2.16}$$

Evaluating the sum of phase factors in a similar fashion for the eight nearest neighbors in a body-centered cubic lattice,

$$\varepsilon_{n,0}(\boldsymbol{k}) = \varepsilon_{n,0} - 8\gamma(\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2})\cos(k_x a/2)\cos(k_y a/2)\cos(k_z a/2)\,. \qquad (18.2.17)$$

Whichever formula is considered, when the energy is expanded about the zone center, an isotropic quadratic expression is obtained for small values of the wave number, much like for free electrons:

$$\varepsilon_{n,0}(\boldsymbol{k}) = \varepsilon_{n,0} + \gamma a^2 \boldsymbol{k}^2\,, \qquad (18.2.18)$$

however, depending on the sign of γ, the parabola can open either upward or downward. As Fig. 18.22 shows, the distortion of the dispersion relation due to the periodic potential of the lattice (i.e., the deviation from the parabolic form) becomes more and more important as the distance from the point $\boldsymbol{k} = 0$ increases.

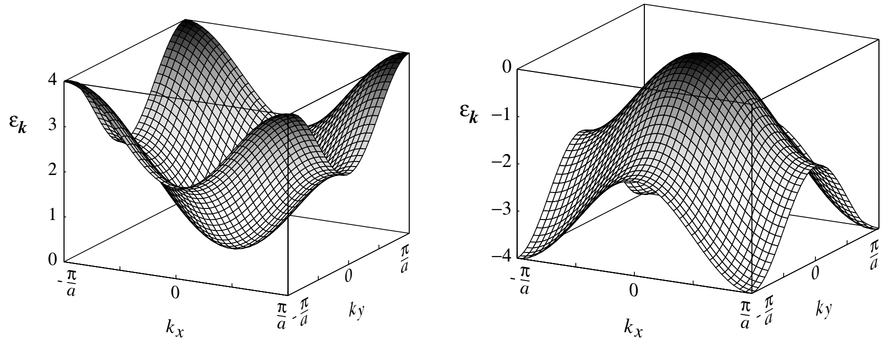


Fig. 18.22. The tight-binding s-band in a simple cubic lattice for a fixed k_z; γ is positive on the left and negative on the right

In the region where the second-order expansion is a good approximation, the constant-energy surfaces are spheres. The behavior of these electrons can be described adequately in terms of an effective scalar mass. Farther from the minimum and maximum energies the constant-energy surfaces are less and less spherical. This is even more obvious in the next two figures. In Fig. 18.23 the lines of constant energy are plotted for the two-dimensional square lattice in the tight-binding approximation, while in Fig. 18.24 the constant-energy surfaces for a simple cubic crystal are shown for two different energies.

In the two-dimensional case, when the band is exactly half filled, the "Fermi sphere" is distorted into a square. A logarithmic singularity appears in the density of states at the corresponding energy, and the concept of effective mass becomes meaningless. For higher band filling the description in terms of holes is more convenient. In the repeated-zone scheme spherical constant-energy surfaces appear around the corners of the Brillouin zone.

The constant-energy surfaces calculated for a face-centered cubic structure are shown in Fig. 17.9. This surface has the same topology as the one

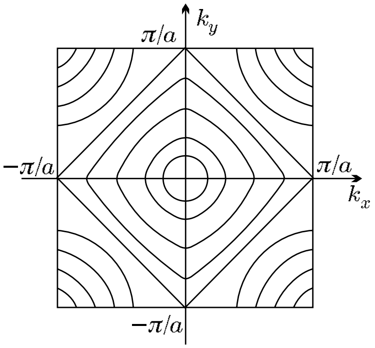


Fig. 18.23. Lines of constant energy for the *s*-band in a two-dimensional square lattice in the tight-binding approximation

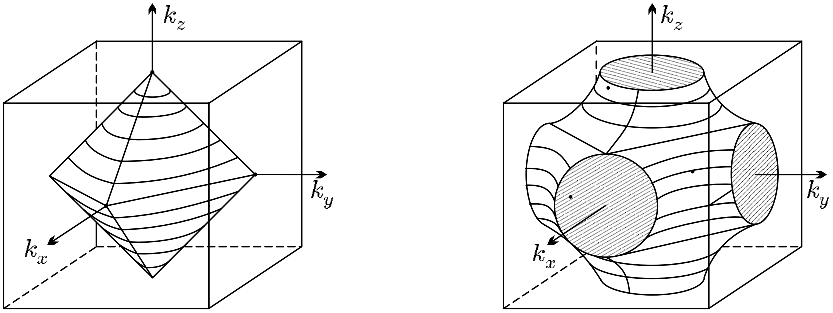


Fig. 18.24. Constant-energy surfaces in a simple cubic lattice in the tight-binding approximation

obtained in the nearly-free-electron approximation (see Fig. 18.20), however the distortion of the "Fermi sphere" is much more pronounced now.

18.2.3 Band of *p*-Electrons

The situation is more complicated when the energies of the band formed by *p*-electrons are considered, since owing to their threefold degeneracy, atomic *p*-states are mixed when bands are formed. For this reason, the atomic wavefunctions are customarily chosen as the linear combinations of the functions

$$\psi_{nlm}(\boldsymbol{r}) = R_{nl}(\boldsymbol{r}) Y_l^m(\theta, \varphi) \tag{18.2.19}$$

obtained in the presence of a spherically symmetric potential; in our particular case ($l = 1$, $m = \pm 1, 0$)

$$w_{n,1}(\boldsymbol{r}) = c_1 \psi_{n,1,1}(\boldsymbol{r}) + c_0 \psi_{n,1,0}(\boldsymbol{r}) + c_{-1} \psi_{n,1,-1}(\boldsymbol{r}) \,. \tag{18.2.20}$$

The corresponding energy is $\varepsilon_{n,1}$. Using this atomic function as an approximate Wannier function, the Schrödinger equation for the one-particle states of the electrons in the crystal is

$$\begin{aligned}\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right]\frac{1}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}w_{n,1}(\boldsymbol{r}-\boldsymbol{R}_j)\\ =\varepsilon_{n,1}(\boldsymbol{k})\frac{1}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}w_{n,1}(\boldsymbol{r}-\boldsymbol{R}_j)\,.\end{aligned}\qquad(18.2.21)$$

Using the expansion (18.2.20) for $w_{n,1}(\boldsymbol{r})$ and the atomic Schrödinger equation for $\psi_{n,1,m}$, we have

$$\begin{aligned}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\sum_{m'=-1}^{1}\left[\varepsilon_{n,1}+U(\boldsymbol{r})-v_{\mathrm{a}}(\boldsymbol{r}-\boldsymbol{R}_j)\right]c_{m'}\psi_{n,1,m'}(\boldsymbol{r}-\boldsymbol{R}_j)\\ =\varepsilon_{n,1}(\boldsymbol{k})\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\sum_{m'=-1}^{1}c_{m'}\psi_{n,1,m'}(\boldsymbol{r}-\boldsymbol{R}_j)\,.\end{aligned}\qquad(18.2.22)$$

Multiplying both sides by $\psi^*_{n,1,m}(\boldsymbol{r}-\boldsymbol{R}_{j'})$, integrating over the crystal volume, and making use of the assumption that the overlap between wavefunctions of different lattice points can be ignored as far as the normalization of the wavefunction is concerned,

$$\varepsilon_{n,1}(\boldsymbol{k})c_m-\sum_{m'=-1}^{1}\sum_l \gamma_{mm'}(\boldsymbol{R}_l)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l}c_{m'}=\varepsilon_{n,1}(\boldsymbol{k})c_m\,,\qquad(18.2.23)$$

where

$$\gamma_{mm'}(\boldsymbol{R}_l)=-\int\psi^*_{n,1,m}(\boldsymbol{r}+\boldsymbol{R}_l)\left[U(\boldsymbol{r})-v_{\mathrm{a}}(\boldsymbol{r})\right]\psi_{n,1,m'}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,.\qquad(18.2.24)$$

The eigenvalue problem leads to a system of linear equations in three variables, which is equivalent to the diagonalization of a 3 × 3 matrix.

The calculation is further simplified when the real functions ψ_{p_x}, ψ_{p_y}, ψ_{p_z} are used instead of the wavefunctions expressed in spherical harmonics. As shown in Fig. 4.10, the three new functions give high densities around the x-, y-, and z-axes, in a cylindrically symmetric geometry. It is readily established from the transformation properties of the functions $\psi_{p_\alpha}(\boldsymbol{r})$ that in simple cubic crystals the coefficients

$$\gamma_{\alpha\beta}(\boldsymbol{\delta}_l)=-\int\psi^*_{p_\alpha}(\boldsymbol{r}+\boldsymbol{\delta}_l)\left[U(\boldsymbol{r})-v_{\mathrm{a}}(\boldsymbol{r})\right]\psi_{p_\beta}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\qquad(18.2.25)$$

obtained for nearest neighbors are diagonal – however the value depends on whether the maxima of the electron density are along or perpendicular to the direction of the neighbors. Introducing the notations

$$\gamma_{\|}=-\int\psi^*_{p_x}(\boldsymbol{r}+a\hat{\boldsymbol{x}})\left[U(\boldsymbol{r})-v_{\mathrm{a}}(\boldsymbol{r})\right]\psi_{p_x}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\qquad(18.2.26)$$

and

$$\gamma_\perp = -\int \psi^*_{p_x}(\boldsymbol{r} + a\hat{\boldsymbol{y}})\,[U(\boldsymbol{r}) - v_{\mathrm{a}}(\boldsymbol{r})]\,\psi_{p_x}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,, \tag{18.2.27}$$

and exploiting the cubic symmetry, the energies for the three p-bands are

$$\begin{aligned}
\varepsilon^{(1)}_{n,1}(\boldsymbol{k}) &= \varepsilon_{n,1} - 2\gamma_{||}\cos k_x a - 2\gamma_\perp[\cos k_y a + \cos k_z a]\,,\\
\varepsilon^{(2)}_{n,1}(\boldsymbol{k}) &= \varepsilon_{n,1} - 2\gamma_{||}\cos k_y a - 2\gamma_\perp[\cos k_x a + \cos k_z a]\,,\\
\varepsilon^{(3)}_{n,1}(\boldsymbol{k}) &= \varepsilon_{n,1} - 2\gamma_{||}\cos k_z a - 2\gamma_\perp[\cos k_x a + \cos k_y a]\,.
\end{aligned} \tag{18.2.28}$$

Note that the three p-bands, when considered individually, do not exhibit the symmetries of the cubic lattice, but as a set they do.

To obtain an estimate for the sign and relative magnitude of the coefficients, the p-orbitals of neighboring atoms are shown schematically in Fig. 18.25.

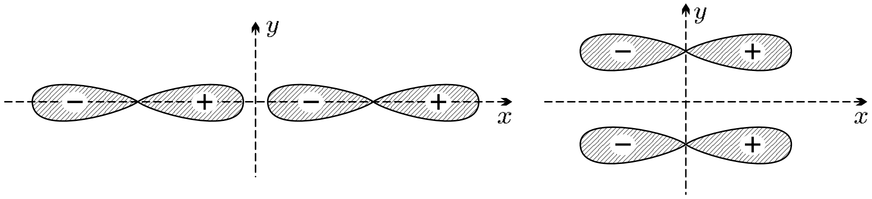


Fig. 18.25. Relative orientation of the p-orbitals in neighboring atoms

As can be inferred from the figure, in the integral for $\gamma_{||}$ the two wavefunctions may overlap in regions where their signs are opposite. Combined with the negative sign in the definition of $\gamma_{||}$ and the negativity of $U(\boldsymbol{r}) - v_{\mathrm{a}}(\boldsymbol{r})$, $\gamma_{||} < 0$. On the other hand, the wavefunctions appear with identical signs in the integrand of $\gamma_\perp$, therefore $\gamma_\perp > 0$. The orientation of the wavefunctions implies $\gamma_\perp < |\gamma_{||}|$, hence the dispersion relation has a maximum in one and minima in two directions at the zone center. The dispersion curves are plotted for a fixed value of k_z in Fig. 18.26.

The p_x- and p_y-bands feature a saddle point in $k_x = k_y = 0$ for a fixed value of k_z; minima and maxima occur at the edge centers of the Brillouin zone. In the p_z-band the minimum is at the center and the maximum is at the vertex of the Brillouin zone.

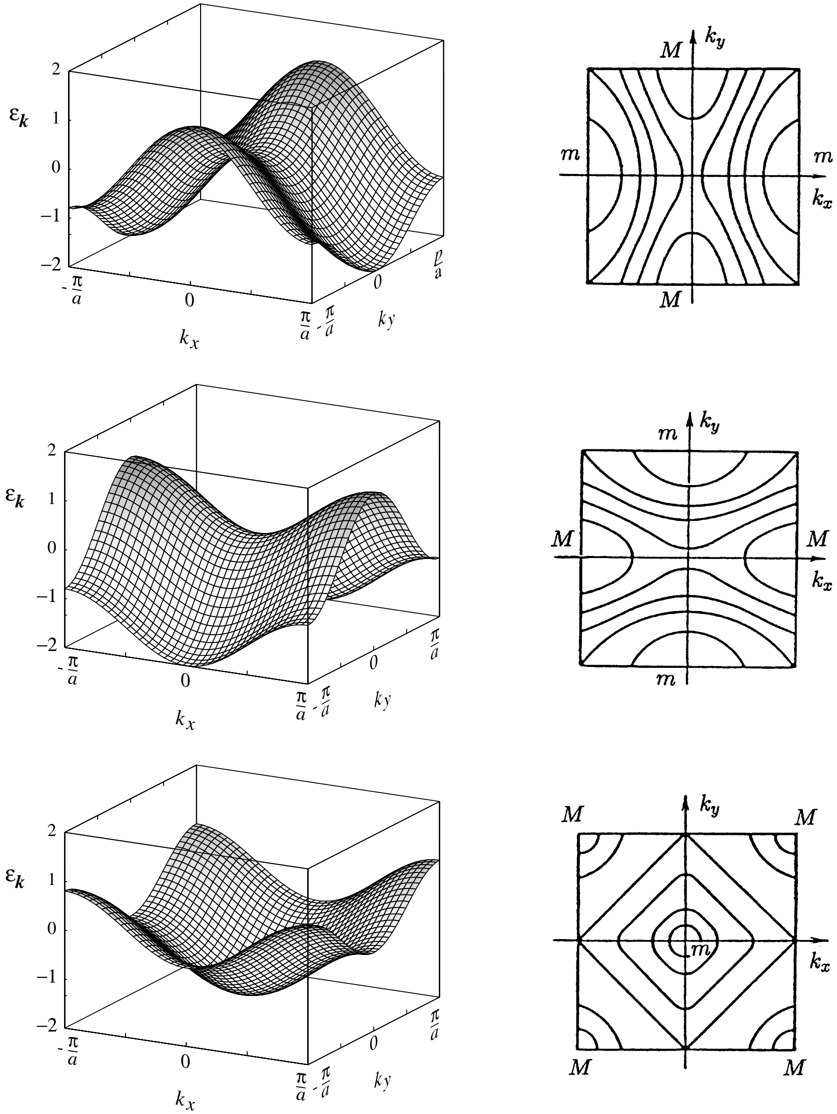


Fig. 18.26. Electron energies in the three p-bands in the tight-binding approximation for a fixed value of k_z, and the constant-energy contours. Minima are marked by m and maxima by M

Further Reading

1. S. L. Altmann, *Band Theory of Solids, An Introduction from the Point of View of Symmetry*, Clarendon Press, Oxford (1991).
2. J. F. Cornwell, *Group Theory and Energy Bands in Solids*, North Holland Publ. Co., Amsterdam (1969).
3. H. Jones, *The Theory of Brillouin Zones and Electronic States in Crystals*, 2nd revised edition, North-Holland Publishing Co., Amsterdam (1975).

19

Methods for Calculating and Measuring the Band Structure

In the previous chapter we examined two simple methods based on opposite approaches to calculate the energies and band structure of one-particle electron states. In the nearly-free-electron model the potential created by the lattice of atoms was considered as a perturbation with respect to the kinetic energy of electrons. In the tight-binding method we started with localized atomic states, and treated the propagation of electrons in the lattice as perturbation. Both methods gave a good qualitative picture about how the allowed energies of Bloch electrons form bands. In the first approach, even though free electrons can have arbitrary energies, certain energies are found to be forbidden in the presence of the potential, while in the second approach discrete atomic energy levels are observed to broaden into bands. Although in certain cases – e.g., for simple metals – the two methods lead to even quantitatively correct results, the kinetic energy of electrons and the potential arising from the interactions with atoms and other electrons are equally important in general, and neither of them can be treated as a perturbation with respect to the other. The accurate calculation of the band structure requires the solution of a difficult numerical problem in which every state – including deep core states – must be taken into account in principle, since even those are broadened into bands.

Two problems arise in connection with the Schrödinger equation (17.1.3). Firstly, the choice of the one-particle potential is dictated by the specific problem, secondly, when $U(\boldsymbol{r})$ is given, a suitable and efficient numerical method is needed to compute the energy eigenvalues quickly and accurately. We shall not deal with the choice of the potential here: it will be deferred to Volume 3 devoted to the study of electron–electron interactions. We shall just note that, since the influence of all other electrons must also be lumped into the potential, the solutions for the wavefunction must be consistent with the electron density used for specifying the potential – that is, self-consistent solutions have to be found.

Assuming that the potential is known, we shall first briefly present various computational methods. As mentioned before, all these methods eventually lead to numerical algorithms, for which dedicated computer codes have been worked out. The discussion of the numerical aspects is far beyond the scope of this book. Our sole purpose is to familiarize the reader with the fundamental principles and concepts of each method. The currently most widely used methods for calculating the band structure rely on the density-functional theory, which will be discussed only in Chapter 30, after a more precise study of electron–electron interactions.

In metals only the electron states close to the Fermi surface are important. Therefore after the introduction of the diverse methods we shall present the band structure of simple metals and the shape of their Fermi surfaces. (The band structure of semiconductors will be the subject of a separate chapter.) At the end of the chapter we shall give a brief overview of the experimental methods that allow the determination of some features of the band structure. The detailed discussion of other experimental methods for determining the parameters of the Fermi surface will be deferred to Chapters 21 and 22.

19.1 Matrix Methods

Some of the methods developed for the computation of the band structure are based on the equivalence of the solution of the Schrödinger equation and that of a matrix eigenvalue problem.

19.1.1 General Formulation of the Problem

The calculation of energy levels in a periodic potential – that is, finding the solutions of the one-particle Schrödinger equation – can be based on the expansion of the wavefunction $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ in a suitable complete but not necessarily orthogonal set of functions $\phi_j(\boldsymbol{k},\boldsymbol{r})$:

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_j c_{nj}(\boldsymbol{k})\phi_j(\boldsymbol{k},\boldsymbol{r})\,, \tag{19.1.1}$$

where the expansion coefficients are determined from the requirement that the state should be an eigenstate of the Hamiltonian. To satisfy Bloch's theorem, the condition

$$\phi_j(\boldsymbol{k},\boldsymbol{r}+\boldsymbol{R}_m) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_m}\phi_j(\boldsymbol{k},\boldsymbol{r}) \tag{19.1.2}$$

is imposed on each member of the set of functions used in the expansion for each lattice vector $\boldsymbol{R}_m$.

Substituting the above expansion into the Schrödinger equation, multiplying it by $\phi_i^*(\boldsymbol{k},\boldsymbol{r})$, and integrating the product over the whole volume of the crystal, the system of homogeneous linear equations

$$\sum_j \left[H_{ij}(\boldsymbol{k}) - \varepsilon_{n\boldsymbol{k}}S_{ij}(\boldsymbol{k})\right] c_{nj}(\boldsymbol{k}) = 0 \tag{19.1.3}$$

arises, where

$$H_{ij}(\boldsymbol{k}) = \int \phi_i^*(\boldsymbol{k},\boldsymbol{r})\mathcal{H}(\boldsymbol{r})\phi_j(\boldsymbol{k},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{19.1.4}$$

and

$$S_{ij}(\boldsymbol{k}) = \int \phi_i^*(\boldsymbol{k},\boldsymbol{r})\phi_j(\boldsymbol{k},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{19.1.5}$$

If the basis functions are not orthogonal, their overlap matrix S_{ij} contains some nonvanishing off-diagonal elements. The eigenfunctions are the solutions of this system of homogeneous linear equations, provided the energy eigenvalues are known. The latter can be determined from the condition that the system of equations for c_{nj} should have a nontrivial solution – that is, the determinant of the matrix $H_{ij}(\boldsymbol{k}) - \varepsilon_{\boldsymbol{k}}S_{ij}(\boldsymbol{k})$ should vanish:

$$\boxed{\det\left[H_{ij}(\boldsymbol{k}) - \varepsilon_{\boldsymbol{k}}S_{ij}(\boldsymbol{k})\right] = 0\,.} \tag{19.1.6}$$

By performing the calculation for each vector $\boldsymbol{k}$ of the Brillouin zone, the set of eigenvalues gives the energies $\varepsilon_{n\boldsymbol{k}}$. If the set of functions is complete then, in principle, an exact solution can be obtained from this expansion – however, it requires the diagonalization of an infinitely large matrix.

To proceed, we shall demonstrate that (19.1.3) can also be considered as the solution of a variational problem. By writing the wavefunction in the form (19.1.1), the expectation value of the energy is

$$\begin{aligned}\langle\psi_{n\boldsymbol{k}}|\mathcal{H}|\psi_{n\boldsymbol{k}}\rangle &= \int \psi_{n\boldsymbol{k}}^*(\boldsymbol{r})\mathcal{H}(\boldsymbol{r})\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \\ &= \sum_{ij} c_{ni}^*(\boldsymbol{k})c_{nj}(\boldsymbol{k})H_{ij}(\boldsymbol{k})\,,\end{aligned} \tag{19.1.7}$$

where $H_{ij}(\boldsymbol{k})$ is defined by (19.1.4). The normalization condition for the wavefunction is

$$\langle\psi_{n\boldsymbol{k}}|\psi_{n\boldsymbol{k}}\rangle = \int \psi_{n\boldsymbol{k}}^*(\boldsymbol{r})\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} = \sum_{ij} c_{ni}^*(\boldsymbol{k})c_{nj}(\boldsymbol{k})S_{ij}(\boldsymbol{k}) = 1\,. \tag{19.1.8}$$

If the minimum of the energy is determined with the constraint that the wavefunction of the state should be normalized, and this condition is taken into account by a Lagrange multiplier, then

$$\sum_{ij} c_{ni}^*(\boldsymbol{k})c_{nj}(\boldsymbol{k})H_{ij}(\boldsymbol{k}) - \lambda\left[\sum_{ij} c_{ni}^*(\boldsymbol{k})c_{nj}(\boldsymbol{k})S_{ij}(\boldsymbol{k}) - 1\right] \tag{19.1.9}$$

has to be minimized with respect to c_{ni}^*. This indeed leads to (19.1.3) when the Lagrange multiplier λ is identified with the energy $\varepsilon_{n\boldsymbol{k}}$.

If the set of equations used in the expansion (19.1.1) is not complete, then this variational approach does not lead to the exact eigenvalues and eigenfunctions. Nonetheless, as an approximation, one may choose the coefficients in such a way that the energy of the state (the expectation value of the Hamiltonian) be a minimum. The accuracy of this method is determined by the choice of functions and the number of terms taken into account in the expansion. The choice of the set of functions should therefore be based on physical reasoning. The methods presented below will differ precisely in the choice of the set of functions and the considerations that justify keeping only a few functions instead of the complete set.

19.1.2 LCAO Method

By choosing the complete set of Wannier states as basis functions, and expanding the wavefunction satisfying Bloch's theorem in terms of them, an exact method can be obtained – at least in principle. It should nevertheless be kept in mind that finding the Wannier states is equivalent to solving the complete eigenvalue problem. An approximate method is obtained by replacing the Wannier functions by functions with similar properties. It was demonstrated in Section 18.2 on the tight-binding approximation that atomic wavefunctions offer a good approximation to the Wannier functions. By way of example we saw how to construct a wavefunction satisfying the Bloch condition using a single s- or three p-states. To generalize this procedure, we shall build the functions $\phi_j(\boldsymbol{k},\boldsymbol{r})$ to be used in (19.1.1) from the atomic wavefunctions $w_j(\boldsymbol{r}-\boldsymbol{R}_m)$ in the form

$$\phi_j(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{N}} \sum_m \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_m} w_j(\boldsymbol{r}-\boldsymbol{R}_m) , \tag{19.1.10}$$

which satisfies the Bloch condition. The label j stands for the set of quantum numbers n, l, and m_l that characterize the atomic states, with each possible value allowed. The band states are thus linear combinations of atomic orbitals – just like molecular orbits were built up of atomic states in Chapter 4. The method is called *LCAO method* in the present context, too, even though we now aim at constructing states that extend over the entire crystal.

In an alternative interpretation of the method one starts with the Wannier representation of the wavefunction $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ and writes the Wannier function $\phi_n(\boldsymbol{r}-\boldsymbol{R}_m)$ as the linear combination of the complete set of atomic orbitals:

$$\phi_n(\boldsymbol{r}-\boldsymbol{R}_m) = \sum_j c_{nj}(\boldsymbol{k}) w_j(\boldsymbol{r}-\boldsymbol{R}_m) . \tag{19.1.11}$$

In the space of these functions the matrix elements are

$$\begin{aligned} H_{ij}(\boldsymbol{k}) &= \int \phi_i^*(\boldsymbol{k},\boldsymbol{r}) \mathcal{H}(\boldsymbol{r}) \phi_j(\boldsymbol{k},\boldsymbol{r}) \,\mathrm{d}\boldsymbol{r} \qquad (19.1.12) \\ &= \frac{1}{N} \sum_{mm'} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_{m'}-\boldsymbol{R}_m)} \int w_i^*(\boldsymbol{r}-\boldsymbol{R}_{m'}) \mathcal{H}(\boldsymbol{r}) w_j(\boldsymbol{r}-\boldsymbol{R}_m) \,\mathrm{d}\boldsymbol{r} . \end{aligned}$$

Of the double sum over the lattice points one can be evaluated since the integral

$$\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_{m'})\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right]w_j(\boldsymbol{r}-\boldsymbol{R}_m)\,\mathrm{d}\boldsymbol{r} \tag{19.1.13}$$

depends only on the difference $\boldsymbol{R}_l=\boldsymbol{R}_{m'}-\boldsymbol{R}_m$ because of the periodicity of $U(\boldsymbol{r})$. Therefore

$$H_{ij}(\boldsymbol{k})=\sum_l \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l}E_{ij}(\boldsymbol{R}_l)\,, \tag{19.1.14}$$

where

$$E_{ij}(\boldsymbol{R}_l)=\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_l)\mathcal{H}(\boldsymbol{r})w_j(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{19.1.15}$$

Similarly, the quantities $S_{ij}(\boldsymbol{k})$ characterizing the overlap between wavefunctions can be written as

$$\begin{aligned}S_{ij}(\boldsymbol{k})&=\int \phi_i^*(\boldsymbol{k},\boldsymbol{r})\phi_j(\boldsymbol{k},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\\&=\frac{1}{N}\sum_{mm'}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_{m'}-\boldsymbol{R}_m)}\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_{m'})w_j(\boldsymbol{r}-\boldsymbol{R}_m)\,\mathrm{d}\boldsymbol{r}\\&=\sum_l \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l}S_{ij}(\boldsymbol{R}_l)\,,\end{aligned} \tag{19.1.16}$$

where

$$S_{ij}(\boldsymbol{R}_l)=\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_l)w_j(\boldsymbol{r})\,d\boldsymbol{r}\,. \tag{19.1.17}$$

Off-diagonal elements appear because the atomic wavefunctions centered on different atoms are not orthogonal.

The quantity $E_{ij}(\boldsymbol{R}_l)$ in the matrix element $H_{ij}(\boldsymbol{k})$ can be further simplified by making use of the Schrödinger equation (18.2.1) of the atomic problem:

$$\begin{aligned}E_{ij}(\boldsymbol{R}_l)&=\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_l)\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2+U(\boldsymbol{r})\right]w_j(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\\&=\varepsilon_j S_{ij}(\boldsymbol{R}_l)+\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_l)\left[U(\boldsymbol{r})-v_\mathrm{a}(\boldsymbol{r})\right]w_j(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,.\end{aligned} \tag{19.1.18}$$

Using the notation

$$\gamma_{ij}(\boldsymbol{R}_l)=-\int w_i^*(\boldsymbol{r}-\boldsymbol{R}_l)\left[U(\boldsymbol{r})-v_\mathrm{a}(\boldsymbol{r})\right]w_j(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,, \tag{19.1.19}$$

we have

$$H_{ij}(\boldsymbol{k})=\varepsilon_j S_{ij}(\boldsymbol{k})-\sum_l \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_l}\gamma_{ij}(\boldsymbol{R}_l)=\varepsilon_j S_{ij}(\boldsymbol{k})-\gamma_{ij}(\boldsymbol{k})\,, \tag{19.1.20}$$

where $\gamma_{ij}(\boldsymbol{k})$ is the Fourier transform of $\gamma_{ij}(\boldsymbol{R}_l)$. Specifying the energy eigenvalues is therefore equivalent to solving the equation

$$\boxed{\det\left[(\varepsilon_j - \varepsilon_{\boldsymbol{k}})S_{ij}(\boldsymbol{k}) - \gamma_{ij}(\boldsymbol{k})\right] = 0\,.} \tag{19.1.21}$$

When the atomic wavefunctions are known, these matrix elements and overlap integrals can be determined. An exact solution would require using the complete set of atomic states. An approximate result is obtained if a finite set of atomic states is considered. Note that by choosing a single atomic state or several degenerate atomic states, the tight-binding approximation is recovered. Obviously, the description is better when a larger number of atomic orbitals are taken. Figure 18.21 illustrating the formation of bands shows that those atomic states mix most whose energies are close together. If there is considerable mixing (hybridization) between s- and p-states that belong to the same principal quantum number, then the matrix elements in a 4×4 block will be important. The solution of the eigenvalue problem is therefore reduced to the diagonalization of such a matrix. For s–d hybridization a 6×6 secular equation must be solved. Even when p-states are also taken into account, the matrix remains small, 9×9.

19.1.3 Plane-Wave Method

Another possibility is to choose plane waves, $\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r})$, instead of the atomic functions as the complete set of functions in the expansion (19.1.1). To satisfy the requirement of completeness, the wave vectors $\boldsymbol{k}$ must not be restricted to the first Brillouin zone: all equivalent vectors $\boldsymbol{k}+\boldsymbol{G}$ must also be allowed. This amounts to using the functions

$$\phi_j(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}}\,. \tag{19.1.22}$$

These functions satisfy the condition (19.1.2), and constitute a complete and orthonormal set:

$$S_{ij}(\boldsymbol{k}) = \frac{1}{V}\int \mathrm{e}^{-\mathrm{i}(\boldsymbol{G}_i-\boldsymbol{G}_j)\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{r} = \delta_{ij}\,. \tag{19.1.23}$$

The matrix elements of the Hamiltonian are

$$\begin{aligned} H_{ij}(\boldsymbol{k}) &= \frac{1}{V}\int \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_i)\cdot\boldsymbol{r}}\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r})\right]\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{r} \\ &= \frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_j)^2\delta_{ij} + U_{ij}\,, \end{aligned} \tag{19.1.24}$$

where

$$U_{ij} = \left\langle \boldsymbol{k}+\boldsymbol{G}_i \middle| U \middle| \boldsymbol{k}+\boldsymbol{G}_j \right\rangle = \frac{1}{V}\int \mathrm{e}^{-\mathrm{i}(\boldsymbol{G}_i-\boldsymbol{G}_j)\cdot\boldsymbol{r}}U(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{19.1.25}$$

is the Fourier transform of the potential. By substituting these formulas into (19.1.6), nontrivial solutions exist if

$$\boxed{\det\left(\left[\frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2-\varepsilon_{\boldsymbol{k}}\right]\delta_{ij}+U_{ij}\right)=0\,.} \tag{19.1.26}$$

The equation for the expansion coefficients in the wavefunction of the solution with energy $\varepsilon_{n\boldsymbol{k}}$ is

$$\left[\frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2-\varepsilon_{n\boldsymbol{k}}\right]c_{ni}(\boldsymbol{k})+\sum_j U_{ij}c_{nj}(\boldsymbol{k})=0\,. \tag{19.1.27}$$

Leaving a volume factor aside, this expression is identical to the general formula (18.1.2) obtained in the nearly-free-electron approximation, provided $c_{n\boldsymbol{k}}(\boldsymbol{G}_j)$ in (18.1.2) is identified with $c_{nj}(\boldsymbol{k})$ above. This is natural, since the choice of a plane-wave basis corresponds to writing the wavefunction as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r})=\frac{1}{\sqrt{V}}\sum_j c_{nj}(\boldsymbol{k})\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}}\,, \tag{19.1.28}$$

which, in turn, is equivalent to expanding the lattice-periodic function $u_{n\boldsymbol{k}}(\boldsymbol{r})$ in the Bloch function into a Fourier series of the reciprocal-lattice vectors as

$$u_{n\boldsymbol{k}}(\boldsymbol{r})=\frac{1}{\sqrt{V}}\sum_j c_{nj}(\boldsymbol{k})\mathrm{e}^{\mathrm{i}\boldsymbol{G}_j\cdot\boldsymbol{r}}\,. \tag{19.1.29}$$

In the nearly-free-electron approximation we solved the system of equations iteratively – and did not go beyond the first iteration step. To understand how degeneracies are lifted, a mere 2 × 2 block was considered. Nonetheless the method can be applied, in principle, to arbitrary potentials, and energies can be calculated exactly by using sufficiently large matrices. This may be necessary even in simple metals, since the method in its above form does not distinguish between the localized core electrons and the mobile ones that account for conduction phenomena: the former are also constructed from plane waves. The calculation of the full band structure may require hundreds or even thousands of plane waves. The diagonalization of matrices of this size would not pose any difficulty to present-day computers, however the slow convergence of numerical techniques encourages the application of improved methods based on physical insight.

19.1.4 Orthogonalized-Plane-Wave Method

In crystalline solids, where all energy eigenstates are organized into bands, the bands formed by relatively strongly bound core states lie much below the Fermi energy and are narrow on account of the small overlap between neighboring atoms. For example, in aluminum, 1s, 2s, and 2p states are considered to belong to the core, and they give rise to narrow bands. Bands that lie closer to the Fermi energy – such as the bands formed by 3s and 3p electrons in aluminum – are broader. In the present chapter we shall call these broader bands *valence bands*, since their electrons participate in chemical bonding. In metals we shall further distinguish partially filled *conduction bands*, as electrons in these are responsible for the metallic conductivity. Note that the terms *conduction band* and *valence band* are used in a slightly different sense in connection with semiconductors.

Both in the plane-wave method and the LCAO method, each band is treated on the same footing, even though the LCAO method, which uses atomic states, is expected to work better for low-lying narrow bands, while the plane-wave method seems to be more adapted to the description of valence-band states. Based on this observation, C. HERRING (1940) proposed a method in which core states and valence-band states are treated differently.

The narrow bands of core electrons – which remain well localized in the crystal, too – can be approximated well by using the atomic wavefunctions $w_\alpha(\boldsymbol{r}-\boldsymbol{R}_m)$ of core states in place of the Wannier functions. In what follows, we shall use the index α only for core states. The Bloch functions associated with them are chosen as

$$\phi_\alpha(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_m \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_m} w_\alpha(\boldsymbol{r}-\boldsymbol{R}_m)\,, \tag{19.1.30}$$

much like in the LCAO approximation. To describe valence-band states using the expansion (19.1.1), another set indexed by j has to be added. We shall choose the plane-wave-like basis functions $\phi_j(\boldsymbol{k},\boldsymbol{r})$ with the requirement that they be orthogonal to the wavefunctions $\phi_\alpha(\boldsymbol{k},\boldsymbol{r})$ made up of core states, that is,

$$\int \phi_\alpha^*(\boldsymbol{k},\boldsymbol{r})\phi_j(\boldsymbol{k},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} = 0\,. \tag{19.1.31}$$

This approach is called the *orthogonalized-plane-wave (OPW) method.*

Our task is significantly simplified: only the functions that belong to the same (equivalent) wave vectors $\boldsymbol{k}$ need to be considered, as those that belong to nonequivalent $\boldsymbol{k}$s are *a priori* orthogonal. Just like in the Gram–Schmidt orthogonalization,[1] the set of functions to be used in (19.1.1) is sought in the form

$$\phi_j(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}} - \sum_\alpha \mu_\alpha(\boldsymbol{k}+\boldsymbol{G}_j)\phi_\alpha(\boldsymbol{k},\boldsymbol{r})\,, \tag{19.1.32}$$

that is, by subtracting a linear combination of core wavefunctions from the plane wave. The coefficients μ_α are determined from the requirement that (19.1.31) should be satisfied. Assuming that the functions $\phi_\alpha(\boldsymbol{k},\boldsymbol{r})$ constructed from the core states are approximately orthogonal, the equation

[1] This procedure allows one to construct an equivalent orthonormalized set of functions from linearly independent elements of a Hilbert space.

$$
\begin{aligned}
\mu_\alpha(\boldsymbol{k}+\boldsymbol{G}_j) &= \frac{1}{\sqrt{V}}\int \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}'}\phi_\alpha^*(\boldsymbol{k},\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}' \\
&= \frac{1}{\sqrt{NV}}\sum_m \int \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot(\boldsymbol{r}'-\boldsymbol{R}_m)} w_\alpha^*(\boldsymbol{r}'-\boldsymbol{R}_m)\,\mathrm{d}\boldsymbol{r}' \\
&= \frac{1}{\sqrt{v}}\int \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}'} w_\alpha^*(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}' \qquad (19.1.33)
\end{aligned}
$$

is obtained, where v is the volume of the primitive cell. In the last step we made use of the property that each cell contributes equally. Substituting this back into (19.1.32), we find

$$
\phi_j(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}} - \sum_\alpha \phi_\alpha(\boldsymbol{k},\boldsymbol{r})\frac{1}{\sqrt{V}}\int \phi_\alpha^*(\boldsymbol{k},\boldsymbol{r}')\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}'}\,\mathrm{d}\boldsymbol{r}'\,. \qquad (19.1.34)
$$

In concise Dirac notation this reads

$$
|\phi_j(\boldsymbol{k})\rangle = |\boldsymbol{k}+\boldsymbol{G}_j\rangle - \sum_\alpha |\phi_\alpha(\boldsymbol{k})\rangle\langle\phi_\alpha(\boldsymbol{k})|\boldsymbol{k}+\boldsymbol{G}_j\rangle = (1-P)|\boldsymbol{k}+\boldsymbol{G}_j\rangle\,, \qquad (19.1.35)
$$

where orthogonality is ensured by the projection operator

$$
P = \sum_\alpha |\phi_\alpha(\boldsymbol{k})\rangle\langle\phi_\alpha(\boldsymbol{k})|\,. \qquad (19.1.36)
$$

Using these OPW functions $\phi_j(\boldsymbol{k},\boldsymbol{r})$ as basis functions in the expansion (19.1.1), and exploiting the properties that plane waves are eigenfunctions of the kinetic energy operator, and the $\phi_\alpha(\boldsymbol{k},\boldsymbol{r})$ are eigenfunctions of the full Hamiltonian with energy ε_α, the equation to be solved is

$$
\begin{aligned}
&\frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2 c_{ni}(\boldsymbol{k}) + \sum_j U_{ij}c_{nj}(\boldsymbol{k}) \\
&\qquad - \sum_j\sum_\alpha \varepsilon_\alpha \mu_\alpha^*(\boldsymbol{k}+\boldsymbol{G}_i)\mu_\alpha(\boldsymbol{k}+\boldsymbol{G}_j)c_{nj}(\boldsymbol{k}) \qquad (19.1.37) \\
&\quad = \varepsilon_{\boldsymbol{k}}c_{ni}(\boldsymbol{k}) - \sum_j\sum_\alpha \varepsilon_{\boldsymbol{k}}\mu_\alpha^*(\boldsymbol{k}+\boldsymbol{G}_i)\mu_\alpha(\boldsymbol{k}+\boldsymbol{G}_j)c_{nj}(\boldsymbol{k})\,.
\end{aligned}
$$

The secular equation, which determines the energies of the nontrivial solutions, can be written in a form that resembles (19.1.26):

$$
\boxed{\det\left(\left[\frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2 - \varepsilon_{\boldsymbol{k}}\right]\delta_{ij} + \mathit{\Gamma}_{ij}\right) = 0\,,} \qquad (19.1.38)
$$

where

$$
\mathit{\Gamma}_{ij} = U_{ij} + \sum_\alpha(\varepsilon_{\boldsymbol{k}} - \varepsilon_\alpha)\mu_\alpha^*(\boldsymbol{k}+\boldsymbol{G}_i)\mu_\alpha(\boldsymbol{k}+\boldsymbol{G}_j)\,. \qquad (19.1.39)
$$

Since in addition to the potential, Γ_{ij} also contains the contribution of core electrons, to establish the same level of accuracy far fewer terms need to be taken into account in the numerical calculations based on this approach than in the plane-wave method. Note, however, that Γ_{ij} itself contains the energy that is to be determined, therefore the equation needs to be solved self-consistently.

As a simple example, consider the case where wavefunctions orthogonalized to the 1s core state are used. Taking the wavefunction of the 1s state of hydrogen as atomic wavefunction,

$$w_{1s}(\boldsymbol{r}) = (1/a_0^3\pi)^{1/2}\mathrm{e}^{-r/a_0}\,. \tag{19.1.40}$$

From (19.1.33), the coefficient $\mu_{1s}(\boldsymbol{k})$ is

$$\begin{aligned}\mu_{1s}(\boldsymbol{k}) = \langle\phi_{1s}(\boldsymbol{k})|\boldsymbol{k}\rangle &= \frac{1}{\sqrt{v}}(1/a_0^3\pi)^{1/2}\int \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}'-r'/a_0}\,\mathrm{d}\boldsymbol{r}' \\ &= \frac{1}{\sqrt{v}}8(\pi/a_0)^{1/2}\frac{a_0^2}{(1+a_0^2\boldsymbol{k}^2)^2}\,,\end{aligned} \tag{19.1.41}$$

while the wavefunction of the orthogonalized plane wave associated with the vector $\boldsymbol{G}_j = 0$ of the reciprocal lattice is

$$\phi(\boldsymbol{k},\boldsymbol{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} - \frac{1}{\sqrt{V}}\sum_m \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_m}\mathrm{e}^{-|\boldsymbol{r}-\boldsymbol{R}_m|/a_0}\frac{8}{(1+a_0^2\boldsymbol{k}^2)^2}\,. \tag{19.1.42}$$

Since this wavefunction has nodes, the convergence of 2s-type states is reached much more easily by using $\phi(\boldsymbol{k},\boldsymbol{r})$ than by using only plane waves in their construction. While this achievement is undoubtedly significant, the most important implication of the OPW method from today's perspective is that it lead to the pseudopotential method.

19.1.5 Pseudopotential Method

In the orthogonalized-plane-wave method relatively good approximate eigenvalues can be obtained for the valence-band states by means of a small number of basis functions that are orthogonal to the core states. As J. C. Phillips and L. Kleinman pointed out in 1959 by comparing (19.1.38) and (19.1.26), the energy eigenvalues obtained in the OPW method with a small number of orthogonalized plane waves can be recovered using the same number of simple plane waves in the plane-wave method if the potential $U(\boldsymbol{r})$ is replaced by an effective potential whose matrix elements are the quantities Γ_{ij}. Choosing once again, by way of example, the wavefunctions orthogonal to the 1s core state, which lead to the coefficients $\mu_{1\mathrm{s}}(\boldsymbol{k})$ given in (19.1.41), and using the Fourier transform of the bare Coulomb potential in (19.1.39) and (19.1.41), we find

$$\Gamma_{ij} = -\frac{4\pi\tilde{e}^2}{|\boldsymbol{G}_i - \boldsymbol{G}_j|^2} + \frac{64\pi a_0^3}{v} \frac{\varepsilon_{\boldsymbol{k}} - \varepsilon_{1s}}{[1 + a_0^2(\boldsymbol{k} + \boldsymbol{G}_i)^2]^2 [1 + a_0^2(\boldsymbol{k} + \boldsymbol{G}_j)^2]^2}, \tag{19.1.43}$$

which clearly shows that the positive correction partially compensates the negative contribution of the bare Coulomb potential, and the effective potential specified by Γ_{ij} is no longer singular.

To put it differently: the description based on the plane-wave method can be made equivalent to the OPW method, leading to identical valence-band energies, if the effects of core electrons are not taken into account in the wavefunction but in a suitably chosen potential. In a certain sense this resembles the renormalization procedure presented in Appendix M of Volume 3, in which those states whose energies are far from the energy of interest are eliminated by absorbing their effects into an interaction, leading to a modified potential.

Treating the problem along the same lines but more generally than in the OPW method, we shall assume that the solutions of the Schrödinger equation are known in a certain, low-lying, energy range:

$$[\mathcal{H}_0 + U(\boldsymbol{r})]\,\phi_\alpha(\boldsymbol{k}, \boldsymbol{r}) = \varepsilon_{\alpha\boldsymbol{k}} \phi_\alpha(\boldsymbol{k}, \boldsymbol{r}), \tag{19.1.44}$$

and that the eigenfunctions can be constructed from the atomic functions $w_\alpha(\boldsymbol{r} - \boldsymbol{R}_m)$ as if they were Wannier functions, just like in the tight-binding approximation. In the energy range above, the eigenfunctions are constructed, according to (19.1.1), from a set of functions $\phi_j(\boldsymbol{k}, \boldsymbol{r})$ satisfying the Bloch condition and the requirement that these functions should be orthogonal to the core states $\phi_\alpha(\boldsymbol{k}, \boldsymbol{r})$.

Another set of functions $\widetilde{\phi}_j(\boldsymbol{k}, \boldsymbol{r})$ is then defined with the yet unknown coefficients $\mu_{j\alpha}(\boldsymbol{k})$:

$$\widetilde{\phi}_j(\boldsymbol{k}, \boldsymbol{r}) = \phi_j(\boldsymbol{k}, \boldsymbol{r}) + \sum_\alpha \mu_{j\alpha}(\boldsymbol{k}) \phi_\alpha(\boldsymbol{k}, \boldsymbol{r}). \tag{19.1.45}$$

The orthogonality of the functions ϕ_j and ϕ_α implies that

$$\mu_{j\alpha}(\boldsymbol{k}) = \int \phi_\alpha^*(\boldsymbol{k}, \boldsymbol{r}') \widetilde{\phi}_j(\boldsymbol{k}, \boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}'. \tag{19.1.46}$$

Expanding the functions ϕ_α on the basis of atomic functions, and exploiting the translational property of the Bloch functions $\widetilde{\phi}_j$,

$$\mu_{j\alpha}(\boldsymbol{k}) = N^{1/2} \int w_\alpha^*(\boldsymbol{r}) \widetilde{\phi}_j(\boldsymbol{k}, \boldsymbol{r})\, \mathrm{d}\boldsymbol{r}. \tag{19.1.47}$$

Expressing $\phi_j(\boldsymbol{k}, \boldsymbol{r})$ from (19.1.45), and substituting it into (19.1.1), we find

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_j c_{nj}(\boldsymbol{k}) \left[\widetilde{\phi}_j(\boldsymbol{k}, \boldsymbol{r}) - \sum_\alpha \mu_{j\alpha}(\boldsymbol{k}) \phi_\alpha(\boldsymbol{k}, \boldsymbol{r}) \right]. \tag{19.1.48}$$

Acting on both sides by the Hamiltonian, and making use of the eigenvalue equation of the core states, a slight rearrangement of the terms leads to the following form of the Schrödinger equation:

$$\begin{aligned}\mathcal{H}\sum_j c_{jn}(\boldsymbol{k})\widetilde{\phi}_j(\boldsymbol{k},\boldsymbol{r})+\sum_j c_{jn}(\boldsymbol{k})\sum_\alpha\left(\varepsilon_{n\boldsymbol{k}}-\varepsilon_{\alpha\boldsymbol{k}}\right)\mu_{j\alpha}(\boldsymbol{k})\phi_\alpha(\boldsymbol{k},\boldsymbol{r})\\ =\varepsilon_{n\boldsymbol{k}}\sum_j c_{jn}(\boldsymbol{k})\widetilde{\phi}_j(\boldsymbol{k},\boldsymbol{r})\,.\end{aligned}\tag{19.1.49}$$

Note that if (19.1.46) is used for $\mu_{j\alpha}(\boldsymbol{k})$, and the *pseudo-wavefunction* is written as

$$\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r})=\sum_j c_{nj}(\boldsymbol{k})\widetilde{\phi}_j(\boldsymbol{k},\boldsymbol{r})\,,\tag{19.1.50}$$

then (19.1.49) is just

$$\mathcal{H}\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r})+\int W(\boldsymbol{r},\boldsymbol{r}')\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'=\varepsilon_{n\boldsymbol{k}}\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r})\,,\tag{19.1.51}$$

where

$$\begin{aligned}W(\boldsymbol{r},\boldsymbol{r}')&=\sum_\alpha\phi_\alpha^*(\boldsymbol{k},\boldsymbol{r}')\left[\varepsilon_{n\boldsymbol{k}}-\varepsilon_{\alpha\boldsymbol{k}}\right]\phi_\alpha(\boldsymbol{k},\boldsymbol{r})\\ &=\sum_\alpha|\phi_\alpha\rangle\left[\varepsilon_{n\boldsymbol{k}}-\varepsilon_{\alpha\boldsymbol{k}}\right]\langle\phi_\alpha|\,.\end{aligned}\tag{19.1.52}$$

Rewriting (19.1.51) as

$$\int\widetilde{\mathcal{H}}(\boldsymbol{r},\boldsymbol{r}')\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'=\varepsilon_{n\boldsymbol{k}}\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r})\,,\tag{19.1.53}$$

where

$$\widetilde{\mathcal{H}}(\boldsymbol{r},\boldsymbol{r}')=\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}_{\boldsymbol{r}}^2+U(\boldsymbol{r})\right]\delta(\boldsymbol{r}-\boldsymbol{r}')+W(\boldsymbol{r},\boldsymbol{r}')\,,\tag{19.1.54}$$

one may say that the energy eigenvalues $\varepsilon_{n\boldsymbol{k}}$ of the original problem can be obtained from a Schrödinger equation with an effective nonlocal potential

$$V(\boldsymbol{r},\boldsymbol{r}')=U(\boldsymbol{r})\delta(\boldsymbol{r}-\boldsymbol{r}')+W(\boldsymbol{r},\boldsymbol{r}')\,,\tag{19.1.55}$$

called the *pseudopotential*, and the eigenfunctions are just the pseudo-wavefunctions $\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r})$.

Compared to the OPW method, the functions $\widetilde{\phi}_j(\boldsymbol{k},\boldsymbol{r})$ in the expansion of $\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r})$ correspond to plane waves, and $\phi_j(\boldsymbol{k},\boldsymbol{r})$ to plane waves orthogonalized to the core states. One may therefore say that if it is enough to take a few orthogonalized plane waves in the OPW method then only the coefficients associated with the same plane waves will be important in the expansion (19.1.50) when the pseudopotential $W(\boldsymbol{r},\boldsymbol{r}')$ is used.

This implies that if the pseudopotential were known, the band structure could be described in terms of a small number of plane waves. To understand this physically, consider the diagonal matrix elements of the pseudopotential. The quantity

$$\langle \phi_{\boldsymbol{k}} | W | \phi_{\boldsymbol{k}} \rangle = \sum_{\alpha} (\varepsilon_{\boldsymbol{k}} - \varepsilon_{\alpha\boldsymbol{k}}) \left| \langle \phi_{\boldsymbol{k}} | \psi_{\alpha\boldsymbol{k}} \rangle \right|^2 \tag{19.1.56}$$

is positive in the energy range of interest, since the energies $\varepsilon_{\alpha\boldsymbol{k}}$ of the bands of core states are lower than the energies of the valence bands. The negative potential of the ions is smoothed out by this positive contribution, therefore instead of the deep atomic potential well electrons move in a relatively weak pseudopotential, almost freely. This is why the conduction electrons of metals can often be described in the free-electron picture.

Besides nonlocality, another important property of the pseudopotential is its dependence on the energy that is to be determined, even though this dependence can be ignored if the energies of the core states are sufficiently far from the energies of the band of interest. We have ample freedom in the choice of the new wavefunction and, together with it, the pseudopotential. In general, the wavefunction should be chosen to be as smooth as possible. As M. H. Cohen and V. Heine pointed out in 1961, the optimal choice for the pseudopotential and the pseudo-wavefunction satisfy the equation

$$V | \widetilde{\psi}_{\boldsymbol{k}} \rangle = (1 - P) U | \widetilde{\psi}_{\boldsymbol{k}} \rangle + \frac{\langle \widetilde{\psi}_{\boldsymbol{k}} | (1 - P) U | \widetilde{\psi}_{\boldsymbol{k}} \rangle}{\langle \widetilde{\psi}_{\boldsymbol{k}} | (1 - P) | \widetilde{\psi}_{\boldsymbol{k}} \rangle} P | \widetilde{\psi}_{\boldsymbol{k}} \rangle \,. \tag{19.1.57}$$

However, this pseudopotential is nonlinear. In the linearized approximation, the plane-wave expansion of the pseudo-wavefunction gives

$$V | \boldsymbol{k} \rangle = U | \boldsymbol{k} \rangle + \sum_{\alpha} \left[\frac{\hbar^2 k^2}{2 m_{\mathrm{e}}} + \langle \boldsymbol{k} | V | \boldsymbol{k} \rangle - \varepsilon_{\alpha} \right] | \alpha \rangle \langle \alpha | \boldsymbol{k} \rangle \,. \tag{19.1.58}$$

In practical applications one often chooses an even simpler form, a smooth, energy-independent, local potential that contains only a few parameters. These are determined from a comparison with some experimental results – for example, from the requirement that the correct value should be recovered for the energy of core states, or that the spectrum of lattice vibrations should lead to the correct value of the sound velocity. This choice is supported by the following arguments: The energy independence can be justified because the investigated energies are relatively far from the energies of the core states, thus the factor $\varepsilon_{n\boldsymbol{k}} - \varepsilon_{\alpha\boldsymbol{k}}$ in W can be taken constant. To understand the requirement of locality, the Schrödinger equation (19.1.51) expressed in terms of the pseudopotential can be formally rewritten as

$$\mathcal{H} \widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}) + W'(\boldsymbol{r}) \widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_{n\boldsymbol{k}} \widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}) \,, \tag{19.1.59}$$

where

$$W'(\boldsymbol{r})\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\alpha} \int \phi^*_{\alpha}(\boldsymbol{k},\boldsymbol{r}') \left[\varepsilon_{n\boldsymbol{k}} - \varepsilon_{\alpha\boldsymbol{k}}\right] \phi_{\alpha}(\boldsymbol{k},\boldsymbol{r})\widetilde{\psi}_{n\boldsymbol{k}}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'. \qquad (19.1.60)$$

Since the functions $\phi_\alpha(\boldsymbol{k},\boldsymbol{r})$ describe core states, when they are given in the Wannier representation there is hardly any overlap between terms that belong to different atoms, thus the potential $W'(\boldsymbol{r})$ takes the form

$$W'(\boldsymbol{r}) = \sum_{\boldsymbol{R}_m} v'_{\mathrm{a}}(\boldsymbol{r} - \boldsymbol{R}_m)\,. \qquad (19.1.61)$$

The rapid decay of the core wavefunctions justifies the approximation that the potential is constant inside the core, while outside it decays as

$$v'_{\mathrm{a}}(\boldsymbol{r}) = A\frac{\mathrm{e}^{-\kappa r}}{r}\,, \qquad (19.1.62)$$

where A and κ have to be determined from the comparison of experimental data with the theoretical values calculated from the band structure. Naturally, other forms are equally possible.

19.2 Variational Methods and Methods Based on Scattering Theory

The previous methods were all based on a set[2] of wavefunctions that satisfy the Bloch condition, and the matrix elements of the Hamiltonian as well as the overlap integrals were determined for these. As an alternative approach, one may first solve the problem of electron states in the region close to the atomic core, inside the Wigner–Seitz cell, in the presence of the potential felt there, and then use variational methods, Green functions, or methods based on scattering theory to determine the band structure of the entire crystal.

19.2.1 Augmented-Plane-Wave Method

In the OPW method the separation between core states characterized by atomic wavefunctions and plane-wave-like valence-band states was based on the energies of the states. In 1937 J. C. Slater proposed using functions of another kind, in which the core part of the wavefunction and the plane-wave-like part in the region between ion cores are treated separately in real space.

Suppose that the potential shows considerable variations only in a small region around the atoms, within a sphere of radius r_{MT} that fits into the Wigner–Seitz cell, and can be approximated by a constant value in the regions between the spheres. This potential is called the *muffin-tin potential.* The equipotential lines of a two-dimensional section are shown in Fig. 19.1.

[2] A complete set in principle, but in practice they contained only a few terms.

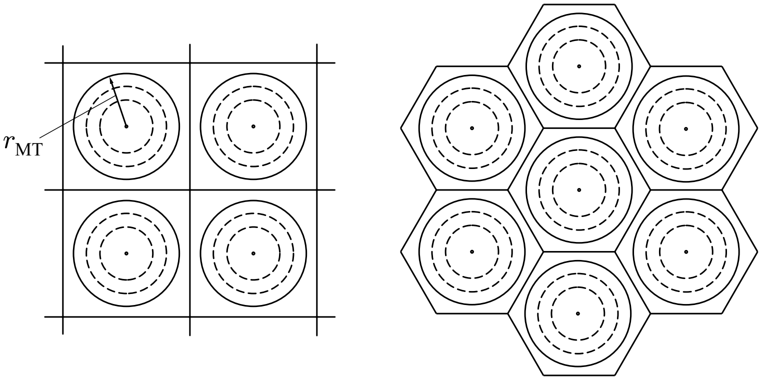


Fig. 19.1. Equipotential lines of the muffin-tin potential in square and hexagonal lattices

In the region between the nonoverlapping muffin-tin spheres, where the potential can be set to zero without loss of generality, the wavefunction can be represented by a plane wave. On the other hand, inside the muffin-tin spheres the wavefunction is obtained as the solution of the Schrödinger equation with the potential.

Consider a single muffin-tin sphere first. Since the potential $v_{\mathrm{a}}(r)$ in its interior is spherically symmetric, the solutions of the Schrödinger equation can be expanded in the spherical harmonics $Y_l^m(\theta,\varphi)$:

$$\phi(\boldsymbol{r}) = \sum_{lm} C_{lm} R_l(\varepsilon, r) Y_l^m(\theta,\varphi)\,, \tag{19.2.1}$$

where θ and φ are the polar and azimuthal angles of the position vector $\boldsymbol{r}$. To determine the amplitude R_l, the Laplacian is written in polar coordinates. Exploiting the properties of spherical harmonics leads to the radial Schrödinger equation

$$-\frac{\hbar^2}{2m_{\mathrm{e}}}\left[\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} - \frac{l(l+1)}{r^2}\right] R_l + v_{\mathrm{a}}(r) R_l = \varepsilon R_l\,, \tag{19.2.2}$$

which has to be solved inside the muffin-tin sphere in the presence of the atomic potential v_{a}.

The coefficients C_{lm} are determined from the requirement that this wavefunction should match smoothly with the plane-wave solution at the surface of the muffin-tin sphere – that is,

$$\phi(\boldsymbol{k},\boldsymbol{r}) = \begin{cases} \sum\limits_{lm} C_{lm}(\boldsymbol{k}) R_l(\varepsilon, r) Y_l^m(\theta,\varphi) & 0 \le r \le r_{\mathrm{MT}}\,, \\ \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r})/\sqrt{V} & r_{\mathrm{MT}} < r \le r_{\mathrm{WS}} \end{cases} \tag{19.2.3}$$

should be continuous at r_{MT}. Such solutions are called *augmented plane waves*, and the method is the *APW method.*

Using expansion (C.4.38) for the plane wave,

$$\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \mathrm{i}^l j_l(kr) Y_l^{m*}(\theta_k, \varphi_k) Y_l^m(\theta, \varphi) , \tag{19.2.4}$$

where θ_k and φ_k are the polar and azimuthal angles of the vector $\boldsymbol{k}$, it is straightforward to show that the $\boldsymbol{k}$-dependent coefficient C_{lm} is

$$C_{lm}(\boldsymbol{k}) = \frac{4\pi}{\sqrt{V}} \mathrm{i}^l Y_l^{m*}(\theta_k, \varphi_k) \frac{j_l(kr_{\mathrm{MT}})}{R_l(\varepsilon, r_{\mathrm{MT}})} . \tag{19.2.5}$$

The wavefunction can also be calculated for a periodic array of muffin-tin spheres, when the lattice-periodic potential is the sum of the potentials of individual muffin-tin spheres:

$$U(\boldsymbol{r}) = \sum_n v_{\mathrm{a}}(\boldsymbol{r} - \boldsymbol{R}_n) . \tag{19.2.6}$$

In the region between muffin-tin spheres the plane wave obviously satisfies the Bloch condition. The same condition can be satisfied within the muffin-tin spheres if the wavefunctions inside individual spheres are summed using suitably chosen phase factors:

$$\phi(\boldsymbol{k}, \boldsymbol{r} + \boldsymbol{R}_n) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_n} \phi(\boldsymbol{k}, \boldsymbol{r}) . \tag{19.2.7}$$

Nonetheless we shall not be concerned with this condition below, since it is sufficient to solve the eigenvalue problem inside a single cell.

These functions are eigenfunctions of the Schrödinger equation (with eigenvalue ε) inside the muffin-tin sphere, but not outside of it, therefore the linear combination of several augmented plane waves needs to be taken. Since the total wavefunction must also satisfy the Bloch condition with a wave vector $\boldsymbol{k}$, we shall choose, just like in the plane-wave and orthogonalized-plane-wave methods, those functions as $\phi_j(\boldsymbol{k}, \boldsymbol{r})$ that can be obtained from $\phi(\boldsymbol{k}, \boldsymbol{r})$ by the substitution $\boldsymbol{k} \to \boldsymbol{k} + \boldsymbol{G}_j$:

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_j c_j(\boldsymbol{k}) \phi_j(\boldsymbol{k}, \boldsymbol{r}) = \sum_j c(\boldsymbol{k} + \boldsymbol{G}_j) \phi(\boldsymbol{k} + \boldsymbol{G}_j, \boldsymbol{r}) . \tag{19.2.8}$$

In what follows, we shall use the notation $\boldsymbol{k}_j = \boldsymbol{k} + \boldsymbol{G}_j$.

At this point we could revert to the matrix method. The difficulty lies in the fact that even though the function defined in (19.2.3) is continuous at the surface of the muffin-tin spheres, its derivative is not. For physical reasons, the continuity of its derivative has to be ensured, too. This requirement can be fulfilled most easily by using variational methods.

It is well known that for the lowest-energy state the Schrödinger equation is equivalent to the statement that the quantity

$$\int \psi_{\boldsymbol{k}}^*(\boldsymbol{r})\,[\mathcal{H}(\boldsymbol{r})-\varepsilon]\,\psi_{\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{19.2.9}$$

has its minimum at the actual wavefunction; ε is a Lagrange multiplier through which the normalization condition of the wavefunction is taken into account, and the minimum of the energy is given by the value of ε at the minimum. We shall now demonstrate that for integrals over the Wigner–Seitz cell this problem is equivalent to the variational problem for the quantity

$$\frac{\hbar^2}{2m_\mathrm{e}}\int\limits_v |\boldsymbol{\nabla}\psi_{\boldsymbol{k}}(\boldsymbol{r})|^2\,\mathrm{d}\boldsymbol{r}+\int\limits_v \psi_{\boldsymbol{k}}^*(\boldsymbol{r})\,[v_\mathrm{a}(\boldsymbol{r})-\varepsilon]\,\psi_{\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{19.2.10}$$

Taking their difference and applying Green's theorem, we obtain

$$\begin{aligned}\int\limits_v \Big[|\boldsymbol{\nabla}\psi_{\boldsymbol{k}}(\boldsymbol{r})|^2+\psi_{\boldsymbol{k}}^*(\boldsymbol{r})\boldsymbol{\nabla}^2\psi_{\boldsymbol{k}}(\boldsymbol{r})\Big]\,\mathrm{d}\boldsymbol{r} &= \int\limits_v \boldsymbol{\nabla}\left[\psi_{\boldsymbol{k}}^*(\boldsymbol{r})\boldsymbol{\nabla}\psi_{\boldsymbol{k}}(\boldsymbol{r})\right]\,\mathrm{d}\boldsymbol{r}\\ &= \int \mathrm{d}\boldsymbol{S}\,\psi_{\boldsymbol{k}}^*(\boldsymbol{r})\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r})}{\partial\boldsymbol{n}(\boldsymbol{r})}\,,\end{aligned} \tag{19.2.11}$$

where the integral is over the surface $\boldsymbol{S}$ of the Wigner–Seitz cell, and $\boldsymbol{n}(\boldsymbol{r})$ is the unit normal of the surface at $\boldsymbol{r}$. Owing to the Bloch condition, on opposite faces of the Wigner–Seitz cell separated by a lattice vector $\boldsymbol{R}_m$,

$$\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R}_m)}{\partial\boldsymbol{n}(\boldsymbol{r}+\boldsymbol{R}_m)}=-\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_m}\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r})}{\partial\boldsymbol{n}(\boldsymbol{r})}\,. \tag{19.2.12}$$

Making use of this relation and the Bloch condition for $\psi_{\boldsymbol{k}}^*$, the right-hand side of (19.2.11) is found to vanish.

When the wavefunction in (19.2.10) is expanded in terms of the augmented plane waves, and the coefficients $c(\boldsymbol{k}_j)$ in (19.2.8) are varied, some tedious algebra, which we shall not reproduce here, leads to the following determinant equation for the eigenvalues:

$$\boxed{\det\left(\left[\frac{\hbar^2}{2m_\mathrm{e}}(\boldsymbol{k}+\boldsymbol{G}_i)^2-\varepsilon\right]\delta_{ij}+\Gamma_{ij}^{\mathrm{APW}}\right)=0\,,} \tag{19.2.13}$$

where

$$\begin{aligned}\Gamma_{ij}^{\mathrm{APW}} = \frac{4\pi r_\mathrm{MT}^2}{V}\frac{\hbar^2}{2m_\mathrm{e}}\bigg\{&-\left[(\boldsymbol{k}+\boldsymbol{G}_i)(\boldsymbol{k}+\boldsymbol{G}_j)-\frac{2m_\mathrm{e}\varepsilon}{\hbar^2}\right]\frac{j_1(|\boldsymbol{G}_i-\boldsymbol{G}_j|r_\mathrm{MT})}{|\boldsymbol{G}_i-\boldsymbol{G}_j|}\\ &+\sum_l(2l+1)P_l(\cos\theta_{ij})j_l(|\boldsymbol{k}+\boldsymbol{G}_i|r_\mathrm{MT})j_l(|\boldsymbol{k}+\boldsymbol{G}_j|r_\mathrm{MT})\frac{R_l'(\varepsilon,r_\mathrm{MT})}{R_l(\varepsilon,r_\mathrm{MT})}\bigg\},\end{aligned} \tag{19.2.14}$$

and θ_{ij} is the angle between $\boldsymbol{k}_i$ and $\boldsymbol{k}_j$. Comparison with (19.1.26) shows that the obtained formula is similar to that of the plane-wave method, but now $\varGamma_{ij}^{\mathrm{APW}}$ appears instead of the matrix elements of the potential. The $\varGamma_{ij}^{\mathrm{APW}}$ depend on the potential only through the values of the radial wavefunction and its derivative at the muffin-tin radius. Since the functions R_l inside the muffin-tin sphere also depend on the energy ε itself, this is not a simple eigenvalue problem. It can be solved by iteration, improving the energy and wavefunction step by step. The more complicated form is more than compensated for by the more rapid convergence of the method compared to previously discussed techniques.

Note that if the primitive cell contains several atoms then the muffin-tin spheres have to be constructed for each of them and the appropriate functions have to be matched at the surface for each sphere.

19.2.2 Green Function or KKR Method

As put forward by J. Korringa (1947), W. Kohn, and N. Rostoker (1954), the band structure in a periodic potential can also be considered from the standpoint that Bloch states are formed as a consequence of multiple scattering events by the periodic potential. The *Korringa–Kohn–Rostocker* or *KKR method* uses scattering theoretical methods to determine the band structure. The method is often referred to as the *Green function method*, since it is customarily formulated in terms of the Green function of electrons moving in the lattice.

The potential of choice is the muffin-tin potential in this method, too, which means that the total potential is written as the sum of spherically symmetric atomic potentials. However, the radius of the sphere is not the radius r_{MT} of the muffin-tin sphere inscribed in the Wigner–Seitz cell; instead it is chosen in such a way that the volume of the sphere be equal to that of the cell. Even though the spheres overlap, they do so only slightly, thus the effects of overlapping are expected to be small. This approach is called the *atomic-sphere approximation* (ASA).

In connection with the scattering of free electrons by impurities, the free-electron Green function was introduced in (16.4.3) with the definition

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 - \varepsilon\right] G(\boldsymbol{r}-\boldsymbol{r}') = -\delta(\boldsymbol{r}-\boldsymbol{r}')\,. \tag{19.2.15}$$

It was shown there that its solution is

$$G(\boldsymbol{r}-\boldsymbol{r}') = -\frac{m_{\mathrm{e}}}{2\pi\hbar^2}\frac{\mathrm{e}^{\mathrm{i}\kappa|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,, \tag{19.2.16}$$

where $\kappa = \sqrt{2m_{\mathrm{e}}\varepsilon/\hbar^2}$, and also that the real part of the right-hand side,

$$G_{\mathrm{s}}(\boldsymbol{r}-\boldsymbol{r}') = -\frac{m_{\mathrm{e}}}{2\pi\hbar^2}\frac{\cos\kappa|\boldsymbol{r}-\boldsymbol{r}'|}{|\boldsymbol{r}-\boldsymbol{r}'|}\,, \tag{19.2.17}$$

is a solution in itself. As we shall see the two choices correspond to traveling- and standing-wave solutions. Expanding them into spherical harmonics and Bessel functions in the region $r < r'$,

$$G(\boldsymbol{r}-\boldsymbol{r}') = \frac{2m_{\mathrm{e}}}{\hbar^2}\kappa \sum_{lm} Y_l^m(\theta,\varphi) Y_l^{m*}(\theta',\varphi') j_l(\kappa r)\left[n_l(\kappa r') - \mathrm{i} j_l(\kappa r')\right], \tag{19.2.18}$$

where θ, φ and θ', φ' are the polar and azimuthal angles of the vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$, respectively, and

$$G_{\mathrm{s}}(\boldsymbol{r}-\boldsymbol{r}') = \frac{2m_{\mathrm{e}}}{\hbar^2}\kappa \sum_{lm} Y_l^m(\theta,\varphi) Y_l^{m*}(\theta',\varphi') j_l(\kappa r) n_l(\kappa r'). \tag{19.2.19}$$

The formula valid in the region $r > r'$ is obtained by exchanging the variables r and r' and the corresponding angular variables.

In the presence of the atomic potentials the Schrödinger equation can be written as

$$\left[-\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 - \varepsilon\right]\psi(\boldsymbol{r}) = -U(\boldsymbol{r})\psi(\boldsymbol{r}). \tag{19.2.20}$$

It is readily seen from this form that the solution satisfies the homogeneous integral equation

$$\psi(\boldsymbol{r}) = \int G(\boldsymbol{r}-\boldsymbol{r}')U(\boldsymbol{r}')\psi(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'. \tag{19.2.21}$$

If the potential is the sum of isolated muffin-tin potentials, and the solution in a periodic potential is characterized by a wave vector $\boldsymbol{k}$,

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_n \int G(\boldsymbol{r}-\boldsymbol{r}')v_{\mathrm{a}}(\boldsymbol{r}'-\boldsymbol{R}_n)\psi_{\boldsymbol{k}}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'. \tag{19.2.22}$$

By changing the variables and making use of the translational properties of the Bloch functions, this can be rewritten as

$$\begin{aligned}\psi_{\boldsymbol{k}}(\boldsymbol{r}) &= \sum_l \int_v G(\boldsymbol{r}-\boldsymbol{r}'-\boldsymbol{R}_l)v_{\mathrm{a}}(\boldsymbol{r}')\psi_{\boldsymbol{k}}(\boldsymbol{r}'+\boldsymbol{R}_l)\,\mathrm{d}\boldsymbol{r}' \\ &= \int_v G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}')v_{\mathrm{a}}(\boldsymbol{r}')\psi_{\boldsymbol{k}}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}',\end{aligned} \tag{19.2.23}$$

where integration is over a single Wigner–Seitz cell (or atomic sphere), and

$$\begin{aligned}G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}') &= \sum_n \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_n} G(\boldsymbol{r}-\boldsymbol{r}'-\boldsymbol{R}_n) \\ &= -\frac{m_{\mathrm{e}}}{2\pi\hbar^2}\sum_n \frac{\exp(\mathrm{i}\kappa|\boldsymbol{r}-\boldsymbol{r}'-\boldsymbol{R}_n|)}{|\boldsymbol{r}-\boldsymbol{r}'-\boldsymbol{R}_n|}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_n}.\end{aligned} \tag{19.2.24}$$

This quantity, which satisfies the relation

$$G(\boldsymbol{k},\boldsymbol{r}+\boldsymbol{R}_m-\boldsymbol{r}') = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_m} G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}') \tag{19.2.25}$$

for translations, is the *structural Green function*, since everything that is characteristic of the structure (i.e., the arrangement of atoms) is lumped into it. Separating the parts that are regular and singular at $r = r'$, the form

$$\begin{aligned} G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}') =& \frac{2m_\mathrm{e}}{\hbar^2} \sum_{lm,l'm'} j_l(\kappa r) \left\{ A_{lm,l'm'} j_{l'}(\kappa r') \right. \\ & \left. + \kappa\delta_{lm,l'm'} n_l(\kappa r')\right\} Y_l^m(\theta,\varphi) Y_{l'}^{m'*}(\theta',\varphi') \end{aligned} \tag{19.2.26}$$

is assumed, as suggested by the free Green function formula (19.2.18). The coefficients will be determined from an alternative representation of the Green function.

Using now plane waves for $\phi_{\boldsymbol{k}}(\boldsymbol{r})$ in the method employed in Section 17.6, the Green function is expanded as

$$G(\boldsymbol{r}-\boldsymbol{r}') = \sum_{\boldsymbol{k}} a_{\boldsymbol{k}}(\boldsymbol{r}')\phi_{\boldsymbol{k}}(\boldsymbol{r}) . \tag{19.2.27}$$

Substituting this formula into the equation for the Green function, and making use of the identity

$$\left[-\frac{\hbar^2}{2m_\mathrm{e}}\boldsymbol{\nabla}^2 - \varepsilon_{\boldsymbol{k}} \right] \phi_{\boldsymbol{k}}(\boldsymbol{r}) = 0 \tag{19.2.28}$$

for plane waves it is straightforward to show that

$$a_{\boldsymbol{k}}(\boldsymbol{r}') = \frac{\phi_{\boldsymbol{k}}^*(\boldsymbol{r}')}{\varepsilon-\varepsilon_{\boldsymbol{k}}} , \tag{19.2.29}$$

hence

$$G(\boldsymbol{r}-\boldsymbol{r}') = \sum_{\boldsymbol{k}} \frac{\phi_{\boldsymbol{k}}^*(\boldsymbol{r}')\phi_{\boldsymbol{k}}(\boldsymbol{r})}{\varepsilon-\varepsilon_{\boldsymbol{k}}} . \tag{19.2.30}$$

The part of the Green function that transforms according to the wave vector $\boldsymbol{k}$ under translations is obtained by summing over the equivalent vectors

$$\boldsymbol{k}_j = \boldsymbol{k} + \boldsymbol{G}_j . \tag{19.2.31}$$

Since the corresponding wavefunctions and energies are

$$\phi_j(\boldsymbol{r}) = \frac{1}{\sqrt{V}} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot\boldsymbol{r}} \qquad \text{and} \qquad \varepsilon_j = \frac{\hbar^2(\boldsymbol{k}+\boldsymbol{G}_j)^2}{2m_\mathrm{e}} , \tag{19.2.32}$$

the solution of the equation for the Green function is

$$G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}') = \frac{1}{V} \sum_j \frac{\mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_j)\cdot(\boldsymbol{r}-\boldsymbol{r}')}}{\varepsilon - \hbar^2(\boldsymbol{k}+\boldsymbol{G}_j)^2/2m_\mathrm{e}} . \tag{19.2.33}$$

The series expansion of the exponential according to (C.4.38) gives

$$G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}') = -\frac{(4\pi)^2}{V}\sum_{\substack{l,m\\ l',m'}}\sum_j \mathrm{i}^{l-l'}\frac{j_l(|\boldsymbol{k}_j|r)j_{l'}(|\boldsymbol{k}_j|r')}{\hbar^2\boldsymbol{k}_j^2/2m_\mathrm{e}-\varepsilon} \times Y_l^m(\theta,\varphi)Y_{l'}^{m'*}(\theta',\varphi')Y_l^{m*}(\theta_j,\varphi_j)Y_{l'}^{m'}(\theta_j,\varphi_j)\,, \tag{19.2.34}$$

where θ, φ and θ', φ' are the polar and azimuthal angles of the vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$, respectively.

From the comparison of the two formulas for the Green function we have

$$A_{lm,l'm'} = \frac{(4\pi)^2}{V}\frac{\mathrm{i}^{l-l'}}{j_l(\kappa r)j_{l'}(\kappa r')}\sum_j \frac{j_l(|\boldsymbol{k}_j|r)j_{l'}(|\boldsymbol{k}_j|r')}{\kappa^2-\boldsymbol{k}_j^2} \times Y_l^{m*}(\theta_j,\varphi_j)Y_{l'}^{m'}(\theta_j,\varphi_j) - \kappa\delta_{ll'}\delta_{mm'}\frac{n_l(\kappa r')}{j_l(\kappa r')}\,. \tag{19.2.35}$$

Through the sum over the vectors of the reciprocal lattice, these coefficients contain information about the structure, which is why they are called *structure constants*. They can be determined once and for all, and they can be looked up in standard references.

When the Green function is known, the wavefunction and the energy eigenvalue can be determined using (19.2.23). Instead of trying to solve these equations in a self-consistent way, we shall demonstrate that they can be derived from a variational problem – much in the same way as in the APW method. Starting with the quantity

$$\Lambda = \int_v \psi_{\boldsymbol{k}}^*(\boldsymbol{r})v_\mathrm{a}(\boldsymbol{r})\psi_{\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} - \int_v\int_{v'} \psi_{\boldsymbol{k}}^*(\boldsymbol{r})v_\mathrm{a}(\boldsymbol{r})G(\boldsymbol{k},\boldsymbol{r}-\boldsymbol{r}')v_\mathrm{a}(\boldsymbol{r}')\psi_{\boldsymbol{k}}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{r}\,, \tag{19.2.36}$$

the requirement

$$\delta\Lambda = 0 \tag{19.2.37}$$

imposed on its variations with respect to $\psi_{\boldsymbol{k}}^*(\boldsymbol{r})$ gives (19.2.23), regardless of whether or not the function satisfies the boundary condition. It can also be shown that Λ itself also vanishes when the exact function $\psi_{\boldsymbol{k}}(\boldsymbol{r})$ is used. Therefore when the equation

$$\Lambda(\psi,\boldsymbol{k},\varepsilon) = 0 \tag{19.2.38}$$

is solved for ε using an approximate function $\psi(\boldsymbol{r})$, the error will be small of the second order. We shall exploit this property to determine the eigenfunctions and eigenvalues from this variational problem.

Adapted to the muffin-tin potential, the wavefunction is expanded in spherical harmonics,

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{lm} C_{lm}(\boldsymbol{k}) R_l(\varepsilon, r) Y_l^m(\theta, \varphi)\,, \tag{19.2.39}$$

where $R_l(\varepsilon, r)$ is the solution of the radial Schrödinger equation. Substituting this into the expression for Λ leads to a formula that is quadratic in $C_{lm}(\boldsymbol{k})$:

$$\Lambda = \sum_{lm,l'm'} C^*_{lm}(\boldsymbol{k}) \Lambda_{lm,l'm'}(\boldsymbol{k}) C_{l'm'}(\boldsymbol{k})\,. \tag{19.2.40}$$

Stationarity requires that

$$\sum_{l'm'} \Lambda_{lm,l'm'}(\boldsymbol{k}) C_{l'm'}(\boldsymbol{k}) = 0 \tag{19.2.41}$$

for every l and m. Nontrivial solutions exist if

$$\det\left(\Lambda_{lm,l'm'}\right) = 0\,. \tag{19.2.42}$$

By determining the matrix elements in this formula, and using the solution of the equation (19.2.15) for the Green function that also satisfies condition (19.2.25), the following form is obtained after some tedious algebra:

$$\det\left(A_{lm,l'm'} + \kappa\delta_{ll'}\delta_{mm'} \frac{L_l(\kappa, r_{\mathrm{MT}}) n_l(\kappa r_{\mathrm{MT}}) - \kappa n'_l(\kappa r_{\mathrm{MT}})}{L_l(\kappa, r_{\mathrm{MT}}) j_l(\kappa r_{\mathrm{MT}}) - \kappa j'_l(\kappa r_{\mathrm{MT}})}\right) = 0\,, \tag{19.2.43}$$

where

$$L_l(\kappa, r_{\mathrm{MT}}) = \frac{1}{R_l}\frac{\mathrm{d}R_l}{\mathrm{d}r}\bigg|_{r=r_{\mathrm{MT}}}\,. \tag{19.2.44}$$

In addition to the spherical Bessel and Neumann functions, this formula also contains their derivatives

$$j'_l(x) = \frac{\mathrm{d}j_l(x)}{\mathrm{d}x} \qquad \text{and} \qquad n'_l(x) = \frac{\mathrm{d}n_l(x)}{\mathrm{d}x}\,. \tag{19.2.45}$$

The potential appears only through the values of the wavefunction and its derivative at the muffin-tin radius.

Note that by using (19.2.35) for $A_{lm,l'm'}$, the equation for the energy eigenvalue can be written in a form analogous to (19.2.13) in the KKR method, too:

$$\boxed{\det\left(\left[\frac{\hbar^2}{2m_{\mathrm{e}}}(\boldsymbol{k} + \boldsymbol{G}_i)^2 - \varepsilon\right]\delta_{ij} + \Gamma_{ij}^{\mathrm{KKR}}\right) = 0\,,} \tag{19.2.46}$$

where

$$\begin{aligned} \Gamma_{ij}^{\mathrm{KKR}} = \frac{4\pi r_{\mathrm{MT}}^2}{V}\frac{2m_{\mathrm{e}}}{\hbar^2}\sum_l (2l+1) P_l(\cos\theta_{ij}) j_l(|\boldsymbol{k}_i| r_{\mathrm{MT}}) j_l(|\boldsymbol{k}_j| r_{\mathrm{MT}}) \\ \times \left[\frac{R'_l(\kappa, r_{\mathrm{MT}})}{R_l(\kappa, r_{\mathrm{MT}})} - \kappa\frac{j'_l(\kappa r_{\mathrm{MT}})}{j_l(\kappa r_{\mathrm{MT}})}\right]\,. \end{aligned} \tag{19.2.47}$$

Comparison with the results obtained with the APW method gives

$$\Gamma_{ij}^{\mathrm{KKR}} = \Gamma_{ij}^{\mathrm{APW}} - \Gamma_{ij}^{0} , \tag{19.2.48}$$

where Γ^0 is the empty-lattice value for Γ^{APW}; in this case the radial function R_l in (19.2.14) is just the spherical Bessel function j_l.

19.2.3 Physical Interpretation of the KKR Method

The results obtained with the Green function method can be given a simple interpretation when the Bloch states are considered to arise from the interference of beams scattered multiply by ions. Before showing this, we rewrite the KKR equations in an equivalent form.

As discussed in Chapter 16 in connection with the state of electrons scattered by an impurity, the scattered partial waves can be characterized by phase shifts δ_l with respect to the incoming partial wave in the region where the potential is negligible. The potential appears only through these δ_l. We shall now show that the quantity

$$\frac{L_l(\kappa, r_{\mathrm{MT}}) n_l(\kappa r_{\mathrm{MT}}) - \kappa n_l'(\kappa r_{\mathrm{MT}})}{L_l(\kappa, r_{\mathrm{MT}}) j_l(\kappa r_{\mathrm{MT}}) - \kappa j_l'(\kappa r_{\mathrm{MT}})} , \tag{19.2.49}$$

which appears in (19.2.43) in addition to the structure constant, is related to the phase shift in a particularly simple manner, and that (19.2.43) is just the self-consistency condition that the wave incident on any muffin-tin sphere should be equal to the sum of the waves scattered by all others.

In the region outside the muffin-tin sphere, where the potential vanishes, the radial Schrödinger equation (19.2.2) can be solved in terms of spherical Bessel and Neumann functions by making use of (C.3.43). The complete solution is sought in the form of the linear combination

$$R_l(\kappa, r) \sim j_l(\kappa r) \cos \delta_l - n_l(\kappa r) \sin \delta_l , \tag{19.2.50}$$

where δ_l is the phase shift of the lth partial wave, since asymptotic forms imply

$$R_l(\kappa, r) \sim \frac{1}{\kappa r} \sin[\kappa r + \delta_l - \pi l/2] . \tag{19.2.51}$$

When this formula, which is valid outside the muffin-tin sphere, is matched at the surface of the sphere to the value obtained from the numerical solution of the Schrödinger equation in the interior, the relation

$$\begin{aligned} L_l(\kappa, r_{\mathrm{MT}}) &= \frac{1}{R_l(\kappa, r_{\mathrm{MT}})} \left. \frac{\mathrm{d}R_l(\kappa, r)}{\mathrm{d}r} \right|_{r=r_{\mathrm{MT}}} \\ &= \kappa \frac{j_l'(\kappa r_{\mathrm{MT}}) \cos \delta_l - n_l'(\kappa r_{\mathrm{MT}}) \sin \delta_l}{j_l(\kappa r_{\mathrm{MT}}) \cos \delta_l - n_l(\kappa r_{\mathrm{MT}}) \sin \delta_l} \end{aligned} \tag{19.2.52}$$

arises, in which the prime stands for the derivative with respect to the argument. Then

$$\tan\delta_l = \frac{L_l(\kappa, r_{\mathrm{MT}}) j_l(\kappa r_{\mathrm{MT}}) - \kappa j_l'(\kappa r_{\mathrm{MT}})}{L_l(\kappa, r_{\mathrm{MT}}) n_l(\kappa r_{\mathrm{MT}}) - \kappa n_l'(\kappa r_{\mathrm{MT}})}\,, \tag{19.2.53}$$

which is precisely the inverse of the quantity which appeared in (19.2.43), therefore the KKR equations can be rewritten as

$$\boxed{\det\left(A_{lm,l'm'} + \kappa\delta_{ll'}\delta_{mm'}\cot\delta_l\right) = 0\,.} \tag{19.2.54}$$

The effects of the potential can be fully absorbed in the phase shifts. This form will now be given an intuitive interpretation.

Had we used the spherical Hankel functions $h_l^{(1)} = j_l + \mathrm{i}n_l$ instead of the spherical Neumann functions in (19.2.50), we would have started with the form

$$R_l(\kappa, r) \sim j_l(\kappa r) + \mathrm{i}\mathrm{e}^{\mathrm{i}\delta_l}\sin\delta_l h_l^{(1)}(\kappa r)\,. \tag{19.2.55}$$

Generalizing this, in the presence of a single scattering potential, the wavefunction can be expanded into spherical Bessel and Hankel functions as

$$\phi(\boldsymbol{r}) = \sum_{lm}\left[a_{lm} j_l(\kappa r) + b_{lm} h_l^{(1)}(\kappa r)\right] Y_l^m(\theta, \varphi)\,, \tag{19.2.56}$$

where the coefficients are related by

$$b_{lm} = \mathrm{i}\mathrm{e}^{\mathrm{i}\delta_l}\sin\delta_l\, a_{lm}\,. \tag{19.2.57}$$

In view of the asymptotic behavior of the spherical Bessel functions at small and large values of r, the first term may be considered as an incoming wave, and the second as an outgoing wave.

For a periodic array of muffin-tin potentials the total wavefunction is chosen as

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\phi(\boldsymbol{r} - \boldsymbol{R}_j) \tag{19.2.58}$$

to meet the Bloch condition. Now consider a point $\boldsymbol{r}$ outside the muffin-tin sphere in the Wigner–Seitz cell of the ith lattice point. The wavefunction can be considered to be composed of incoming and outgoing terms. The system is in a stationary state when the incoming wave in the ith cell is equal to the sum of the outgoing waves from the other ($j \neq i$) cells. For the cell around the origin this implies

$$\begin{aligned} &\sum_{lm} a_{lm} j_l(\kappa|\boldsymbol{r}|) Y_l^m(\theta_i, \varphi_i) \\ &\qquad = \sum_{j\neq i}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j}\sum_{l'm'} b_{l'm'} h_l^{(1)}(\kappa|\boldsymbol{r} - \boldsymbol{R}_j|) Y_{l'}^{m'}(\theta_j, \varphi_j)\,, \end{aligned} \tag{19.2.59}$$

where θ_j and φ_j are the angular variables of the vector $\boldsymbol{r} - \boldsymbol{R}_j$. Using the common notation f for spherical Bessel, Neumann, and Hankel functions, and making use of their addition theorems,

$$f_l(r'')Y_l^m(\theta'',\varphi'') = \sum_{\substack{l',l'' \\ m'}} \mathrm{i}^{l'+l''-l}(-1)^m(2l'+1)(2l''+1)C(l,l',l'';m,m') \times j_{l'}(r)Y_{l'}^{m'}(\theta,\varphi)f_{l''}(r')Y_{l''}^{m-m'}(\theta',\varphi')\,, \tag{19.2.60}$$

where $\boldsymbol{r}'' = \boldsymbol{r} + \boldsymbol{r}'$ (while θ'' and φ'' are the corresponding angular variables), and C can be expressed in terms of Wigner $3j$ symbols or Clebsch–Gordan coefficients. Rewriting the right-hand side of (19.2.59), and equating the coefficient of $j_l(r)Y_l^m(\theta,\varphi)$ on the two sides leads to an equation of the form

$$a_{lm} = \sum_{l'm'} G_{lm,l'm'}b_{l'm'}\,. \tag{19.2.61}$$

Exploiting (19.2.57), the homogeneous system of equations

$$\sum_{l'm'}\left[G_{lm,l'm'} + \mathrm{i}\frac{\mathrm{e}^{-\mathrm{i}\delta_l}}{\sin\delta_l}\delta_{ll'}\delta_{mm'}\right]b_{l'm'} = 0 \tag{19.2.62}$$

is obtained for the coefficients b_{lm}. Introducing $A_{lm,l'm'}$ defined by

$$G_{lm,l'm'} = \frac{\mathrm{i}}{\kappa}A_{lm,l'm'} - \delta_{ll'}\delta_{mm'}\,, \tag{19.2.63}$$

and making use of the specific form of the coefficients $G_{lm,l'm'}$, it can be demonstrated that this leads to a system of equations that is indeed equivalent to the KKR equation (19.2.54). It is also clear from the result that the structure constant $A_{lm,l'm'}$ depends on the structure alone, and that the potential appears only through the phase shifts. The convergence of the method is determined by the phase shifts. In general, it is sufficient to focus on the partial waves $l = 0$, 1, 2, 3, leading to a 16 × 16 determinant.

19.2.4 LMTO Method

In the matrix methods we chose a complete set of basis functions that were independent of the energy to be determined, and had to compute the eigenvalues of the matrix made up of the matrix elements in this basis.

On the other hand, we used energy-dependent partial waves in the muffin-tin method, and the equations for the one-particle energies were obtained from the matching conditions. The resulting system of equations is nonlinear in the energy, and therefore its numerical solution is more difficult than for a fixed basis. This difficulty can be overcome by applying a linear method based on energy-independent muffin-tin orbitals, the LMTO (linear muffin-tin orbital) method.

The wavefunction $\psi_{\boldsymbol{k}}(\boldsymbol{r})$ is once again constructed from the solutions obtained for individual muffin-tin spheres, as

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{lm} B_{lm}(\boldsymbol{k}) \sum_{j} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} \phi_{lm}(\boldsymbol{r} - \boldsymbol{R}_j)\,, \tag{19.2.64}$$

with a suitable choice of the muffin-tin orbitals $\phi_{lm}(\boldsymbol{r})$. Owing to the spherical symmetry of the potential, the angular dependence can be specified in terms of spherical harmonics, while in the radial part, inside the radius r_{AS} of the atomic sphere a suitably chosen smooth function is added to the function $R_l(\varepsilon, r)$ obtained from the solution of the radial Schrödinger equation:

$$\phi_{lm}(\boldsymbol{r}) = \mathrm{i}^l Y_l^m(\theta, \varphi) \left[R_l(\varepsilon, r) + p_l(\varepsilon) \left(\frac{r}{r_{\mathrm{AS}}} \right)^l \right] \qquad r < r_{\mathrm{AS}}\,. \tag{19.2.65}$$

Outside the atomic sphere, where the potential vanishes, the radial wavefunction is now not chosen as a plane wave (like in the APW method) but as $(r_{\mathrm{AS}}/r)^{l+1}$, the solution of the Laplace equation with zero kinetic energy that vanishes at infinity, since it satisfies

$$\boldsymbol{\nabla}^2 \left(\frac{r_{\mathrm{AS}}}{r} \right)^{l+1} \equiv \left[\frac{\partial^2}{\partial r^2} + \frac{2}{r} \frac{\partial}{\partial r} \right] \left(\frac{r_{\mathrm{AS}}}{r} \right)^{l+1} = \frac{l(l+1)}{r^2} \left(\frac{r_{\mathrm{AS}}}{r} \right)^{l+1}\,, \tag{19.2.66}$$

hence

$$\phi_{lm}(\boldsymbol{r}) = \mathrm{i}^l Y_l^m(\theta, \varphi) \left(\frac{r_{\mathrm{AS}}}{r} \right)^{l+1} \qquad r > r_{\mathrm{AS}}. \tag{19.2.67}$$

The coefficient $p_l(\varepsilon)$ is determined from the condition that the wavefunction should be continuous across the surface of the sphere,

$$p_l(\varepsilon) = \frac{D_l(\varepsilon) + l + 1}{D_l(\varepsilon) - l}\,, \tag{19.2.68}$$

where

$$D_l(\varepsilon) = \frac{r_{\mathrm{AS}}}{R_l(\varepsilon, r_{\mathrm{AS}})} \frac{\partial R_l(\varepsilon, r)}{\partial r} \bigg|_{r = r_{\mathrm{AS}}}\,. \tag{19.2.69}$$

Having fixed the form of the wavefunction, the following procedure could be used to evaluate the single-particle energies. The correct value of the radial wavefunction inside the muffin-tin sphere is known to be $R_l(\varepsilon, r)$. An extra term was added to this inside the muffin-tin sphere. Moreover, the external wavefunctions associated with other muffin-tin spheres also reach into this region. To recover the correct solution, the two must cancel, that is,

$$\begin{aligned} \sum_{lm} B_{lm}(\boldsymbol{k}) \bigg[& p_l(\varepsilon) \mathrm{i}^l Y_l^m(\theta, \varphi) \left(\frac{r}{r_{\mathrm{AS}}} \right)^l \\ & + \sum_{j \neq 0} \mathrm{i}^l Y_l^m(\theta_j, \varphi_j) \left(\frac{r_{\mathrm{AS}}}{|\boldsymbol{r} - \boldsymbol{R}_j|} \right)^{l+1} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} \bigg] = 0\,, \end{aligned} \tag{19.2.70}$$

where θ_j and φ_j are the angular variables of the vector $\boldsymbol{r}-\boldsymbol{R}_j$. Exploiting the known properties of the spherical harmonics, the second term can be expanded about $\boldsymbol{r}=0$ as

$$\begin{aligned}\sum_{j\neq 0} \mathrm{i}^l Y_l^m(\theta_j,\varphi_j)\left(\frac{r_{\mathrm{AS}}}{|\boldsymbol{r}-\boldsymbol{R}_j|}\right)^{l+1} & \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{R}_j} \\ &= -\sum_{l'm'} \frac{S_{lm,l'm'}(\boldsymbol{k})}{2(2l'+1)}\left(\frac{r}{r_{\mathrm{AS}}}\right)^{l'} \mathrm{i}^{l'} Y_{l'}^{m'}(\theta,\varphi)\,.\end{aligned} \tag{19.2.71}$$

The quantities $S_{lm,l'm'}(\boldsymbol{k})$ defined by this formula are called *canonical structure constants*. Canonical, as they depend on the structure alone and not the energy, therefore they can be tabulated for various points of the Brillouin zone.

Substituting this back into (19.2.70), we find

$$\sum_{ll'mm'} B_{lm}(\boldsymbol{k})\left[p_l(\varepsilon)\delta_{ll'}\delta_{mm'} - \frac{S_{lm,l'm'}(\boldsymbol{k})}{2(2l'+1)}\right]\left(\frac{r}{r_{\mathrm{AS}}}\right)^{l'} \mathrm{i}^{l'} Y_{l'}^{m'}(\theta,\varphi) = 0\,, \tag{19.2.72}$$

and so

$$\sum_{lm} B_{lm}(\boldsymbol{k})\big[2(2l+1)p_l(\varepsilon)\delta_{ll'}\delta_{mm'} - S_{lm,l'm'}(\boldsymbol{k})\big] = 0\,. \tag{19.2.73}$$

The system of homogeneous linear equations for the coefficients $B_{lm}(\boldsymbol{k})$ has nontrivial solutions if

$$\det\Big[2(2l+1)p_l(\varepsilon)\delta_{ll'}\delta_{mm'} - S_{lm,l'm'}(\boldsymbol{k})\Big] = 0\,. \tag{19.2.74}$$

The solutions of this equation are the energies for the wave vector $\boldsymbol{k}$. As mentioned above, the structure constant is energy-independent, and all information about the potential is contained in $p_l(\varepsilon)$ through the logarithmic derivative $D_l(\varepsilon)$.

This procedure is in the same spirit as the one in the KKR method. However, there are two essential differences. Firstly, instead of spherical Bessel functions with energy-dependent arguments, functions associated with zero energy are used outside the muffin-tin sphere. Secondly, the structure constant is also energy-independent. Nonetheless the procedure requires the solution of a system of equations that is nonlinear in the energy. As demonstrated by O. K. Andersen (1975), relatively good accuracy can be achieved by employing yet another simplification: fixing the energy variable in $R_l(\varepsilon,r)$ – and thus in $p_l(\varepsilon)$ – at a particular value (for example the Fermi energy), and using the arising energy-independent wavefunctions $\phi_{lm}(\boldsymbol{r})$ as the set of basis functions in the matrix method to evaluate the energy eigenvalues. The resulting equations provide one of the most efficient methods for calculating the band structure.

19.3 Band Structure and Fermi Surface of Simple Metals

As indicated by the foregoing discussion, band-structure calculations demand considerable numerical effort in general. Without going into details, we shall present the calculated band structure and Fermi surface of mono-, di-, tri-, and tetravalent simple metals, and compare them with the experimental results. As we shall see, the empty-lattice approximation often gives surprisingly good pictures of the real Fermi surfaces of simple metals. We shall also observe the failure of this simple method in transition metals, where *d*-electrons contribute considerably to the metallic properties. We shall present only briefly some characteristic features of their band structure.

Images of the calculated Fermi surfaces of metallic elements are available on various dedicated websites, for example www.phys.ufl.edu/fermisurface/.

19.3.1 Monovalent Metals

The elements of groups 1 (IA) and 11 (IB) of the periodic table – alkali metals and noble metals – are monovalent. Their electronic and crystalline structure are summarized in Table 19.1.

Table 19.1. The electronic and crystalline structure of monovalent metals

Element	Electronic structure	Crystalline structure	Element	Electronic structure	Crystalline structure
Li	$1s^2\,2s^1$	bcc			
Na	[Ne] $3s^1$	bcc			
K	[Ar] $4s^1$	bcc	Cu	[Ar] $3d^{10}\,4s^1$	fcc
Rb	[Kr] $5s^1$	bcc	Ag	[Kr] $4d^{10}\,5s^1$	fcc
Cs	[Xe] $6s^1$	bcc	Au	[Xe] $4f^{14}\,5d^{10}\,6s^1$	fcc

Alkali metals crystallize in body-centered cubic structure – even though for lithium and sodium this phase becomes stable only above $T = 77\,\mathrm{K}$ and $T = 23\,\mathrm{K}$, respectively. Hydrogen is different, as its structure in the solid phase contains two atoms per primitive cell – which explains why it is an insulator rather than a metal. Under high pressure, solid hydrogen also becomes metallic, but the properties of this phase are relatively little known.

Figure 19.2 shows the calculated band structure of sodium in certain special directions of the Brillouin zone of the bcc lattice. Comparison with the result for the empty bcc lattice shown in Fig. 18.4 indicates that accidental degeneracies are lifted at the zone boundary (and also at the zone center for higher-lying bands), and it is also there that the dispersion relation exhibits the largest deviations from the form obtained in the empty-lattice approximation, nevertheless the effects of the potential are essentially weak. In agreement with this observation, the constant-energy surfaces, which are spherical

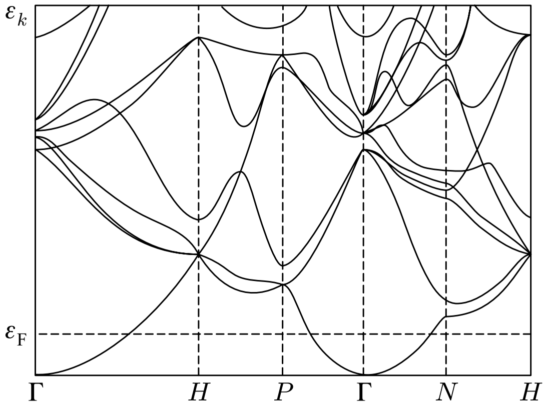


Fig. 19.2. The band structure of sodium, calculated by the LCAO method [Reprinted with permission from W. Y. Ching and J. Callaway, *Phys. Rev. B* **11**, 1324 (1975). ©1975 by the American Physical Society]

in an empty lattice, undergo substantial distortion only in the vicinity of the zone boundary when the periodic potential is turned on. Therefore barring exceptional cases – namely, when the Fermi surface is close to the zone boundary –, the nearly-free-electron model can be a suitable starting point for the calculation of the Fermi surface.

Let us compare the radius of the Fermi sphere obtained in the empty-lattice approximation with the dimensions of the Brillouin zone in a bcc lattice. According to (18.1.17), the Fermi momentum in monovalent metals is

$$k_F = 0.620(2\pi/a)\,. \tag{19.3.1}$$

Among the points located on the boundaries of the rhombic dodecahedral Brillouin zone of the bcc lattice, the face center $N = (2\pi/a)(\frac{1}{2}, \frac{1}{2}, 0)$ is closest to the zone center. Its distance from Γ is

$$\overline{\Gamma N} = (2\pi/a)\sqrt{(\tfrac{1}{2})^2 + (\tfrac{1}{2})^2} = 0.707(2\pi/a)\,. \tag{19.3.2}$$

This is larger than k_F, thus the Fermi sphere is entirely inside the first Brillouin zone in the empty-lattice approximation, as shown in Fig. 18.10. The same is illustrated in two sections of the Brillouin zone in Figs. 19.3(a) and (b).

As k_F is over 10% smaller than $\overline{\Gamma N}$, the measured Fermi surfaces are almost perfectly spherical in alkali metals where the potential is weak: deviations are on the order of a few percent. The distorted Fermi sphere is completely inside the Brillouin zone, as shown in Fig. 19.3(c) for sodium. Consequently, the metallic properties of alkali metals can be very well explained in the free-electron model, just the electron mass needs to be replaced by an effective mass. This effective mass appears in the low-temperature specific heat of electrons, as well as in the cyclotron resonance and the de Haas–van Alphen effect (to be discussed later). The effective masses derived from the experi-

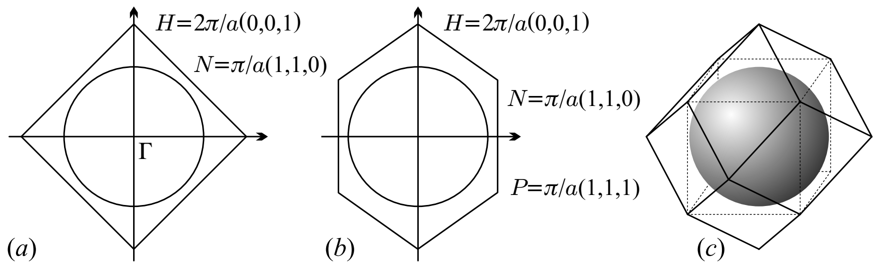


Fig. 19.3. (*a*) and (*b*): Fermi sphere of radius k_F in two sections of the Brillouin zone. (*c*): Fermi surface for the conduction electrons of sodium

mental data on specific heat and the motion of electrons in a magnetic field are listed in Table 19.2.

Table 19.2. The ratio of the Bloch electron effective mass to the electron mass for alkali metals. m^*_ds is the density-of-states effective mass derived from low-temperature specific-heat data, while m_c is determined either from the cyclotron resonance or the de Haas–van Alphen effect. For comparison, the calculated effective mass is also given

Element	$m^*_\mathrm{ds}/m_\mathrm{e}$	$m_\mathrm{c}/m_\mathrm{e}$	$m^*_\mathrm{calc}/m_\mathrm{e}$
Li	2.168	1.8	1.66
Na	1.210	1.24	1.00
K	1.234	1.22	1.09
Rb	1.226	1.20	1.21
Cs	1.355	1.44	1.76

The picture is more complicated for noble metals. Over and above the electrons on the closed core shells, which are irrelevant from the viewpoint of metallic properties, 11 electrons have to be accommodated, which requires six bands. In the tight-binding approximation these are mixed from *d*- and *s*-states. In the LCAO method these six states would be used as basis. Mixing is generally so strong that one can no longer speak of pure *s*- and *d*-type bands. States close to the Fermi energy are dominantly *s*-type, nonetheless the Harrison construction fails to account adequately for the Fermi surface of noble metals.

To understand this, we shall compare the free-electron Fermi wave number with the dimensions of the Brillouin zone in this case, too. According to (18.1.18), for monovalent metals with fcc structure

$$k_\mathrm{F} = 0.782(2\pi/a)\,. \tag{19.3.3}$$

Among the points of the zone boundary, $L = (2\pi/a)(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ is closest to the center, at a distance of $0.866(2\pi/a)$. Thus the Fermi sphere is entirely inside the first Brillouin zone in the empty-lattice approximation, as illustrated in Fig. 18.11. Figure 19.4(*a*) shows the location of the Fermi sphere in a section of the truncated-octahedron-shaped Brillouin zone.

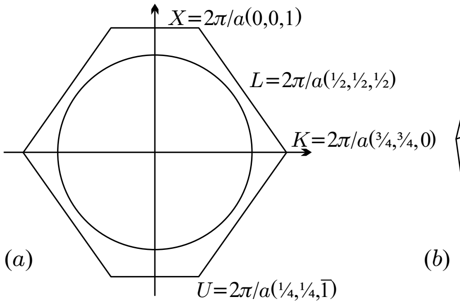


Fig. 19.4. (*a*) Fermi sphere of radius k_{F} in a section of the first Brillouin zone of the fcc lattice. (*b*) Fermi surface for copper

In the presence of a weak periodic potential the Fermi surface is once again expected to be almost spherical, with small humps toward points L. However, as the free-electron Fermi sphere is now closer to the zone boundary than in bcc lattices, a ten percent distortion of the Fermi sphere – in which d-electrons may play a prime role – is sufficient for that the Fermi surface touch the zone boundary. Measurements confirm that the spherical shape is indeed distorted in such a way in noble metals. The Fermi surface is perpendicular to the hexagonal faces in the neighborhood of the eight equivalent points L, as illustrated in Fig. 19.4(*b*). This Fermi surface is similar to that shown in Fig. 18.20. In the repeated-zone scheme the Fermi surface is not made up of disjoint spheres: spherical regions are connected by “necks” along the [111] directions. As we shall see, this has a profound influence on the motion of electrons in magnetic fields.

The calculated band structure of copper is shown in Fig. 19.5. Note that due to the “necks” of the Fermi surface along the [111] directions the conduction-electron band remains below the Fermi energy along the whole line connecting Γ with the center L of the hexagonal face of the Brillouin zone. The Fermi energy is reached only at $\boldsymbol{k}$ vectors lying in the hexagonal face between points L and W.

Similar band structures are obtained for other noble metals, and the Fermi surfaces are also similar: spherical regions are connected by “necks” in the extended-zone scheme. As we shall see in Chapter 21, the measured shapes of the Fermi surfaces are in excellent agreement with this picture.

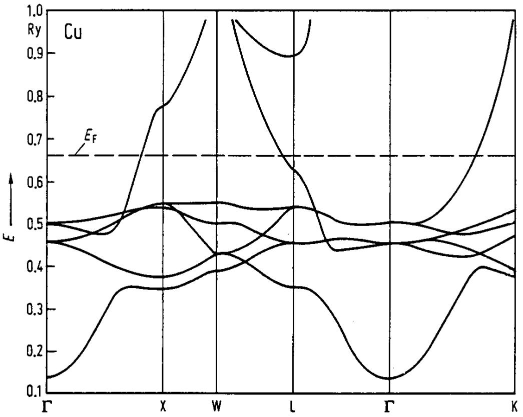


Fig. 19.5. The band structure of copper along some special lines of the Brillouin zone, calculated using the APW method [Reprinted with permission from G. A. Burdick, *Phys. Rev.* **129**, 138 (1963). ©1963 by the American Physical Society]

19.3.2 Divalent Metals

The elements of groups 2 (IIA) and 12 (IIB) of the periodic table – alkaline-earth metals and the elements of the zinc group – are divalent. Their electronic and crystalline structure are summarized in Table 19.3.

Table 19.3. The electronic and crystalline structure of divalent metals

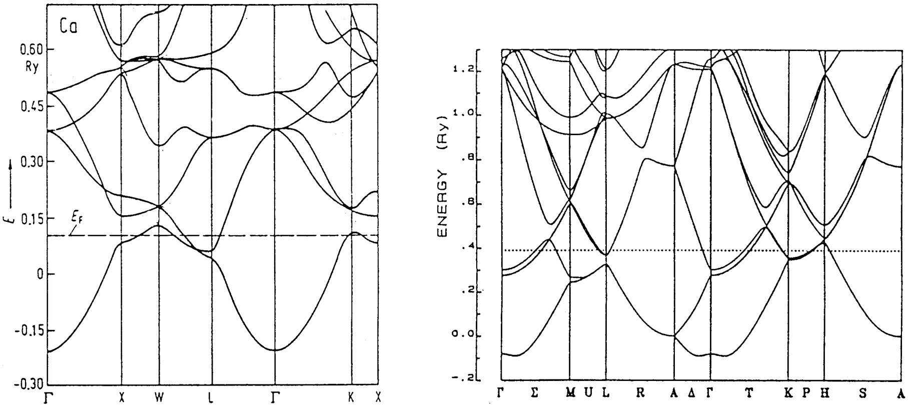
Element	Electronic structure	Crystalline structure	Element	Electronic structure	Crystalline structure
Be	$1s^2\,2s^2$	hcp			
Mg	[Ne] $3s^2$	hcp			
Ca	[Ar] $4s^2$	fcc	Zn	[Ar] $3d^{10}\,4s^2$	hcp
Sr	[Kr] $5s^2$	fcc	Cd	[Kr] $4d^{10}\,5s^2$	hcp
Ba	[Xe] $6s^2$	bcc	Hg	[Xe] $4f^{14}\,5d^{10}\,6s^2$	rhombohedral

Even though the outermost shell contains two electrons in the atomic configuration, these materials are all metals. This can be understood most easily in the nearly-free-electron model. In a bcc lattice $k_{\mathrm{F}} = 0.782\,(2\pi/a)$ for two electrons per atom, which is larger than the distance $\overline{\Gamma N}$, while in an fcc lattice $k_{\mathrm{F}} = 0.985\,(2\pi/a)$, which is larger than the distance $\overline{\Gamma L}$. In both cases the free-electron Fermi sphere extends beyond the first Brillouin zone. Thus, both the first and second Brillouin zones are only partially filled.

The Fermi surfaces obtained in the empty-lattice approximation for bcc and fcc structures are shown in Figs. 18.10 and 18.11. Note that the Fermi surface in the first Brillouin zone is not drawn around Γ but around H and

X since in this representation the states that are farthest from Γ (and, consequently, unoccupied) form connected regions around the edges joining points H and P (for bcc), and edges joining points W (for fcc). In the first band the so-called "monster" surrounds hole states, while the piece in the second Brillouin zone corresponds to electron states.

Owing to the periodic potential of the lattice, the spherical Fermi surface of free electrons becomes distorted. Nevertheless for small perturbations the splitting at the zone boundary is not too large, therefore the allowed energies of the bands overlap, even though identical energies are associated with different wave vectors in the two bands. This is illustrated in Fig. 19.6(*a*) for the calculated band structure of calcium, which crystallizes in an fcc lattice. The calculated Fermi surface is very similar to the one obtained from the Harrison construction.



(*a*) (*b*)

Fig. 19.6. Calculated band structure of (*a*) calcium and (*b*) magnesium

The situation is similar for elements that crystallize in hcp structure. As illustrated in Fig. 19.6(*b*), the bands that cross the Fermi energy overlap there, too. The Fermi surface is rather complicated; its empty-lattice approximation is shown in Fig. 19.7. The first Brillouin zone contains a "monster", while the states in the second Brillouin zone lead to a cigar- of lens-shaped Fermi surface.

19.3.3 Trivalent Metals

Of the trivalent metals in group 13 (IIIA) of the periodic table, the band structure calculated in the LCAO method for aluminum, which crystallizes in fcc structure, was illustrated in Fig. 17.3. Although several bands cross each other and get hybridized close to the Fermi surface, the shape of the Fermi

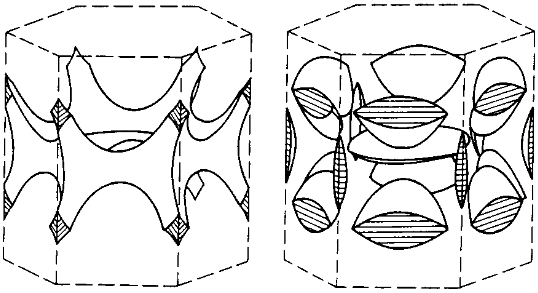


Fig. 19.7. Fermi surface of divalent metals in an empty hcp lattice [Reprinted with permission from W. A. Harrison, *Phys. Rev.* **118**, 1190 (1960). ©1960 by the American Physical Society]

surface is obtained fairly accurately in the nearly-free-electron approximation. As shown in Fig. 18.11, the first Brillouin zone is completely filled. In the second band, the occupied states are located close to the square and hexagonal faces of the truncated octahedron, leading to a hole-type Fermi surface. A "monster" appears in the third band, and tiny "pockets" in the fourth. The potential of the lattice modifies the Fermi surface in such a way that electrons are transferred from the fourth band to the second and third bands, and the monster in the third band is transformed into a set of rings, as illustrated in Fig. 19.8.

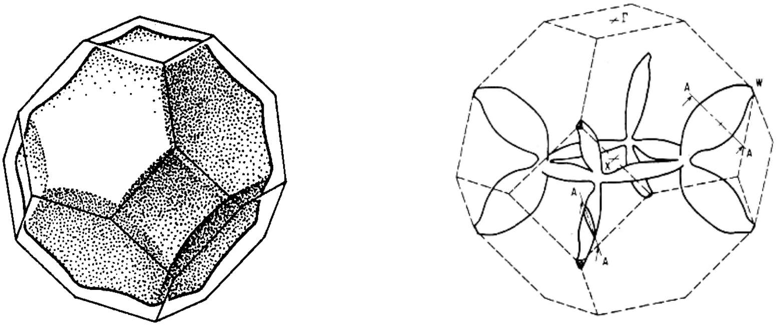
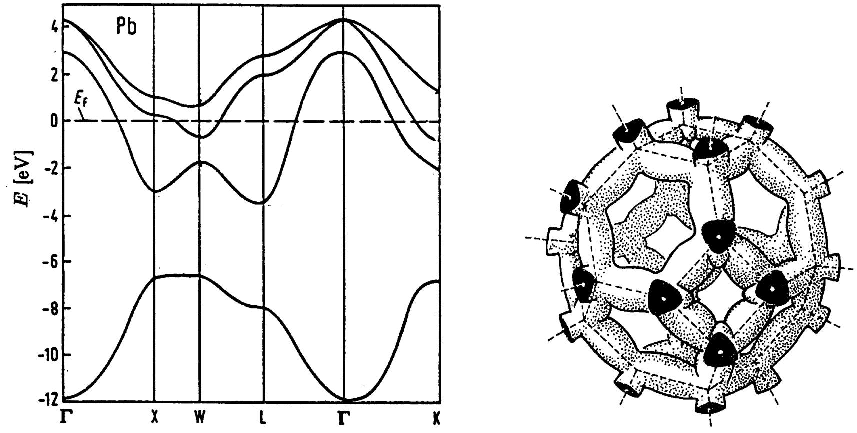


Fig. 19.8. Fermi surfaces in the second and third Brillouin zones for aluminum [Reprinted with permission from W. A. Harrison, *Phys. Rev.* **116**, 555 (1959). ©1959 by the American Physical Society]

19.3.4 Tetravalent Elements

The band structure of two prominent representatives of the carbon group [group 14 (IVA)], silicon and germanium, will be studied in detail in Chapter 20. We shall see that they are semiconductors: their Fermi energy lies in a gap. Below we shall briefly overview the band structure and Fermi surface of lead, which crystallizes in fcc structure and exhibits metallic properties (see Figure 19.9(*a*)).



(*a*) (*b*)

Fig. 19.9. (*a*): Calculated band structure for lead. (*b*): Fermi surface in the third Brillouin zone [Reprinted with permission from J. R. Anderson and A. V. Gold, *Phys. Rev.* **139**, A1459 (1965). ©1965 by the American Physical Society]

The metallic character is the consequence of the Fermi energy lying in two bands. The Fermi surface in Fig. 18.11 obtained for tetravalent metals in the empty-lattice approximation is composed of similar pieces as for trivalent metals, just the "monster" is somewhat bulkier, and the "pockets" are larger. This picture is in good agreement with the measured Fermi surface shown in Fig. 19.9(*b*).

19.3.5 Band Structure of Transition Metals and Rare-Earth Metals

The elements located between group 2 (IIA; alkaline-earth metals) and group 13 (IIIA; boron group) – that is, the elements of groups 3 through 12 – are called transition metals. In their atomic state a *d*-shell (the 3d-, 4d-, or 5d-shell) is partially filled – or even when it is completely filled, as in groups 11 and 12, these states do not lie deep. Therefore, in solids the bands formed by *d*-electrons are close to the Fermi energy, and the latter can cross these bands. Another consequence is that, apart from the elements of groups 11 and 12, the properties of transition metals are principally determined by the *d*-electrons, in contrast to the elements studied above. Using the nearly-free-electron approximation for describing these bands is totally unjustified, and their Fermi surfaces bear hardly any resemblance to those obtained with the Harrison construction. Figure 19.10 shows the Fermi surfaces for molybdenum and tungsten obtained in the Lomer model.[3]

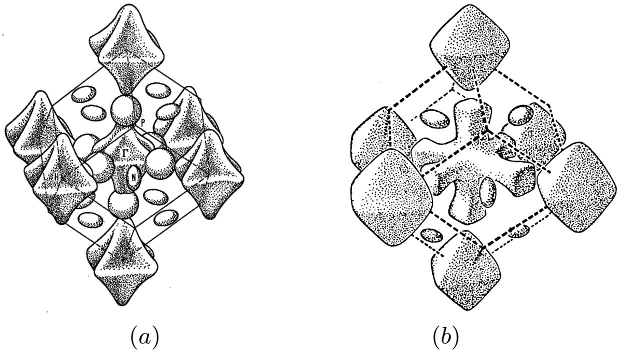


Fig. 19.10. Fermi surface proposed for (*a*) molybdenum and (*b*) tungsten

In transition metals the width of the *d*-band is usually smaller than the usual width of the conduction band in simple metals, and this narrow region must contain sufficiently many states to accommodate ten electrons. The density of states is therefore higher than usual, and this value appears in the low-temperature specific heat as well as the Pauli susceptibility.

A particular difficulty appears in the band-structure calculation of magnetically ordered transition metals. This manifests itself through the spin dependence of the energy eigenvalues: spin-up and spin-down electrons do not fill up the same energy levels, and the net magnetization is the result of the unequal number of occupied spin-up and spin-down states. To account for this, the band structures have to be calculated separately albeit self-consistently; this requires more complex methods than those discussed above. By way of example, the calculated band structure of ferromagnetic iron is shown in Fig. 19.11.

In rare-earth metals the 4f level is partially filled in the atomic configuration, therefore the 4f-band is expected to lie close to the Fermi energy, just like the *d*-band of transition metals. But this is not the case: electrons around the Fermi energy have hardly any 4f character: the latter are fairly well localized in a rather narrow band deep below the Fermi energy. This result cannot be consistently interpreted in the one-particle picture. The inapplicability of this picture to rare-earth metals is due to electron–electron correlations. The study of such correlations remains one of the most exciting subjects of solid-state physics to date. We shall examine certain aspects in Volume 3.

[3] W. M. Lomer, 1962.

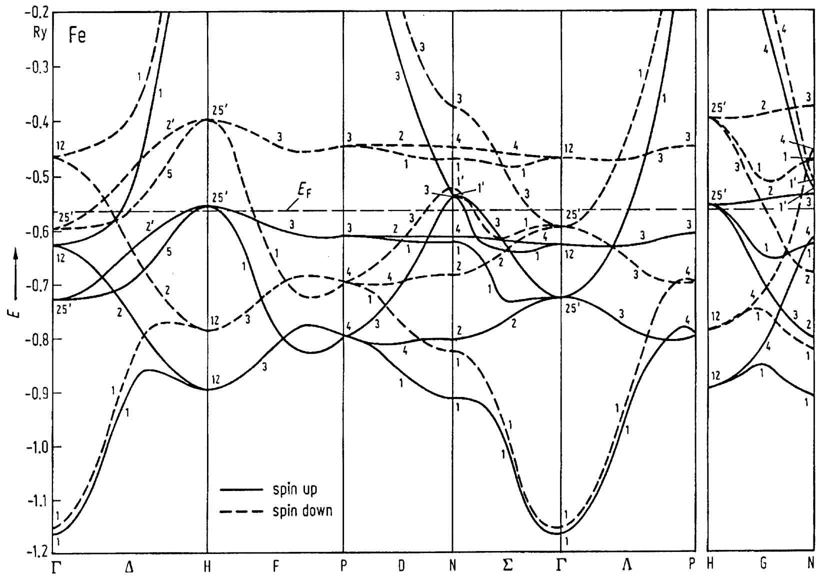


Fig. 19.11. The band structure of ferromagnetic iron calculated by the LCAO method [Reprinted with permission from J. Callaway and C. S. Wang, *Phys. Rev. B* **16**, 2095 (1977). ©1977 by the American Physical Society]

19.4 Experimental Study of the Band Structure

Diverse experimental methods are available for measuring the energy spectrum of Bloch electrons, their band structure, or the shape of the Fermi surface. Some of them allow the direct determination of the energy versus wave vector relation, at least over a part of the Brillouin zone. Other methods provide information about the topology and characteristic parameters of the Fermi surface. The latter are based in part on the property that the shape of the Fermi surface can be inferred from the motion of electrons in a magnetic field. We shall discuss these methods in some detail in later chapters. Below we shall give a brief account of some other techniques.

19.4.1 Positron Annihilation and Compton Scattering

When a beam of high-energy positrons is incident on a solid, the positrons lose most of their kinetic energy because of the interactions with the electrons in the solid. They are slowed down rather quickly (in about 10^{-12} seconds) to thermal energies (about 25 meV). Then they annihilate with the conduction electrons in about 10^{-10} seconds, emitting two γ photons. Assuming that the momentum of the positron is negligible compared to that of the conduction electron, the conservation of energy (applied to the relativistic expression of energy) and the conservation of momentum lead to the following equations:

$$\begin{aligned} m_\mathrm{e}c^2 + c\sqrt{(m_\mathrm{e}c)^2 + \boldsymbol{p}_\mathrm{e}^2} &= \hbar(\omega_1 + \omega_2)\,, \\ \boldsymbol{p}_\mathrm{e} &= \hbar(\boldsymbol{k}_1 + \boldsymbol{k}_2)\,. \end{aligned} \tag{19.4.1}$$

The energy of the photons cannot be less then the electron rest energy, about 0.5 MeV. Since the kinetic energy of the electron is only a few eV, it can be neglected in the energy balance. This leads to

$$k_1 \approx k_2 \approx \frac{m_\mathrm{e}c^2}{\hbar c} \tag{19.4.2}$$

for the photon wave number. The ratio of p_e and $\hbar k_i$ is also small, on the order of 10^{-3}. Consequently, the two annihilation photons are emitted in almost opposite directions, making an angle $\pi - \theta$, where θ is likewise small. This is shown in Fig. 19.12.

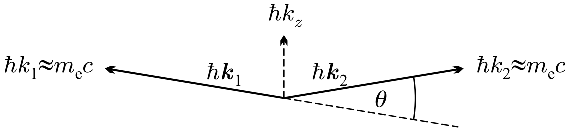

$\hbar k_z$
$\hbar k_1 \approx m_\mathrm{e} c$
$\hbar \boldsymbol{k}_1$
$\hbar \boldsymbol{k}_2$
$\hbar k_2 \approx m_\mathrm{e} c$
θ

Fig. 19.12. A pair of photons emitted in nearly opposite directions in a positron–electron annihilation process

Using (19.4.2) in the equation of momentum conservation, the component of the wave vector of the annihilated electron along the angular bisector of the directions of the two emitted photons is

$$k_z = 2\frac{m_\mathrm{e}c}{\hbar}\sin\left(\tfrac{1}{2}\theta\right) \approx \frac{m_\mathrm{e}c}{\hbar}\theta\,. \tag{19.4.3}$$

This component can thus be determined from the measured angle between two photons in coincidence. For a spherical Fermi surface the maximum value of k_z is k_F. This corresponds to a cut-off angle θ_max: the angle between the directions of the two emerging photons cannot be larger than this value. By measuring θ_max, the Fermi momentum can be determined.

In the usual measurement setup the two other components of the electron wave vector are not measured separately but are integrated out. Therefore below the critical angle the intensity of the emitted γ radiation shows a quadratic angular dependence. The angular correlation observed in experiments is different: anisotropic, as shown in Fig. 19.13. This provides information about the shape of the Fermi surface.

Similar results are obtained in the approach based on the Compton scattering[4] of X-rays by the solid. When hitting an electron at rest, the γ photon transfers a part of its energy to the electron, thus the scattered photon will be

[4] Arthur Holly Compton (1892–1962) was awarded the Nobel Prize in 1927 “for his discovery of the effect named after him”.

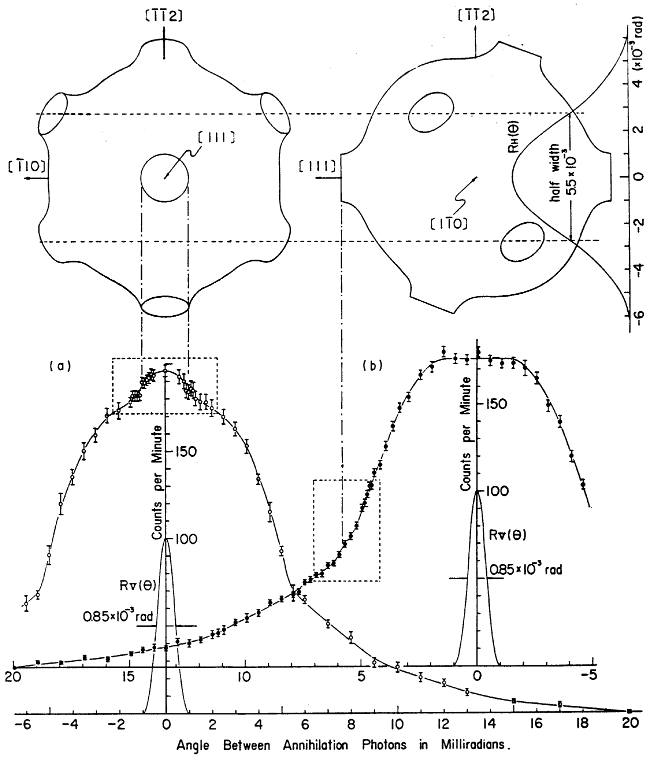


Fig. 19.13. Measurement results of angular correlation in electron–positron annihilation in a copper single crystal for two different orientations of the sample. A sketch of the Fermi surface is shown in the upper part [K. Fujiwara and O. Sueoka, *J. Phys. Soc. Japan* **21**, 1947 (1966)]

of lower energy, i.e., of larger wavelength.[5] From the conservation of energy and momentum, the change in the photon wavelength is

$$\Delta\lambda = \frac{h}{m_e c}(1-\cos\theta) = \frac{2h}{m_e c}\sin^2\left(\tfrac{1}{2}\theta\right), \tag{19.4.4}$$

where θ is the scattering angle (the angle between the directions of the incoming and outgoing photons), and $\lambda_C = h/m_e c$ is the Compton wavelength. If the incident beam is monochromatic, the photons scattered in a particular direction are found to be monoenergetic (i.e., of the same wavelength).

This is not observed in experiments performed on metals, since the conduction electrons, by which the photons are scattered, cannot be considered to be at rest. Owing to the motion of the electrons, an additional Doppler

[5] In reality, Compton scattering is a two-step process. The electron absorbs the incident photon, and then emits a photon of different energy – or the photon emission may just as well precede the absorption of the incident photon.

shift occurs. This causes a broadening of the lines of well-defined energies in a way that depends on the momentum distribution of the electrons. The shape of the Fermi surface can be inferred from the Compton line shape (profile), that is, the Doppler broadening of the Compton line.

19.4.2 Photoelectron Spectroscopy

In Compton scattering, only a small part of the energy of the incoming X-ray photon is transferred to an electron, the larger part is taken away by the emitted X-ray photon, and the wavelength of the latter is measured. However, the photon may transfer all its energy to the electron, and if the electron's energy exceeds the work function (photoemission threshold), which is usually on the order of a few eV, electron emission occurs. This is the photoelectric effect, which was first observed by H. Hertz in 1887. Its correct interpretation in terms of the quantized nature of light was given in 1905 by A. Einstein. By measuring the kinetic energy of the emitted electron one can infer the band structure. Figure 19.14 shows a simple band structure with the corresponding photoemission spectrum.

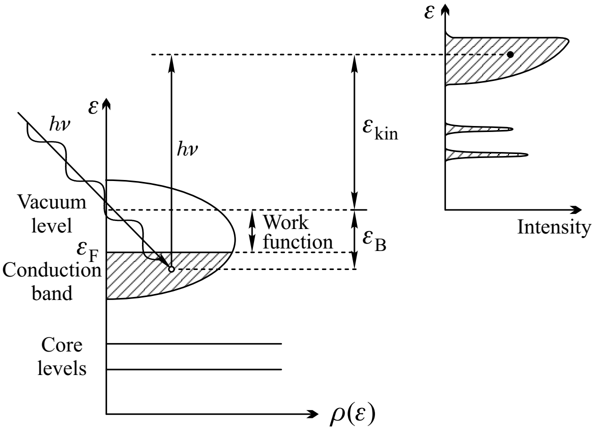


Fig. 19.14. Connection between the band structure of electrons (schematic representation on the left-hand side) and the photoelectron spectrum (right-hand side)

Due in great part to K. Siegbahn's work[6] photoelectron (or photoemission) spectroscopy (PES) became one of the most widely used methods for the experimental study of electron states in solids. The same method applied to ultraviolet radiation in the range 10–50 eV is called *ultraviolet photoelectron (photoemission) spectroscopy* (UPS). It is suited above all to studying the

[6] Kai Manne Börje Siegbahn (1918–2007) was awarded the Nobel Prize in 1981 "for his contribution to the development of high-resolution electron spectroscopy".

states of the valence band. When soft X-rays (of energies in the keV range) are used, the method is called *X-ray photoelectron (photoemission) spectroscopy* (XPS). Because of the high energies involved, this is particularly adapted to the determination of the energies of deep core states,[7] nevertheless valence-band states can also be studied using high-resolution devices.

Identical processes occur in both cases. When an electron of energy ε_{i} occupying a level below the Fermi energy absorbs a photon of energy $\hbar\omega$, it moves up to the excited state of energy

$$\varepsilon_{\mathrm{f}} = \varepsilon_{\mathrm{i}} + \hbar\omega \,, \tag{19.4.5}$$

which is so much above the vacuum energy level that the electron can leave the solid. Referring the energies to the Fermi energy, the kinetic energy of the emitted electron is

$$\varepsilon_{\mathrm{kin}} = \varepsilon_{\mathrm{f}} - \Phi \,, \tag{19.4.6}$$

where Φ is the work function, i.e., the distance of the vacuum energy level from the Fermi energy. From the conservation of energy,

$$\varepsilon_{\mathrm{kin}} = \hbar\omega - \Phi - |\varepsilon_{\mathrm{i}}| \,. \tag{19.4.7}$$

Since the number of electrons emitted at a given kinetic energy is proportional to the density of states at the corresponding initial state, this method provides experimental information about the electronic density of states. Figure 19.15 shows the density of states for copper, the XPS spectrum derived from it, and the measured spectrum.

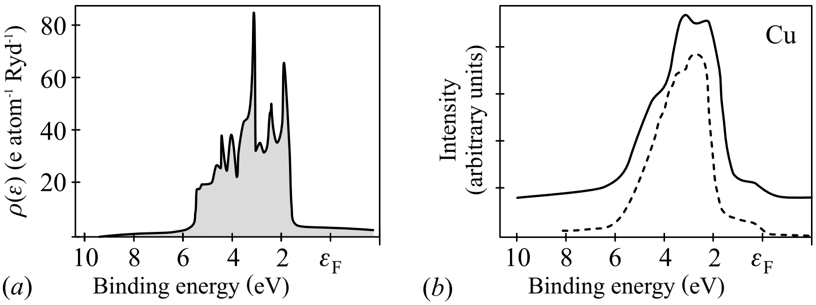


Fig. 19.15. (*a*) The density of occupied states for copper. (*b*) The XPS spectrum derived from this (solid line) and the measured spectrum (dashed line) [M. Lähdeniemi et al., *J. Phys. F: Metal Phys.* **11**, 1531 (1981)]

[7] Besides being specific to the element, the energies of these levels also depend on the chemical environment, i.e., the type of bonding. Therefore XPS is also suitable for the chemical analysis of surfaces, which is why it is often referred to as ESCA, electron spectroscopy for chemical analysis.

The reverse process is used in *inverse photoemission spectroscopy* (IPES). When a high-energy electron beam hits the sample, an electron can be trapped in an unoccupied state above the Fermi energy, while the energy difference is taken away by an emitted photon. This phenomenon is called *bremsstrahlung* (braking radiation). For this reason, the method is also called *bremsstrahlung isochromat spectroscopy* (BIS). It provides direct information about the electronic density of states above the Fermi energy.

The techniques presented above measure the spectrum against energy: the distribution is integrated over the angular variables. The advent of high-intensity synchrotron radiation sources opened the way to sufficiently accurate measurements of the angular distribution of electrons emitted at specified energies. This method, called *angle-resolved photoemission spectroscopy* (ARPES), gives more details of the electronic band structure. Figure 19.16 shows the photoemission spectrum of GaAs for various values of the azimuthal angle and two different polar angles.

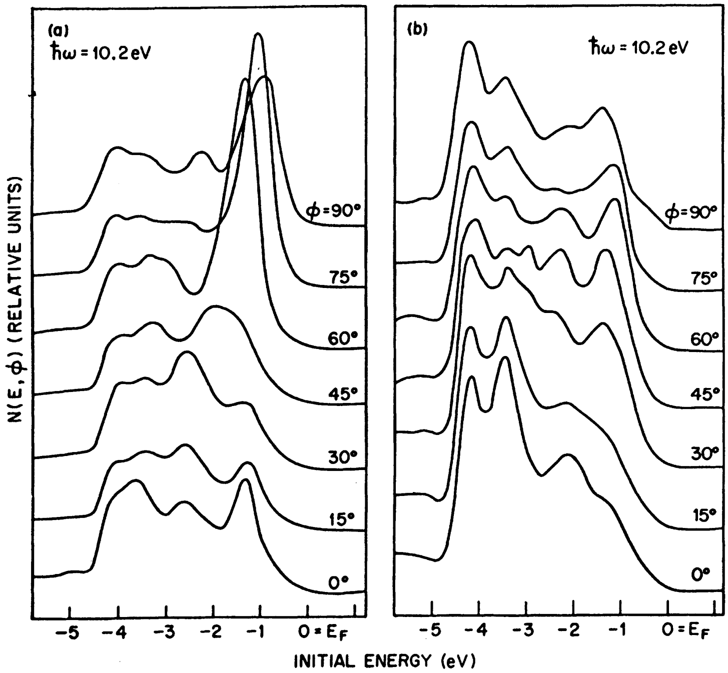


Fig. 19.16. Variations in the photoemission spectrum of GaAs with the azimuthal angle, for two different polar angles [Reprinted with permission from N. V. Smith and M. M. Traum, *Phys. Rev. Lett.* **31**, 1247 (1973). ©1973 by the American Physical Society]

A complete theoretical description of ARPES can be given by using the Green function technique of the many-body problem.[8] To obtain a qualitative picture, we shall denote the initial energy inside the solid by ε_{i}, the energy of the emitted electron by ε_{f}, and the corresponding wave vectors by $\boldsymbol{k}_{\mathrm{i}}$ and $\boldsymbol{k}_{\mathrm{f}}$. The conservation of energy and momentum[9] in the photoelectric process implies

$$\varepsilon_{\mathrm{i}} + \hbar\omega = \varepsilon_{\mathrm{f}} \,, \qquad \boldsymbol{k}_{\mathrm{i}}^{\|} = \boldsymbol{k}_{\mathrm{f}}^{\|} + \boldsymbol{G}^{\|} \,. \tag{19.4.8}$$

In the second equation we exploited the property that below 100 eV, a typical energy used in ARPES measurements, the wave number of the photon is much smaller than that of the electron, and momentum conservation (which is valid up to a reciprocal-lattice vector) holds only for the component parallel to the surface: as translational symmetry is broken, the conservation of momentum does not apply to the perpendicular component. By measuring the kinetic energy of the emitted electron, the energy of its initial state inside the solid can be reconstructed by using the relation

$$\varepsilon_{\mathrm{kin}} = \varepsilon_{\mathrm{i}} + \hbar\omega - \varepsilon_{\mathrm{vac}} \,. \tag{19.4.9}$$

On the other hand, for electrons emitted in a given direction characterized by the polar angle θ measured from the surface normal and the azimuthal angle φ with respect to a preferred crystallographic direction of the crystal, the relations

$$k_{\mathrm{f}x} = \frac{(2m_{\mathrm{e}}\varepsilon_{\mathrm{kin}})^{1/2}}{\hbar} \sin\theta \cos\varphi \,, \qquad k_{\mathrm{f}y} = \frac{(2m_{\mathrm{e}}\varepsilon_{\mathrm{kin}})^{1/2}}{\hbar} \sin\theta \sin\varphi \tag{19.4.10}$$

can be used to calculate the components of $\boldsymbol{k}$ that lie in the surface. It is more complicated to determine the component perpendicular to the surface: it requires comparison with other measurements or assuming a nearly-free-electron-like behavior. The dispersion relation of electrons can be determined from spectra measured in different directions. The band structure obtained in this way for copper is shown in Fig. 19.17 for two directions of the Brillouin zone.

The problem related to the perpendicular component of the wave vector $\boldsymbol{k}$ does not arise in quasi-two-dimensional layered systems. High-T_{c} superconductors are typical examples, therefore ARPES is ideally suited to the study of their electronic structure.

Electron states in solids can also be investigated using the reflection or absorption of photons in the optical region. We shall examine this possibility in Chapter 25.

[8] The intensity of photoemission is proportional to the spectral function – that is, the imaginary part of the single-particle Green function. Thus ARPES provides direct insight into the electronic structure of the many-body system.

[9] More precisely, the conservation of crystal momentum is required, otherwise the photon could not be absorbed by the crystal.

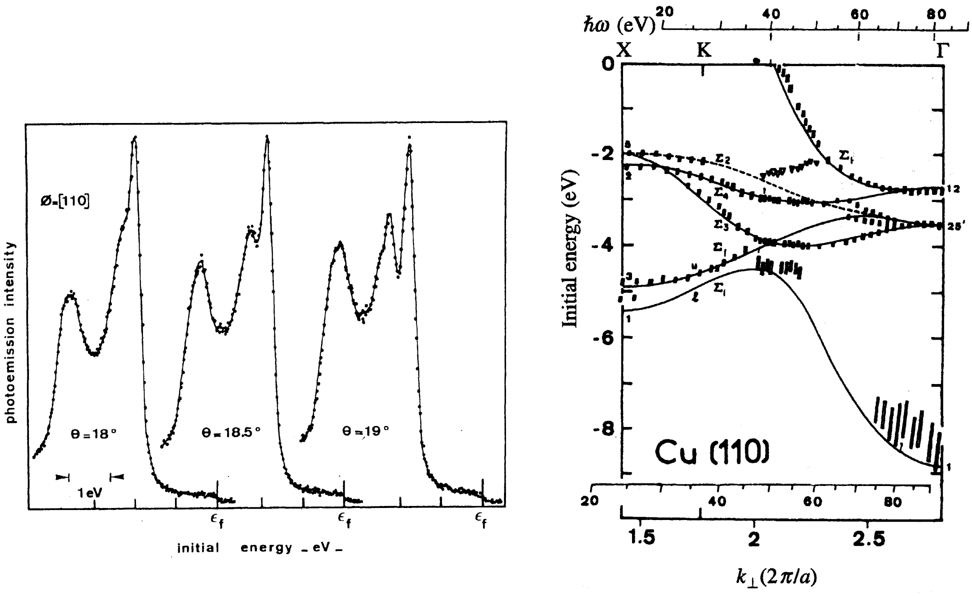


Fig. 19.17. The energy distribution of photoelectrons emitted by the (100) surface of copper for three polar angles, and the band structure of copper along the line $\Gamma K X$ of the Brillouin zone determined from ARPES data [Reprinted with permission from P. Thiry et al., *Phys. Rev. Lett.* **43**, 82 (1979). ©1979 by the American Physical Society]

Further Reading

1. J. Callaway, *Energy Band Theory*, Academic Press, New York (1964).
2. G. C. Fletcher, *The Electron Band Theory of Solids*, North-Holland Publishing Company, Amsterdam (1971).
3. W. A. Harrison, *Pseudopotentials in the Theory of Metals*, Benjamin, Reading, Mass. (1966).
4. R. M. Martin, *Electronic Structure, Basic Theory and Practical Methods*, Cambridge University Press, Cambridge (2004).
5. D. A. Papaconstantopoulos, *Handbook of the Band Structure of Elemental Solids*, Plenum Press, New York (1986).
6. J. M. Ziman, *The Calculation of Bloch Functions*, in Solid State Physics, Vol. 26. p. 1., Academic Press, New York (1971).

20

Electronic Structure of Semiconductors

Since the invention of the transistor in 1947–1948, and especially the start of the mass production of integrated circuits and microprocessors, semiconductor devices have played an ever increasing role in modern information technology as well as many other applied fields. The optical properties of semiconductors, which set them distinctly apart of metals, are also exploited in a wide range of applications. In addition to being very important for materials science, the study of semiconductors is of great interest for fundamental research as well, as a great number of new phenomena can be observed in them. For example, the discovery of the quantum Hall effect arose from the possibility of creating semiconductor heterojunctions in which the electron gas is practically confined to a two-dimensional region next to the interface. This opened the way to the study of the properties rooted in the two-dimensional character of the system. Besides, very high purity materials can be fabricated from semiconductors, which is a prerequisite to studying certain physical phenomena.

To derive the physical properties of semiconductors from first principles, we have to familiarize ourselves with the characteristic properties of their electronic structure. The methods of band-structure calculation presented in the previous chapter can be equally applied to metals and insulators, so they can just as well be used for the description of the electronic structure of semiconductors. However, the devices designed to exploit the physical properties of semiconductors practically never use pure crystals: the number of charge carriers is controlled by doping the semiconductor components with impurities. Therefore we shall also study the states formed around impurities, and how the energy spectrum is modified by them. The phenomena required to understand the operation of semiconductor devices, as well as the particular conditions arising close to the interfaces and in inhomogeneous semiconductor structures will be presented in Chapter 27, after the discussion of transport phenomena.

20.1 Semiconductor Materials

As mentioned earlier, a characteristic property of semiconductors is that their resistivity falls between that of metals and insulators. An even more interesting feature is the temperature dependence of resistivity. In contrast to metals and semimetals, it increases exponentially with decreasing temperature in pure semiconductors: $\varrho \propto \exp(\varepsilon_0/k_\mathrm{B}T)$. The Hall coefficient R_H is positive in several cases, which can be interpreted by assuming that the principal charge carriers in these materials are not electrons but holes. It is also known from the behavior of the Hall coefficient that, in contrast to metals, the number of carriers depends strongly on temperature.

This is in perfect agreement with a band structure in which the bands that are filled completely in the ground state are separated from the completely empty bands by a forbidden region, a gap, whose width ε_g is finite but not much larger than the thermal energy $k_\mathrm{B}T$. In semiconductors the highest band that is completely filled in the ground state is called the *valence band*, while the lowest completely empty band is called the *conduction band*. Current can flow only when charge carriers – electrons in the conduction band and holes in the valence band – are generated by thermal excitation.

As we shall see, the probability of generating carriers by thermal excitation is proportional to $\exp(-\varepsilon_\mathrm{g}/2k_\mathrm{B}T)$ because of the finite gap. If the band gap is around $\varepsilon_\mathrm{g} \sim 5\,\mathrm{eV}$, this probability is about e^{-100} or 10^{-45} at room temperature ($k_\mathrm{B}T \sim 0.025\,\mathrm{eV}$). Since the total electron density is on the order of 10^{23} per cm^3, the conduction band contains practically no electrons. However, when the gap is only $1\,\mathrm{eV}$, the excitation probability is 10^{-9} because of the exponential dependence, and so the density of thermally excited electrons is $10^{14}/\mathrm{cm}^3$. Such a number of mobile charges gives rise to observable phenomena.

The resistivity of semiconductors is very sensitive to the presence of impurities. For this reason, pure stoichiometric semiconductors are of much smaller practical importance than doped ones. Nonetheless we shall start our discussion with pure materials.

20.1.1 Elemental Semiconductors

As mentioned in Chapter 17, elemental semiconductors are located in the even-numbered groups of the periodic table, to the right of transition metals. Certain elements of the group IIIA (boron group)[1] and the group VA (nitrogen group) also occur in compound semiconductors. Table 20.1 shows the electronic configuration of the outermost shell in the atomic state of the elements of interest.

Elemental semiconductors are found in group IVA (carbon group), and group VIA (oxygen group, chalcogens). The first element of the carbon group,

[1] Following the conventions of semiconductor physics, we shall often use only the traditional designation for the groups of the periodic table in this chapter.

Table 20.1. Elements occurring in semiconductor materials, and the configuration of their outermost shell in the atomic state

Group IIB	Group IIIA	Group IVA	Group VA	Group VIA
	B $2s^2\,2p^1$	C $2s^2\,2p^2$	N $2s^2\,2p^3$	O $2s^2\,2p^4$
	Al $3s^2\,3p^1$	Si $3s^2\,3p^2$	P $3s^2\,3p^3$	S $3s^2\,3p^4$
Zn $3d^{10}\,4s^2$	Ga $4s^2\,4p^1$	Ge $4s^2\,4p^2$	As $4s^2\,4p^3$	Se $4s^2\,4p^4$
Cd $4d^{10}\,5s^2$	In $5s^2\,5p^1$	Sn $5s^2\,5p^2$	Sb $5s^2\,5p^3$	Te $5s^2\,5p^4$
Hg $5d^{10}\,6s^2$	Tl $6s^2\,6p^1$	Pb $6s^2\,6p^2$	Bi $6s^2\,6p^3$	Po $6s^2\,6p^4$

carbon, has several allotropes. Even though not a semiconductor, diamond is of great interest here as it features the same type of bonding and structure as the other, semiconducting, elements of the group. The sp^3 hybrid states given in (4.4.52) of the outermost *s*- and *p*-electrons form four covalent bonds in the directions of the vertices of a regular tetrahedron. This leads to the diamond lattice shown in Fig. 7.16. As discussed in Chapter 4, in covalently bonded materials the density of electrons is highest in the region between the two atoms. This was illustrated for germanium in Fig. 4.5, where a section of the spatial distribution of the valence electrons was shown in the vicinity of the line joining neighboring atoms.

Of the elements of group IVA, the band structure of diamond and gray tin (α-Sn) – both determined by the LCAO method – were shown in Fig. 17.10. The band structure of silicon and germanium will be discussed in detail and illustrated later (Figs. 20.2 and 20.5). In each case, there are further narrow and completely filled bands below the shown bands. The lowest four of the shown bands – which are partially degenerate along certain high-symmetry directions – are formed by the *s*- and *p*-electrons that participate in covalent bonding. In the ground state these bands are completely filled, since the primitive cell contains two electrons with four valence electrons each. Except for α-Sn, these four bands are separated by a finite gap from the higher-lying ones (that are completely empty in the ground state).

Pure diamond is an insulator as its energy gap is 5.48 eV. However, when doped, it exhibits the characteristic properties of semiconductors. The two next elements of the carbon group, silicon (Si) and germanium (Ge) are good semiconductors even in their pure form; their energy gap is close to 1 eV. The measured data show a slight nonetheless clear temperature dependence. This is due to the variations of the lattice constant, which modify the overlap between the electron clouds of neighboring atoms, leading to an inevitable shift of the band energies. The energy gap measured at room temperature and the value extrapolated to $T = 0$ from low-temperature measurements are listed in Table 20.2.

The fourth element of this group is tin. Among its several allotropes gray tin is also a semiconductor, but it has hardly any practical importance. The

Table 20.2. Energy gap at room temperature and at low temperatures for group IVA elements

Element	$\varepsilon_{\mathrm{g}}(300\,\mathrm{K})$ (eV)	$\varepsilon_{\mathrm{g}}(T=0)$ (eV)
C	5.48	5.4
Si	1.110	1.170
Ge	0.664	0.744

energy gap is not listed in the table because there is no real gap in the band structure. Nonetheless, as mentioned in Chapter 17, it behaves practically as a semiconductor, since the density of states is very low at the bottom of the conduction band, therefore electrons excited thermally at room temperature do not occupy these levels but a local minimum of the conduction band. Although this minimum is located 0.1 eV higher, it has a larger density of states.

It is readily seen from the band structure and the values given in the table that the gap decreases toward the high end of the column. Tin is followed by lead, a metal whose band structure was shown in Fig. 19.9.

Among the elements of the oxygen group, selenium (Se) and tellurium (Te) are the only semiconductors. The gap is 1.8 eV for Se and 0.33 eV for Te. In these materials the outermost *s*- and *p*-electrons can form two covalent bonds located along a helix, as shown in Fig. 7.24(*a*). The relatively weak van der Waals forces between the chains are strong enough to ensure that selenium and tellurium behave as three-dimensional materials.

20.1.2 Compound Semiconductors

Among covalently bonded compounds quite a few are semiconductors. For that each *s*- and *p*-electron form saturated covalent bonds, the compound has to be built up of cations and anions that give, on the average, four electrons to the tetrahedrally coordinated bonds. This can be ensured in several different ways. The simplest possibility is to build a compound of two elements of the carbon group (group IVA) – or else an element of the boron group (group IIIA) can be combined with one of the nitrogen group (group VA), or an element of the zinc group (group IIB) with one of the oxygen group (group VIA).

Among the compounds of the elements of the carbon group, silicon carbide (SiC, also called carborundum) is particularly noteworthy. Its energy gap is 2.42 eV. In stoichiometric composition it is a good insulator, but a small excess of carbon or other dopants turn it into a good semiconductor. For applications, III–V ($\mathrm{A^{III}B^{V}}$) and II–VI ($\mathrm{A^{II}B^{VI}}$) semiconductors are more important. As their names show, these are compounds of elements in groups IIIA and VA, and IIB and VIA. Table 20.3 shows the room-temperature energy gap for a small selection of them.

Table 20.3. Energy gap for III–V, II–VI, and I–VII semiconductors at room temperature

III–V compound	ε_{g} (eV)	II–VI compound	ε_{g} (eV)	I–VII compound	ε_{g} (eV)
AlSb	1.63	ZnO	3.20	AgF	2.8
GaP	2.27	ZnS	3.56	AgCl	3.25
GaAs	1.43	ZnSe	2.67	AgBr	2.68
GaSb	0.71	CdS	2.50	AgI	3.02
InP	1.26	CdSe	1.75	CuCl	3.39
InAs	0.36	CdTe	1.43	CuBr	3.07
InSb	0.18	HgS	2.27	CuI	3.11

A large number of III–V and II–VI semiconductors crystallize in the sphalerite structure. The prototype of this structure, sphalerite (zinc sulfide) is a semiconductor itself. As shown in Fig. 7.16, this structure can be derived from the diamond structure by placing one kind of atom at the vertices and face centers of a fcc lattice, and the other kind of atom at the center of the four small cubes. The structure can also be considered to be made up of two interpenetrating fcc sublattices displaced by a quarter of the space diagonal. Thus each atom is surrounded tetrahedrally by four of the other kind. Here, too, *s*- and *p*-electrons participate in bonding, however the bond is not purely covalent on account of the different electronegativities of the two atoms.

In $\mathrm{A}^{\mathrm{III}}\mathrm{B}^{\mathrm{V}}$ semiconductors the ion cores A^{3+} and B^{5+} stripped of their *s*- and *p*-electrons do not attract electrons in the same way. Unlike in germanium (Fig. 4.5), the wavefunction of the electrons forming the covalent bond – and hence the density of the electrons – is not symmetric with respect to the position of the two ions, as shown in Fig. 4.7 for GaP: it is biassed toward the B^{5+} ions. This gives a slight ionic character to the bond.

The asymmetry is even more pronounced for II–VI semiconductors, and the density maximum of the binding electrons is now even closer to the element of group VIA, as illustrated in Fig. 4.7 for ZnSe. The bond is thus more strongly ionic in character. The elements of the carbon group, for which no such asymmetry occurs, are called nonpolar semiconductors, while III–V and II–VI compounds are polar semiconductors.

Even more strongly polar are the compounds of the elements of groups IB and VIIA. Some halides of noble metals behave as semiconductors, even though their gap is rather large. Copper compounds feature tetrahedrally coordinated bonds, but silver compounds do not: they crystallize in the sodium chloride structure.

It is worth comparing the properties of these compounds with those of alkali halides (formed by elements of groups IA and VIIA). In the latter the difference between the electronegativities of the two constituents is so large that purely ionic bonds are formed instead of polarized covalent bonds. Consequently the structure is also different, NaCl- or CsCl-type, and their gap (listed in Table 20.4 for a few alkali halides) is much larger than in any previously mentioned case.

Table 20.4. Energy gap in some alkali halides

Compound	ε_g (eV)	Compound	ε_g (eV)	Compound	ε_g (eV)
LiF	13.7	LiCl	9.4	LiBr	7.6
NaF	11.5	NaCl	8.7	NaBr	7.5
KF	10.8	KCl	8.4	KBr	7.4
RbF	10.3	RbCl	8.2	RbBr	7.4
CsF	9.9	CsCl	8.3	CsBr	7.3

In addition to those listed above, there exist further compound semiconductors with non-tetrahedrally coordinated covalent bonds. Tin and lead (group IVA) form semiconducting materials with elements of the oxygen group (group VIA) that crystallize in the NaCl structure. Elements of group IVA may also form compound semiconductors with alkaline-earth metals (group IIA) in the composition II_2–IV ($\mathrm{A}_2^{\mathrm{II}}\mathrm{B}^{\mathrm{IV}}$). Semiconductor compounds with a non-1:1 composition can also be formed by IB and VIA, or IIB and VA elements. The best known examples are Cu_2O ($\varepsilon_g = 2.17\,\mathrm{eV}$) and TiO_2 ($\varepsilon_g = 3.03\,\mathrm{eV}$). The gap is much narrower in some other compounds, as listed in Table 20.5.

Table 20.5. Energy gap in some compound semiconductors at low temperatures

Compound	ε_g (eV)	Compound	ε_g (eV)
SnS	1.09	Mg_2Si	0.77
SnSe	0.95	Mg_2Ge	0.74
SnTe	0.36	Mg_2Sn	0.36
PbO	2.07	Cu_2O	2.17
PbS	0.29	Ag_2S	0.85
PbSe	0.14	Zn_3As_2	0.86
PbTe	0.19	Cd_3P_2	0.50

Semiconducting properties are also observed in various nonstoichiometric compounds and even amorphous materials.

20.2 Band Structure of Pure Semiconductors

According to the foregoing, in the ground state of semiconductors the completely filled valence band is separated from the completely empty conduction band by a narrow gap. We shall first examine the band structure of the two best known semiconductors, silicon and germanium, focusing on how the energy gap is formed and how the bands closest to the chemical potential can be characterized. To understand the band structure, we shall follow the steps outlined in Chapter 18: start with the empty-lattice approximation, and determine how degeneracy is lifted and how gaps appear in the nearly-free-electron approximation. We shall then present the theoretical results based on more accurate calculations, and the experimental results.

20.2.1 Electronic Structure in the Diamond Lattice

We have seen that the elemental semiconductors of the carbon group crystallize in a diamond structure. Since the Bravais lattice is face-centered cubic, the Brillouin zone is the truncated octahedron depicted in Fig. 7.11. Below we shall repeatedly make reference to the special points and lines of the Brillouin zone, therefore we shall recall them in Fig. 20.1(a).

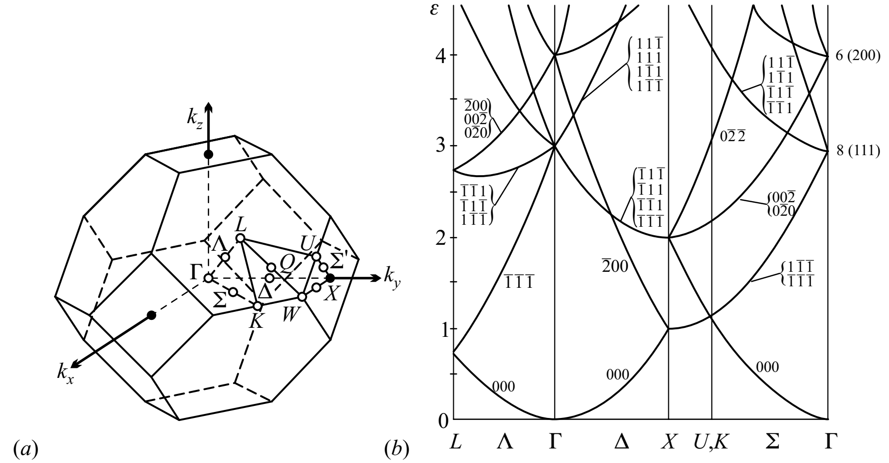


Fig. 20.1. (a) Brillouin zone of the face-centered cubic lattice of the diamond structure with the special points and lines. (b) Band structure in the empty-lattice approximation, with the energy given in units of $(\hbar^2/2m_\mathrm{e})(2\pi/a)^2$. The triplets hkl next to each branch refer to the corresponding reciprocal-lattice vector $(2\pi/a)(h,k,l)$

The figure also shows the band structure calculated in the empty-lattice approximation, along the four high-symmetry directions of the Brillouin zone, as usual. The lines Δ, Λ, and Σ join the center $\Gamma = (0,0,0)$ with the center $X = (2\pi/a)(0,0,1)$ of a square face, the center $L = (2\pi/a)(\frac{1}{2},\frac{1}{2},\frac{1}{2})$ of a hexagonal face, and an edge center $K = (2\pi/a)(\frac{3}{4},\frac{3}{4},0)$ of a hexagonal face, respectively, while another line joins point $X' = (2\pi/a)(0,1,0)$ (which is equivalent to X) and point $U = (2\pi/a)(\frac{1}{4},1,\frac{1}{4})$ (which is equivalent to K).

The periodic potential modifies this band structure. As far as the behavior of semiconductors is concerned, the most important is to understand what happens at and close to the zone center Γ where an eightfold degenerate state is found above the lowest nondegenerate level. According to the discussion in Chapter 18, the wavefunctions of these eight states can be written as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G}_i)\cdot\boldsymbol{r}} \tag{20.2.1}$$

in the empty-lattice approximation, where $\boldsymbol{G}_i = (2\pi/a)\,(\pm1,\pm1,\pm1)$. To determine the extent to which this eightfold degeneracy is removed in $\boldsymbol{k} = 0$, the method described in Chapter 18 for the lifting of accidental degeneracies is applied to the diamond lattice.

The little group of point Γ – i.e., the group of those symmetry operations that take Γ into itself or an equivalent point – is the 48-element group O_h. Using the character table of irreducible representations given in Appendix D of Volume 1, the eight-dimensional representation over the eight functions above can be reduced to two one-dimensional (Γ_1 and Γ_2') and two three-dimensional irreducible representations (Γ_{15} and Γ_{25}'):

$$\Gamma = \Gamma_1 + \Gamma_2' + \Gamma_{15} + \Gamma_{25}' \,. \tag{20.2.2}$$

It is also straightforward to find the wavefunctions that transform according to these irreducible representations as linear combinations of the functions $\mathrm{e}^{\mathrm{i}\boldsymbol{G}_i\cdot\boldsymbol{r}}$. The representation Γ_1 is associated with the symmetric combination

$$\begin{aligned}\psi_{\Gamma_1}(\boldsymbol{r}) = \frac{1}{8}\Big[&\mathrm{e}^{2\pi\mathrm{i}(x+y+z)/a} + \mathrm{e}^{2\pi\mathrm{i}(x+y-z)/a} + \mathrm{e}^{2\pi\mathrm{i}(x-y+z)/a} + \mathrm{e}^{2\pi\mathrm{i}(-x+y+z)/a} + \\ &+ \mathrm{e}^{-2\pi\mathrm{i}(x+y+z)/a} + \mathrm{e}^{-2\pi\mathrm{i}(x+y-z)/a} + \mathrm{e}^{-2\pi\mathrm{i}(x-y+z)/a} + \mathrm{e}^{-2\pi\mathrm{i}(-x+y+z)/a}\Big] \\ &= \cos(2\pi x/a)\cos(2\pi y/a)\cos(2\pi z/a)\,. \end{aligned} \tag{20.2.3}$$

The combination associated with Γ_2' is

$$\psi_{\Gamma_2'}(\boldsymbol{r}) = \sin(2\pi x/a)\sin(2\pi y/a)\sin(2\pi z/a)\,, \tag{20.2.4}$$

while the combinations associated with the three-dimensional representations are

$$\begin{aligned}\psi^{(1)}_{\Gamma_{15}}(\boldsymbol{r}) &= \sin(2\pi x/a)\cos(2\pi y/a)\cos(2\pi z/a)\,, \\ \psi^{(2)}_{\Gamma_{15}}(\boldsymbol{r}) &= \cos(2\pi x/a)\sin(2\pi y/a)\cos(2\pi z/a)\,, \\ \psi^{(3)}_{\Gamma_{15}}(\boldsymbol{r}) &= \cos(2\pi x/a)\cos(2\pi y/a)\sin(2\pi z/a)\,, \end{aligned} \tag{20.2.5}$$

and

$$\begin{aligned}\psi^{(1)}_{\Gamma'_{25}}(\boldsymbol{r}) &= \cos(2\pi x/a)\sin(2\pi y/a)\sin(2\pi z/a)\,,\\ \psi^{(2)}_{\Gamma'_{25}}(\boldsymbol{r}) &= \sin(2\pi x/a)\cos(2\pi y/a)\sin(2\pi z/a)\,,\\ \psi^{(3)}_{\Gamma'_{25}}(\boldsymbol{r}) &= \sin(2\pi x/a)\sin(2\pi y/a)\cos(2\pi z/a)\,.\end{aligned} \tag{20.2.6}$$

Hence the eightfold degeneracy is lifted in such a way that two triply degenerate states, of symmetry Γ_{15} and Γ'_{25}, and two nondegenerate states, of symmetry Γ_1 and Γ'_2, arise. Apart from exceptional cases, their energies are different.

We can also examine what happens to the electron states along the lines Δ and Λ close to Γ. To this end we have to make use of the compatibility relations between the irreducible representations that belong to point Γ and lines Δ and Λ, which can be directly established from the character tables. Table 20.6 contains these relations for the relevant representations.

Table 20.6. Compatibility relations between irreducible representations for point Γ and points Δ and Λ

Γ_1	Γ'_2	Γ_{15}	Γ'_{25}
Δ_1	Δ'_2	$\Delta_1\,\Delta_5$	$\Delta'_2\,\Delta_5$
Λ_1	Λ_1	$\Lambda_1\,\Lambda_3$	$\Lambda_1\,\Lambda_3$

It is immediately seen from these relations that the energies of the triply degenerate states associated with the representations Γ_{15} and Γ'_{25} are split further along these lines, to a nondegenerate and a doubly degenerate level. Nothing more specific can be said using symmetry considerations alone – except for one thing. The nondegenerate state starting from Γ'_{25} necessarily becomes degenerate with the corresponding state of the lowest band in point X – that is, the double degeneracy obtained in the free-electron model is not lifted there by the periodic potential, while it is lifted in point L.

The band structure obtained in this way is in good agreement with the band structure of diamond and gray tin (Fig. 17.10), as well as silicon and germanium (to be presented later). Once accidental degeneracies are lifted, the state of symmetry Γ'_{25} is found to be the lowest in Γ. Since the basis contains two atoms, eight electrons have to be accommodated in the bands, therefore, taking spin degeneracy into account, the bands containing the triply degenerate state of symmetry Γ'_{25} (as well as the band containing the states associated with the vector $\boldsymbol{G} = 0$) are completely filled in the ground state, whereas higher-lying bands (of symmetry Γ_{15}, Γ'_2, and Γ_1 in point Γ) are left completely empty. The gap between occupied and unoccupied bands is so large in diamond that it behaves as an insulator. This insulator character is not changed even when one takes into account further splitting due to spin–orbit coupling (which is weak because of the low atomic number). The role of spin–orbit coupling will be discussed later for silicon and germanium.

20.2.2 Band Structure of Silicon

The band structure of silicon calculated without taking spin–orbit coupling into account is shown in Fig. 20.2. The pattern is very similar to what can be expected from the previous general considerations and the example of diamond. Here, too, the bands containing the energy level of symmetry Γ'_{25} are the valence bands, and they are completely filled. The bands containing the state of symmetry Γ_{15} are the lowest-lying empty bands. They are the conduction bands. The gap between the state of symmetry Γ'_{25} at the top of the valence band and the $\boldsymbol{k} = 0$ state of symmetry Γ_{15} in the conduction band is 2.5 eV.

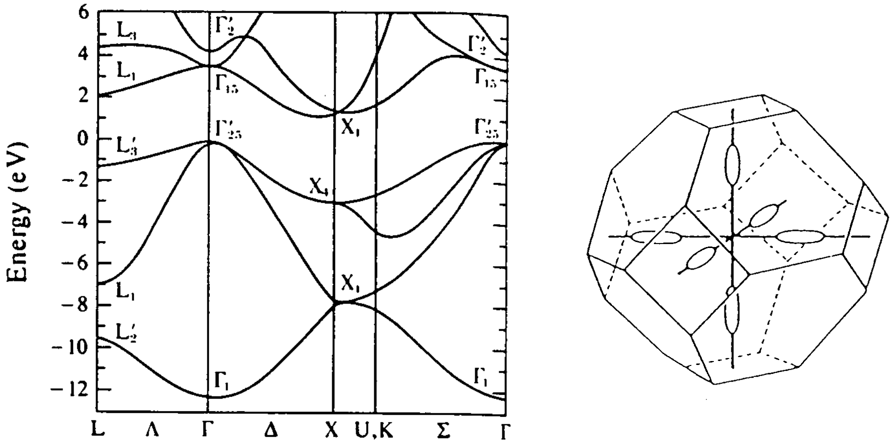


Fig. 20.2. Energy bands of silicon and low-energy constant-energy surfaces of the conduction band [Reprinted with permission from J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **10**, 5095 (1974). ©1974 by the American Physical Society]

However, the experimental value for the gap, measured by means of the thermal excitation of electrons, is much lower: 1.11 eV at room temperature. The reason for this discrepancy is that the minimum of the conduction band is not at $\boldsymbol{k} = 0$ but at $\boldsymbol{k}_{\mathrm{c}} = (2\pi/a)(0, 0, \pm 0.85)$ (and four other symmetrically located points along the k_x- and k_y-axes). The dispersion relation can be approximated by a quadratic expression around these minima. Because of the rotational symmetry around the k_z-axis, the form

$$\varepsilon_{\boldsymbol{k}} = \hbar^2 \left[\frac{k_x^2 + k_y^2}{2m^*_{n\perp}} + \frac{(k_z - k_{cz})^2}{2m^*_{n\|}} \right] \tag{20.2.7}$$

is expected with two different effective masses. Electrons promoted into the conduction band fill these valleys. Cyclotron resonance, which will be discussed in the next chapter, is the method of choice for determining these effective masses. According to measurements,

$$m^*_{n\|} = 0.916\, m_e, \qquad m^*_{n\perp} = 0.191\, m_e\,. \tag{20.2.8}$$

Owing to the huge difference between the components of the effective-mass tensor, the constant-energy surfaces of the conduction band form six prolate ellipsoids of rotation in the Brillouin zone. These are shown on the right-hand side of Fig. 20.2.

Since electrons feel an appreciable field gradient close to the nucleus, spin–orbit interaction gives rise to further splitting inside the valence band. To specify it, it has to be borne in mind that the spin variable transforms according to the irreducible representation $D_{1/2}$ of the rotation group, therefore the wavefunction that contains both spatial and spin variables transforms according to the direct product of Γ'_{25} and $D_{1/2}$ for the states at the top of the valence band. It follows directly from (6.1.19) that the triply (or, together with spin, sixfold) degenerate state of symmetry Γ'_{25} is split in point Γ into a fourfold degenerate state that transforms according to the irreducible representation Γ_8 of the double group and a doubly degenerate state that transforms according to Γ_7. The magnitude of spin–orbit splitting is about 0.044 eV. In the points $\boldsymbol{k} \neq 0$ the fourfold degenerate level is split further; however, as the point group of the diamond structure contains inversion symmetry, spin degeneracy is preserved. The schematic pattern of energy levels is shown in Fig. 20.3.

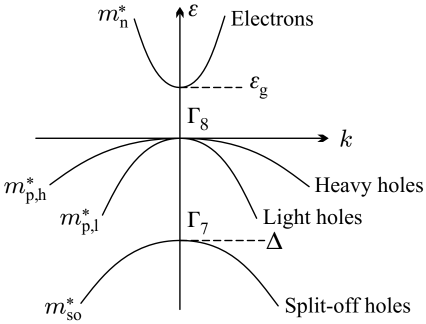


Fig. 20.3. Splitting of the state of symmetry Γ'_{25} close to point Γ, due to spin–orbit coupling. Indices show the irreducible representations of the double group according to which each band transforms in Γ

The energies obtained by taking spin–orbit interaction into account can no longer be described by a quadratic function at the top of the valence band; the form

$$\begin{aligned}
\varepsilon_{\mathrm{p}1}(\boldsymbol{k}) &= -\frac{\hbar^2}{2m_{\mathrm{e}}}\left\{Ak^2 - \left[B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2)\right]^{1/2}\right\},\\
\varepsilon_{\mathrm{p}2}(\boldsymbol{k}) &= -\frac{\hbar^2}{2m_{\mathrm{e}}}\left\{Ak^2 + \left[B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2)\right]^{1/2}\right\}, \qquad (20.2.9)\\
\varepsilon_{\mathrm{p}3}(\boldsymbol{k}) &= -\varDelta - A\frac{\hbar^2}{2m_{\mathrm{e}}}k^2
\end{aligned}$$

have to be used instead. Measurements give

$$A = 4.27\,, \qquad B = 0.63\,, \qquad C = 4.93\,, \qquad (20.2.10)$$

for silicon, and $\varDelta = 0.044\,\mathrm{eV}$, as mentioned above. The corresponding constant-energy surfaces are not ellipsoidal: they look like spheres with humps along the $\langle 100\rangle$ or $\langle 111\rangle$ directions, as illustrated in Fig. 20.4.

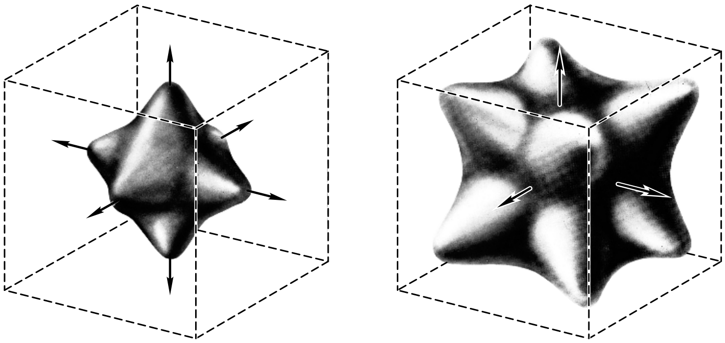


Fig. 20.4. Constant-energy surfaces at the top of the valence band of silicon for heavy and light holes

Despite the nonquadratic energy versus wave vector relation for the first two bands, it is customary to speak of effective hole masses, by which values averaged over all directions are meant. A heavy- and a light-hole mass are obtained:

$$m^*_{\mathrm{p,h}} = 0.537\,m_{\mathrm{e}}\,, \qquad m^*_{\mathrm{p,l}} = 0.153\,m_{\mathrm{e}}\,. \qquad (20.2.11)$$

In the third band called split-off band holes of effective mass $m^*_{\mathrm{so}} = 0.234\,m_{\mathrm{e}}$ appear.

20.2.3 Band Structure of Germanium

The calculated band structure of germanium is shown in Fig. 20.5. By and large, it is also in agreement with the qualitative picture presented above for the diamond lattice, which was obtained by lifting the degeneracies in the empty-lattice approximation.

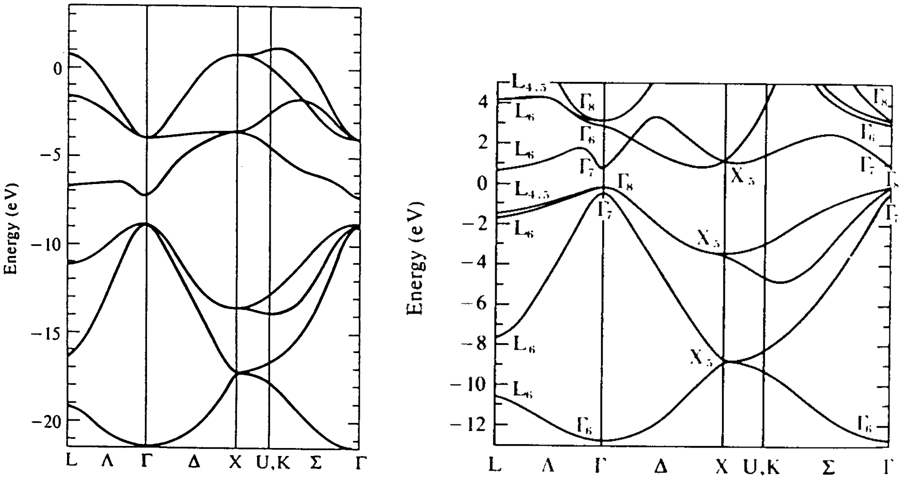


Fig. 20.5. The energy bands of germanium. The bands calculated by the LCAO method are displayed on the left, while the results of a pseudopotential calculation that takes spin–orbit coupling into account are shown on the right. In the latter case the letters next to each band refer to the corresponding irreducible representation of the double group [Reprinted with permission from J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**, 556 (1976). ©1976 by the American Physical Society]

When the spin–orbit interaction is neglected, the eightfold degenerate level in point Γ splits to four levels of unequal energies in the same way as in silicon. However, the order of these levels, which cannot be determined from symmetry considerations alone, is different. Above the triply degenerate level of symmetry Γ'_{25} the nondegenerate level of symmetry Γ'_2 is located. The state of symmetry Γ_{15} lies above this. However, just like in silicon, the gap, which determines the thermal properties, is not equal to the splitting between the Γ'_{25} and Γ'_2 levels in point Γ, since the wave vector is not the same for the highest-lying state in the valence band and the lowest-lying state in the conduction band.

As Fig. 20.5 shows, the minimum of the conduction band is at the edge of the Brillouin zone, in $\boldsymbol{k}_{\mathrm{c}} = L = (2\pi/a)(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ (plus three symmetrically located points).[2] Figure 20.6 shows those regions in $\boldsymbol{k}$-space where thermally excited conduction electrons are located. Just like silicon, germanium is a multivalley semiconductor.

Close to the bottom of the conduction band, the energy can be approximated by a quadratic form. By choosing the direction [111] as one of the

[2] This is because the eight points $(2\pi/a)(\pm\frac{1}{2}, \pm\frac{1}{2}, \pm\frac{1}{2})$ form four equivalent pairs.

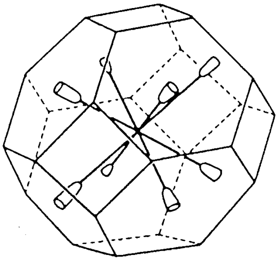


Fig. 20.6. Constant-energy surfaces at the bottom of the conduction band for germanium

principal axes, the effective-mass tensor is found to be diagonal, with a longitudinal and a transverse mass. Their experimental values determined by cyclotron resonance are

$$m^*_{n\|} = 1.588\, m_{\mathrm{e}} \qquad \text{and} \qquad m^*_{n\perp} = 0.082\, m_{\mathrm{e}}\,. \tag{20.2.12}$$

Spin–orbit interaction gives rise to further splitting at the top of the valence band in germanium, too. This is illustrated on the right-hand side of Fig. 20.5, where the irreducible representations of the double group are used for indexing states. The parameter values for germanium in (20.2.9), the formula for the electron energies, are

$$A = 13.38\,, \qquad B = 8.57\,, \qquad C = 12.78\,. \tag{20.2.13}$$

The masses of heavy and light holes, obtained by averaging over directions, are

$$m^*_{\mathrm{p,h}} = 0.28\, m_{\mathrm{e}}\,, \qquad m^*_{\mathrm{p,l}} = 0.044\, m_{\mathrm{e}}\,. \tag{20.2.14}$$

The third energy level is shifted by $\Delta = 0.295\,\mathrm{eV}$, which is almost an order of magnitude larger than for silicon. The effective mass of holes in the split-off band is $m^*_{\mathrm{so}} = 0.095\, m_{\mathrm{e}}$.

20.2.4 Band Structure of Compound Semiconductors

The band structure of compound semiconductors that crystallize in the sphalerite structure is very similar to those of silicon and germanium. This is illustrated in Fig. 20.7 for two cases – both calculated by taking spin–orbit interaction into account.

The most important difference with silicon and germanium is that the minimum of the conduction band is at Γ. In its vicinity, electron states can be characterized by a scalar effective mass. The highest energy in the valence band also occurs at Γ. Here, too, the triple (or, together with spin, sixfold) degeneracy is split by spin–orbit interaction to a lower-lying nondegenerate level and a doubly degenerate level that splits further into two levels away

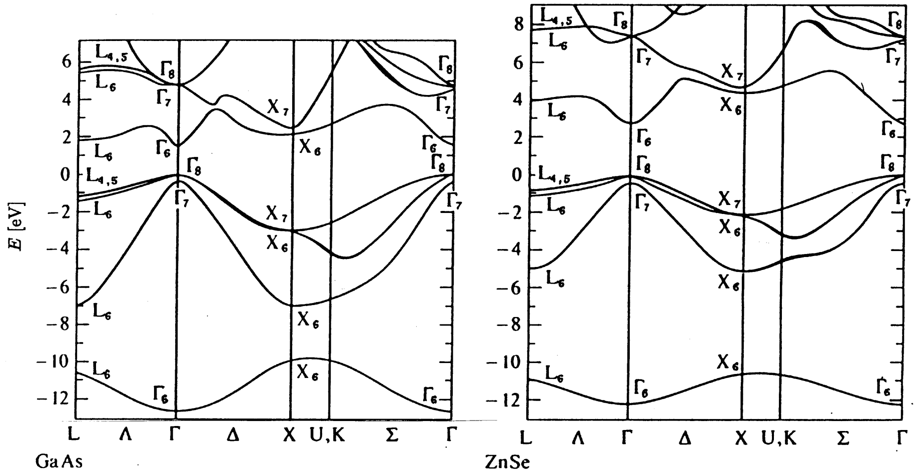


Fig. 20.7. The energy bands of GaAs and ZnSe [Reprinted with permission from J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**, 556 (1976). ©1976 by the American Physical Society]

from Γ: one corresponds to a low- and the other to a high-effective-mass hole. When spin–orbit interaction is taken into account more accurately, a more complicated band structure arises. This is because inversion symmetry is broken in the sphalerite structure, and thus the double degeneracy due to spin is lifted. This phenomenon, called the *Dresselhaus splitting*, was discussed in Chapter 17. The conduction band containing the $\boldsymbol{k} = 0$ state of symmetry Γ_6 is split and a tiny asymmetry – a small k^3 term – appears in the energy. Nevertheless, the full band structure still possesses the symmetry $\varepsilon_{\boldsymbol{k}} = \varepsilon_{-\boldsymbol{k}}$. Similar things apply to the valence bands that contain the points associated with the irreducible representations Γ_7 and Γ_8. However, the asymmetry is so weak that it is imperceptible in Fig. 20.7.

Table 20.7 shows the effective mass for electron and hole states in some compound semiconductors – which are all direct-gap semiconductors, that is, the lowest-energy state in the conduction band and the highest-energy state in the valence band are associated with the same wave vector $\boldsymbol{k}$. The effective mass is seen to be much lower than the electron mass. It is not difficult to show that this is not by accident but the direct consequence of the formation of the gap. To do so, we have to go back to the nearly-free-electron calculation of how the two energy levels, which are degenerate in the absence a periodic potential, are split at the center or edges of the Brillouin zone.

It was shown in Chapter 18 in connection with (18.1.29) that the free-electron energies $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i}$ and $\varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j}$ are equal for those vectors $\boldsymbol{k}$ for which $\boldsymbol{k}+\boldsymbol{G}_i$ and $\boldsymbol{k}+\boldsymbol{G}_j$ are on the same Bragg plane that perpendicularly bisects the vector $\boldsymbol{G}_i - \boldsymbol{G}_j$. It then follows that the vector

Table 20.7. Effective masses in the valence and conduction bands for some compound semiconductors

Crystal	$m_{\mathrm{n}}^*/m_{\mathrm{e}}$	$m_{\mathrm{p,h}}^*/m_{\mathrm{e}}$	$m_{\mathrm{p,l}}^*/m_{\mathrm{e}}$	$m_{\mathrm{so}}^*/m_{\mathrm{e}}$
GaAs	0.066	0.47	0.07	0.15
GaSb	0.042	0.35	0.05	0.12
InP	0.077	0.56	0.12	0.12
InAs	0.024	0.43	0.026	0.14
InSb	0.014	0.39	0.016	0.43

$$\boldsymbol{k}_{\|} = \boldsymbol{k} + \tfrac{1}{2}(\boldsymbol{G}_i + \boldsymbol{G}_j) \tag{20.2.15}$$

lies in the Bragg plane, and is perpendicular to the vector $\boldsymbol{G}_i - \boldsymbol{G}_j$. To determine the effective masses, consider now the vectors

$$\boldsymbol{k} = \boldsymbol{k}_{\|} + \boldsymbol{k}_{\perp} - \tfrac{1}{2}(\boldsymbol{G}_i + \boldsymbol{G}_j) \tag{20.2.16}$$

away from the Bragg plane, where $\boldsymbol{k}_{\|}$ lies in the Bragg plane and $\boldsymbol{k}_{\perp}$ is perpendicular to it. Then

$$\boldsymbol{k} + \boldsymbol{G}_i = \boldsymbol{k}_0 + \boldsymbol{k}_{\|} + \boldsymbol{k}_{\perp} \qquad \text{and} \qquad \boldsymbol{k} + \boldsymbol{G}_j = -\boldsymbol{k}_0 + \boldsymbol{k}_{\|} + \boldsymbol{k}_{\perp}\,, \tag{20.2.17}$$

where the vector $\boldsymbol{k}_0 = \frac{1}{2}(\boldsymbol{G}_i - \boldsymbol{G}_j)$ is perpendicular to $\boldsymbol{k}_{\|}$. Using a quadratic dispersion relation for the unperturbed states,

$$\begin{aligned} \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_i} =& \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\boldsymbol{k}_{\|}^2 + \boldsymbol{k}_0^2 + 2\boldsymbol{k}_0 \cdot \boldsymbol{k}_{\perp} + \boldsymbol{k}_{\perp}^2\right), \\ \varepsilon^{(0)}_{\boldsymbol{k}+\boldsymbol{G}_j} =& \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\boldsymbol{k}_{\|}^2 + \boldsymbol{k}_0^2 - 2\boldsymbol{k}_0 \cdot \boldsymbol{k}_{\perp} + \boldsymbol{k}_{\perp}^2\right). \end{aligned} \tag{20.2.18}$$

Substituting these into (18.1.34), the energies in the presence of a periodic potential are

$$\varepsilon_{\boldsymbol{k}} = \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\boldsymbol{k}_{\|}^2 + \boldsymbol{k}_0^2 + \boldsymbol{k}_{\perp}^2\right) \pm \left[4\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{k}_0^2\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{k}_{\perp}^2 + |U(2\boldsymbol{k}_0)|^2\right]^{1/2}. \tag{20.2.19}$$

The vector $\boldsymbol{k}$ is close to the Bragg plane provided $|\boldsymbol{k}_{\perp}|$ is small. Expansion for small values of $\boldsymbol{k}_{\perp}$ yields

$$\begin{aligned} \varepsilon^{(+)}_{\boldsymbol{k}} =& \varepsilon^{(0)}_{\boldsymbol{k}_0} + \frac{\hbar^2\boldsymbol{k}_{\|}^2}{2m_{\mathrm{e}}} + |U(2\boldsymbol{k}_0)| + \frac{\hbar^2\boldsymbol{k}_{\perp}^2}{2m_{\mathrm{e}}}\left[\frac{2\varepsilon^{(0)}_{\boldsymbol{k}_0}}{|U(2\boldsymbol{k}_0)|} + 1\right], \\ \varepsilon^{(-)}_{\boldsymbol{k}} =& \varepsilon^{(0)}_{\boldsymbol{k}_0} + \frac{\hbar^2\boldsymbol{k}_{\|}^2}{2m_{\mathrm{e}}} - |U(2\boldsymbol{k}_0)| - \frac{\hbar^2\boldsymbol{k}_{\perp}^2}{2m_{\mathrm{e}}}\left[\frac{2\varepsilon^{(0)}_{\boldsymbol{k}_0}}{|U(2\boldsymbol{k}_0)|} - 1\right]. \end{aligned} \tag{20.2.20}$$

It is immediately seen that the motion parallel to the Bragg plane is not affected by the periodic potential in this approximation: this element of the effective-mass tensor remains the same. However, the masses associated with the motion perpendicular to the Bragg plane are modified, renormalized:

$$m^* = \pm \frac{m_\mathrm{e}}{2\varepsilon^{(0)}_{\boldsymbol{k}_0}/|U(2\boldsymbol{k}_0)| \mp 1} \,. \tag{20.2.21}$$

Since the magnitude of the potential is usually smaller than the energy of the level to be split, a positive and a negative effective mass are obtained, and both are smaller than the electron mass.

20.2.5 Indirect- and Direct-Gap Semiconductors

In silicon and germanium, the maximum of the valence band was seen to be at $\boldsymbol{k}_\mathrm{v} = 0$, while the minimum of the conduction band at another wave vector, $\boldsymbol{k}_\mathrm{c} \neq 0$. Such semiconductors are called *indirect-gap semiconductors.* When the top of the valence band and the bottom of the conduction band are at the same wave vector, $\boldsymbol{k}_\mathrm{c} = \boldsymbol{k}_\mathrm{v}$, we speak of *direct-gap semiconductors.* GaAs and InSb are two notable representatives of the second group. The distinction is important because in indirect-gap semiconductors it is not immaterial how the gap is measured.

The thermal properties depend strongly on the number of thermally excited electrons in the conduction band. This, in turn, is determined in part by the smallest energy required for the excitation of electrons. The same energy appears as the activation energy in the temperature dependence of semiconductors in the region where the most important factor in the variation of resistivity is the variation in the number of thermally excited carriers.

However, it is not this energy that is observed in indirect-gap semiconductors in optical measurements. By shining light on a semiconductor, an electron in the valence band may absorb a photon and may be promoted to the conduction band. Since the wave number of the light inducing the transition is typically $k \sim 10^6\,\mathrm{m}^{-1}$, which is much smaller than the typical wave number of electrons calculated from the size of the Brillouin zone (of order $10^{10}\,\mathrm{m}^{-1}$), the wave number of the electron practically does not change in the transition. In the $\boldsymbol{k}$-space plot of the optical transition between the valence band and the conduction band, the electron is located vertically above the hole, as in Fig. 20.8(a).

By varying the energy of the incident light, absorption occurs above a photon energy threshold, where excitation into the conduction band becomes possible. This absorption threshold is the width of the direct (or optical) gap. The indirect band gap, which determines thermal properties, cannot be measured directly in optical absorption measurements. However, there exist higher-order processes in which the electron excited into the conduction band is scattered into the minimum of the conduction band upon the emission of a

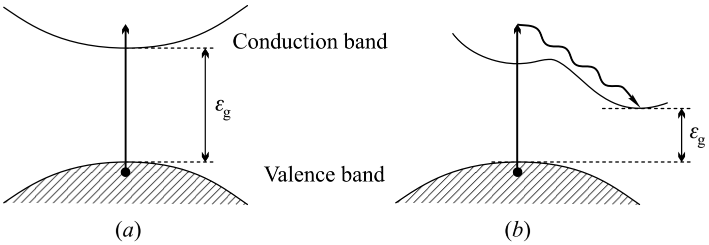


Fig. 20.8. Optical transitions in semiconductors: (*a*) direct transition, (*b*) indirect transition

phonon, as illustrated in the process in Fig. 20.8(*b*). If the phonon spectrum is known, the indirect band gap can be determined.

20.3 Electrons and Holes in Intrinsic Semiconductors

In the previous sections of this chapter we dealt with the determination of the energy of one-particle states – i.e., the band structure. In the ground state the valence band is completely filled and the conduction band is empty. The interesting properties of semiconductors are due partly to the fact that the energy gap is relatively small, thus charge carriers can be generated in the conduction and valence bands even by thermal excitation. The electrical conductivity of pure (intrinsic) semiconductors – which are void of impurities and dopants – is, to a large extent, determined precisely by the number of such carriers. It is relatively easy to evaluate it theoretically, since if the band structure is known, the occupation of the one-particle states at finite temperatures can be calculated using the methods of statistical physics. Assuming a simple band structure, we shall determine the number of charge carriers below, and also examine the position of the chemical potential with respect to the conduction and valence bands.

20.3.1 Number of Thermally Excited Charge Carriers

Thermal excitation of electrons is usually possible only from the top of the valence band to the bottom of the conduction band. Therefore the number of thermally excited charge carriers can be calculated without knowing the full band structure. To simplify the calculation, we shall assume that the states that lie close to the bottom of the conduction band and to the top of the valence band can be characterized by the scalar effective masses m_n^* and m_p^*, respectively. Then, by generalizing (16.2.54), we shall use the formulas

$$\rho_{\mathrm{c}}(\varepsilon) = \frac{1}{2\pi^2}\left(\frac{2m_{\mathrm{n}}^*}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon-\varepsilon_{\mathrm{c}}}\,, \tag{20.3.1-a}$$

$$\rho_{\mathrm{v}}(\varepsilon) = \frac{1}{2\pi^2}\left(\frac{2m_{\mathrm{p}}^*}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon_{\mathrm{v}}-\varepsilon} \tag{20.3.1-b}$$

for the density of states in the conduction and valence bands, where ε_{c} is the energy of the bottom of the conduction band and ε_{v} is that of the top of the valence band.

These formulas can be straightforwardly generalized to the cases when the longitudinal and transverse components of the effective-mass tensor are different, and when the possibility of having light and heavy holes is taken into account. If the one-particle energies in the conduction band can be characterized approximately by the elements of an effective-mass tensor m_{ij}^*, then, according to (17.4.42), the effective mass is replaced by

$$m_{\mathrm{ds}}^* = \left[\det\left(m_{ij}^*\right)\right]^{1/3} \tag{20.3.2}$$

in the density of states. When the conduction band contains ν equivalent valleys (as in silicon and germanium), the density of states calculated for a single valley has to be multiplied by ν. This is equivalent to saying that the mass in the density of states is $\nu^{2/3}$ times the effective mass determined for a single-valley configuration.

By taking the average of the transverse and longitudinal effective masses, the contribution of a single valley can be taken into account by an effective mass of $m_{\mathrm{therm}} = 0.32\,m_{\mathrm{e}}$ in silicon. Since the electrons in each of the six valleys contribute equally to the density of states, one has to use the density-of-states mass

$$m_{\mathrm{n,ds}}^* = 6^{2/3} m_{\mathrm{therm}} = 1.18\,m_{\mathrm{e}}\,. \tag{20.3.3}$$

The contribution of a single valley can be taken into account by an effective mass of $m_{\mathrm{therm}} = 0.22\,m_{\mathrm{e}}$ in germanium. Since the bottom of the conduction band consists of four equivalent valleys, the correct density-of-states mass is

$$m_{\mathrm{n,ds}}^* = 4^{2/3} m_{\mathrm{therm}} = 0.55\,m_{\mathrm{e}}\,. \tag{20.3.4}$$

To determine the total density of states in the valence band, the density of states for the heavy and light holes (of mass $m_{\mathrm{p,h}}^*$ and $m_{\mathrm{p,l}}^*$) must be added, which is equivalent to saying that holes of effective mass

$$m_{\mathrm{p,ds}}^* = \left({m_{\mathrm{p,h}}^*}^{3/2} + {m_{\mathrm{p,l}}^*}^{3/2}\right)^{2/3} \tag{20.3.5}$$

form a single band. For silicon,

$$m_{\mathrm{p,ds}}^* = 0.59\,m_{\mathrm{e}}\,, \tag{20.3.6}$$

whereas for germanium

$$m^*_{\mathrm{p,ds}} = 0.37\, m_{\mathrm{e}}\,. \tag{20.3.7}$$

In what follows, the simple notations m^*_{n} and m^*_{p} will refer to these density-of-state masses.

Using the Fermi–Dirac distribution for the thermal occupation of the conduction-band states, the density of electrons in the conduction band is

$$n(T) = \int\limits_{\varepsilon_{\mathrm{c}}}^{\infty} \rho_{\mathrm{c}}(\varepsilon) \frac{1}{\mathrm{e}^{(\varepsilon-\mu)/k_{\mathrm{B}}T}+1}\,\mathrm{d}\varepsilon \tag{20.3.8}$$

in thermal equilibrium, where μ is the yet-unknown value of the chemical potential.

Analogously, the density of electrons in the valence band is

$$n_{\mathrm{v}}(T) = \int\limits_{-\infty}^{\varepsilon_{\mathrm{v}}} \rho_{\mathrm{v}}(\varepsilon) \frac{1}{\mathrm{e}^{(\varepsilon-\mu)/k_{\mathrm{B}}T}+1}\,\mathrm{d}\varepsilon\,. \tag{20.3.9}$$

Since the valence band is almost completely filled, it is more practical to consider the holes there instead of the electrons, and to specify the number density of thermally excited states. Denoting the density of holes by p,

$$p(T) = \int\limits_{-\infty}^{\varepsilon_{\mathrm{v}}} \rho_{\mathrm{v}}(\varepsilon) \left[1 - \frac{1}{\mathrm{e}^{(\varepsilon-\mu)/k_{\mathrm{B}}T}+1}\right]\mathrm{d}\varepsilon\,. \tag{20.3.10}$$

After a simple rearrangement,

$$p(T) = \int\limits_{-\infty}^{\varepsilon_{\mathrm{v}}} \rho_{\mathrm{v}}(\varepsilon) \frac{1}{\mathrm{e}^{(\mu-\varepsilon)/k_{\mathrm{B}}T}+1}\,\mathrm{d}\varepsilon\,. \tag{20.3.11}$$

The negative sign in the energy dependence (relative to the usual Fermi–Dirac distribution) is the consequence of the fact that holes are obtained from electrons through the transformation $\varepsilon - \mu \leftrightarrow -(\varepsilon - \mu)$. The density of states and the thermal population of the states at finite temperature are illustrated in Fig. 20.9 for both bands. The third part of the figure shows the occupation of electron and hole states – that is, the distribution of thermally excited electrons (holes) in the conduction (valence) band is plotted against energy.

Substituting the density of states formula (20.3.1) into (20.3.8), the density of thermally excited electrons is

$$\begin{aligned} n(T) &= \frac{1}{2\pi^2}\left(\frac{2m^*_{\mathrm{n}}}{\hbar^2}\right)^{3/2} \int\limits_{\varepsilon_{\mathrm{c}}}^{\infty} \frac{\sqrt{\varepsilon-\varepsilon_{\mathrm{c}}}}{\mathrm{e}^{(\varepsilon-\mu)/k_{\mathrm{B}}T}+1}\,\mathrm{d}\varepsilon \\ &= 2\left(\frac{m^*_{\mathrm{n}}k_{\mathrm{B}}T}{2\pi\hbar^2}\right)^{3/2} F_{1/2}\left(-(\varepsilon_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T\right), \end{aligned} \tag{20.3.12}$$

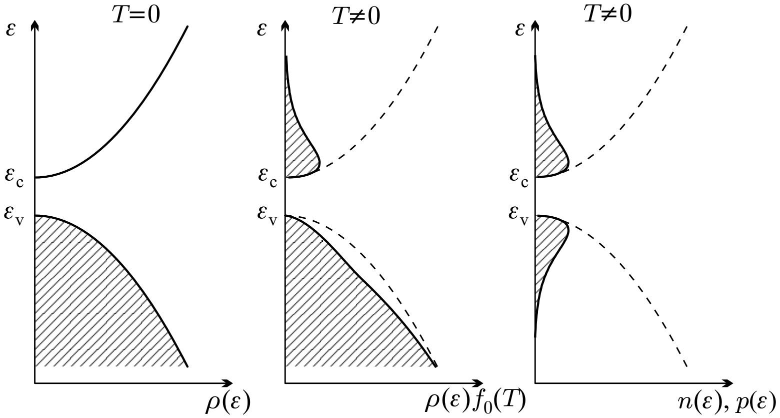


Fig. 20.9. The density of states in the valence and conduction bands, and the thermal occupation of states at a finite temperature, assuming equal effective masses. On the right-hand side the occupation of hole (rather than electron) states is shown in the valence band

where $F_{1/2}(x)$ is the Fermi integral of index $j = 1/2$ defined in (16.2.62). While $F_{1/2}(x)$ could be approximated by its low-temperature asymptotic form for metals, the asymptotic expression in the opposite limit can be usually employed for semiconductors. The value of the chemical potential μ has not been established yet, but it is already known that at $T = 0$ it is somewhere inside the energy gap separating filled and empty states. This means that in semiconductors there are no electrons whose energy is the same as the Fermi energy – thus the Fermi surface is absent.

We shall see that the chemical potential is located around mid-gap. Since the gap in semiconductors is usually much larger than the thermal energy at room temperature, $k_\mathrm{B}T \approx 0.025\,\mathrm{eV}$, we shall assume that

$$\varepsilon_\mathrm{c} - \mu \gg k_\mathrm{B}T\,, \qquad \mu - \varepsilon_\mathrm{v} \gg k_\mathrm{B}T\,. \tag{20.3.13}$$

In this limit the quantum mechanical Fermi–Dirac distribution can be approximated by the classical Maxwell–Boltzmann distribution:

$$\frac{1}{\mathrm{e}^{(\varepsilon-\mu)/k_\mathrm{B}T}+1} \sim \mathrm{e}^{-(\varepsilon-\mu)/k_\mathrm{B}T}, \qquad \text{if } \varepsilon > \varepsilon_\mathrm{c}\,, \tag{20.3.14-a}$$

$$\frac{1}{\mathrm{e}^{(\mu-\varepsilon)/k_\mathrm{B}T}+1} \sim \mathrm{e}^{-(\mu-\varepsilon)/k_\mathrm{B}T}, \qquad \text{if } \varepsilon < \varepsilon_\mathrm{v}\,. \tag{20.3.14-b}$$

Below we shall deal only with this so-called *nondegenerate* case. When condition (20.3.13) is not met – that is, quantum statistics cannot be adequately approximated by classical statistics –, we speak of a *degenerate semiconductor.* The calculations are somewhat more tedious in this case but they do not present any difficulty of principle.

Substituting the classical distribution function (20.3.14-a) into (20.3.8), we have

$$n(T) = \int\limits_{\varepsilon_{\mathrm{c}}}^{\infty} \rho_{\mathrm{c}}(\varepsilon) \mathrm{e}^{-(\varepsilon-\mu)/k_{\mathrm{B}}T} \,\mathrm{d}\varepsilon \,. \tag{20.3.15}$$

Since the density of states in the conduction band depends on $\varepsilon - \varepsilon_{\mathrm{c}}$ alone, by introducing the quantity

$$N_{\mathrm{c}}(T) = \int\limits_{\varepsilon_{\mathrm{c}}}^{\infty} \rho_{\mathrm{c}}(\varepsilon) \mathrm{e}^{-(\varepsilon-\varepsilon_{\mathrm{c}})/k_{\mathrm{B}}T} \,\mathrm{d}\varepsilon \,, \tag{20.3.16}$$

the density (concentration) of electrons can be written as

$$n(T) = N_{\mathrm{c}}(T) \mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T} \,. \tag{20.3.17}$$

The value of $N_{\mathrm{c}}(T)$ depends on the specific form of the density of states. Making use of the formula for the density of states for electrons of effective mass m_{n}^*, (20.3.1), and exploiting (C.2.1), the integral (20.3.16) can be evaluated:

$$N_{\mathrm{c}}(T) = 2 \left(\frac{m_{\mathrm{n}}^* k_{\mathrm{B}} T}{2\pi\hbar^2} \right)^{3/2} . \tag{20.3.18}$$

The formula for the density of conduction electrons is therefore equivalent to (20.3.12), with the Fermi integral approximated by its high-temperature asymptotic form, (C.2.24). It should be noted that by substituting the actual values of the physical constants, $N_{\mathrm{c}}(T)$ can be rewritten as

$$N_{\mathrm{c}}(T) = 2.5 \left(\frac{m_{\mathrm{n}}^*}{m_{\mathrm{e}}} \right)^{3/2} \left(\frac{T}{300\,\mathrm{K}} \right)^{3/2} \times 10^{19}/\mathrm{cm}^3. \tag{20.3.19}$$

It is readily seen that in nondegenerate semiconductors the density of charge carriers cannot exceed 10^{18} to $10^{19}/\mathrm{cm}^3$.

Analogously, the density of holes in the valence band is

$$p(T) = P_{\mathrm{v}}(T) \mathrm{e}^{-(\mu-\varepsilon_{\mathrm{v}})/k_{\mathrm{B}}T} \,, \tag{20.3.20}$$

where

$$P_{\mathrm{v}}(T) = \int\limits_{-\infty}^{\varepsilon_{\mathrm{v}}} \rho_{\mathrm{v}}(\varepsilon) \mathrm{e}^{-(\varepsilon_{\mathrm{v}}-\varepsilon)/k_{\mathrm{B}}T} \,\mathrm{d}\varepsilon \,. \tag{20.3.21}$$

Assuming an effective hole mass of m_{p}^* and substituting the density of states (20.3.1) into the integral,

$$P_{\mathrm{v}}(T) = 2 \left(\frac{m_{\mathrm{p}}^* k_{\mathrm{B}} T}{2\pi\hbar^2} \right)^{3/2} . \tag{20.3.22}$$

In pure, undoped semiconductors, where all thermally excited electrons in the conduction band come from the valence band, the number of holes generated in the valence band must be equal to the number of electrons in the conduction band: $n(T) = p(T)$. This common value, denoted by $n_{\mathrm{i}}(T)$, is called the *intrinsic carrier density* or *intrinsic carrier concentration.* To evaluate it, it should be noted that in thermal equilibrium the product of the number of electrons in the conduction band and the number of holes in the valence band,

$$n(T)\,p(T) = N_{\mathrm{c}}(T)\,P_{\mathrm{v}}(T)\,\mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\varepsilon_{\mathrm{v}})/k_{\mathrm{B}}T}, \tag{20.3.23}$$

is independent of the value of the chemical potential, thus

$$n_{\mathrm{i}}(T) = \sqrt{N_{\mathrm{c}}(T)P_{\mathrm{v}}(T)}\mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\varepsilon_{\mathrm{v}})/2k_{\mathrm{B}}T} = \sqrt{N_{\mathrm{c}}(T)P_{\mathrm{v}}(T)}\mathrm{e}^{-\varepsilon_{\mathrm{g}}/2k_{\mathrm{B}}T}. \tag{20.3.24}$$

The exponent in the formula for the density of the thermally excited charge carriers contains the half of the gap, confirming the statement made in the introduction of the chapter. This is formally due to the mid-gap location of the chemical potential at low temperatures, therefore the number of excitations depends on the energy measured from that point. A more intuitive interpretation is based on the picture of the thermal generation of electron–hole pairs with electrons in the conduction band and holes in the valence band. Since the pair creation energy is the same as the gap, half of it is attributed to the electron and half to the hole.

By substituting (20.3.18) and (20.3.22) into (20.3.24), the formula

$$n_{\mathrm{i}}(T) = 2\left(\frac{k_{\mathrm{B}}T}{2\pi\hbar^2}\right)^{3/2}(m_{\mathrm{n}}^* m_{\mathrm{p}}^*)^{3/4}\mathrm{e}^{-\varepsilon_{\mathrm{g}}/2k_{\mathrm{B}}T} \tag{20.3.25}$$

for the intrinsic carrier density contains the known parameters of semiconductors. It should be emphasized once again: to obtain numerical values, the masses in the expression for the density of states have to be used. Applying (20.3.19) and its counterpart to P_{v}, this can be rewritten as

$$n_{\mathrm{i}}(T) = 2.5\left(\frac{m_{\mathrm{n}}^*}{m_{\mathrm{e}}}\right)^{3/4}\left(\frac{m_{\mathrm{p}}^*}{m_{\mathrm{e}}}\right)^{3/4}\left(\frac{T}{300\,\mathrm{K}}\right)^{3/2}\mathrm{e}^{-\varepsilon_{\mathrm{g}}/2k_{\mathrm{B}}T}\times 10^{19}/\mathrm{cm}^3\,. \tag{20.3.26}$$

From the known values of the effective mass and the energy gap, the intrinsic carrier density of silicon at room temperature is $n_{\mathrm{i}}^{\mathrm{Si}}(300\,\mathrm{K}) = 1.02\times 10^{10}/\mathrm{cm}^3$, while it is larger in germanium on account of the smaller gap: $n_{\mathrm{i}}^{\mathrm{Ge}}(300\,\mathrm{K}) = 2.33\times 10^{13}/\mathrm{cm}^3$.

The derivation of (20.3.23) was based on the single assumption that the edges of the conduction and valence bands are both far from the chemical potential on the scale of the thermal energy, therefore classical statistics can be used. Thus the same equation applies to doped semiconductors, too. Even though $n(T) \neq n_{\mathrm{i}}(T)$ and $p(T) \neq n_{\mathrm{i}}(T)$, since the impurity atoms give rise to extra states between the conduction band and the valence band, and therefore the chemical potential is shifted, nonetheless the product of $n(T)$ and $p(T)$ is independent of the chemical potential, and the relationship

$$n(T)\, p(T) = n_{\mathrm{i}}^2(T) \tag{20.3.27}$$

holds for doped semiconductors as well – with the intrinsic carrier density n_i given in (20.3.25). This formula is called the *law of mass action.*

20.3.2 Temperature Dependence of the Chemical Potential

In intrinsic semiconductors the value of the chemical potential can be determined from the requirement that the number of charge carriers should be the same in the valence and conduction bands:

$$N_{\mathrm{c}}(T)\mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T} = P_{\mathrm{v}}(T)\mathrm{e}^{-(\mu-\varepsilon_{\mathrm{v}})/k_{\mathrm{B}}T}. \tag{20.3.28}$$

Rearrangement of the terms gives

$$\frac{P_{\mathrm{v}}(T)}{N_{\mathrm{c}}(T)} = \exp\left[\frac{2\mu - (\varepsilon_{\mathrm{c}}+\varepsilon_{\mathrm{v}})}{k_{\mathrm{B}}T}\right], \tag{20.3.29}$$

hence the chemical potential is

$$\mu = \tfrac{1}{2}(\varepsilon_{\mathrm{c}}+\varepsilon_{\mathrm{v}}) + \tfrac{1}{2}k_{\mathrm{B}}T \ln \frac{P_{\mathrm{v}}(T)}{N_{\mathrm{c}}(T)}\,. \tag{20.3.30}$$

At $T = 0$ the second term vanishes, and the chemical potential is indeed located precisely in the middle of the gap. Substituting (20.3.18) and (20.3.22) into the second term, which gives the corrections at finite temperatures, we have

$$\mu = \tfrac{1}{2}(\varepsilon_{\mathrm{c}}+\varepsilon_{\mathrm{v}}) + \tfrac{3}{4}k_{\mathrm{B}}T \ln \frac{m_{\mathrm{p}}^*}{m_{\mathrm{n}}^*}\,. \tag{20.3.31}$$

Thus, unlike in metals, the chemical potential varies linearly with temperature in semiconductors. The sign of the variations depends on the ratio of the effective masses in the two bands. The chemical potential shifts toward the band whose effective mass is lower, but the deviation from the middle of the gap is tiny even at room temperature, and can be practically neglected.

Introducing the notation μ_{i} for the chemical potential of pure semiconductors, its connection with n_{i} can be given in two equivalent ways:

$$n_{\mathrm{i}} = N_{\mathrm{c}}(T)\mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\mu_{\mathrm{i}})/k_{\mathrm{B}}T} \tag{20.3.32}$$

or

$$n_{\mathrm{i}} = P_{\mathrm{v}}(T)\mathrm{e}^{-(\mu_{\mathrm{i}}-\varepsilon_{\mathrm{v}})/k_{\mathrm{B}}T}. \tag{20.3.33}$$

20.4 Electronic Structure of Doped Semiconductors

In the previous sections we studied the electronic structure of semiconductors of stoichiometric composition, in which carriers are generated by exciting electrons from the valence band into the conduction band across the gap, leaving behind holes. The number of carriers can be controlled only by varying the temperature. By deviating from the stoichiometric composition, e.g., by doping a semiconductor with impurities, the concentration of electrons and holes can be greatly increased. Semiconductors in which a substantial proportion of the carriers are provided by impurities are called *extrinsic* or *doped semiconductors*. Present-day silicon-based semiconductor technology is capable of producing staggeringly high purity crystals with about one dopant per 10^{12} Si atoms – that is, the impurity concentration is about 10^{10} atoms/cm^3. The majority of the impurity atoms are not ionized, and are therefore inactive in the sense that they do not affect the number of electrons or holes responsible for conduction. More important are those active impurities that change the number of mobile carriers. The most important role is played by those dopants that can easily give away or take up an electron. Typically, elemental semiconductors in the carbon group (group IVA) – e.g., silicon – are doped by substituting some of the group IVA (Si) atoms by an element of group IIIA (e.g., gallium) or group VA (e.g., arsenic).[3]

Elements of group VA have one more electron on their outermost p-shell than silicon. Only four s- and p-electrons are required to form the covalent bonds. To ensure charge neutrality, each dopant donates an additional electron to the system. Such impurities are therefore called *donors*. Figure 20.10(a) shows the electrons in the tetrahedrally coordinated covalent bonds and the extra electron of the donor.

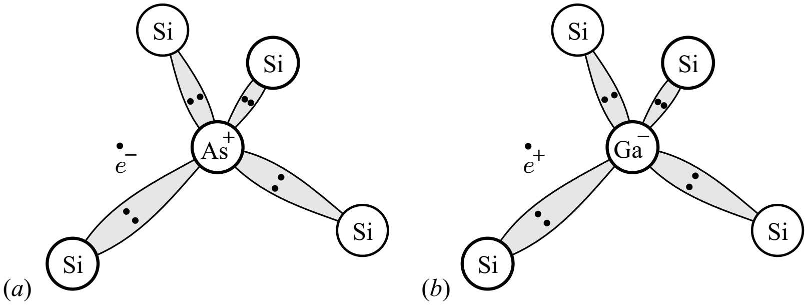


Fig. 20.10. Electron excess (a) and electron deficiency (b) around an arsenic and a gallium atom in silicon

[3] Since the number of thermally excited charge carriers in pure silicon is as low as 10^{10}/cm^3 even at room temperature, the behavior is practically determined by the electrons of the dopants alone.

One would naively expect that the extra electron of the donor occupies a state in the conduction band. But this is wrong. As discussed in Chapter 17, bound states may appear around an impurity, below or above the band, depending on the attractive or repulsive character of the interaction. Since the interaction between the ionized donor and the extra electron is attractive, one would expect that the band structure is modified by donor atoms in such a way that bound states are formed below the conduction band, accommodating the extra electrons of the donor atoms in the ground state. We shall demonstrate that the binding energy is often sufficiently low for that thermal excitation of these states be easily possible, thus the extra electrons also contribute to conduction.

On the other hand, the outermost shell of elements of group IIIA is one electron short compared to silicon. With spatially well localized covalent bonds in mind this implies that in samples doped with such impurities the number of electrons per impurity atom is one less than what would be necessary to form valence bonds. The formation of the chemical bond requires an additional electron, creating an electron deficiency, an apparently positively charged hole, as shown in Fig. 20.10(b). Such impurities are called *acceptors*.

One would expect that the electron deficiencies brought about by the covalent bonds in the presence of acceptors give rise to empty states, holes, in the valence band. This is not the case, as the states in the valence band are reorganized by the impurity potential. When the bonds are formed, the acceptor can be considered as a negatively charged ion, which therefore repels other electrons. The repulsive potential pushes a state outside the continuum, above the band. Since the impurity does not change the total number of possible electron states, there remain one less state in the band as without the impurity. Thus, in spite of the electron deficiency, the valence band is completely filled in the ground state, and there is one electron per impurity atom in the bound state above the band. Since this level could accommodate two electrons, one may say that acceptors add weakly bound holes to the system.

20.4.1 Energy of Donor and Acceptor Levels

Our introductory remarks showed that bound states can be formed around ionized impurities. The method for evaluating the binding energy is known, at least in principle. Based on a simple physical picture, we shall try to estimate these energies below.

Figure 20.10(a) is a schematic representation of the spatial distribution of electrons in the four tetrahedrally coordinated covalent bonds between an arsenic atom and its four silicon neighbors. Since the two cores are different, the bond is slightly polarized. This is felt above all by the excess electron that does not participate in the bonds. Since the charge of the arsenic core is higher than that of the silicon core, the excess electron may be weakly bound to the arsenic ion. As a rough estimate – which will turn out not to be too rough – the sample can be considered electrically neutral and uniform far from the impurity, while the neighborhood of the arsenic ion can be viewed as if a charge $+e$ were located at the impurity in a medium of dielectric constant ϵ_r, and a single electron of charge $-e$ were moving in its field.

This model is very similar to a hydrogen atom, where the electron is in a bound state in the proton's Coulomb field, and the binding energy of the lowest-lying state – the ionization potential – is 13.6 eV. This energy is much larger than the energy gap of semiconductors. If the binding energy were just as high, then this bound state would be completely irrelevant to the properties of semiconductors. However, the energy is much lower in reality: about 0.05 eV for electrons donated by an arsenic atom in a silicon matrix. The reason for this is twofold:

1. The electron moves around the impurity in a dielectric, not in vacuum. In contrast to metals, the static dielectric constant, which is due essentially to core electrons, is rather large in semiconductors because of the finite gap: $\epsilon_\mathrm{r} = 11.7$ in silicon and $\epsilon_\mathrm{r} = 16.0$ in germanium. Similar values, around 10 to 20, are found for the majority of compound semiconductors. Therefore the electron does not feel the bare Coulomb potential $\tilde{e}^2/r$ but a weaker one, $\tilde{e}^2/\epsilon_\mathrm{r} r$. One could say that the other electrons screen the Coulomb potential of the charge $+e$ of the donor atom. The effects of dielectric screening can be taken into account by replacing e by $e/\sqrt{\epsilon_\mathrm{r}}$ in each formula.
2. The motion of the electron in the crystal lattice is characterized by an effective mass m^*. As we have seen, this is often much smaller than the electron mass. It also reduces the binding energy.

If the effects of the periodic potential are taken into account by an effective mass and the influence of the other electrons by a dielectric constant, then the energy of an electron moving in the field of an impurity atom can be calculated from the Schrödinger equation

$$\left\{-\frac{\hbar^2}{2m^*}\boldsymbol{\nabla}^2 - \frac{\tilde{e}^2}{\epsilon_\mathrm{r} r}\right\}\psi(\boldsymbol{r}) = \varepsilon\psi(\boldsymbol{r})\,. \tag{20.4.1}$$

The problem of an electron around a donor in a semiconductor is thus analogous to the quantum mechanical problem of the hydrogen atom, provided the substitutions

$$\tilde{e} \to \frac{\tilde{e}}{\sqrt{\epsilon_\mathrm{r}}} \qquad \text{and} \qquad m_\mathrm{e} \to m^* \tag{20.4.2}$$

are made. The dimensions of electron orbits in the hydrogen atom are known to be on the order of the Bohr radius a_0. With the above substitutions, we now have

$$r_0 = \frac{\hbar^2}{m^*(\tilde{e}^2/\epsilon_\mathrm{r})} = \frac{m_\mathrm{e}}{m^*}\,\epsilon_\mathrm{r}\,a_0 \tag{20.4.3}$$

for the spatial extent of the bound state. Using the known experimental values for silicon and germanium leads to the estimates $r_0^\mathrm{Si} \approx 30\,\mathrm{Å}$ and $r_0^\mathrm{Ge} \approx 80\,\mathrm{Å}$, which are at least an order of magnitude larger than the lattice constant. At such distances the assumption of the uniform background and the use of the static dielectric constant are both acceptable, giving an a posteriori justification of our starting point.

The counterpart of the lowest energy in the hydrogen atom,

$$\varepsilon_0 = m_\mathrm{e}\tilde{e}^4/2\hbar^2 = 13.6\,\mathrm{eV}\,, \tag{20.4.4}$$

is now

$$\varepsilon_\mathrm{b} = \frac{m^*(\tilde{e}^2/\epsilon_\mathrm{r})^2}{2\hbar^2} = \frac{m^*}{m_\mathrm{e}}\frac{1}{\epsilon_\mathrm{r}^2}\frac{m_\mathrm{e}\tilde{e}^4}{2\hbar^2} = \frac{m^*}{m_\mathrm{e}}\frac{1}{\epsilon_r^2} \times 13.6\,\mathrm{eV}\,. \tag{20.4.5}$$

Using values that are typical for m^* and ϵ_r in semiconductors, this binding energy is much smaller than the gap: the estimated value is 20 meV for silicon and 5.5 meV for germanium. These energies are measured from the continuum – that is, the bottom of the conduction band in our case. Thus, below but fairly close to the conduction band, a weakly bound state of energy

$$\varepsilon_\mathrm{d} = \varepsilon_\mathrm{c} - \varepsilon_\mathrm{b} \tag{20.4.6}$$

appears. It is called a *donor level.* The density of states containing this new level is shown in Fig. 20.11(a).

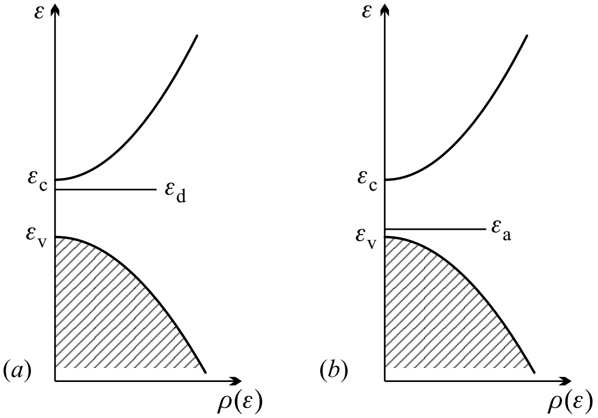


Fig. 20.11. Impurity levels in the gap: (a) donor level; (b) acceptor level

According to the above estimate, the binding energy of the donor level depends only on the properties of the matrix but not on the donor atom that gives rise to it. The distances of donor levels from the bottom of the conduction band (i.e., the binding energies $\varepsilon_\mathrm{b} = \varepsilon_\mathrm{c} - \varepsilon_\mathrm{d}$) are listed in Table 20.8 for various donor atoms in silicon and germanium. These experimental data are in order-of-magnitude agreement with the estimated values, but depend clearly, albeit weakly, on the impurity.

Table 20.8. Binding energies (in meV) of donor levels in silicon and germanium

	P	As	Sb	Bi
Si	45.3	53.7	42.7	70.6
Ge	12.9	14.2	10.3	12.8

The analogy with the hydrogen atom implies the existence of a whole series of bound states at energies $\varepsilon_{\text{b}}/n^2$, which can indeed be observed in optical experiments. However, they do not affect those properties of semiconductors that are the most important in practical applications. As we shall see, owing to the smallness of the binding energy, almost all of the donors are ionized except at low temperatures. Therefore we shall consider a single bound state (one donor level) per donor atom below.

There are as many donor levels in the system as donor atoms in the sample. In the ground state the extra electron provided by the donor does not occupy a conduction-band state but a donor-level state, and there is one electron on each donor level. According to our estimates, the distance of the donor levels from the bottom of the conduction band is much smaller than the gap, therefore these electrons are much easier to excite thermally than those in the valence band. Electrons on donor levels thus constitute an important source of charge carriers. Those semiconductors in which the carriers are dominantly conduction electrons coming from donor levels – and hence there are more electrons in the conduction band than holes in the valence band – are called *n-type semiconductors.*

The case when an acceptor atom of group IIIA is embedded in silicon or germanium can be treated in perfect analogy with the foregoing. The missing electron can be pictured as a positively charged hole of mass m^*, moving around the negatively charged impurity atom in a neutral background of dielectric constant ϵ_{r}. Since the energy changes sign in the electron–hole transformation, we now have a weakly bound hole state, an *acceptor level* above the valence band. Its energy is denoted by ε_{a}. The distances of acceptor levels from the top of the valence band (i.e., the binding energies $\varepsilon_{\text{b}} = \varepsilon_{\text{a}} - \varepsilon_{\text{v}}$) are listed in Table 20.9 for various configurations.

Table 20.9. Binding energies (in meV) of acceptor levels in silicon and germanium

	B	Al	Ga	In	Tl
Si	45.0	68.5	71	155	245
Ge	10.8	11.1	11.3	12.0	13.5

The density of states containing the acceptor level is shown in Fig. 20.11(*b*). In the ground state the acceptor level contains a single electron (that is, a single hole). Since its distance from the top of the valence band is much smaller than the band gap, it is much easier to excite electrons of the valence band to these levels than into the conduction band. The thermal generation of mobile holes in the valence band is therefore relatively easy. Those semiconductors in which the charge carriers are dominantly holes in the valence band generated by the excitation of electrons to acceptor levels are called *p-type semiconductors*.

The situation is very similar when an $\mathrm{A^{III}B^{V}}$ semiconductor is doped by an element of group IVA, silicon or germanium. If the substituted atom is trivalent (A), the group IVA atom behaves as a donor; if it is pentavalent (B), it behaves as an acceptor. The binding energy is of the same order of magnitude as before, as shown in Table 20.10 for some compound semiconductors.

Table 20.10. Binding energies (in meV) of acceptor and donor levels due to silicon or germanium doping in some compound semiconductors

Compound	ε_{b} Acceptor		ε_{b} Donor	
	Si	Ge	Si	Ge
GaP	210	265	85	204
GaAs	34.8	40.4	5.84	5.88

The bound states due to impurities are not always so close to the edge of the conduction or valence band. In silicon or germanium doped with transition metals, the bound states are deep inside the gap. For example, in Si doped with Zn, Fe, and Mn, the bound states are 0.32, 0.39, and 0.5 eV above the top of the valence band. In contrast to the shallow levels, the analogy with the hydrogen atom does not work for the energies of these *deep levels*: the details of the band structure must be taken into account. These levels may play an important role in the nonequilibrium processes in semiconductors, since electrons that become trapped there can no longer be excited thermally.

20.5 Doped Semiconductors at Finite Temperatures

Having determined the energy spectrum of doped semiconductors, let us now examine what happens in these semiconductors at finite temperatures. We shall consider a sample that contains n_{d} donors and n_{a} acceptors per unit volume. Henceforth, when speaking of the number of atoms or electrons, we shall always refer to their density (number per unit volume).

In addition to the electrons necessary for forming the valence bonds, each donor gives an additional electron to the system that can easily become mobile. In contrast, acceptor atoms are one electron short to form all covalent bonds, so each of them binds an otherwise mobile electron into the fourth bond. If the number (per unit volume) of electrons on the outermost shells is n_{e}, which can just fill the valence band before doping, then the number of electrons that are relevant to semiconductor behavior (electrons in the valence and conduction bands as well as on the donor and acceptor levels) is $n_{\mathrm{e}} + n_{\mathrm{d}} - n_{\mathrm{a}}$ after doping. The positive and negative charges of the impurity ions ensure the overall charge neutrality of the system. Below, we shall examine how these electrons populate the levels.

20.5.1 Condition of Charge Neutrality

It was demonstrated in the previous section that a hydrogen-like spectrum appears around each donor atom, nonetheless, for simplicity, it is justified to take into consideration a single level, of energy ε_{d}. In the absence of acceptors, n_{e} of the $n_{\mathrm{e}} + n_{\mathrm{d}}$ electrons fill the valence band in the ground state, and each of the remaining n_{d} electrons occupies a separate donor level, thereby neutralizing the charge of the ion.

At finite temperature some of the electrons are excited into higher-energy states. The probability that an electron hops to a donor level that is already occupied by another electron is much smaller than the probability that it is promoted to the conduction band because the distance of the donor levels from the bottom of the conduction band (i.e., the binding energy) is much smaller than the energy of the strong Coulomb repulsion between two electrons on the same donor level. We shall therefore assume that donor levels cannot accommodate two electrons. Denoting the number (density) of neutral donor atoms (with one electron on the donor level) by n_{d}^0, and that of the positively charged, ionized donors (with no electron on the donor level) by n_{d}^+,

$$n_{\mathrm{d}} = n_{\mathrm{d}}^0 + n_{\mathrm{d}}^+ . \tag{20.5.1}$$

Since all electrons that are detached from the ionized donors end up in the conduction band, as long as the states of the valence band are not excited thermally, the condition for charge neutrality is

$$n = n_{\mathrm{d}}^+ , \tag{20.5.2}$$

where n is the number of electrons in the conduction band. At higher temperatures, valence-band electrons can also be excited into the conduction band, leaving behind an empty state, a hole in the valence band. (According to our earlier assumption, this requires less energy than the excitation of an electron to a donor level that is already occupied by another electron.) Denoting the number of holes in the valence band by p, the condition for charge neutrality is now

$$n = n_{\mathrm{d}}^{+} + p\,. \tag{20.5.3}$$

A somewhat different formulation is used for acceptors. Because of the initial electron deficiency of the acceptor atom, the number of available electrons is $n_{\mathrm{e}} - n_{\mathrm{a}}$. At the same time, the valence band is known to be modified by the potential of the acceptors, and can accommodate $n_{\mathrm{e}} - 2n_{\mathrm{a}}$ electrons per unit volume (when spin is also taken into account). These states are all occupied in the ground state. Each of the remaining n_{a} electrons occupies a separate acceptor level, ensuring the charge neutrality of the acceptor atom.

In the excited state an acceptor level may become either doubly occupied by binding another electron of the opposite spin or empty. The former, negatively ionized state is easily realized because the attractive interaction between the acceptor and the electrons can partially overcome the Coulomb repulsion between electrons. Owing to the proximity of the acceptor levels to the top of the valence band, valence-band electrons can be thermally excited to singly occupied, neutral acceptor levels, leaving behind a hole in the valence band. The energy of a positively ionized, empty acceptor level is much higher, so this state can be neglected in calculations. We shall therefore assume that each acceptor level is occupied by at least one electron. Denoting the number of neutral acceptor atoms (with one electron on the acceptor level) by n_{a}^{0}, and that of the negatively charged, ionized acceptor atoms (with two electrons on the acceptor level) by n_{a}^{-},

$$n_{\mathrm{a}} = n_{\mathrm{a}}^{0} + n_{\mathrm{a}}^{-}\,. \tag{20.5.4}$$

Acceptors are ionized by capturing electrons from the valence band. Since at low temperatures valence-band electrons can be excited only to acceptor levels, the number of holes p in the valence band must be equal to n_{a}^{-}:

$$n_{\mathrm{a}}^{-} = p\,. \tag{20.5.5}$$

Excitation into the conduction band requires somewhat higher energies. When such excitations occur, some of the electrons leaving the valence band ionize acceptors, while others populate the conduction band, so the equation for charge neutrality reads

$$n + n_{\mathrm{a}}^{-} = p\,. \tag{20.5.6}$$

When donors and acceptors are simultaneously present, then out of the $n_{\mathrm{e}} + n_{\mathrm{d}} - n_{\mathrm{a}}$ electrons $n_{\mathrm{e}} - 2n_{\mathrm{a}}$ fill the valence band in the ground state. The number of remaining electrons is $n_{\mathrm{d}} + n_{\mathrm{a}}$, and acceptors can accommodate $2n_{\mathrm{a}}$. Therefore, when $n_{\mathrm{d}} < n_{\mathrm{a}}$, n_{d} acceptor levels will be doubly and $n_{\mathrm{a}} - n_{\mathrm{d}}$ singly occupied. Acceptors are negatively ionized in the former and neutral in the latter case. Donor levels remain empty, and thus donors are positively charged (ionized).

$$n_{\mathrm{a}}^{-} = n_{\mathrm{d}}, \qquad n_{\mathrm{a}}^{0} = n_{\mathrm{a}} - n_{\mathrm{d}}, \qquad n_{\mathrm{d}}^{+} = n_{\mathrm{d}}, \qquad n_{\mathrm{d}}^{0} = 0\,. \tag{20.5.7}$$

The condition for overall neutrality requires that the number of negatively charged acceptors be equal to the number of positively charged donors:

$$n_\mathrm{a}^- = n_\mathrm{d}^+ . \tag{20.5.8}$$

At finite temperatures conduction-band states can also be excited, and so holes appear in the valence band. Charge neutrality now requires that the number of electrons in the conduction band plus negative (ionized) acceptors be equal to the number of positive (ionized) donors and holes in the valence band:

$$n + n_\mathrm{a}^- = n_\mathrm{d}^+ + p . \tag{20.5.9}$$

The same equation applies when $n_\mathrm{d} > n_\mathrm{a}$. This formula plays a central role in the description of the behavior of semiconductors. Before turning to an application, the evaluation of the chemical potential, we shall examine each term separately.

20.5.2 Thermal Population of Donor and Acceptor Levels

The formulas (20.3.17) and (20.3.20) obtained for the number of electrons in the conduction band and the number of holes in the valence band may be used in this situation, too, as their derivation was based solely on the thermal occupation of one-particle states. The suitable choice of the chemical potential will make the only difference.

However, the Fermi–Dirac distribution function valid for the occupation of independent states cannot be automatically applied to the thermal population of donor and acceptor levels; the relevant formulas have to be derived from the general principles of statistical mechanics. When electron spins are also taken into account, four states are possible. The donor level can be either empty, singly occupied (with either spin orientation, leading to two states), or doubly occupied (by electrons of opposite spin). The Fermi–Dirac distribution function would apply to donor levels if the energy of the doubly occupied state were twice the energy of the singly occupied state. However, on account of the Coulomb repulsion between the two electrons, it is much higher than that, therefore it is energetically more favorable for the second electron to occupy a state in the conduction band. Donor atoms can therefore be in one of three possible states: the donor level is either empty, or singly occupied by a spin-up or spin-down electron. In a grand canonical ensemble, the probabilities of these states are determined by the Boltzmann factors

$$\mathrm{e}^{-\beta\varepsilon_0}/Z \qquad \text{and} \qquad \mathrm{e}^{-\beta(\varepsilon_0+\varepsilon_\mathrm{d}-\mu)}/Z . \tag{20.5.10}$$

The distribution function f_d of donors – that is, the mean number of electrons per donor atom – is

$$\begin{aligned} f_\mathrm{d}(\varepsilon_\mathrm{d}) &= \frac{\sum n_i \mathrm{e}^{-\beta(E_i-\mu n_i)}}{\sum \mathrm{e}^{-\beta(E_i-\mu n_i)}} = \frac{2\mathrm{e}^{-(\varepsilon_0+\varepsilon_\mathrm{d}-\mu)/k_\mathrm{B}T}}{\mathrm{e}^{-\varepsilon_0/k_\mathrm{B}T} + 2\mathrm{e}^{-(\varepsilon_0+\varepsilon_\mathrm{d}-\mu)/k_\mathrm{B}T}} \\ &= \frac{1}{\frac{1}{2}\mathrm{e}^{(\varepsilon_\mathrm{d}-\mu)/k_\mathrm{B}T}+1} . \end{aligned} \tag{20.5.11}$$

This formula also gives the probability that the donor atom is neutral (i.e., it is occupied by one electron). If the number of donor atoms is n_d, then the number of neutral ones is

$$n_\mathrm{d}^0 = n_\mathrm{d} f_\mathrm{d}(\varepsilon_\mathrm{d}) = \frac{n_\mathrm{d}}{\frac{1}{2}\mathrm{e}^{(\varepsilon_\mathrm{d}-\mu)/k_\mathrm{B}T}+1}\,. \tag{20.5.12}$$

(20.5.1) implies that the average number of ionized donors is

$$n_\mathrm{d}^+ = n_\mathrm{d} - \frac{n_\mathrm{d}}{\frac{1}{2}\mathrm{e}^{(\varepsilon_\mathrm{d}-\mu)/k_\mathrm{B}T}+1} = \frac{n_\mathrm{d}}{1+2\mathrm{e}^{(\mu-\varepsilon_\mathrm{d})/k_\mathrm{B}T}}\,. \tag{20.5.13}$$

The situation is, in a sense, the opposite for acceptors. In the neutral ground state the acceptor level is occupied by one electron. Owing to the two spin orientations, the ground state is doubly degenerate. The excited state, in which the acceptor level has two electrons of opposite spin, is not degenerate. The distribution function $f_\mathrm{a}(\varepsilon_\mathrm{a})$ of acceptors is defined in such a way that it gives the expected number of electrons on the acceptor level:

$$\begin{aligned} f_\mathrm{a}(\varepsilon_\mathrm{a}) &= \frac{2\mathrm{e}^{-(\varepsilon_0-\mu)/k_\mathrm{B}T}+2\mathrm{e}^{-(\varepsilon_0+\varepsilon_\mathrm{a}-2\mu)/k_\mathrm{B}T}}{2\mathrm{e}^{-(\varepsilon_0-\mu)/k_\mathrm{B}T}+\mathrm{e}^{-(\varepsilon_0+\varepsilon_\mathrm{a}-2\mu)/k_\mathrm{B}T}} \\ &= \frac{1+\mathrm{e}^{(\mu-\varepsilon_\mathrm{a})/k_\mathrm{B}T}}{1+\frac{1}{2}\mathrm{e}^{(\mu-\varepsilon_\mathrm{a})/k_\mathrm{B}T}}\,. \end{aligned} \tag{20.5.14}$$

Since there is at least one electron on the acceptor level (one in the neutral state and two in the ionized state), the probability that the acceptor is ionized is

$$f_\mathrm{a}(\varepsilon_\mathrm{a}) - 1 = \frac{1+\mathrm{e}^{(\mu-\varepsilon_\mathrm{a})/k_\mathrm{B}T}}{1+\frac{1}{2}\mathrm{e}^{(\mu-\varepsilon_\mathrm{a})/k_\mathrm{B}T}} - 1 = \frac{1}{1+2\mathrm{e}^{(\varepsilon_\mathrm{a}-\mu)/k_\mathrm{B}T}}\,, \tag{20.5.15}$$

and the number of ionized acceptors is

$$n_\mathrm{a}^- = \frac{n_\mathrm{a}}{1+2\mathrm{e}^{(\varepsilon_\mathrm{a}-\mu)/k_\mathrm{B}T}}\,. \tag{20.5.16}$$

This expression is perfectly analogous to the one specifying the number of ionized donors, but energies are now measured in the opposite direction, downward from the chemical potential. In this upside-down picture electrons appear as holes, and vice versa.

20.5.3 Number of Carriers in Doped Semiconductors

The density of charge carriers – that is, the number of electrons in the conduction band plus holes in the valence band per unit volume – has a very strong influence on the properties of doped semiconductors. The carrier concentrations can be calculated by exploiting (20.3.17) and (20.3.20). However, it is often more convenient to use the formulas implied by (20.3.32) and (20.3.33),

$$n = n_{\mathrm{i}}\,\mathrm{e}^{(\mu-\mu_{\mathrm{i}})/k_{\mathrm{B}}T}, \qquad p = n_{\mathrm{i}}\,\mathrm{e}^{(\mu_{\mathrm{i}}-\mu)/k_{\mathrm{B}}T}\,. \tag{20.5.17}$$

These relations clearly show that the effects of doping appears only through the chemical potential. An upward shift of the chemical potential due to doping gives rise to an exponential increase in the number of carriers in the conduction band, and a similar decrease in the valence band. Variations in the opposite sense are observed when the chemical potential decreases due to doping.

In the foregoing we have derived all the formulas that are necessary for determining the chemical potential. Substituting (20.3.17), (20.3.20), (20.5.13), and (20.5.16) into the charge-neutrality condition (20.5.9), we have

$$\begin{aligned} N_{\mathrm{c}}(T)\mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T} &+ \frac{n_{\mathrm{a}}}{1+2\mathrm{e}^{(\varepsilon_{\mathrm{a}}-\mu)/k_{\mathrm{B}}T}} \\ &= \frac{n_{\mathrm{d}}}{1+2\mathrm{e}^{(\mu-\varepsilon_{\mathrm{d}})/k_{\mathrm{B}}T}} + P_{\mathrm{v}}(T)\mathrm{e}^{-(\mu-\varepsilon_{\mathrm{v}})/k_{\mathrm{B}}T}. \end{aligned} \tag{20.5.18}$$

By solving this equation, we shall now calculate the chemical potential and carrier concentration in different situations.

n-Type Semiconductors Containing Only Donors

We shall first deal with the case of semiconductors that contain only donor atoms. The neutrality condition (20.5.9) is then reduced to (20.5.3). At low temperatures electrons are excited into the conduction band principally from the donor levels, and hardly from the valence band, thus $p \ll n$. Neglecting p in the neutrality condition, and writing out the two other terms in full,

$$N_{\mathrm{c}}(T)\mathrm{e}^{-(\varepsilon_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T} = \frac{n_{\mathrm{d}}}{1+2\mathrm{e}^{(\mu-\varepsilon_{\mathrm{d}})/k_{\mathrm{B}}T}}\,. \tag{20.5.19}$$

This can be rewritten as a quadratic equation for $x = \exp(\mu/k_{\mathrm{B}}T)$. Its solution gives

$$\mu = k_{\mathrm{B}}T \ln\left\{\frac{1}{4}\mathrm{e}^{\varepsilon_{\mathrm{d}}/k_{\mathrm{B}}T}\left[\sqrt{1+\frac{8n_{\mathrm{d}}}{N_{\mathrm{c}}(T)}\mathrm{e}^{(\varepsilon_{\mathrm{c}}-\varepsilon_{\mathrm{d}})/k_{\mathrm{B}}T}}-1\right]\right\}. \tag{20.5.20}$$

At very low temperatures, where

$$\frac{8n_{\mathrm{d}}}{N_{\mathrm{c}}(T)}\mathrm{e}^{(\varepsilon_{\mathrm{c}}-\varepsilon_{\mathrm{d}})/k_{\mathrm{B}}T} \gg 1\,, \tag{20.5.21}$$

the chemical potential is

$$\begin{aligned} \mu &= k_{\mathrm{B}}T \ln\left[\frac{1}{4}\mathrm{e}^{\varepsilon_{\mathrm{d}}/k_{\mathrm{B}}T}\sqrt{\frac{8n_{\mathrm{d}}}{N_{\mathrm{c}}(T)}}\mathrm{e}^{(\varepsilon_{\mathrm{c}}-\varepsilon_{\mathrm{d}})/2k_{\mathrm{B}}T}\right] \\ &= \tfrac{1}{2}(\varepsilon_{\mathrm{c}}+\varepsilon_{\mathrm{d}}) + \tfrac{1}{2}k_{\mathrm{B}}T \ln\frac{n_{\mathrm{d}}}{2N_{\mathrm{c}}(T)}\,. \end{aligned} \tag{20.5.22}$$

As $N_\mathrm{c}(T)$ is proportional to the 3/2th power of the temperature, the second term vanishes in the $T \to 0$ limit. Bearing resemblance on intrinsic semiconductors, the chemical potential is located halfway between the bottom of the completely empty conduction band and the completely filled donor level.

Thus, at very low temperatures, where (20.5.21) is satisfied, the number of conduction electrons is obtained by substituting the above formula of the chemical potential into (20.3.17):

$$n = N_\mathrm{c}(T)\mathrm{e}^{-(\varepsilon_\mathrm{c}-\mu)/k_\mathrm{B}T} = \sqrt{\frac{N_\mathrm{c}(T)n_\mathrm{d}}{2}}\mathrm{e}^{-(\varepsilon_\mathrm{c}-\varepsilon_\mathrm{d})/2k_\mathrm{B}T}. \tag{20.5.23}$$

The temperature dependence is dominated by the exponential factor. Since the density of carriers is exponentially small, this temperature range is called the *freeze-out range* or *partial-ionization range.*

The chemical potential shows an initial increase with increasing temperature. Unless the number of donors is sufficiently large, this increase stops before the chemical potential could reach the bottom of the conduction band. If this condition is not met, then classical statistics can no longer be used for determining the number of electrons in the conduction band, and quantum statistics has to be applied instead. Such semiconductors are called *degenerate.*

Slightly above the freeze-out range, where $N_\mathrm{c}(T) \gg 8n_\mathrm{d}$ and thus

$$\frac{8n_\mathrm{d}}{N_\mathrm{c}(T)}\mathrm{e}^{(\varepsilon_\mathrm{c}-\varepsilon_\mathrm{d})/k_\mathrm{B}T} \ll 1\,, \tag{20.5.24}$$

the chemical potential for nondegenerate semiconductors can be obtained by a series expansion of (20.5.20) leading to

$$\begin{aligned}\mu &= k_\mathrm{B}T\ln\left\{\frac{1}{4}\mathrm{e}^{\varepsilon_\mathrm{d}/k_\mathrm{B}T}\left[1+\frac{4n_\mathrm{d}}{N_\mathrm{c}(T)}\mathrm{e}^{(\varepsilon_\mathrm{c}-\varepsilon_\mathrm{d})/k_\mathrm{B}T}-1\right]\right\}\\ &= k_\mathrm{B}T\ln\left(\frac{n_\mathrm{d}}{N_\mathrm{c}(T)}\mathrm{e}^{\varepsilon_\mathrm{c}/k_\mathrm{B}T}\right) = \varepsilon_\mathrm{c} + k_\mathrm{B}T\ln\frac{n_\mathrm{d}}{N_\mathrm{c}(T)}\,,\end{aligned} \tag{20.5.25}$$

while the number of electrons in the conduction band is

$$n = N_\mathrm{c}(T)\mathrm{e}^{-(\varepsilon_\mathrm{c}-\mu)/k_\mathrm{B}T} = N_\mathrm{c}(T)\mathrm{e}^{\ln n_\mathrm{d}/N_\mathrm{c}(T)} = n_\mathrm{d}\,. \tag{20.5.26}$$

Because of the condition $N_\mathrm{c}(T) \gg 8n_\mathrm{d}$, the chemical potential decreases with increasing temperature in this range. By comparing the equation $n = n_\mathrm{d}$ with (20.5.3), and exploiting that $p \approx 0$, almost all donor impurities are seen to be ionized: their extra electrons are in the conduction band now. This medium-temperature range is often referred to as the *saturation range* or *extrinsic range*, since the carrier concentration is determined by the number of impurities.

More careful calculations show that a few electrons of the valence band are already excited to the conduction band, nonetheless the number of holes left behind in the valence band is still low. The holes in the valence band are therefore called *minority carriers* In n-type semiconductors electrons in the conduction band are the *majority carriers.*

To evaluate the number of minority carriers, the previous calculation has to be refined, as p can no longer be neglected. Assuming that all donors are ionized, we have

$$n = p + n_\mathrm{d} \,. \tag{20.5.27}$$

Another relation is obtained from the law of mass action. As mentioned in connection with pure semiconductors, (20.3.27) relating p and n is valid in the presence of impurities, too. The physically sensible solution of the two equations is

$$\begin{aligned} n &= \tfrac{1}{2} n_\mathrm{d} \left[\sqrt{1 + (2n_\mathrm{i}/n_\mathrm{d})^2} + 1 \right] , \\ p &= \tfrac{1}{2} n_\mathrm{d} \left[\sqrt{1 + (2n_\mathrm{i}/n_\mathrm{d})^2} - 1 \right] , \end{aligned} \tag{20.5.28}$$

where n_i is determined by (20.3.24). In the temperature range where the carriers coming from donors dominate, $n_\mathrm{i} \ll n_\mathrm{d}$. As mentioned above, the leading order of the series expansion gives $n = n_\mathrm{d}$ for the number of carriers. Keeping second-order terms, too, we have

$$n = n_\mathrm{d} + n_\mathrm{i}^2/n_\mathrm{d} \,, \qquad p = n_\mathrm{i}^2/n_\mathrm{d} \,. \tag{20.5.29}$$

Figure 20.12 shows the thermal population of electron and hole states in an n-type semiconductor. Comparison with Fig. 20.9 shows that the number of electrons in the conduction band has increased, while the number of holes in the valence band has decreased.

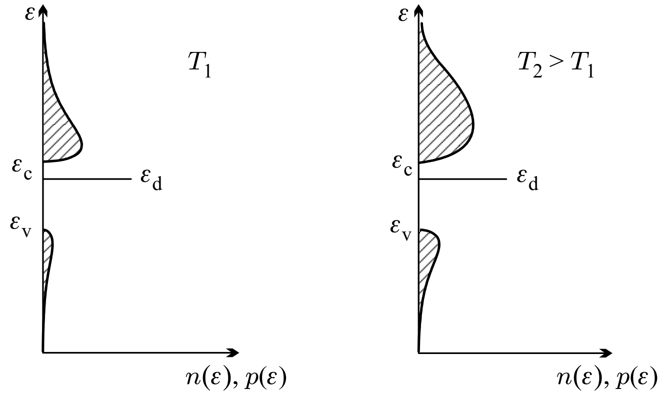


Fig. 20.12. The thermal distribution of electron and hole states in an n-type semiconductor at two different temperatures

A more accurate formula is obtained for the chemical potential when μ is expressed from (20.3.17) and (20.5.28) is used for n:

$$\mu = \varepsilon_\mathrm{c} + k_\mathrm{B} T \ln \left[\frac{n_\mathrm{d}}{2N_\mathrm{c}(T)} \left(1 + \sqrt{1 + \frac{4n_\mathrm{i}^2}{n_\mathrm{d}^2}} \right) \right] . \qquad (20.5.30)$$

At even higher temperatures carriers excited thermally from the valence band may start to dominate: $n_\mathrm{i} \gg n_\mathrm{d}$. Then (20.5.28) implies

$$n \approx n_\mathrm{i} + \tfrac{1}{2} n_\mathrm{d} \, , \qquad p \approx n_\mathrm{i} - \tfrac{1}{2} n_\mathrm{d} \, . \qquad (20.5.31)$$

Taking, once again, the chemical potential from (20.3.17) and making use of the previous result,

$$\mu = \varepsilon_\mathrm{c} + k_\mathrm{B} T \ln \frac{n}{N_\mathrm{c}(T)} \approx \varepsilon_\mathrm{c} + k_\mathrm{B} T \ln \frac{n_\mathrm{i}}{N_\mathrm{c}(T)} . \qquad (20.5.32)$$

Using (20.3.24) for n_i, we find

$$\mu = \tfrac{1}{2}(\varepsilon_\mathrm{c} + \varepsilon_\mathrm{v}) + \tfrac{1}{2} k_\mathrm{B} T \ln \frac{P_\mathrm{v}(T)}{N_\mathrm{c}(T)} \, , \qquad (20.5.33)$$

which is the same result as for intrinsic semiconductors. This shows that for semiconductors doped by donors there is a range at sufficiently high temperatures where the contribution of donors to the carrier concentration is negligible. Since $n \approx p$, the overwhelming majority of the electrons in the conduction band come from the valence band. Once again, the number of electrons increases exponentially with temperature, however the exponent is not $(\varepsilon_\mathrm{c} - \varepsilon_\mathrm{d})/2$, as at low temperatures, but $(\varepsilon_\mathrm{c} - \varepsilon_\mathrm{v})/2 = \varepsilon_\mathrm{g}/2$. The semiconductor behaves as if it were intrinsic. This high-temperature range is known as the *intrinsic range*. The threshold temperature of this range increases with the concentration of donor atoms. The temperature dependence of the chemical potential and the number of electrons in the conduction band are schematically summarized in Fig. 20.13.

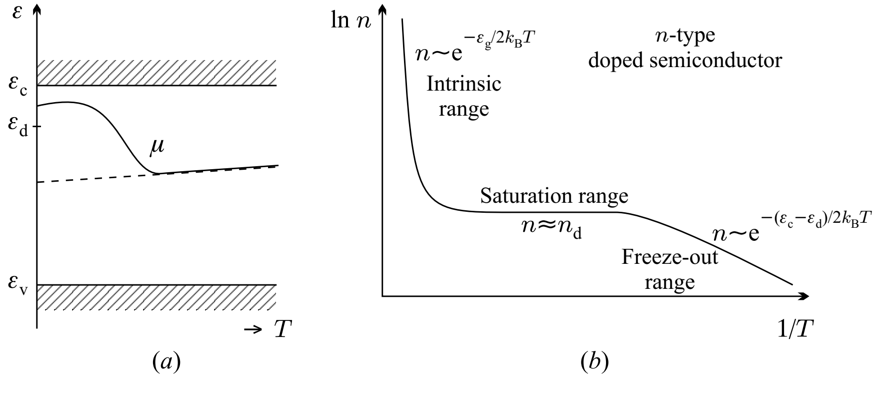


Fig. 20.13. (a) Temperature dependence of the chemical potential in the presence of donor atoms. The dashed line shows the variations of the chemical potential in an intrinsic semiconductor. (b) Temperature dependence of the number of electrons in the conduction band

p-Type Semiconductors Containing Only Acceptors

Analogous calculations can be performed for p-type semiconductors that contain only acceptors. Starting with the equation of charge neutrality,

$$n + n_{\mathrm{a}}^{-} = p\,, \tag{20.5.34}$$

the results can be written down immediately, by exploiting the electron–hole symmetry i.e., the correspondence between electron and hole states in n- and p-type semiconductors. These considerations lead to Fig. 20.14, in which the chemical potential and the number of holes in the valence band are plotted against temperature.

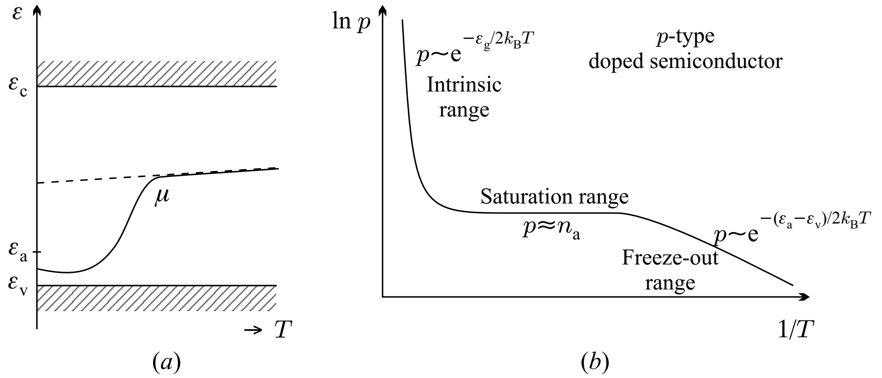


Fig. 20.14. (a) Temperature dependence of the chemical potential in a semiconductor doped by acceptor atoms. (b) Temperature dependence of the number of holes in the valence band

At low temperatures, where no electrons are excited to the conduction band, the chemical potential is usually written as

$$\mu = -k_{\mathrm{B}}T \ln\left\{\frac{1}{4}\mathrm{e}^{-\varepsilon_{\mathrm{a}}/k_{\mathrm{B}}T}\left[\sqrt{1+\frac{8n_{\mathrm{a}}}{P_{\mathrm{v}}(T)}\mathrm{e}^{-(\varepsilon_{\mathrm{v}}-\varepsilon_{\mathrm{a}})/k_{\mathrm{B}}T}}-1\right]\right\}, \tag{20.5.35}$$

in analogy with (20.5.20). In the $T \to 0$ limit, where $P_{\mathrm{v}}(T) \ll n_{\mathrm{a}}$, the chemical potential is halfway between the top of the valence band and the acceptor level. At low temperatures the dominant process is the excitation of electrons from the valence band to the acceptor level – that is, viewed from another perspective, the dominant carriers are holes excited into the valence band from the acceptor level. Here, too, this is known as the freeze-out (or partial-ionization) range.

In the intermediate temperature range each acceptor is ionized, $n_\mathrm{a}^- \approx n_\mathrm{a}$, but $n \approx 0$, and thus $p \approx n_\mathrm{a}$. This is saturation or extrinsic range. More accurate calculations yield

$$\begin{aligned} p &= \tfrac{1}{2} n_\mathrm{a} \left[\sqrt{1 + (2n_\mathrm{i}/n_\mathrm{a})^2} + 1 \right] \approx n_\mathrm{a} + n_\mathrm{i}^2/n_\mathrm{a} \,, \\ n &= \tfrac{1}{2} n_\mathrm{a} \left[\sqrt{1 + (2n_\mathrm{i}/n_\mathrm{a})^2} - 1 \right] \approx n_\mathrm{i}^2/n_\mathrm{a} \,. \end{aligned} \tag{20.5.36}$$

At even higher temperatures the number of electrons excited into the conduction band may exceed the number of impurities. Hence, above some characteristic temperature that depends on the acceptor concentration, the semiconductor behaves as if it did not contain any impurities. This is the intrinsic conduction range. Figure 20.15 shows the temperature dependence of the chemical potential for various donor and acceptor concentrations.

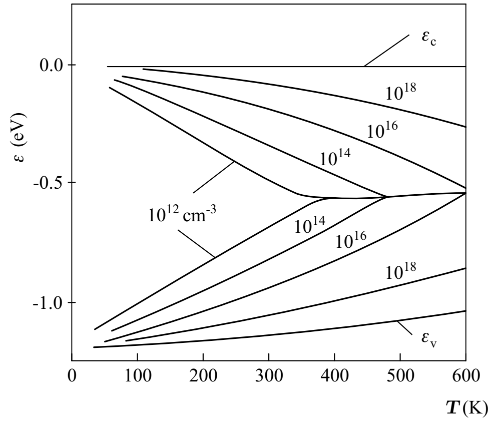


Fig. 20.15. Temperature dependence of the chemical potential in silicon for various donor and acceptor concentrations

Semiconductors Containing Donors and Acceptors

In principle, the number of charge carriers and the chemical potential may also be determined from (20.5.9) for samples containing both donors and acceptors. Two cases must be considered: $n_\mathrm{d} > n_\mathrm{a}$ and $n_\mathrm{d} < n_\mathrm{a}$. The sample is an *n-type* semiconductor in the first case, and a *p-type* semiconductor in the second.

When $n_\mathrm{d} > n_\mathrm{a}$, the valence band is completely filled in the ground state and each acceptor level is doubly occupied, hence all acceptors are ionized. The remaining $n_\mathrm{d} - n_\mathrm{a}$ electrons occupy donor levels, but these levels are only partially filled. The chemical potential is obviously the energy of the donor level: $\mu = \varepsilon_\mathrm{d}$.

At finite temperatures the full equation (20.5.9) has to be considered. The introduction of the variable $x = \exp(\mu/k_\mathrm{B}T)$ leads to a quartic equation. However, the equation is considerably simplified in four characteristic temperature ranges, and the solution is straightforward. We shall now consider each of them.

Once again, in the low-temperature range, where the chemical potential remains close to ε_d, practically only the electrons occupying donor levels are excited into the conduction band, i.e., $p \approx 0$ and $n_\mathrm{a}^- \approx n_\mathrm{a}$. The neutrality condition is then reduced to

$$N_\mathrm{c}(T)\mathrm{e}^{-(\varepsilon_\mathrm{c}-\mu)/k_\mathrm{B}T} + n_\mathrm{a} = \frac{n_\mathrm{d}}{1+2\mathrm{e}^{(\mu-\varepsilon_\mathrm{d})/k_\mathrm{B}T}}\,. \tag{20.5.37}$$

This formula is further simplified at very low temperatures where only a very small number of electrons are excited into the conduction band. By neglecting the first term on the left-hand side compared to n_a,

$$n_\mathrm{a} = \frac{n_\mathrm{d}}{1+2\mathrm{e}^{(\mu-\varepsilon_\mathrm{d})/k_\mathrm{B}T}}\,. \tag{20.5.38}$$

Rearrangement of the terms gives

$$\mu = \varepsilon_\mathrm{d} + k_\mathrm{B}T \ln\frac{n_\mathrm{d}-n_\mathrm{a}}{2n_\mathrm{a}}\,. \tag{20.5.39}$$

Substituting this into (20.3.17), the number of electrons in the conduction band is

$$n = \frac{N_\mathrm{c}(T)(n_\mathrm{d}-n_\mathrm{a})}{2n_\mathrm{a}}\mathrm{e}^{-(\varepsilon_\mathrm{c}-\varepsilon_\mathrm{d})/k_\mathrm{B}T}. \tag{20.5.40}$$

At slightly higher but still low temperatures – where the thermal energy $k_\mathrm{B}T$ is still smaller than the distance of the chemical potential from the donor levels ($k_\mathrm{B}T \ll |\varepsilon_\mathrm{d}-\mu|$) – only a small proportion of the donors become ionized and practically none of the electrons in the valence band are excited into the conduction band. If $n_\mathrm{d} \gg n_\mathrm{a}$, there is a temperature range where the number of electrons promoted to the conduction band from the donor levels exceeds the number of ionized acceptors, therefore n_a can be neglected in (20.5.37). Exploiting the condition $\varepsilon_\mathrm{d}-\mu \gg k_\mathrm{B}T$ in the distribution function for donors,

$$N_\mathrm{c}(T)\mathrm{e}^{-(\varepsilon_\mathrm{c}-\mu)/k_\mathrm{B}T} = \tfrac{1}{2}n_\mathrm{d}\mathrm{e}^{-(\mu-\varepsilon_\mathrm{d})/k_\mathrm{B}T}, \tag{20.5.41}$$

and so

$$\mu = \tfrac{1}{2}(\varepsilon_\mathrm{c}+\varepsilon_\mathrm{d}) + \tfrac{1}{2}k_\mathrm{B}T\ln\frac{n_\mathrm{d}}{2N_\mathrm{c}(T)}. \tag{20.5.42}$$

Using this form of the chemical potential to evaluate the number of charge carriers in the conduction band:

$$n = \sqrt{\frac{N_\mathrm{c}(T)n_\mathrm{d}}{2}}\mathrm{e}^{-(\varepsilon_\mathrm{c}-\varepsilon_\mathrm{d})/2k_\mathrm{B}T}. \tag{20.5.43}$$

This is the same as (20.5.23), the low-temperature formula obtained for semiconductors containing only donor atoms. Our previous considerations show that if acceptors are also present in a small quantity, then they play a role only at very low temperatures. They modify the exponent in the exponential temperature dependence of the number of charge carriers to $(\varepsilon_{\mathrm{c}} - \varepsilon_{\mathrm{d}})/k_{\mathrm{B}}T$. Only at slightly higher temperatures does the value characteristic of the freeze-out range of n-type semiconductors, $(\varepsilon_{\mathrm{c}} - \varepsilon_{\mathrm{d}})/2k_{\mathrm{B}}T$, appear in the exponent.

At even higher temperatures the samples that contain a small number of acceptors in addition to donors behave like a material that contains only donors – however n_{d} is replaced by $n_{\mathrm{d}} - n_{\mathrm{a}}$ in every previously derived formula. There is a saturation range in this case, too, where – in addition to the already ionized acceptors – each donor becomes ionized: $n_{\mathrm{d}}^{+} \approx n_{\mathrm{d}}$. The difference between the number of electrons in the conduction band and the number of holes in the valence band is then

$$\Delta n = n - p = n_{\mathrm{d}} - n_{\mathrm{a}} \,. \tag{20.5.44}$$

The law of mass action is valid here, too, and takes the form

$$n\,p = n_{\mathrm{i}}^2 \,. \tag{20.5.45}$$

The two equations yield

$$\begin{aligned} n &= \tfrac{1}{2}\left[(n_{\mathrm{d}} - n_{\mathrm{a}})^2 + 4n_{\mathrm{i}}^2\right]^{1/2} + \tfrac{1}{2}(n_{\mathrm{d}} - n_{\mathrm{a}}) \,, \\ p &= \tfrac{1}{2}\left[(n_{\mathrm{d}} - n_{\mathrm{a}})^2 + 4n_{\mathrm{i}}^2\right]^{1/2} - \tfrac{1}{2}(n_{\mathrm{d}} - n_{\mathrm{a}}) \,. \end{aligned} \tag{20.5.46}$$

In the saturation range the carriers supplied by the impurities dominate, $\Delta n \gg n_{\mathrm{i}}$, thus in leading order

$$n \approx n_{\mathrm{d}} - n_{\mathrm{a}} \,, \qquad p \approx \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}} - n_{\mathrm{a}}} \,. \tag{20.5.47}$$

In this intermediate temperature range the number of majority carriers is independent of the temperature. Even though the formula for the number of carriers contains the difference $n_{\mathrm{d}} - n_{\mathrm{a}}$ – that is, donors are partially compensated by acceptors –, the overall behavior of the semiconductor depends on the total number of donors and acceptors $n_{\mathrm{d}} + n_{\mathrm{a}}$, too, as they also act as scattering and recombination centers.

Finally, at even higher temperatures the intrinsic conduction range is reached. Figure 20.16 shows the temperature dependence of the chemical potential and the number of electrons excited into the conduction band in the four discussed temperature ranges.

In the opposite case, when acceptors are in excess of donors ($n_{\mathrm{a}} > n_{\mathrm{d}}$), the chemical potential is just ε_{a} in the ground state. Apart from the range of very low temperatures the same behavior is obtained as for p-type semiconductors containing only acceptors, with the single difference that now

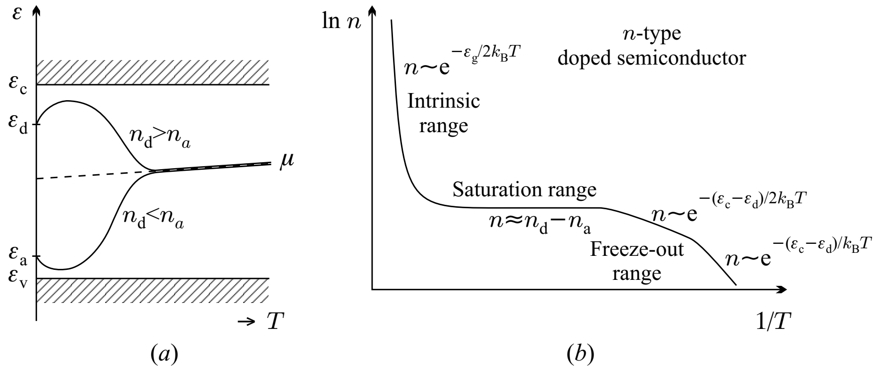


Fig. 20.16. (*a*) Temperature dependence of the chemical potential for n-type ($n_d > n_a$) and p-type ($n_a > n_d$) semiconductors. The dashed line shows the temperature dependence for a pure semiconductor. (*b*) Temperature dependence of the number of electrons excited into the conduction band for an n-type semiconductor

$$n \approx \frac{n_i^2}{n_a - n_d}, \qquad p \approx n_a - n_d \tag{20.5.48}$$

in the saturation range.

The above-discussed temperature dependence of the number of charge carriers plays an important role in the transport properties of homogeneous semiconductors. The temperature dependence of the conductivity is dominated by the variation of the number of charge carriers in the low- and high-temperature regions. The conductivity increases sharply with temperature due to the exponential increase of the carrier density. In the medium-temperature saturation range, where the carrier concentration is temperature independent, the conductivity decreases with temperature.

Further Reading

1. M. Balkanski and R. E. Wallis, *Semiconductor Physics and Applications*, Oxford University Press, Oxford (2000).
2. K. W. Böer, *Survey of Semiconductor Physics*, Vol. 1. *Electrons and Other Particles in Bulk Semiconductors*, Van Nostrand Reinhold, New York (1990).
3. *Handbook on Semiconductors*, Completely Revised and Enlarged Edition, Series Editor T. S. Moss, Volume 1. *Basic Properties of Semiconductors*, Volume Editor P. T. Landsberg, North-Holland, Amsterdam (1992).
4. B. Sapoval and C. Hermann, *Physics of Semiconductors*, 2nd printing, Springer-Verlag, Berlin (2003).
5. K. Seeger, *Semiconductor Physics: An Introduction*, 9th Edition, Springer-Verlag, Berlin (2004).

6. P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors, Physics and Materials Properties*, Springer-Verlag, Berlin (1996).
7. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice Hall, Inc., Englewood Cliffs (1989).

21

Semiclassical Dynamics of Electrons

Having specified the stationary electron states, we can now turn to another issue that is even more interesting in several respects: What happens to the electrons in solids when the sample is placed in an applied electromagnetic field? This question does not arise only in the context of conduction (or transport) phenomena but also when the motion of electrons in a constant magnetic field is studied or when one tries to infer the processes occurring inside the material from the optical properties.

In the presence of an electromagnetic field the states of Bloch electrons cannot be calculated exactly. High-frequency fields can induce interband transitions, and similar transitions can also occur in sufficiently strong uniform electric fields, as we shall see at the end of this chapter. To describe the processes in applied fields, we shall use the so-called semiclassical approximation. In the present chapter we shall introduce this method, and present its limitations as well.

As we shall see, the wave vector of electrons in a uniform magnetic field moves along $\boldsymbol{k}$-space orbits of constant energy in a plane perpendicular to the magnetic field. Since electron states outside the narrow region of width $k_{\mathrm{B}}T$ around the Fermi energy are fully occupied or completely empty, the motion in a magnetic field provides information about the properties of electrons near the Fermi surface, which play a primary role in the behavior of metals. Therefore we shall also briefly discuss some experimental methods that use the motion in a magnetic field to determine the main features of the Fermi surface. Other consequences of semiclassical dynamics, for example the transport properties in solids, will be studied in more detail in Chapter 24.

21.1 Basics of Semiclassical Dynamics

When studying the motion of free electrons in a uniform electric field in Chapter 16, we demonstrated that if the electric field is specified in terms of a scalar potential then the behavior of the electronic wavefunctions under translations can still be characterized by a wave vector $\boldsymbol{k}$, however $\boldsymbol{k}$ will be time dependent. The time dependence is given by (16.3.11). Rewritten in differential form,

$$\hbar\dot{\boldsymbol{k}}(t) = -e\boldsymbol{E}\,. \tag{21.1.1}$$

Formally, this is the same as the classical equation of motion, since for free electrons $\hbar\boldsymbol{k}$ is the momentum and $-e\boldsymbol{E}$ is the accelerating force of the electric field. We shall examine how this is modified for Bloch electrons, as the crystal momentum, $\hbar\boldsymbol{k}$, which is related to the invariance under discrete translations, is not the same as the momentum, and the motion of electrons is affected not only by the applied field but also by the periodic potential. As we shall see, in spite of these differences, (21.1.1) is valid for Bloch electrons, too, in the semiclassical approximation.

In this approach our primary concern is not the eigenvalues and eigenfunctions of the Hamiltonian

$$\mathcal{H} = \frac{1}{2m_{\mathrm{e}}}\left[\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e\boldsymbol{A}(\boldsymbol{r})\right]^2 + U(\boldsymbol{r}) - e\varphi(\boldsymbol{r}) \tag{21.1.2}$$

but the motion of a wave packet that can be represented as the superposition of Bloch states that belong in the same band:

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r},t) = \sum_{\boldsymbol{k}'} g(\boldsymbol{k}')\psi_{n\boldsymbol{k}'}(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}\varepsilon_{n\boldsymbol{k}'}t/\hbar}\,. \tag{21.1.3}$$

The wave packet can be reasonably characterized by a wave vector $\boldsymbol{k}$ if the sum over $\boldsymbol{k}'$ is limited to a sufficiently small region around $\boldsymbol{k}$ whose width $\Delta\boldsymbol{k} = \boldsymbol{k}' - \boldsymbol{k}$ satisfies $|\Delta\boldsymbol{k}| \ll |\boldsymbol{k}|$. As is well known from quantum mechanics, the spatial extent of the wave packet depends on the size of this $\boldsymbol{k}$-space region. For wave packets obtained by the superposition of plane waves the uncertainties of the wave vector and position are related by

$$|\Delta\boldsymbol{k}| \cdot |\Delta\boldsymbol{r}| \sim 1\,. \tag{21.1.4}$$

It will be shown in Section 21.4 on the limitations of the semiclassical approximation that the same applies to wave packets constructed from Bloch states. Keeping this limitation in mind, one may speak of a Bloch electron at position $\boldsymbol{r}$ with crystal momentum $\hbar\boldsymbol{k}$.

By interpreting the motion of electrons as the propagation of wave packets, a semiclassical description can be obtained. This requires a proper definition of the velocity and mass of Bloch electrons, and then establishing the equation of motion in terms of them. We shall deal with these issues below, and return to the limitations of the semiclassical approximation in the last section of the chapter.

21.1.1 Velocity of Bloch Electrons

As established in optics, the group velocity of a wave packet made up of waves of dispersion relation $\omega(\boldsymbol{k})$ is

$$\boldsymbol{v}(\boldsymbol{k}) = \frac{\partial \omega(\boldsymbol{k})}{\partial \boldsymbol{k}} \,. \tag{21.1.5}$$

Since the energy ε of quantum particles and the frequency ω that is characteristic of the variations of the wavefunctions with time are related by $\varepsilon = \hbar\omega$, the expected formula for the velocity of Bloch electrons with a general dispersion relation $\varepsilon_{n\boldsymbol{k}}$ is

$$\boxed{\boldsymbol{v}_{n\boldsymbol{k}} = \frac{1}{\hbar}\frac{\partial \varepsilon_{n\boldsymbol{k}}}{\partial \boldsymbol{k}} \,.} \tag{21.1.6}$$

Below we shall demonstrate more rigorously that this is indeed correct by calculating the expectation value of the velocity operator for an electron in a Bloch state.

According to the customary quantum mechanical formulas, the velocity operator, as the time derivative of the position operator, can be expressed as a commutator with the Hamiltonian:

$$\boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar}\left[\mathcal{H}, \boldsymbol{r}\right]. \tag{21.1.7}$$

Since the periodic potential of the crystal in

$$\mathcal{H} = -\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}^2 + U(\boldsymbol{r}) \tag{21.1.8}$$

depends on the position coordinates alone, the evaluation of the commutator for electrons moving in the lattice potential leads to the same velocity operator,

$$\boldsymbol{v} = \frac{\boldsymbol{p}}{m_{\mathrm{e}}} = \frac{\hbar}{\mathrm{i}m_{\mathrm{e}}}\boldsymbol{\nabla} \,, \tag{21.1.9}$$

as (16.2.21), the operator for free electrons. The expectation value of the velocity in the Bloch state $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is

$$\boldsymbol{v}_{n\boldsymbol{k}} = \langle \psi_{n\boldsymbol{k}} | \boldsymbol{v} | \psi_{n\boldsymbol{k}} \rangle = \int \psi^*_{n\boldsymbol{k}}(\boldsymbol{r}) \frac{\hbar}{\mathrm{i}m_{\mathrm{e}}} \boldsymbol{\nabla}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} \,. \tag{21.1.10}$$

In order to relate this to the energy, we shall consider the Schrödinger equation that determines the energies and the Bloch functions. More precisely, (17.1.17), the equation obtained for the lattice-periodic function $u_{\boldsymbol{k}}(\boldsymbol{r})$ is used with an additional band index. In the latter equation $\mathcal{H}_{\boldsymbol{k}}$ is defined by (17.1.16). Applying the equation to a state whose wave number $\boldsymbol{k} + \delta\boldsymbol{k}$ is slightly different from $\boldsymbol{k}$, the substitution $\boldsymbol{k} \to \boldsymbol{k} + \delta\boldsymbol{k}$ gives

$$\mathcal{H}_{\boldsymbol{k}+\delta\boldsymbol{k}} u_{n,\boldsymbol{k}+\delta\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_{n,\boldsymbol{k}+\delta\boldsymbol{k}} u_{n,\boldsymbol{k}+\delta\boldsymbol{k}}(\boldsymbol{r}) \,, \tag{21.1.11}$$

where

$$\begin{aligned} \mathcal{H}_{\boldsymbol{k}+\delta\boldsymbol{k}} &= \frac{\hbar^2}{2m_{\mathrm{e}}}\left(\frac{1}{\mathrm{i}}\boldsymbol{\nabla} + \boldsymbol{k} + \delta\boldsymbol{k}\right)^2 + U(\boldsymbol{r}) \\ &= \mathcal{H}_{\boldsymbol{k}} + \frac{\hbar^2}{m_{\mathrm{e}}}\left(\frac{1}{\mathrm{i}}\boldsymbol{\nabla} + \boldsymbol{k}\right)\delta\boldsymbol{k} + \frac{\hbar^2}{2m_{\mathrm{e}}}(\delta\boldsymbol{k})^2 \,. \end{aligned} \tag{21.1.12}$$

If $\delta\boldsymbol{k}$ is small, $\mathcal{H}_{\boldsymbol{k}}$ can be formally considered as the Hamiltonian of an unperturbed system, and the other terms as weak perturbations; $\delta\boldsymbol{k}$ is then the small parameter of the perturbation expansion. When looking for the leading-order term of the variation of the energy, the terms that are of the second and higher orders in $\delta\boldsymbol{k}$ can be neglected in (21.1.12), and a first-order perturbation expansion can be employed using the term linear in $\delta\boldsymbol{k}$. According to (G.1.10), the first-order energy correction is

$$\delta\varepsilon_{n\boldsymbol{k}} = \int u^*_{n\boldsymbol{k}}(\boldsymbol{r})\frac{\hbar^2}{m_{\mathrm{e}}}\left(\frac{1}{\mathrm{i}}\boldsymbol{\nabla} + \boldsymbol{k}\right)\delta\boldsymbol{k}\, u_{n\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{21.1.13}$$

By writing the variation of the energy as $\delta\varepsilon_{n\boldsymbol{k}} = (\partial\varepsilon_{n\boldsymbol{k}}/\partial\boldsymbol{k})\delta\boldsymbol{k}$, and equating the coefficients of $\delta\boldsymbol{k}$ on the two sides,

$$\frac{\partial\varepsilon_{n\boldsymbol{k}}}{\partial\boldsymbol{k}} = \int u^*_{n\boldsymbol{k}}(\boldsymbol{r})\frac{\hbar^2}{m_{\mathrm{e}}}\left(\frac{1}{\mathrm{i}}\boldsymbol{\nabla} + \boldsymbol{k}\right) u_{n\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{21.1.14}$$

is obtained for the gradient of the dispersion curve. Switching back from the lattice-periodic functions $u_{n\boldsymbol{k}}(\boldsymbol{r})$ to the full Bloch function, and by inserting the phase factor $\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r})$, we have

$$\frac{\partial\varepsilon_{n\boldsymbol{k}}}{\partial\boldsymbol{k}} = \frac{\hbar^2}{m_{\mathrm{e}}}\int \psi^*_{n\boldsymbol{k}}(\boldsymbol{r})\frac{1}{\mathrm{i}}\boldsymbol{\nabla}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,. \tag{21.1.15}$$

Apart from a factor of $\hbar$, the right-hand side is just the expectation value of the velocity, as given in (21.1.10), thus

$$\frac{\partial\varepsilon_{n\boldsymbol{k}}}{\partial\boldsymbol{k}} = \hbar\boldsymbol{v}_{n\boldsymbol{k}}\,, \tag{21.1.16}$$

that is, (21.1.6) is indeed the velocity of Bloch electrons in solids.

Let us now apply this result to a simple situation: to the electron states in a simple cubic lattice obtained in the tight-binding approximation. The left-hand side of Fig. 21.1 shows the energy spectrum in a high-symmetry

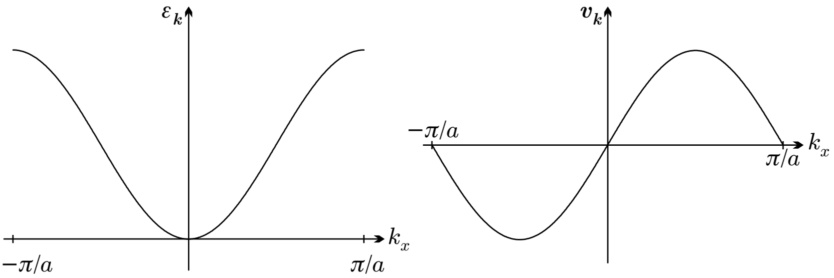


Fig. 21.1. Dispersion curve for electrons in the tight-binding approximation and their velocity in a cubic crystal along a high-symmetry direction of the Brillouin zone

direction of the Brillouin zone. (The choice of a high-symmetry direction is important as only then does the gradient of the dispersion relation point in the same direction.) The right-hand side of the figure shows the velocity determined from (21.1.6). The velocity vanishes at the center and boundaries of the Brillouin zone, while it takes both positive and negative values between them.

21.1.2 Semiclassical Equation of Motion

Once the velocity of Bloch electrons is known, it is straightforward to establish the $\boldsymbol{k}$-space equation of motion for the variations of the wave vector with time. If the wave packet made up of Bloch electrons in the nth band moves at a velocity $\boldsymbol{v}_{n\boldsymbol{k}}$, the work done in time $\mathrm{d}t$ on an electron in a uniform electric field $\boldsymbol{E}$ is

$$\mathrm{d}W = -e\boldsymbol{E} \cdot \boldsymbol{v}_{n\boldsymbol{k}}\, \mathrm{d}t\,. \tag{21.1.17}$$

This has to be the same as the variation of the one-particle energy

$$\frac{\mathrm{d}\varepsilon_{n\boldsymbol{k}}}{\mathrm{d}t}\mathrm{d}t\,. \tag{21.1.18}$$

The variation of the energy is due, in part, to the variation of the wave vector $\boldsymbol{k}$. Another possible source of the variation is the interband transition of the Bloch electron. As we shall demonstrate later, interband transitions can be neglected, thus the variation in the energy due to the change of the wave vector is

$$\mathrm{d}W = \frac{\partial\varepsilon_{n\boldsymbol{k}}}{\partial\boldsymbol{k}}\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t}\mathrm{d}t = \hbar\boldsymbol{v}_{n\boldsymbol{k}}\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t}\mathrm{d}t\,. \tag{21.1.19}$$

Comparison of the two formulas for $\mathrm{d}W$ shows that the free-electron expression

$$\boxed{\hbar\dot{\boldsymbol{k}} = -e\boldsymbol{E}} \tag{21.1.20}$$

is valid for Bloch electrons, too. Its solution is

$$\hbar\boldsymbol{k}(t) = \hbar\boldsymbol{k}(0) - e\boldsymbol{E}t\,. \tag{21.1.21}$$

The uniform variation of the wave vector can be valid only as long as $\boldsymbol{k}$ is inside the Brillouin zone. Since the wave vector is determined only up to a reciprocal-lattice vector, when $\boldsymbol{k}(t)$ reaches the zone boundary it jumps to the equivalent vector $\boldsymbol{k} + \boldsymbol{G}$ on the opposite side of the zone, from where the uniform variation continues. Thus $\boldsymbol{k}$ shows in fact periodic variation with time. This is plotted in Fig. 21.2 for a high-symmetry direction of the Brillouin zone.

Similar oscillations are obtained from

$$\dot{\boldsymbol{r}}(t) = \boldsymbol{v}_n[\boldsymbol{k}(t)]\,, \tag{21.1.22}$$

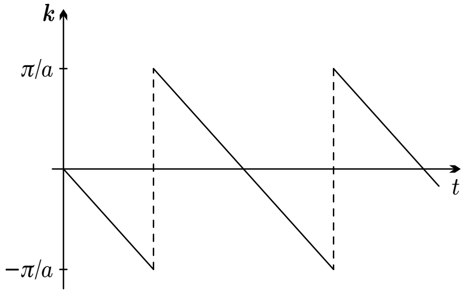


Fig. 21.2. Periodic variation of the wave vector of a Bloch electron in a uniform electric field, along a high-symmetry direction of the Brillouin zone

the equation that determines the motion of electrons in real space, when due account is taken of the property that the velocity itself oscillates as $\boldsymbol{k}$ runs over the Brillouin zone. As shown in Fig. 21.3, the electron starting from the state $\boldsymbol{k} = 0$ first accelerates, then decelerates, and finally turns around. Therefore in a uniform electric field, under ideal circumstances, electrons oscillate.

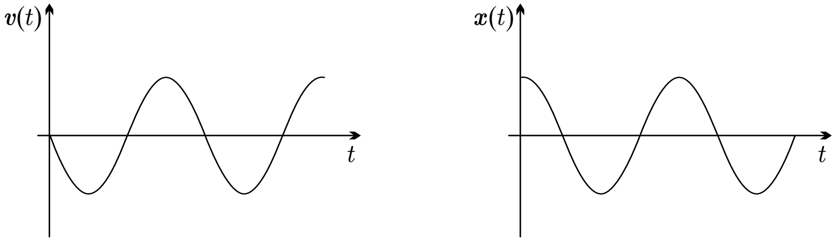


Fig. 21.3. Periodic variations in the velocity and position of a Bloch electron in a uniform electric field

In reality, the situation is more complicated because of collisions. As has been mentioned in connection with the Drude model, electrons fly freely for an average time τ, accelerating as determined by the equation of motion, and then collide, losing the energy taken from the electric field. Taking typical relaxation times for metals, in customarily applied electric fields the variation of the wave vector is typically much smaller than the size of the Brillouin zone, thus these oscillations cannot be observed. Electrons moving in the periodic field of the lattice carry a direct current. This will be discussed in detail in Chapter 24.

It should be emphasized that the foregoing implies only that if the Bloch electron is in the state of wave vector $\boldsymbol{k}_0$ in the n_0th band at $t = 0$ – that is, its translational properties are determined by the equation

$$\psi_{n_0\boldsymbol{k}_0}(\boldsymbol{r} + \boldsymbol{t}_m) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}_0\cdot\boldsymbol{t}_m}\psi_{n_0\boldsymbol{k}_0}(\boldsymbol{r}) \,, \tag{21.1.23}$$

then after time t the wavefunction satisfies

$$\psi(\boldsymbol{r}+\boldsymbol{t}_m,t)=\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{t}_m}\psi(\boldsymbol{r},t)\,, \tag{21.1.24}$$

where $\boldsymbol{k}$ can be determined from (21.1.21). This new state is usually not an eigenstate of the Hamiltonian but a superposition of states whose wave vectors $\boldsymbol{k}$ are in different bands. Electric fields can thus induce interband transitions. As we shall see, the applicability of semiclassical dynamics is limited by this very fact: the approximation is valid only as long as interband transitions can be neglected.

In the presence of a magnetic field the considerations based on the variation of energy are not useful, since the electron energy is conserved in uniform magnetic fields. One possibility is to use the Schrödinger equation for the time evolution of the electronic wavefunction $\psi_{n_0\boldsymbol{k}_0}(\boldsymbol{r})$ at time $t=0$ to obtain ψ at a later time $\mathrm{d}t$:

$$\psi(\boldsymbol{r},\mathrm{d}t)=\mathrm{e}^{-\mathrm{i}\mathcal{H}\mathrm{d}t/\hbar}\psi_{n_0\boldsymbol{k}_0}(\boldsymbol{r})\,, \tag{21.1.25}$$

where the kinetic-energy term of the Hamiltonian is now written in terms of the kinetic (rather than the canonical) momentum, which contains the vector potential as well, and the variation of the wavefunction under translations is determined for $\psi(\boldsymbol{r},\mathrm{d}t)$. The operator for a translation through $\boldsymbol{t}_m$ is known to be

$$T_{\boldsymbol{t}_m}=\mathrm{e}^{-\boldsymbol{t}_m\cdot\boldsymbol{\nabla}}\,, \tag{21.1.26}$$

since the straightforward series expansion of the exponent leads to

$$T_{\boldsymbol{t}_m}\psi(\boldsymbol{r})=\mathrm{e}^{-\boldsymbol{t}_m\cdot\boldsymbol{\nabla}}\psi(\boldsymbol{r})=\psi(\boldsymbol{r}-\boldsymbol{t}_m)\,. \tag{21.1.27}$$

When this operator is applied to the wavefunction at time $\mathrm{d}t$, then

$$T_{\boldsymbol{t}_m}\psi(\boldsymbol{r},\mathrm{d}t)=\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{t}_m}\psi(\boldsymbol{r},\mathrm{d}t) \tag{21.1.28}$$

can no longer be satisfied exactly with a magnetic-field-dependent $\boldsymbol{k}$. It can nonetheless be demonstrated that when the terms that are quadratic in the magnetic field are neglected, the behavior under translations is characterized by the wave vector

$$\boldsymbol{k}=\boldsymbol{k}_0-\frac{e}{\hbar}\boldsymbol{v}_{n_0\boldsymbol{k}_0}\times\boldsymbol{B}\mathrm{d}t\,, \tag{21.1.29}$$

where $\boldsymbol{v}_{n\boldsymbol{k}}$ is the velocity of the Bloch electron. The second term on the right-hand side contains the Lorentz force acting on a particle of charge $-e$ in a magnetic field $\boldsymbol{B}$. Its presence can be understood intuitively. Since in many respects the wave packet behaves as a classical particle of momentum $\hbar\boldsymbol{k}$, it is not surprising that the equation governing the variation of the wave vector of the wave packet is the same as the classical equation of motion for an electron. Consequently, we shall assume that the $\boldsymbol{k}$-space equation of motion

$$\boxed{\hbar\dot{\boldsymbol{k}}=-e\big[\boldsymbol{E}(\boldsymbol{r},t)+\boldsymbol{v}_{n\boldsymbol{k}}\times\boldsymbol{B}(\boldsymbol{r},t)\big]} \tag{21.1.30}$$

is valid for time- and position-dependent fields as well.

21.1.3 Effective-Mass Tensor

The effective mass of Bloch electrons was defined through the dispersion relation in Chapter 17, even though physical meaning was attached to this concept only close to the bottom of the band (or to the top for holes). Below we shall examine whether the dynamics of electrons is characterized by the same effective mass. To this end, we first determine the acceleration of the electrons by differentiating the velocity formula (21.1.6) with respect to time:

$$\dot{\boldsymbol{v}}_{n\boldsymbol{k}} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\hbar}\frac{\partial \varepsilon_{n\boldsymbol{k}}}{\partial \boldsymbol{k}}\right). \tag{21.1.31}$$

For notational simplicity, we shall often suppress the band index, keeping in mind its implicit presence. Since the energy depends on time only through $\boldsymbol{k}$, the rules of implicit differentiation give

$$\dot{\boldsymbol{v}}_{\boldsymbol{k}} = \frac{\partial}{\partial \boldsymbol{k}}\left(\frac{1}{\hbar}\frac{\partial \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}}\right)\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}^2}\hbar\dot{\boldsymbol{k}}\,. \tag{21.1.32}$$

As $\hbar\dot{\boldsymbol{k}}$ is the force $\boldsymbol{F}$ on the electron, comparison with the classical equation of motion

$$\dot{\boldsymbol{v}} = \frac{1}{M}\boldsymbol{F} \tag{21.1.33}$$

for a particle of mass M gives the same formula,

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}^2}\,, \tag{21.1.34}$$

for the effective mass governing the dynamics of an electron moving through the periodic field of the lattice as (17.4.9), the defining equation of the effective mass, obtained from the curvature of the dispersion relation.

In more general cases, where the dispersion relation of the electrons is not isotropic, their motion can be described in terms of an effective-mass tensor. Defining the tensor through the generalization of (21.1.33),

$$\dot{v}_\alpha = \left(\frac{1}{M^*}\right)_{\alpha\beta} F_\beta\,, \tag{21.1.35}$$

we have

$$\boxed{\left(\frac{1}{M^*}\right)_{\alpha\beta} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{\boldsymbol{k}}}{\partial k_\alpha \partial k_\beta}\,, \qquad \alpha,\, \beta = x,\, y,\, z} \tag{21.1.36}$$

for the inverse dynamical effective-mass tensor. Obviously, this is the same as (17.4.12). The effective mass can thus be defined for electrons associated with any point of the Brillouin zone. The concept proves really useful when the effective mass is the same for a reasonably large group of electrons.

21.1.4 Motion of Electrons and Holes

It was mentioned in connection with the density of states that the dispersion relation can usually be approximated by a quadratic form at the bottom of the band, which facilitates the definition of the effective mass there. At the top of the band it is more practical to consider empty states as holes. Below we shall examine their motion in an applied electromagnetic field.

The semiclassical equations governing the dynamics of electrons are the same whether or not the state is occupied. In the phase space spanned by vectors $\boldsymbol{k}$ and $\boldsymbol{r}$ the path traced out by the wave packet is independent of the occupation. As long as collisions can be neglected, occupied states evolve into occupied, and unoccupied into unoccupied states. When an unoccupied state is treated as a hole, the equation of motion for the variation of the wave vector of the electron

$$\hbar\dot{\boldsymbol{k}}_{\mathrm{e}} = -e\left[\boldsymbol{E}(\boldsymbol{r},t) + \boldsymbol{v}_{\mathrm{e}n}(\boldsymbol{k}_{\mathrm{e}}) \times \boldsymbol{B}(\boldsymbol{r},t)\right], \tag{21.1.37}$$

(in which even the velocity has an explicit "electron" label) can be rewritten in terms of the hole wave vector $\boldsymbol{k}_{\mathrm{h}} = -\boldsymbol{k}_{\mathrm{e}}$ and the hole velocity defined by

$$\boldsymbol{v}_{\mathrm{h}n}(\boldsymbol{k}_{\mathrm{h}}) = \frac{1}{\hbar}\frac{\partial\varepsilon_{\mathrm{h}n}(\boldsymbol{k}_{\mathrm{h}})}{\partial\boldsymbol{k}_{\mathrm{h}}}. \tag{21.1.38}$$

As (17.4.18) implies, $\boldsymbol{v}_{\mathrm{h}n}(\boldsymbol{k}_{\mathrm{h}}) = \boldsymbol{v}_{\mathrm{e}n}(\boldsymbol{k}_{\mathrm{e}})$, so

$$\hbar\dot{\boldsymbol{k}}_{\mathrm{h}} = +e\left[\boldsymbol{E}(\boldsymbol{r},t) + \boldsymbol{v}_{\mathrm{h}n}(\boldsymbol{k}_{\mathrm{h}}) \times \boldsymbol{B}(\boldsymbol{r},t)\right]. \tag{21.1.39}$$

This can be interpreted by attributing a positive charge to the hole. We shall also show that the current carried by holes can also be calculated using the same elementary picture: particles of charge $+e$ move at a velocity $\boldsymbol{v}_{\mathrm{h}n}(\boldsymbol{k}_{\mathrm{h}})$.

If the charge carriers occupy several bands, the total current is obtained by summing the contributions of the electrons in each band. We shall first demonstrate that completely filled bands do not contribute to the conductivity, so only partially filled bands need to be considered. This is at the heart of our previous assertion that the prerequisite for metallic behavior is at least one incompletely filled band.

For any lattice-periodic function $f(\boldsymbol{r})$ the integral over the primitive cell (of volume v) is invariant under the translation of the argument by an arbitrary vector $\boldsymbol{r}'$:

$$\int\limits_{v} \mathrm{d}\boldsymbol{r} f(\boldsymbol{r}) = \int\limits_{v} \mathrm{d}\boldsymbol{r} f(\boldsymbol{r}+\boldsymbol{r}'), \tag{21.1.40}$$

hence

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{r}'}\int\limits_{v} \mathrm{d}\boldsymbol{r} f(\boldsymbol{r}+\boldsymbol{r}') &= \int\limits_{v} \mathrm{d}\boldsymbol{r}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{r}'} f(\boldsymbol{r}+\boldsymbol{r}') = \int\limits_{v} \mathrm{d}\boldsymbol{r}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{r}} f(\boldsymbol{r}+\boldsymbol{r}') \\ &= \int\limits_{v} \mathrm{d}\boldsymbol{r}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{r}} f(\boldsymbol{r}) = 0.\end{aligned} \tag{21.1.41}$$

A similar statement is true for functions that are periodic in the reciprocal lattice, provided integration is over the primitive cell of the reciprocal lattice or the equivalent Brillouin zone.

To apply this to the electric and energy currents, consider their semiclassical formulas. The current density carried by the electrons in the nth band is

$$\boldsymbol{j} = -e\frac{1}{V}\sum_{\boldsymbol{k},\sigma} \boldsymbol{v}_{n\boldsymbol{k}} = -e\int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\frac{1}{\hbar}\frac{\partial \varepsilon_{n\boldsymbol{k}}}{\partial \boldsymbol{k}}\,, \tag{21.1.42}$$

while the energy-current density is

$$\boldsymbol{j}_\varepsilon = \frac{1}{V}\sum_{\boldsymbol{k},\sigma} \varepsilon_{n\boldsymbol{k}}\boldsymbol{v}_{n\boldsymbol{k}} = \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\varepsilon_{n\boldsymbol{k}}\frac{1}{\hbar}\frac{\partial \varepsilon_{n\boldsymbol{k}}}{\partial \boldsymbol{k}} = \frac{1}{2}\int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\frac{1}{\hbar}\frac{\partial \varepsilon^2_{n\boldsymbol{k}}}{\partial \boldsymbol{k}}\,, \tag{21.1.43}$$

where the integrals are over the occupied states. Since $\varepsilon_{n\boldsymbol{k}}$ and its square are lattice-periodic in the reciprocal space, the electric and energy currents both vanish for completely filled bands. It is therefore sufficient to consider bands close to the Fermi energy only; those deep below are completely filled and can therefore be ignored.

If the band is partially filled, the $\boldsymbol{k}$-space sum can be decomposed into two sums: over occupied and unoccupied states. According to the foregoing,

$$\int\limits_{\text{filled}} \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{n\boldsymbol{k}} + \int\limits_{\text{empty}} \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{n\boldsymbol{k}} = 0\,. \tag{21.1.44}$$

The electric current carried by occupied states can then be rewritten in terms of the empty states as

$$\boldsymbol{j} = -e\int\limits_{\text{filled}} \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{n\boldsymbol{k}} = +e\int\limits_{\text{empty}} \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{n\boldsymbol{k}}\,. \tag{21.1.45}$$

The electric current can thus be interpreted as arising from positively charged holes. This observation will be particularly useful for semiconductors, where the contribution of the almost completely filled valence band is best treated in terms of holes.

21.2 Bloch Electrons in Uniform Magnetic Fields

The study of the motion of electrons in a uniform magnetic field provides fundamental information about the effective mass and other important properties of the Bloch electrons whose wave vectors make up the Fermi surface, and about the shape of the Fermi surface itself. We shall examine this motion in the semiclassical approximation.

21.2.1 Motion in Reciprocal and Real Spaces

The variation of the wave vector $\boldsymbol{k}$ in a uniform magnetic field is governed by the equation

$$\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = -\frac{e}{\hbar}\boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}\,, \tag{21.2.1}$$

where the velocity is calculated from the dispersion relation of the electrons in the customary way:

$$\boldsymbol{v}_{\boldsymbol{k}} = \frac{1}{\hbar}\frac{\partial \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}}\,. \tag{21.2.2}$$

For simplicity, we shall suppress band indices.

A direct consequence of (21.2.1) is that the variation of the wave vector with time is perpendicular to the magnetic field, so $\boldsymbol{k}_{\|}$, the component of $\boldsymbol{k}$ along the magnetic field, is conserved. Another conserved quantity is the energy. Taking the time derivative of the energy of electrons and making use of the property that the energy depends on time only through the wave vector, we have

$$\frac{\mathrm{d}\varepsilon_{\boldsymbol{k}}}{\mathrm{d}t} = \frac{\partial \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}} \cdot \frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = -e\boldsymbol{v}_{\boldsymbol{k}} \cdot (\boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}) = 0\,. \tag{21.2.3}$$

Just like free electrons, Bloch electrons cannot absorb any energy from the magnetic field either. In the semiclassical approximation the wave vector of the Bloch electron has to remain on a constant-energy surface in $\boldsymbol{k}$-space – more specifically, on its intersection with a plane perpendicular to the magnetic field. This is illustrated in Fig. 21.4, which shows the $\boldsymbol{k}$-space orbits for different Fermi surfaces in a simple cubic lattice. The almost spherical Fermi surface on the left-hand side corresponds to nearly free electrons, while the one on the right-hand side depicts an electron system for which the band is more than half filled and that can be modeled in the tight-binding approximation. If the magnetic field is along the z-axis, the $\boldsymbol{k}$ vector moves on the Fermi surface in the direction marked by arrows.

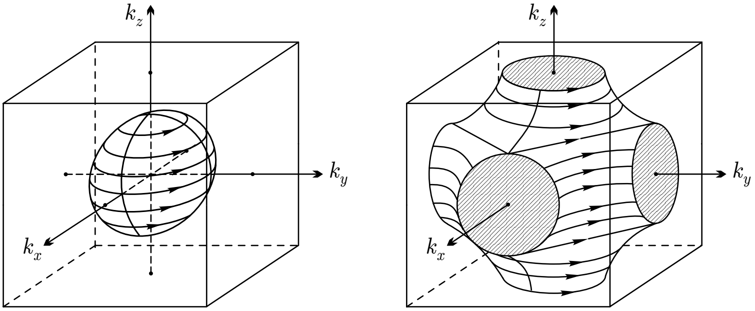


Fig. 21.4. $\boldsymbol{k}$-space orbits on two Fermi surfaces in the presence of a uniform magnetic field applied along the z-axis

In an applied magnetic field the trajectory in real space[1] has a similar shape to the orbit in $\boldsymbol{k}$-space. Writing the equation for $\boldsymbol{k}$ as

$$\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = -\frac{e}{\hbar}\left(\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} \times \boldsymbol{B}\right), \tag{21.2.4}$$

integration yields

$$\boldsymbol{k}(t) - \boldsymbol{k}(0) = -\frac{e}{\hbar}\left[\boldsymbol{r}(t) - \boldsymbol{r}(0)\right] \times \boldsymbol{B}. \tag{21.2.5}$$

Because of the vector product, only the component of $\boldsymbol{r}(t) - \boldsymbol{r}(0)$ that is perpendicular to the z-directed magnetic field contributes to the right-hand side:

$$\left[\boldsymbol{r}(t) - \boldsymbol{r}(0)\right]_{\perp} \times \hat{\boldsymbol{z}} = -\frac{\hbar}{eB}\left[\boldsymbol{k}(t) - \boldsymbol{k}(0)\right], \tag{21.2.6}$$

where $\hat{\boldsymbol{z}}$ is the unit vector along the z-axis. In component form:

$$\begin{aligned} x(t) - x(0) &= \frac{\hbar}{eB}\left[k_y(t) - k_y(0)\right], \\ y(t) - y(0) &= -\frac{\hbar}{eB}\left[k_x(t) - k_x(0)\right]. \end{aligned} \tag{21.2.7}$$

Consequently, the projection of the real-space trajectory of the electrons to the xy-plane follows the $\boldsymbol{k}$-space motion, however:

1. The real-space motion is scaled up by a factor of $l_0^2 = \hbar/eB$ with respect to the $\boldsymbol{k}$-space motion;[2]
2. Multiplication by the unit vector $\hat{\boldsymbol{z}}$ gives rise to a $\pi/2$ phase shift between real- and $\boldsymbol{k}$-space motions.

This is shown in Fig. 21.5. Since the shape of the real-space trajectory is similar to that of the $\boldsymbol{k}$-space orbit, and the latter is determined by the shape of the constant-energy surfaces, the trajectory will be circular (as in the free-electron case) only when the constant-energy surfaces are spheres – i.e., the energy can be given in terms of a scalar effective mass. If the situation is more complex, elliptical or even more complicated trajectories are obtained.

[1] To describe the motion of electrons, some authors use the word *orbit* both in $\boldsymbol{k}$-space and real space. Others distinguish the $\boldsymbol{k}$-space and real-space motions by using *orbit* and *trajectory*, respectively. When this distinction is important, we shall adopt the latter choice.

[2] As will be demonstrated in the next chapter, in magnetic fields applied customarily in experiments the magnetic length l_0 is of order 10^{-6} cm. Since the magnitude of the wave vector $\boldsymbol{k}$ is on the order of the linear dimension of the Brillouin zone for electrons moving on the Fermi surface, the dimensions of the real-space trajectory are much larger than the atomic dimensions.

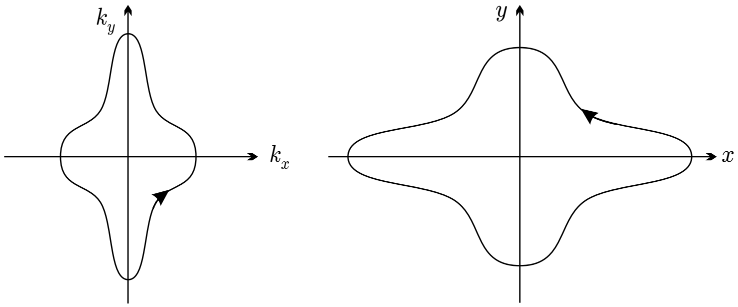


Fig. 21.5. The motion in $\boldsymbol{k}$-space and the projection of the real-space trajectory to the plane perpendicular to the magnetic field

21.2.2 Open and Closed Orbits in Magnetic Fields

The distortions of the Fermi "sphere" in a square lattice due to the weak periodic potential were analyzed in Chapter 18. Figure 18.18 showed that the Fermi surface is almost circular when the number of electrons is small, whereas it is made up of several sheets when the number of electrons is large. Very similar figures are obtained for the $k_z = 0$ section of the Fermi surface in a simple cubic lattice when the band filling is sufficiently high for that some electrons occupy states beyond the first Brillouin zone. This is illustrated in Fig. 21.6(a).

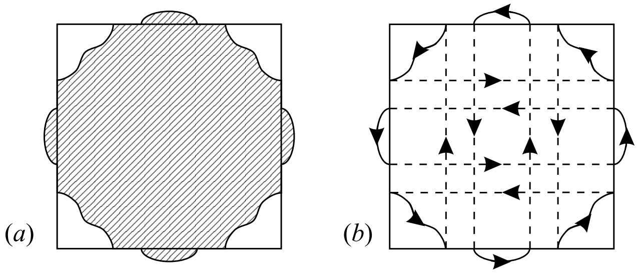


Fig. 21.6. (a) Section of the Fermi surface of nearly free electrons for a relatively high band filling, represented in the extended-zone scheme. (b) The $\boldsymbol{k}$-space orbit of electrons on the Fermi surface in an applied magnetic field

In a uniform magnetic field along the z-axis, the $\boldsymbol{k}$ vector traverses traverses the intersection of the Fermi surface with the k_z = constant plane. This is depicted in the extended-zone scheme in Fig. 21.6(b). Since the Fermi surface is discontinuous at the zone boundaries, when electrons reach them they continue their motion from the equivalent point across the zone, therefore they do not trace out the entire intersection just a small portion of it.

This is more conspicuous when the $\boldsymbol{k}$-space motion of electrons is studied in the repeated-zone scheme.

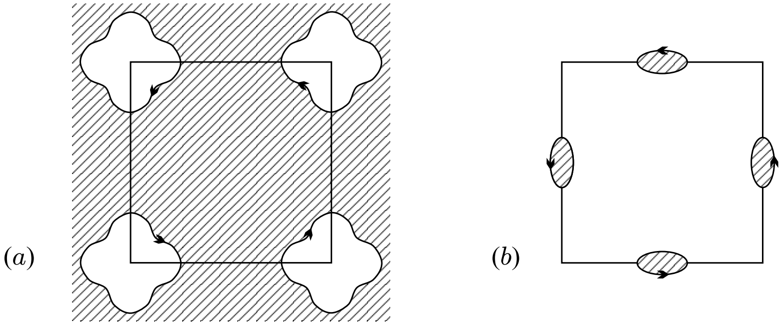


Fig. 21.7. The repeated-zone-scheme representation of those portions of the Fermi surface shown in the previous figure that belong in the first and second bands

When the portions of the Fermi sphere are translated through a reciprocal-lattice vector, as shown in Figs. 21.7(a) and (b), the portions that were initially in the first and second Brillouin zones make up a connected Fermi surface, or a closed orbit when its section is considered. The portions of the Fermi surface that belong in the second band make up a curve whose interior contains occupied states, and in a magnetic field perpendicular to the orbit electrons move along the orbit counterclockwise, just as free electrons do. The situation is different for the portions that belong in the first band: they surround empty states, so this is a hole-type Fermi surface, and electrons trace out the orbit clockwise.

Depending on the shape of the Fermi surface, the orbits may not close in on themselves even in the repeated-zone scheme, or they may be closed in certain directions but open in others. Such an open orbit is illustrated in Fig. 21.8. Electrons then move on orbits that run to infinity.

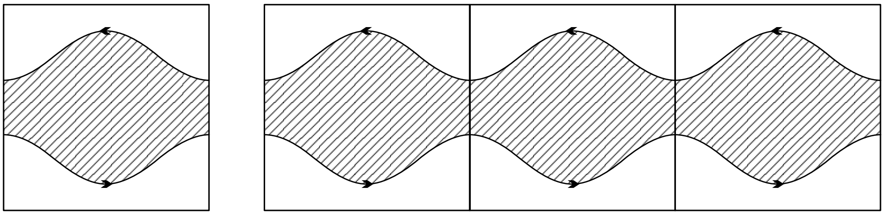


Fig. 21.8. Open orbits on the Fermi surface in the reduced- and repeated-zone schemes

By varying the direction of the magnetic field, the shape of the sections of the Fermi surface – and thus the shape of the electrons' orbits – can be changed. This property can be used for the experimental determination of the shape of the Fermi surface.

21.2.3 Cyclotron Frequency in a Closed Orbit

For closed orbits one may calculate the period T_c that is required for the $\boldsymbol{k}$ vector of an electron of energy ε to traverse the intersection of the surface of constant energy ε with a plane perpendicular to the magnetic field. By choosing, as usual, the z-axis along the direction of the magnetic field B, k_z is conserved. Since the element of arc in the $(k_x,\, k_y)$ plane can be written as

$$\mathrm{d}l = \sqrt{(\mathrm{d}k_x)^2 + (\mathrm{d}k_y)^2}\,, \tag{21.2.8}$$

the time rate of change for the arc length along the $\boldsymbol{k}$-space orbit is

$$\frac{\mathrm{d}l}{\mathrm{d}t} = \sqrt{\left(\frac{\mathrm{d}k_x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}k_y}{\mathrm{d}t}\right)^2}\,. \tag{21.2.9}$$

As

$$\frac{\mathrm{d}k_x}{\mathrm{d}t} = -\frac{e}{\hbar} v_y B\,, \qquad \frac{\mathrm{d}k_y}{\mathrm{d}t} = \frac{e}{\hbar} v_x B\,, \tag{21.2.10}$$

we have

$$\frac{\mathrm{d}l}{\mathrm{d}t} = \frac{eB}{\hbar}\sqrt{v_x^2 + v_y^2} = \frac{eB}{\hbar} v_\perp\,, \tag{21.2.11}$$

where $v_\perp$ is the velocity component that is perpendicular to the field direction. Rearrangement then gives

$$\mathrm{d}t = \frac{\hbar}{eB}\frac{\mathrm{d}l}{v_\perp}\,. \tag{21.2.12}$$

For a closed orbit C the period T_c is the circular integral around the orbit:

$$T_\mathrm{c} = \frac{\hbar}{eB}\oint\limits_C \frac{\mathrm{d}l}{v_\perp}\,. \tag{21.2.13}$$

This expression can be related to the area $\mathcal{A}$ enclosed by the orbit in $\boldsymbol{k}$-space. To demonstrate this, consider the constant k_z section of the surfaces of constant energy ε and $\varepsilon + \mathrm{d}\varepsilon$.

As illustrated in Fig. 21.9, the area between the lines of constant energy ε and $\varepsilon + \mathrm{d}\varepsilon$ is

$$\mathrm{d}\mathcal{A} = \oint \mathrm{d}l\, \mathrm{d}k_\perp\,. \tag{21.2.14}$$

The $\boldsymbol{k}$-space distance $\mathrm{d}k_\perp$ of the lines of constant energy can be expressed in terms of the energy difference $\mathrm{d}\varepsilon$ through the relation

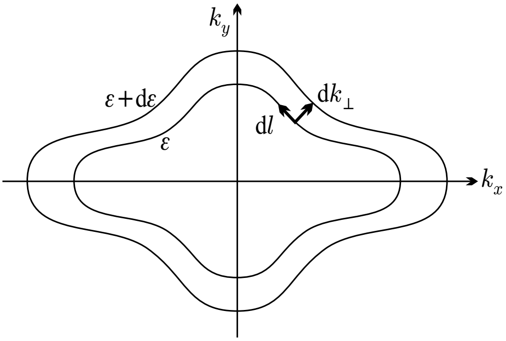


Fig. 21.9. Sections of the surfaces of constant energy ε and $\varepsilon + \mathrm{d}\varepsilon$ perpendicular to the uniform z-directed magnetic field, at a particular value of k_z

$$\mathrm{d}k_\perp = \frac{\mathrm{d}k_\perp}{\mathrm{d}\varepsilon}\mathrm{d}\varepsilon = \frac{\mathrm{d}\varepsilon}{\mathrm{d}\varepsilon/\mathrm{d}k_\perp} = \frac{\mathrm{d}\varepsilon}{\hbar v_\perp}\,. \tag{21.2.15}$$

Substituting this into (21.2.14),

$$\mathrm{d}\mathcal{A} = \oint \frac{\mathrm{d}l}{\hbar v_\perp}\,\mathrm{d}\varepsilon\,, \tag{21.2.16}$$

and so

$$\frac{\partial \mathcal{A}}{\partial \varepsilon} = \oint \frac{\mathrm{d}l}{\hbar v_\perp}\,. \tag{21.2.17}$$

This is the same integral as in expression (21.2.13) for the period, therefore

$$\boxed{T_\mathrm{c} = \frac{\hbar^2}{eB}\frac{\partial \mathcal{A}}{\partial \varepsilon}\,.} \tag{21.2.18}$$

The frequency of the periodic motion is

$$\nu_\mathrm{c} = \frac{eB}{\hbar^2}\left(\frac{\partial \mathcal{A}}{\partial \varepsilon}\right)^{-1}, \tag{21.2.19}$$

so its angular frequency is just

$$\boxed{\omega_\mathrm{c} = \frac{2\pi eB}{\hbar^2}\left(\frac{\partial \mathcal{A}}{\partial \varepsilon}\right)^{-1}.} \tag{21.2.20}$$

As this is the (angular) frequency of the periodic motion of an electron in an applied magnetic field, ω_c is called the *cyclotron frequency* – even though rigorously speaking the term *cyclotron angular frequency* would be more appropriate. For a magnetic field direction that is fixed relative to the Fermi surface, the cyclotron frequency is generally different in different sections of the Fermi surface (i.e., for electrons with different $k_{\|}$, where $k_{\|}$ is the projection of the wave vector to the direction of the magnetic field). We shall see that this quantity can be measured in experiments when a substantial fraction of all electrons on the Fermi surface move at the same cyclotron frequency. This occurs for stationary (extremal) sections of the Fermi surface – that is for those sections for which the energy derivative of the area of the cross section of the Fermi surface with the plane perpendicular to the magnetic field varies only slowly as the position of the plane is changed. Consequently, the cyclotron frequencies associated with the maximal and minimal cross sections can be measured. Since in general ω_{c} depends on the magnetic field direction, the shape of the Fermi surface can be determined from the direction dependence of the cyclotron frequency.

21.2.4 Cyclotron Mass

Writing the one-particle energy for free electrons in terms of the wave-vector components that are parallel (k_z) and perpendicular ($k_\perp$) to the magnetic field, the surface of constant energy ε is obtained as the related pairs (k_z, $k_\perp$) that satisfy the equation

$$\varepsilon = \frac{\hbar^2}{2m_{\mathrm{e}}}(k_z^2 + k_\perp^2)\,. \tag{21.2.21}$$

The $\boldsymbol{k}$-space area of its cross section with the plane perpendicular to z at height k_z is

$$\mathcal{A}(\varepsilon, k_z) = k_\perp^2 \pi = \frac{2\pi m_{\mathrm{e}}}{\hbar^2}\varepsilon - k_z^2 \pi\,. \tag{21.2.22}$$

This implies

$$\frac{\partial \mathcal{A}}{\partial \varepsilon} = \frac{2\pi m_{\mathrm{e}}}{\hbar^2}\,. \tag{21.2.23}$$

Substituting this into the cyclotron frequency formula (21.2.20),

$$\omega_{\mathrm{c}} = \frac{eB}{m_{\mathrm{e}}} \tag{21.2.24}$$

is obtained for free electrons.

Analogously, the period and cyclotron frequency are customarily written in the free-electron-like form

$$T_{\mathrm{c}} = \frac{2\pi m_{\mathrm{c}}}{eB}\,, \qquad \omega_{\mathrm{c}} = \frac{eB}{m_{\mathrm{c}}} \tag{21.2.25}$$

for more general one-particle energy spectra, too, where the *cyclotron mass* m_{c} is defined by

$$m_{\mathrm{c}} = \frac{\hbar^2}{2\pi}\frac{\partial \mathcal{A}}{\partial \varepsilon}\,. \tag{21.2.26}$$

By repeating the steps of the calculation of the free-electron case, it is immediately established that when the energy of Bloch electrons can be characterized by a scalar effective mass m^* then the cyclotron mass is the same as the effective mass obtained from the band structure:

$$m_{\mathrm{c}} = m^* . \tag{21.2.27}$$

For more general dispersion relation of the Bloch electrons, the constant-energy surfaces are ellipsoids of general orientation in the vicinity of the minima and maxima, where the dispersion relation can be approximated by quadratic expressions, and the intersections of the Fermi surface with planes perpendicular to the magnetic field are ellipses. As shown in Fig. 21.10, the area enclosed by the cyclotron orbit depends on the particular choice of the point on the Fermi surface. However, just like for spherical Fermi surfaces, the period of the motion of the electron and, through it, the cyclotron frequency are independent of the height parameter (formerly denoted by k_z) of the section, it depends only on the orientation of the applied magnetic field relative to the principal axes of the ellipsoidal Fermi surface. The cyclotron mass is a certain average of the components of the effective-mass tensor, namely, when the magnetic field is along the z-axis,

$$m_{\mathrm{c}} = \left(\frac{\det \mathsf{M}^*_{ij}}{\mathsf{M}^*_{zz}} \right)^{1/2} . \tag{21.2.28}$$

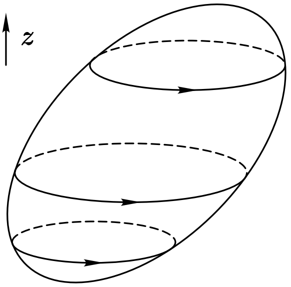


Fig. 21.10. Elliptic cyclotron orbits on an ellipsoidal Fermi surface

The most straightforward way to derive this is solving the semiclassical equation of motion for electrons described by the dispersion relation (17.4.13) in the presence of a z-directed magnetic field. We shall apply the same method in the next subsection in a more refined calculation that takes into account the finite lifetime of the electrons due to collisions.

For notational simplicity, we shall shift the origin to $\boldsymbol{k}_0$. Calculating the velocity in the semiclassical equation of motion from the dispersion relation, we have

$$\begin{aligned}
\frac{\mathrm{d}k_x}{\mathrm{d}t} &= -e\left[\left(\frac{1}{M^*}\right)_{xy} k_x + \left(\frac{1}{M^*}\right)_{yy} k_y + \left(\frac{1}{M^*}\right)_{yz} k_z\right] B_z\,,\\
\frac{\mathrm{d}k_y}{\mathrm{d}t} &= e\left[\left(\frac{1}{M^*}\right)_{xx} k_x + \left(\frac{1}{M^*}\right)_{xy} k_y + \left(\frac{1}{M^*}\right)_{xz} k_z\right] B_z\,, \qquad (21.2.29)\\
\frac{\mathrm{d}k_z}{\mathrm{d}t} &= 0\,.
\end{aligned}$$

As the motion is presumably periodic, we seek solutions of the form $\boldsymbol{k}\mathrm{e}^{-\mathrm{i}\omega_{\mathrm{c}}t}$. The above formulas then lead to the set of homogeneous equations

$$\begin{aligned}
\left[\mathrm{i}\omega_{\mathrm{c}} - e\left(\frac{1}{M^*}\right)_{xy} B_z\right] k_x - e\left(\frac{1}{M^*}\right)_{yy} B_z k_y - e\left(\frac{1}{M^*}\right)_{yz} B_z k_z &= 0\,,\\
e\left(\frac{1}{M^*}\right)_{xx} k_x + \left[\mathrm{i}\omega_{\mathrm{c}} + e\left(\frac{1}{M^*}\right)_{xy} B_z\right] k_y + e\left(\frac{1}{M^*}\right)_{xz} B_z k_z &= 0\,, \qquad (21.2.30)\\
\mathrm{i}\omega_{\mathrm{c}} k_z &= 0\,.
\end{aligned}$$

Nontrivial solutions exist when the determinant formed from the coefficients vanishes:

$$\begin{vmatrix}
\mathrm{i}\omega_{\mathrm{c}} - e\left(\frac{1}{M^*}\right)_{xy} B_z & -e\left(\frac{1}{M^*}\right)_{yy} B_z & -e\left(\frac{1}{M^*}\right)_{yz} B_z \\
e\left(\frac{1}{M^*}\right)_{xx} B_z & \mathrm{i}\omega_{\mathrm{c}} + e\left(\frac{1}{M^*}\right)_{xy} B_z & e\left(\frac{1}{M^*}\right)_{xz} B_z \\
0 & 0 & \mathrm{i}\omega_{\mathrm{c}}
\end{vmatrix} = 0\,. \qquad (21.2.31)$$

One solution is, obviously, $\omega_{\mathrm{c}} = 0$. The physically interesting solution is obtained from the equation

$$\omega_{\mathrm{c}}^2 = e^2 B_z^2 \left[\left(\frac{1}{M^*}\right)_{xx}\left(\frac{1}{M^*}\right)_{yy} - \left(\frac{1}{M^*}\right)_{xy}^2\right]. \qquad (21.2.32)$$

By defining the cyclotron mass in the customary way, and changing to the effective-mass tensor M^* from the inverse effective-mass tensor M^{*-1}, (21.2.28) is indeed recovered.

It is worth writing down the result for the case when the coordinate axes are chosen along the principal axes of the ellipsoidal Fermi surface but the magnetic induction does not point in a high-symmetry direction. The effective-mass tensor is then diagonal, so the dispersion relation can be written as

$$\varepsilon_{\boldsymbol{k}} = \frac{\hbar^2 k_1^2}{2m_1^*} + \frac{\hbar^2 k_2^2}{2m_2^*} + \frac{\hbar^2 k_3^2}{2m_3^*}\,, \qquad (21.2.33)$$

and, according to our assumptions, the masses are positive along each of the three directions. Specifying the projections of the magnetic field along the principal axes by the direction cosines α_1, α_2, and α_3:

$$B_1 = B\alpha_1 \,, \qquad B_2 = B\alpha_2 \,, \qquad B_3 = B\alpha_3 \,. \tag{21.2.34}$$

The equations of motion for the components along the principal axes are

$$\begin{aligned} \frac{\mathrm{d}k_1}{\mathrm{d}t} &= -e\frac{k_2}{m_2^*}B_3 + e\frac{k_3}{m_3^*}B_2 \,, \\ \frac{\mathrm{d}k_2}{\mathrm{d}t} &= -e\frac{k_3}{m_3^*}B_1 + e\frac{k_1}{m_1^*}B_3 \,, \\ \frac{\mathrm{d}k_3}{\mathrm{d}t} &= -e\frac{k_1}{m_1^*}B_2 + e\frac{k_2}{m_2^*}B_1 \,. \end{aligned} \tag{21.2.35}$$

Seeking solutions of the form $\boldsymbol{k}\mathrm{e}^{-\mathrm{i}\omega_\mathrm{c}t}$,

$$\begin{aligned} \mathrm{i}\omega_\mathrm{c}k_1 - \frac{eB_3}{m_2^*}k_2 + \frac{eB_2}{m_3^*}k_3 &= 0 \,, \\ \mathrm{i}\omega_\mathrm{c}k_2 - \frac{eB_1}{m_3^*}k_3 + \frac{eB_3}{m_1^*}k_1 &= 0 \,, \\ \mathrm{i}\omega_\mathrm{c}k_3 - \frac{eB_2}{m_1^*}k_1 + \frac{eB_1}{m_2^*}k_2 &= 0 \end{aligned} \tag{21.2.36}$$

is obtained, hence the condition for the existence of nontrivial solutions is

$$\begin{vmatrix} \mathrm{i}\omega_\mathrm{c} & -\dfrac{eB_3}{m_2^*} & +\dfrac{eB_2}{m_3^*} \\ +\dfrac{eB_3}{m_1^*} & \mathrm{i}\omega_\mathrm{c} & -\dfrac{eB_1}{m_3^*} \\ -\dfrac{eB_2}{m_1^*} & +\dfrac{eB_1}{m_2^*} & \mathrm{i}\omega_\mathrm{c} \end{vmatrix} = 0 \,. \tag{21.2.37}$$

The expansion of the determinant leads to a cubic equation in ω_c:

$$\begin{aligned} \mathrm{i}\omega_\mathrm{c}\left[-\omega_\mathrm{c}^2 + \frac{(eB_1)^2}{m_2^* m_3^*}\right] + \frac{eB_3}{m_2^*}\left[\mathrm{i}\omega_\mathrm{c}\frac{eB_3}{m_1^*} - \frac{e^2B_1B_2}{m_1^*m_3^*}\right] \\ + \frac{eB_2}{m_3^*}\left[\mathrm{i}\omega_\mathrm{c}\frac{eB_2}{m_1^*} + \frac{e^2B_1B_3}{m_1^*m_2^*}\right] = 0 \,. \end{aligned} \tag{21.2.38}$$

It is immediately seen that the first solution is $\omega_\mathrm{c} = 0$. The two others are given by

$$\begin{aligned} \omega_\mathrm{c}^2 &= e^2\left[\frac{B_1^2}{m_2^*m_3^*} + \frac{B_2^2}{m_1^*m_3^*} + \frac{B_3^2}{m_1^*m_2^*}\right] \\ &= (eB)^2\frac{m_1^*\alpha_1^2 + m_2^*\alpha_2^2 + m_3^*\alpha_3^2}{m_1^*m_2^*m_3^*} \,. \end{aligned} \tag{21.2.39}$$

The connection between the cyclotron mass and the cyclotron frequency, given in the second equation of (21.2.25), leads to

$$\frac{1}{m_{\mathrm{c}}} = \sqrt{\frac{m_1^* \alpha_1^2 + m_2^* \alpha_2^2 + m_3^* \alpha_3^2}{m_1^* m_2^* m_3^*}} \,. \tag{21.2.40}$$

In the special case when the energy is given by

$$\varepsilon_{\boldsymbol{k}} = \frac{\hbar^2}{2m_\perp^*}(k_x^2 + k_y^2) + \frac{\hbar^2 k_z^2}{2m_\parallel^*} \,, \tag{21.2.41}$$

that is, a longitudinal and a transverse mass are distinguished (as in semiconductors), and the magnetic field makes an angle θ with the z-axis, the cyclotron mass is

$$\frac{1}{m_{\mathrm{c}}} = \sqrt{\frac{\cos^2\theta}{m_\perp^{*\,2}} + \frac{\sin^2\theta}{m_\perp^* m_\parallel^*}} \,. \tag{21.2.42}$$

The cyclotron mass is thus a particular average of the components of the effective-mass tensor.

21.2.5 Cyclotron Resonance

As discussed in the previous subsection, the angular frequency ω_{c} of the periodic motion in a uniform magnetic field is directly related to the cyclotron mass. On the other hand, when the dependence of the cyclotron frequency on the magnetic field direction is known, the components of the effective-mass tensor can be derived. The phenomenon of *cyclotron resonance* offers a straightforward possibility to measure them, and thus, through the effective-mass tensor, to characterize the Fermi surface.

In the measurement setup the sample is placed in a uniform magnetic field B, and an additional weak electric field of frequency ω is applied in a perpendicular direction. In a uniform magnetic field electrons on the Fermi surface are known to move in a helical path whose projection on the plane perpendicular to the field depends on the shape of the Fermi surface, and the frequency of the periodic projected motion is the cyclotron frequency. If the frequency of the alternating electric field is the same, the electron moves in phase with the electric field and can thus gains energy from it. By varying the frequency of the applied field or the magnetic field strength – i.e., the cyclotron frequency –, absorption occurs only when the condition $\omega = \omega_{\mathrm{c}}$ is met.

In real samples electrons moving in the cyclotron orbit undergo collisions and fall out of phase. Therefore even at the frequency $\omega = \omega_{\mathrm{c}}$ they cannot absorb as much energy as a completely free electron gas; on the other hand, absorption is observed not only at $\omega = \omega_{\mathrm{c}}$ but also when the frequency of the electric field differs slightly from ω_{c}. If the collisions are sufficiently infrequent

– that is, the collision time is much larger than the period, $\tau \gg T_{\mathrm{c}}$ ($\omega_{\mathrm{c}}\tau \gg 1$) – then the absorption shows a resonance around ω_{c}. This phenomenon is called *cyclotron resonance* or *diamagnetic resonance.*

To understand the experimental findings, we shall again discuss the motion of electrons in the semiclassical approximation, however, amend (21.1.30), which describes the motion in $\boldsymbol{k}$-space, by a term that accounts for collisions. Collisions are assumed to lead to a finite relaxation time, just as in the Drude model. Consequently, a term $-\boldsymbol{k}/\tau$ appears in the equation of motion, which indicates that in the absence of an applied field the wave vector would tend to $\boldsymbol{k} = 0$ with a relaxation time τ:

$$\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = -\frac{e}{\hbar}\left(\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}\right) - \frac{\boldsymbol{k}}{\tau}\,. \tag{21.2.43}$$

Rewriting the equation in the system of coordinates spanned by the principal axes of the effective-mass tensor:

$$\begin{aligned}
\frac{\mathrm{d}k_1}{\mathrm{d}t} &= -\frac{e}{\hbar}E_1 - e\frac{k_2}{m_2^*}B_3 + e\frac{k_3}{m_3^*}B_2 - \frac{k_1}{\tau}\,,\\
\frac{\mathrm{d}k_2}{\mathrm{d}t} &= -\frac{e}{\hbar}E_2 - e\frac{k_3}{m_3^*}B_1 + e\frac{k_1}{m_1^*}B_3 - \frac{k_2}{\tau}\,,\\
\frac{\mathrm{d}k_3}{\mathrm{d}t} &= -\frac{e}{\hbar}E_3 - e\frac{k_1}{m_1^*}B_2 + e\frac{k_2}{m_2^*}B_1 - \frac{k_3}{\tau}\,.
\end{aligned} \tag{21.2.44}$$

Assuming that in an electric field $\boldsymbol{E} = \boldsymbol{E}_0\mathrm{e}^{-\mathrm{i}\omega t}$ of frequency ω the solution will also be periodic with the same frequency, the equation of motion reads

$$-\,\mathrm{i}\left(\omega + \frac{\mathrm{i}}{\tau}\right)k_1 = -\frac{e}{\hbar}E_1 - e\frac{k_2}{m_2^*}B_3 + e\frac{k_3}{m_3^*}B_2\,, \tag{21.2.45-a}$$

$$-\mathrm{i}\left(\omega + \frac{\mathrm{i}}{\tau}\right)k_2 = -\frac{e}{\hbar}E_2 - e\frac{k_3}{m_3^*}B_1 + e\frac{k_1}{m_1^*}B_3\,, \tag{21.2.45-b}$$

$$-\mathrm{i}\left(\omega + \frac{\mathrm{i}}{\tau}\right)k_3 = -\frac{e}{\hbar}E_3 - e\frac{k_1}{m_1^*}B_2 + e\frac{k_2}{m_2^*}B_1\,. \tag{21.2.45-c}$$

By solving the set of equations that is now inhomogeneous because of the presence of the electric field, and exploiting the property that $\boldsymbol{k}$ varies in the plane perpendicular to $\boldsymbol{B}$,

$$\hbar k_1 = \frac{1}{-(\omega + \mathrm{i}/\tau)^2 + \omega_{\mathrm{c}}^2}\left[\mathrm{i}eE_1\left(\omega + \frac{\mathrm{i}}{\tau}\right) + eE_2\frac{eB_3}{m_2^*} - eE_3\frac{eB_2}{m_3^*}\right] \tag{21.2.46}$$

is obtained, where ω_{c} is given by (21.2.39) – in other words, it is the cyclotron frequency associated with the cyclotron mass specified by (21.2.40). Similar equations are valid for the other components, too. Denoting the electron density by n_{e}, the component of the current carried by the electrons along the 1-axis is

$$j_1 = -en_\mathrm{e}v_1 = -\frac{en_\mathrm{e}}{m_1^*}\hbar k_1 \,. \tag{21.2.47}$$

The σ_{11} element of the complex conductivity tensor is then

$$\sigma_{11} = \frac{e^2\,n_\mathrm{e}}{m_1^*}\frac{-\mathrm{i}\omega + 1/\tau}{-(\omega + \mathrm{i}/\tau)^2 + \omega_\mathrm{c}^2} = \sigma_0\frac{1 - \mathrm{i}\omega\tau}{1 + (\omega_\mathrm{c}^2 - \omega^2)\tau^2 - 2\mathrm{i}\omega\tau} \,, \tag{21.2.48}$$

where $\sigma_0 = n_\mathrm{e}e^2\tau/m_1^*$ is the conductivity of electrons of effective mass m_1^* in the Drude model. Figure 21.11 shows how the real part of σ_{11} depends on the strength of the magnetic field for different values of $\omega\tau$. When $\omega\tau \gg 1$, a sharp resonance occurs at $\omega = \omega_\mathrm{c}$, while if $\omega\tau < 1$, the curve is completely featureless. This is in agreement with the previous physical picture that electrons can absorb energy from the field and a resonance can be built up only when the electrons are accelerated by the field during several periods – that is, the mean time τ between collisions is much larger than the period.

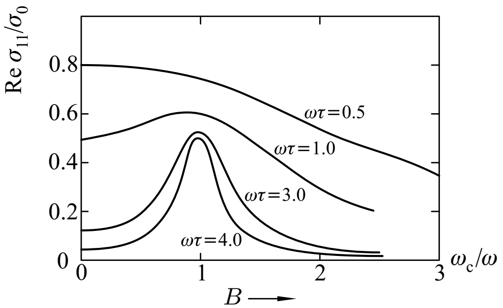


Fig. 21.11. The dependence of absorption on the strength of the magnetic field for various values of $\omega\tau$. The variation of the field is expressed in terms of the variation of ω_c

The quantum mechanical treatment of the electron states and transitions would lead to the same results as the semiclassical approach above. We shall demonstrate in the next chapter that in a uniform magnetic field the part of the kinetic energy that comes from the motion perpendicular to the magnetic field gives rise to discrete energy levels spaced at a regular distance $\hbar\omega_\mathrm{c}$. Considering the electric field of frequency ω as a perturbation, it can lead to a transition between two such levels if its frequency is the same as ω_c. When this condition is met, absorption may occur. When electrons undergo collisions, energy levels become broadened in the quantum mechanical picture, and the absorption peak is consequently broadened, too.

In experiments, the surface resistance of the sample, the absorption of the electromagnetic field, or the reflection is measured. In fields of order 10^{-1} tesla the cyclotron frequency is in the microwave region (GHz frequencies or cm wavelengths). When the energy absorbed from the microwave field is measured, resonances are observed. The method is particularly useful for semiconductors, where the dispersion relation can be relatively well approximated by a quadratic expression at the bottom of the conduction band and the top of the valence band. The longitudinal and transverse effective masses can then be determined easily. The results of the first successful measurements, performed on semiconducting germanium, are shown in Fig. 21.12.

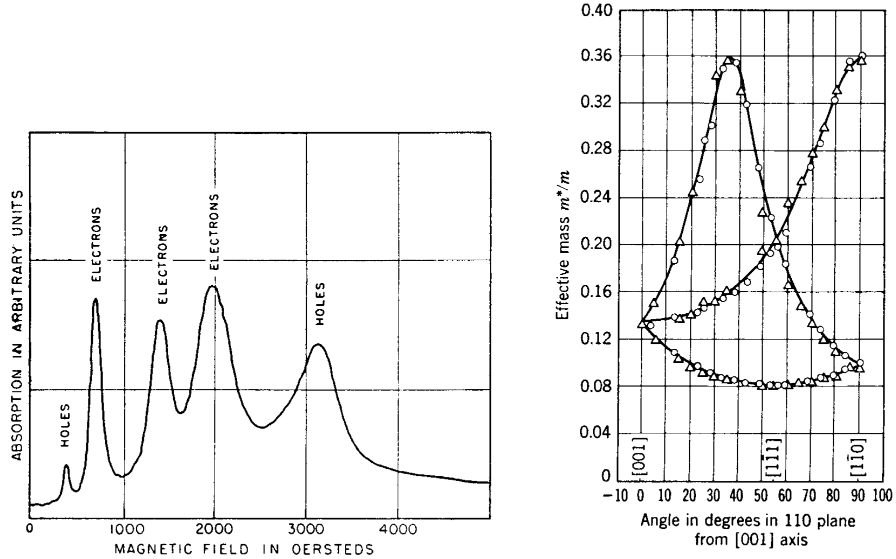


Fig. 21.12. Cyclotron resonance peaks in germanium due to electrons and holes, and the dependence of the cyclotron mass on the field direction [Reprinted with permission from G. Dresselhaus et al., *Phys. Rev.* **98**, 368 (1955). ©1955 by the American Physical Society]

21.2.6 Azbel–Kaner Resonance

The previous calculation and the experimental setup used for cyclotron-resonance measurements in semiconductors with a time-dependent but spatially uniform electric field cannot be used for the determination of the effective mass of Bloch electrons in metals. This is because our assumption that the electron feels the accelerating electric field all along its periodic orbit is invalid in metals. The high-frequency electric field used for the detection of the periodic motion in the magnetic field penetrates to a depth δ into the sample.[3] The skin depth is typically on the order of 10^{-5} to 10^{-6} cm, whereas in a field of $B \sim 1$ tesla electrons move in a circular orbit of radius $r_{\mathrm{c}} \sim 10^{-3}$ cm, thus

[3] This is the skin depth observed in the skin effect.

in the largest part of their orbits they do not feel the electric field. Nonetheless M. I. Azbel and E. A. Kaner demonstrated in 1956 that in a suitably chosen geometry resonance may also occur in metals. This phenomenon is called the *Azbel–Kaner resonance.*

If the applied magnetic field is parallel to the surface and the field direction is chosen as the x-axis, then the electrons move in a helical path. They propagate freely in the z direction, and trace out a circle in the yz-plane. They can stay close to the surface only during a small fraction of their motion; the rest of the trajectory is far from the surface.

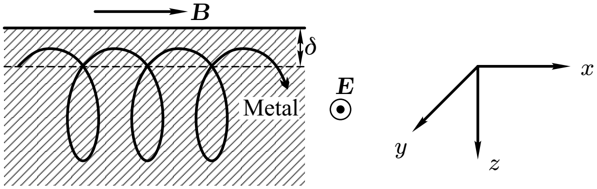


Fig. 21.13. The geometry of the Azbel–Kaner resonance

The applied accelerating electric field is along the y-axis – that is, also parallel to the surface but perpendicular to $\boldsymbol{B}$ (crossed-field setup). Then those electrons that approach the surface by δ or less can absorb energy from the electric field. Absorption will be resonance-like if the electric field is always in the same phase when the electron returns to the skin layer after a cycle. The condition for this is that the period T_c of the circular motion of the electron should be an integral multiple of the period T_E of the alternating electric field:

$$T_\mathrm{c} = nT_E \,. \tag{21.2.49}$$

Then resonance occurs for those frequencies ω that satisfy

$$\omega = n\,\omega_\mathrm{c} \,. \tag{21.2.50}$$

Customarily, it is the magnetic field strength rather than the frequency of the electric field that is varied in measurements of the Azbel–Kaner resonance. Expressed in terms of B, resonance occurs for those values B_n that satisfy

$$\frac{1}{B_n} = n\frac{2\pi e}{\hbar^2\omega}\left(\frac{\partial\mathcal{A}}{\partial\varepsilon}\right)^{-1} \,. \tag{21.2.51}$$

While in the ordinary setup cyclotron resonance occurs only at a single value of the magnetic field or frequency, in the Azbel–Kaner resonance a whole set of absorption peaks are observed. An example is shown in Fig. 21.14. Note that the reciprocals of the corresponding values of the magnetic field ($1/B$) are regularly spaced. In this case, too, the resonance is sharp only when the relaxation time of electrons is much larger than the period of their motion, i.e.,

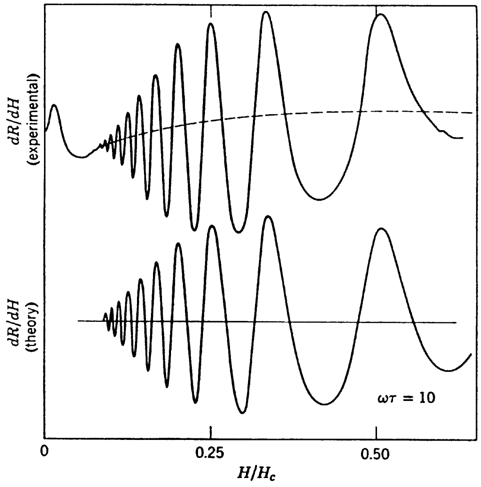


Fig. 21.14. Azbel–Kaner resonance in copper, compared to the theoretical results of Azbel and Kaner for the derivative of the resonance absorption, for $\omega\tau = 10$ [Reprinted with permission from A. F. Kip et al., *Phys. Rev.* **124**, 359 (1961). ©1961 by the American Physical Society]

$\omega_c\tau \gg 1$. When the field strength is reduced, ω_c becomes smaller, too, and the resonance disappears when $\omega_c\tau \sim 1$. The figure also shows the disappearance of the resonance.

For ellipsoidal Fermi surfaces the frequency is known to depend on the relative orientation of the magnetic field and the principal axes of the ellipsoid but not on the particular section of the Fermi surface on which electrons move. For nonellipsoidal Fermi surfaces, like the one shown in Fig. 21.15(*a*), the frequency is different for each section, and only a small number of electrons contribute to each of them. Measurable effects are due to electrons in those parts of the Fermi surface where the period varies little over a relatively wide region – that is, where the area of the cross section has an extremum. Two

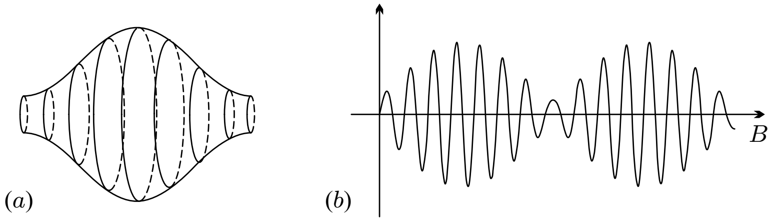


Fig. 21.15. (*a*) Nonellipsoidal Fermi surface with two extremal cross sections. (*b*) The corresponding Azbel–Kaner resonance featuring two characteristic frequencies

such loci are found on the Fermi surface in Fig. 21.15(a): a minimal and a maximal cross section. Consequently there are two characteristic dominant frequencies. This leads to the resonance curve plotted in Fig. 21.15(b).

21.2.7 Magnetoacoustic Geometric Oscillations

Instead of an electromagnetic field, ultrasound can also be used to obtain information about the shape and size of the Fermi surface. This is because in the semiclassical limit real-space electron trajectories are similar to the $\boldsymbol{k}$-space orbits, just the scales are different.

Consider a sample placed in a z-directed uniform magnetic field in which transverse lattice vibrations propagating perpendicularly to the field (in the x direction) are generated by ultrasound with a frequency in the MHz region. The transverse vibration of the ions in the crystal creates an oscillating electric field whose wavelength λ is much larger than the atomic dimensions. If the cyclotron frequency associated with the magnetic field is much larger than the frequency of the ultrasound, then the electric field of the ultrasound remains practically unchanged while the electron makes a full cycle in the magnetic field. If globally – that is, summed over the entire trajectory – the electron experiences a net accelerating force, it can absorb energy. This is expected to occur when the electron is accelerated maximally in the turning points of the real-space orbit. This is shown schematically in Fig. 21.16.

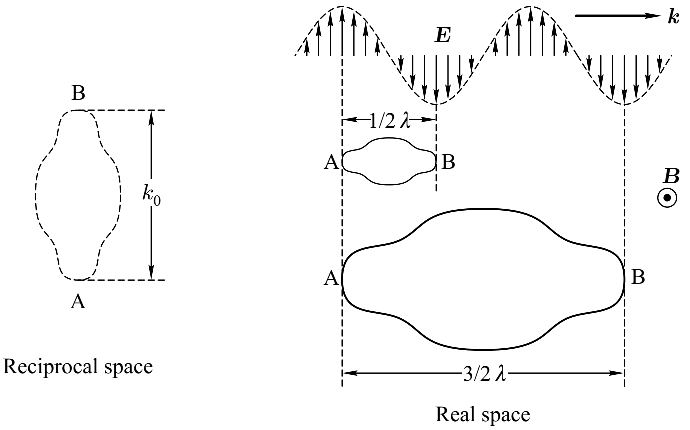


Fig. 21.16. Semiclassical orbit of an electron in reciprocal space, and the interaction with the electric field created by the ultrasound in real space

Observable sound attenuation requires that the electron complete several cycles between two subsequent collisions, that is, $\omega_c\tau \gg 1$. This occurs only in sufficiently pure samples.

According to this simple picture an electron can absorb energy from the ultrasound if the extremal diameter of the real-space orbit in the direction of sound propagation is a half-integral multiple of the wavelength of the sound:

$$d_{\text{ext}} = (n + 1/2)\lambda\,. \tag{21.2.52}$$

More rigorous calculations lead to a slightly different location of the absorption maximum than suggested by this simple physical picture. Assuming that the Fermi surface is spherical, electrons move in circular orbits in real space, too, whose radius is $\hbar/eB$ times the radius of the $\boldsymbol{k}$-space orbit. In the extremal cross section of the Fermi surface, the equation for the real-space orbit of an electron moving at a frequency ω_{c} is

$$x(t) = r\cos\omega_{\text{c}}t\,, \qquad y(t) = r\sin\omega_{\text{c}}t\,, \tag{21.2.53}$$

where

$$r = \frac{\hbar}{eB}k_{\text{F}}\,. \tag{21.2.54}$$

The energy absorbed in one cycle by the electron from the electric field $E_y\mathrm{e}^{\mathrm{i}qx}$ created in the direction perpendicular to the propagation of the sound wave is

$$I \sim \int\limits_0^{T_{\text{c}}} \boldsymbol{E}(t)\cdot\boldsymbol{v}(t)\,\mathrm{d}t = \int\limits_0^{T_{\text{c}}} E_y\mathrm{e}^{\mathrm{i}qx(t)}v_y\,\mathrm{d}t\,. \tag{21.2.55}$$

Using the formula $v_y(t) = v_0\cos\omega_{\text{c}}t$ for the velocity, the previous integral can be expressed in terms of the Bessel function J_1 as

$$\begin{aligned} I &\sim \int\limits_0^{T_{\text{c}}} E_y\mathrm{e}^{\mathrm{i}qr\cos\omega_{\text{c}}t}v_0\cos\omega_{\text{c}}t\,\mathrm{d}t \\ &= \frac{E_yv_0}{\omega_{\text{c}}}\int\limits_0^{2\pi}\mathrm{e}^{\mathrm{i}qr\cos\xi}\cos\xi\,\mathrm{d}\xi = \frac{E_yv_0}{\omega_{\text{c}}}2\pi J_1(qr)\,. \end{aligned} \tag{21.2.56}$$

When q is expressed by the wavelength of the sound wave and its real-space radius by k_{F}, relatively wide maxima appear at

$$\frac{\hbar k_{\text{F}}}{eB}\frac{1}{\lambda} = 1.22,\ 2.23,\ 3.24,\ \ldots, n + 5/8\,. \tag{21.2.57}$$

The real- and reciprocal-space orbits are similar in shape in more general cases, too, and their relative scale factor is $\hbar/eB$, therefore the dimensions of the Fermi surface can be inferred from the size of the real-space orbit in such cases as well. Since the most important contributions are due to electrons whose wave vector moves along the section of the Fermi surface of extremal diameter, the latter can be immediately determined from these measurements using the known wavelength of the sound wave.

Figure 21.17 shows the experimental results for high-purity magnesium. Measurements in different directions provide information about various sections of the Fermi surface. The oscillations shown in the figure correspond to different sections of the arm of the monster-shaped Fermi surface (see Fig. 19.7) of divalent metals with hcp structure, such as magnesium.

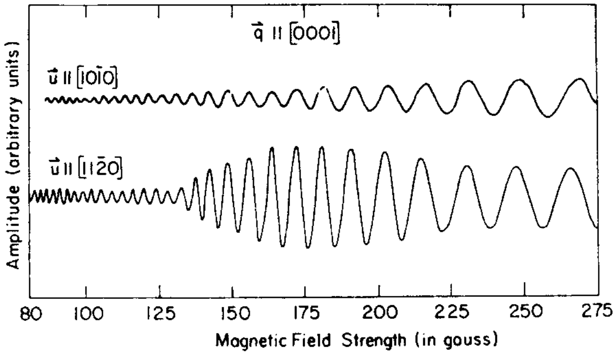


Fig. 21.17. Magnetoacoustic geometric resonance oscillations measured in high-purity magnesium [Reprinted with permission from J. B. Ketterson and R. W. Stark, *Phys. Rev.* **156**, 748 (1967). ©1967 by the American Physical Society]

Note that in stronger magnetic fields very sharp peaks can be observed in the ultrasonic attenuation. Figure 21.18 shows such experimental results for gallium. The correct description of the phenomenon and the specification of what information can be obtained for the Fermi surface from the location of the absorption peaks requires the quantum mechanical description of electrons

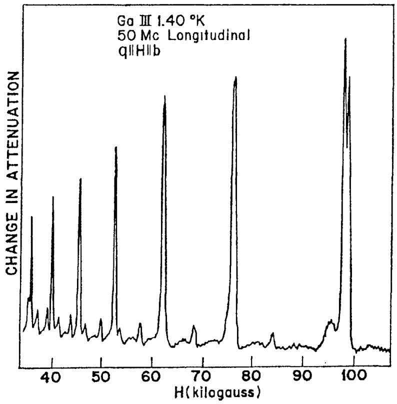


Fig. 21.18. Giant quantum oscillations in the ultrasonic absorption spectrum of gallium [Reprinted with permission from V. Shapira and B. Lax, *Phys. Rev.* **138**, A 1191 (1965). ©1965 by the American Physical Society]

in strong magnetic fields. That is why the phenomenon is called *giant quantum oscillation*.

21.3 Size Effects

It was assumed in the foregoing that the thickness of the sample is much larger than the diameter of the cyclotron orbit. If this is not the case, new methods become available for determining the properties of the Fermi surface. Since these experimental methods are based on the requirement that the size of the sample and the cyclotron orbit be somehow matched, they are collectively called *size effects*.

21.3.1 Extinction of the Resonance in Thin Samples

In the analysis of the Azbel–Kaner resonance we assumed that electrons repeatedly return to the skin layer during their periodic motion. When the applied magnetic field is strong and the dimensions of the (real-space) orbit are small, this indeed occurs with a high probability. However, as the magnetic field becomes weaker, the orbit becomes larger and larger, and at a certain point its diameter becomes comparable to the thickness of the sample. The electron may then be reflected by the other surface and leave the cyclotron orbit, leading to the disappearance of the resonance, as pointed out by E. A. Kaner in 1958. Such a situation is illustrated in Fig. 21.19. Plotted as a function of $1/B$, resonances follow each other at regular distances up to a point, where they disappear. When the thickness of the sample is known, the Fermi momentum can be determined from the value of the field in the point where the resonances disappear.

The same geometry is used as in the description of the Azbel–Kaner resonance: the homogeneous magnetic field is directed along the x-axis and the high-frequency electric field along the y-axis. The only difference is that the sample is now of finite width (d) in the z-direction. Consider the situation shown in Fig. 21.20, when the nth cyclotron orbit, which satisfies the condition

$$n\frac{eB}{m_{\mathrm{c}}} = \omega\,, \tag{21.3.1}$$

is just inside the sample – that is, the nth resonance can be observed but not the $n+1$th.

We shall now integrate the equation of motion in the y-direction,

$$\hbar\frac{\mathrm{d}k_y}{\mathrm{d}t} = -ev_zB\,, \tag{21.3.2}$$

over half a period, from the top of the orbit to its bottom; the displacement of the electron is then just the diameter d_n of the orbit:

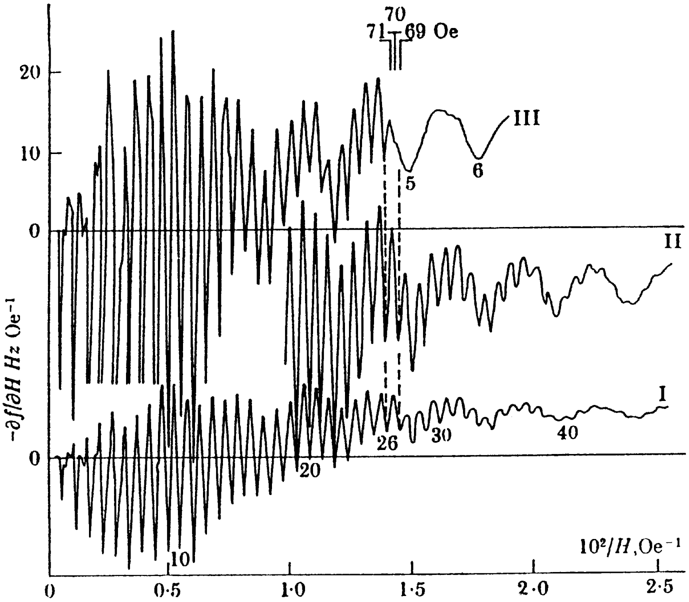


Fig. 21.19. Azbel–Kaner resonances recorded on single crystals of tin of a few mm thickness. As the field strength decreases, the resonance disappears after the 26th peak in the thinnest sample [M. S. Khaikin, *Soviet Physics JETP*, **14**, 1260 (1962)]

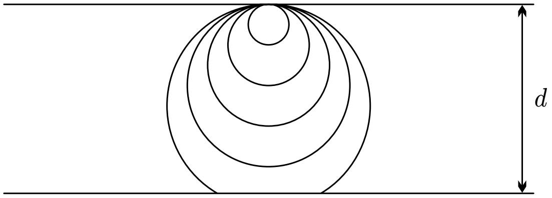


Fig. 21.20. Cyclotron orbits in a sample of finite width in a perpendicular magnetic field. The circles are associated with Azbel–Kaner resonances of different indices

$$k_y(T/2) - k_y(0) = -\frac{eB}{\hbar}\int\limits_0^{T/2} v_z(t')\mathrm{d}t' = -\frac{eB}{\hbar}d_n\,. \qquad (21.3.3)$$

By eliminating the magnetic field using the resonance condition (21.3.1), we have

$$k_y(T/2) - k_y(0) = -\frac{m_{\mathrm{c}}\,\omega}{n\hbar}\,d_n\,. \qquad (21.3.4)$$

If the resonances disappear at a sufficiently large value of n, d_n can be practically identified with the thickness of the sample, so $k_y(T/2) - k_y(0)$ can be determined from the measurements. For electrons that contribute to the resonance both $k_y(0)$ and $k_y(T/2)$ are the quasimomenta associated with extremal cross sections of the Fermi surface, just they are oppositely directed. Therefore this method allows the measurement of the Fermi momentum.

21.3.2 Radiofrequency Size Effect

Setting up a resonance does not require the application of GHz microwaves. It can also be achieved using a RF field of a few MHz, polarized perpendicular to the magnetic field, provided the frequency is sufficiently high for that the skin depth be small compared to the sample thickness. This phenomenon is called the *Gantmakher effect.*[4]

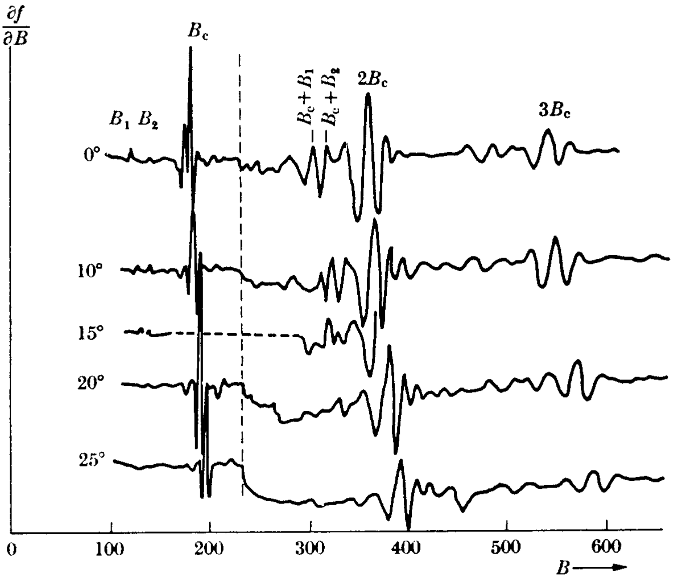


Fig. 21.21. Anomalies of the surface impedance in a high-purity single crystal of tin of thickness 0.4 mm at a frequency of 3.06 MHz, measured at several angles between the direction of the uniform magnetic field and the [001] crystallographic direction [V. F. Gantmakher, *Soviet Physics JETP*, **16**, 247 (1963)]

When the electron on the helical path returns to the skin layer after a full cycle, it will feel that, because of the low frequency, the phase of the RF field has practically not changed. The electron can thus absorb energy coherently. This is true as long as the real-space diameter of the cyclotron orbit is smaller than the thickness of the sample, that is,

[4] V. F. Gantmakher, 1962.

$$\frac{\hbar}{eB}\Delta k < d\,. \tag{21.3.5}$$

As the magnetic field becomes weaker, the radius of the semiclassical orbit increases. Anomaly appears in the absorption when the real-space orbit associated with an extremal diameter of the Fermi surface no longer fits in the sample. The extremal diameter of the Fermi surface can then be determined from the thickness of the sample and the appropriate field B_c using the formula

$$\Delta k_\mathrm{c} = \frac{eB_\mathrm{c}}{\hbar}d\,. \tag{21.3.6}$$

As shown in Fig. 21.21, the anomaly does not appear at a single well-defined value B_c but also at its integral multiples. This can be illustrated by the following picture: The electrons that pass close to the surface of the sample, and get accelerated there, penetrate into the sample, and turn back in depth $\hbar\Delta k_\mathrm{c}/eB$. Around this depth they create a thin layer, whose width is the same as the skin depth, in which the current density is high, and this can accelerate other electrons traversing the layer. This geometry is shown in Fig. 21.22.

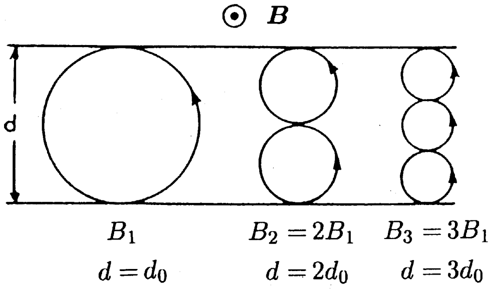


Fig. 21.22. Orbits of electrons accelerated inside the sample by the electric field of the primary electrons

When the trajectory of such secondary electrons reaches the other side of the slab, electromagnetic radiation is emanated from there, and the impedance increases. This transfer of energy through the thin sample may occur in multiple steps as well. Consequently, anomaly is observed at magnetic fields $B_n = nB_\mathrm{c}$ for which the extremal diameter d_0 is just $1/n$ of the sample thickness. This phenomenon provides one of the best ways to determine the Fermi momentum.

21.4 Limitations of the Semiclassical Description

Throughout the previous sections we assumed the applicability of the semiclassical description based on a single wave packet. We shall now discuss the limitations of this approach, examine the conditions that the wavelength, frequency, and amplitude of the electromagnetic field must satisfy. Finally, we shall briefly show how the interband transitions – which are completely neglected in the semiclassical treatment – appear at higher field strengths.

21.4.1 Conditions of the Applicability of Semiclassical Dynamics

The spatial extent of a wave packet

$$\phi(\boldsymbol{r},t) = \sum_{\boldsymbol{k}'} g(\boldsymbol{k}') \exp\left[\mathrm{i}\left(\boldsymbol{k}' \cdot \boldsymbol{r} - \varepsilon_{n\boldsymbol{k}'} t/\hbar\right)\right] \tag{21.4.1}$$

made up of plane waves is known to be determined by the spread of the $\boldsymbol{k}'$-sum in momentum space. If $g(\boldsymbol{k}')$ is substantially different from zero inside a sphere of radius Δk around a vector $\boldsymbol{k}$, then the state is localized within a region of linear dimension $\Delta r \sim 1/\Delta k$ in real space.

When Bloch states are combined to form wave packets, as in (21.1.3), the Bloch form of the wavefunction leads to

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r},t) = \sum_{\boldsymbol{k}'} g(\boldsymbol{k}') u_{n\boldsymbol{k}'}(\boldsymbol{r}) \exp\left[\mathrm{i}\left(\boldsymbol{k}' \cdot \boldsymbol{r} - \varepsilon_{n\boldsymbol{k}'} t/\hbar\right)\right] . \tag{21.4.2}$$

According to our assumption, $g(\boldsymbol{k}')$ is substantially different from zero only inside a sphere of radius Δk around $\boldsymbol{k}$. If, moreover, $u_{n\boldsymbol{k}'}(\boldsymbol{r})$ varies slowly with $\boldsymbol{k}'$ in the same region, then we can use the form

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r},t) \approx u_{n\boldsymbol{k}}(\boldsymbol{r}) \sum_{\boldsymbol{k}'} g(\boldsymbol{k}') \exp\left[\mathrm{i}\left(\boldsymbol{k}' \cdot \boldsymbol{r} - \varepsilon_{n\boldsymbol{k}'} t/\hbar\right)\right] , \tag{21.4.3}$$

so the wavefunction is now written as a combination of plane waves with coefficients $g(\boldsymbol{k}')u_{n\boldsymbol{k}}(\boldsymbol{r})$. The state is thus confined to a region of linear dimension $\Delta r \sim 1/\Delta k$ in real space.

If we wish to associate this wave packet with a $\boldsymbol{k}$ vector inside the Brillouin zone then the wave packet must be made up of states $\boldsymbol{k}'$ for which Δk is much smaller than the size of the zone – that is, the condition $\Delta k \ll 1/a$ has to be satisfied, where a is the lattice constant. Then $\Delta r \gg a$, so the size of the wave packet is much larger than the lattice constant: it extends over many primitive cells.

As far as the dynamics of electrons is concerned, the wave packet can be considered point-like provided the variation of the applied field is slow on the scale of the wave packet. For external fields of the form $\boldsymbol{E}_0 \exp[\mathrm{i}(\boldsymbol{q} \cdot \boldsymbol{r} - \omega t)]$ this implies $|\boldsymbol{q}| \ll \Delta k$, that is, the condition is met by long-wavelength applied fields.

Limits can be set for the field strength E and the frequency ω of the field. These are derived from the requirement that interband transitions should be either forbidden or negligible. By absorbing a photon from the field, the electron can jump to a higher band. To prevent this, the photon energy needs to be smaller than the band gap ε_{g}:

$$\hbar\omega < \varepsilon_{\mathrm{g}} \,. \tag{21.4.4}$$

The condition for the field strength can be derived from the time-dependent Schrödinger equation. Taking the electric field into account through the scalar potential $-\boldsymbol{E}\cdot\boldsymbol{r}$, we have

$$-\frac{\hbar}{\mathrm{i}}\frac{\partial\psi(\boldsymbol{r},t)}{\partial t} = \left[\mathcal{H}_0 + e\boldsymbol{E}\cdot\boldsymbol{r}\right]\psi(\boldsymbol{r},t)\,, \tag{21.4.5}$$

where $\mathcal{H}_0$ contains the kinetic-energy operator and the periodic potential.

Now consider a Bloch electron whose wavefunction would be

$$\psi_{\boldsymbol{k}}(\boldsymbol{r},t) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_{\boldsymbol{k}}(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}\varepsilon_{\boldsymbol{k}}t/\hbar} \tag{21.4.6}$$

in the absence of an electric field, and suppose that the effects of the applied field can all be lumped into the variation of the wave vector with time. If the state is characterized by a wave vector $\boldsymbol{k}$ at time $t=0$ then its time evolution around $t=0$ is governed by

$$-\frac{\hbar}{\mathrm{i}}\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r},t)}{\partial t} = \left[\varepsilon_{\boldsymbol{k}} + e\boldsymbol{E}\cdot\boldsymbol{r}\right]\psi_{\boldsymbol{k}}(\boldsymbol{r},t)\,. \tag{21.4.7}$$

On the other hand, the wave vector itself also shows explicit time dependence. Making use of the Bloch form of the wavefunction, we have

$$\begin{aligned}
-\frac{\hbar}{\mathrm{i}}\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r},t)}{\partial t} &= -\frac{\hbar}{\mathrm{i}}\left(\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r},t)}{\partial t}\right)_{\boldsymbol{k}} - \frac{\hbar}{\mathrm{i}}\left(\frac{\partial\psi_{\boldsymbol{k}}(\boldsymbol{r},t)}{\partial \boldsymbol{k}}\right)_{t}\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} \\
&= \varepsilon_{\boldsymbol{k}}\psi_{\boldsymbol{k}}(\boldsymbol{r},t) + \frac{1}{\mathrm{i}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}e\boldsymbol{E}\cdot\left(\mathrm{i}\boldsymbol{r} + \frac{\partial}{\partial\boldsymbol{k}}\right)u_{\boldsymbol{k}}(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}\varepsilon_{\boldsymbol{k}}t/\hbar} \\
&= \left(\varepsilon_{\boldsymbol{k}} + e\boldsymbol{E}\cdot\boldsymbol{r} + \frac{1}{\mathrm{i}}e\boldsymbol{E}\cdot\frac{\partial\ln u_{\boldsymbol{k}}(\boldsymbol{r})}{\partial\boldsymbol{k}}\right)\psi_{\boldsymbol{k}}(\boldsymbol{r},t)\,.
\end{aligned} \tag{21.4.8}$$

Compared to the previous formula, an additional term is present. The semiclassical approximation can be applied when the contribution of this term is negligible.

It was established in the nearly-free-electron approximation that the leading correction to the band gap is proportional to the periodic potential. Assuming that the periodic potential opens a gap at the zone boundary, the distortion of the dispersion relation is appreciable in a region of width Δk for which

$$\varepsilon_{\mathrm{g}} \approx \frac{\partial\varepsilon_k}{\partial k}\Delta k\,. \tag{21.4.9}$$

Inside this region, the deviations from the plane-wave form are important: the change in $u_{\boldsymbol{k}}$ may be comparable to $u_{\boldsymbol{k}}$ itself, and so

$$\frac{\partial \ln u_{\boldsymbol{k}}(\boldsymbol{r})}{\partial \boldsymbol{k}} \approx \frac{1}{\Delta k} \approx \frac{\hbar^2 k}{m_{\mathrm{e}} \varepsilon_{\mathrm{g}}} \,. \tag{21.4.10}$$

The contribution of the new term in (21.4.8) is therefore

$$eE \frac{\hbar^2 k}{m_{\mathrm{e}} \varepsilon_{\mathrm{g}}} \,. \tag{21.4.11}$$

It can be neglected if it is smaller than the band gap,

$$eE \frac{\hbar^2 k}{m_{\mathrm{e}} \varepsilon_{\mathrm{g}}} \ll \varepsilon_{\mathrm{g}} \,. \tag{21.4.12}$$

Apart from factors of order unity, for wave vectors on the zone boundary we have

$$k \approx 1/a \,, \qquad \frac{\hbar^2 k^2}{2 m_{\mathrm{e}}} \approx \varepsilon_{\mathrm{F}} \,, \tag{21.4.13}$$

where a is the lattice constant. The interband transitions induced by the electric field can thus be neglected when

$$eEa \ll \frac{\varepsilon_{\mathrm{g}}^2}{\varepsilon_{\mathrm{F}}} \,. \tag{21.4.14}$$

The limitations imposed on the magnetic field can be treated analogously. Then the condition

$$e\boldsymbol{v} \times \boldsymbol{B} \cdot \frac{\partial \ln u_{\boldsymbol{k}}(\boldsymbol{r})}{\partial \boldsymbol{k}} \ll \varepsilon_{\mathrm{g}} \tag{21.4.15}$$

has to be met. Using the previous formulas, this is equivalent to

$$\hbar \omega_{\mathrm{c}} \ll \frac{\varepsilon_{\mathrm{g}}^2}{\varepsilon_{\mathrm{F}}} \,. \tag{21.4.16}$$

21.4.2 Electric and Magnetic Breakdown

The above condition for the magnitude of the electric field is always met in metals. When both the current density and the resistivity are chosen large, $10^2\,\mathrm{A/cm^2}$ and $100\,\mu\Omega\,\mathrm{cm}$, respectively, the electric field strength is $E = \rho j \sim 10^{-2}\,\mathrm{V/cm}$. The variation of the energy over distances that are comparable to the atomic spacing $a \sim 10^{-8}\,\mathrm{cm}$ is $eEa \sim 10^{-10}\,\mathrm{eV}$. Since the Fermi energy ε_{F} is on the order of an eV, interband transitions are highly improbable for any reasonable value of the separation between bands, since the condition is met for any gap in excess of $10^{-5}\,\mathrm{eV}$. The electric field can be several orders of magnitude stronger in insulators and semiconductors than in metals, therefore an applied electric field can stimulate interband transitions in them. This phenomenon is called electric breakdown. As C. Zener (1934) pointed out, it can be interpreted as a tunneling across the potential barrier arising from the gap between the two bands. For this reason the phenomenon is also called *Zener tunneling.*

To estimate the probability of the transition by tunneling, it should be noted that the energy of electron states is shifted by $e\boldsymbol{E}\cdot\boldsymbol{r}$, and thus becomes position-dependent in the presence of an electric field. The state at the top of the lower band can tunnel to a state at the bottom of the upper band at a distance x_0 if their energies are equal. The distance x_0 is determined from the condition $eEx_0 = \varepsilon_{\mathrm{g}}$. The tunneling probability for a particle of mass m^* and energy ε through a potential barrier $U(x)$ is

$$P \propto \exp\left\{ -\frac{2}{\hbar}\int\limits_0^{x_0} \sqrt{2m^*[U(x)-\varepsilon]}\,\mathrm{d}x \right\}, \tag{21.4.17}$$

which leads to the estimate

$$P \propto \exp\left[-C\frac{\sqrt{2m^*}}{\hbar}\frac{\varepsilon_{\mathrm{g}}^{3/2}}{eE} \right] \tag{21.4.18}$$

with a coefficient C of order unity. Applying this to a state close to the zone boundary, and making use of (21.4.13), the tunneling probability can be rewritten as

$$P \propto \exp\left[-C\frac{\varepsilon_{\mathrm{g}}}{eEa}\left(\frac{\varepsilon_{\mathrm{g}}}{\varepsilon_{\mathrm{F}}}\right)^{1/2} \right]. \tag{21.4.19}$$

More precise calculations lead to the result

$$P \propto \exp\left[-C\frac{\varepsilon_{\mathrm{g}}^2}{eEa\varepsilon_{\mathrm{F}}} \right]. \tag{21.4.20}$$

Comparison with (21.4.14) shows that if the electric field strength does not meet the condition of the applicability of the semiclassical approximation – that is, eEa becomes comparable to $\varepsilon_{\mathrm{g}}^2/\varepsilon_{\mathrm{F}}$ –, then Zener tunneling may appear.

Repeating the calculation in the presence of a magnetic field, the tunneling probability is found to be

$$P \propto \exp\left[-C\frac{\varepsilon_{\mathrm{g}}^2}{\hbar\omega_{\mathrm{c}}\varepsilon_{\mathrm{F}}} \right]. \tag{21.4.21}$$

Once again, when the condition for the applicability of the semiclassical approximation is not met, the electron may tunnel to a nearby state in another band (where nearby refers to the distance in real space as well as reciprocal space). In a magnetic field of $B \sim 1$ tesla, $\hbar\omega_{\mathrm{c}} \sim 10^{-4}\,\mathrm{eV}$. Therefore magnetic breakdown may occur even if the band gap is as small as $10^{-2}\,\mathrm{eV}$. Note that this can occur even at the low temperatures used in cyclotron resonance measurements, where thermal excitation can hardly induce interband transitions across such a gap.

Looking back at the cases shown in Figs. 21.6 and 21.7, we can conclude that if an electron reaches the zone boundary and does not continue its semiclassical trajectory from the equivalent point across the zone but jumps to a nearby state in another band, then, on account of these jumps, it moves as if it hardly felt the periodic potential. The orbit will be a slightly deformed circle, almost like for free electrons; it will encircle a larger area in $\boldsymbol{k}$-space than the semiclassical orbits, and the period will show an abrupt increase.

Further Reading

1. A. A. Abrikosov, *Fundamentals of the Theory of Metals*, North Holland, Amsterdam (1988).
2. A. P. Cracknell and K. C. Wong, *The Fermi Surface, Its Concept, Determination and Use in the Physics of Metals*, Clarendon Press, Oxford (1973).
3. I. M. Lifshitz, M. Ya. Azbel' and M. I. Kaganov, *Electron Theory of Metals*, Consultants Bureau, New York (1973).
4. W. Mercouroff, *La surface de Fermi des métaux*, Collection de Monographies de Physique, Masson et C$^{\text{ie}}$, Éditeurs, Paris (1967).

22

Electrons in Strong Magnetic Fields

The semiclassical treatment of the dynamics of electrons is justified only in relatively weak magnetic fields. Using present-day technology it is quite easy to produce strong fields in which the conditions derived in the previous chapter are not met. In such uniform static magnetic fields interband transitions can occur. Even more importantly, the wavefunctions of electron states meet the condition (21.1.28) only in relatively weak magnetic fields: in stronger fields the state can no longer be characterized by a wave vector $\boldsymbol{k}$ as a quantum number, and we have to solve the complete quantum mechanical problem. As mentioned earlier, this problem cannot be solved exactly in general. However, when the periodic lattice potential can be ignored, an exact solution becomes possible. Therefore we shall first calculate the energy spectrum of a free-electron gas in a magnetic field, and then try to generalize the results to Bloch electrons. Using the one-particle spectrum, we shall determine the ground-state energy of the electron gas as well as its finite-temperature free energy. Both of them show oscillations as functions of the magnetic field, and this can lead to similar oscillations in other physical quantities, too. Their measurement can provide insight into the properties of the electron system.

22.1 Free Electrons in a Magnetic Field

Just like in Chapter 16, we consider an electron gas confined to a rectangular box of sides L_x, L_y, and L_z that is subject to periodic boundary conditions, but this time we place the system in an applied magnetic field $\boldsymbol{B}$. Since the quantum mechanical problem was first solved by L. D. LANDAU in 1930, the electronic energy levels in a magnetic field are called *Landau levels*.

22.1.1 One-Particle Energy Spectrum

We shall again specify the magnetic field $\boldsymbol{B}$ in terms of a vector potential $\boldsymbol{A}$. To simplify calculations, we shall neglect the interaction of the electron spin with the magnetic field.[1] To determine the one-particle energy spectrum, the Schrödinger equation

$$\frac{1}{2m_\mathrm{e}}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+e\boldsymbol{A}\right)^2\psi(\boldsymbol{r})=\varepsilon\psi(\boldsymbol{r}) \tag{22.1.1}$$

needs to be solved. In the Landau gauge the z-directed magnetic field can be derived from the vector potential $\boldsymbol{A}=(0,\,Bx,\,0)$, since $\boldsymbol{B}=\operatorname{curl}\boldsymbol{A}$. Another common choice is the symmetric gauge, in which $\boldsymbol{A}=\frac{1}{2}\boldsymbol{B}\times\boldsymbol{r}=\frac{1}{2}(-By,\,Bx,\,0)$. Later we shall use this gauge, too.

Writing the momentum operator in component form, the Hamiltonian in Landau gauge reads

$$\begin{aligned}\mathcal{H}&=\frac{1}{2m_\mathrm{e}}\left[p_x^2+(p_y+eBx)^2+p_z^2\right]\\&=-\frac{\hbar^2}{2m_\mathrm{e}}\frac{\partial^2}{\partial x^2}-\frac{\hbar^2}{2m_\mathrm{e}}\left(\frac{\partial}{\partial y}+\mathrm{i}\frac{eB}{\hbar}x\right)^2-\frac{\hbar^2}{2m_\mathrm{e}}\frac{\partial^2}{\partial z^2}\,.\end{aligned} \tag{22.1.2}$$

Even though the magnetic field breaks the invariance under arbitrary translations, invariance along the y- and z-directions is preserved by our choice of the gauge. Consequently, the two corresponding momentum components are conserved,

$$\dot{p}_y=\frac{\mathrm{i}}{\hbar}\left[\mathcal{H},p_y\right]=0\,,\qquad \dot{p}_z=\frac{\mathrm{i}}{\hbar}\left[\mathcal{H},p_z\right]=0\,. \tag{22.1.3}$$

Following LANDAU, we shall use the ansatz

$$\psi(x,y,z)=u(x)\mathrm{e}^{\mathrm{i}k_y y}\mathrm{e}^{\mathrm{i}k_z z}\,. \tag{22.1.4}$$

Inserting it into the Schrödinger equation, and using (22.1.2) for the Hamiltonian, the equation for $u(x)$ is

$$-\frac{\hbar^2}{2m_\mathrm{e}}\frac{\mathrm{d}^2u(x)}{\mathrm{d}x^2}+\frac{\hbar^2}{2m_\mathrm{e}}\left(k_y+\frac{eB}{\hbar}x\right)^2u(x)+\frac{\hbar^2k_z^2}{2m_\mathrm{e}}u(x)=\varepsilon u(x)\,, \tag{22.1.5}$$

which can be rearranged as

$$-\frac{\hbar^2}{2m_\mathrm{e}}\frac{\mathrm{d}^2u(x)}{\mathrm{d}x^2}+\frac{1}{2}m_\mathrm{e}\left(\frac{eB}{m_\mathrm{e}}\right)^2\left(x+\frac{\hbar}{eB}k_y\right)^2u(x)=\left(\varepsilon-\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\right)u(x)\,. \tag{22.1.6}$$

The combination eB/m_e can be recognized as the cyclotron frequency ω_c of free electrons. By introducing the notation

$$x_0=-\frac{\hbar}{eB}k_y=-\frac{\hbar}{m_\mathrm{e}\omega_\mathrm{c}}k_y\,, \tag{22.1.7}$$

the equation reads

[1] Later we shall return to the role of spin.

$$-\frac{\hbar^2}{2m_\mathrm{e}}\frac{\mathrm{d}^2u(x)}{\mathrm{d}x^2}+\frac{1}{2}m_\mathrm{e}\omega_\mathrm{c}^2(x-x_0)^2u(x)=\left(\varepsilon-\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\right)u(x)\,. \tag{22.1.8}$$

This is the Schrödinger equation of a linear harmonic oscillator that oscillates about x_0 with an angular frequency ω_c. It is well known from the quantum mechanical treatment of oscillators that the wavefunction of the state of quantum number n of an oscillator centered at x_0 can be written in terms of the Hermite polynomial H_n that satisfies (C.4.1):

$$u_n(x)=\frac{1}{\pi^{1/4}l_0^{1/2}\sqrt{2^n n!}}H_n\left(\frac{x-x_0}{l_0}\right)\mathrm{e}^{-(x-x_0)^2/2l_0^2}\,, \tag{22.1.9}$$

where the magnetic length

$$l_0=\sqrt{\frac{\hbar}{m_\mathrm{e}\omega_\mathrm{c}}}=\sqrt{\frac{\hbar}{eB}} \tag{22.1.10}$$

introduced on page 250 characterizes the spatial variations of the wavefunction. It can be given a simple intuitive meaning by realizing that in the semiclassical approximation an electron of energy $\hbar\omega_\mathrm{c}/2$ moves in a circular orbit of radius l_0 in real space. As a characteristic length l_0 is also frequently used in quantum mechanics. By substituting the numerical values of $\hbar$ and e, and expressing the magnetic field in teslas, we immediately get

$$l_0=\frac{25.66\,\mathrm{nm}}{\sqrt{B[T]}}\,. \tag{22.1.11}$$

This length is on the order of 10^{-6} cm in magnetic fields applied customarily in measurements, and is thus much larger than atomic distances.

Another celebrated result of quantum mechanics states that the energy eigenvalues of the oscillator are quantized in units of $\hbar\omega_\mathrm{c}$, with a zero-point energy of $\frac{1}{2}\hbar\omega_\mathrm{c}$. Identifying the energy formula on the right-hand side of (22.1.8) with the oscillator energy,

$$\varepsilon-\frac{\hbar^2k_z^2}{2m_\mathrm{e}}=\left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\,, \tag{22.1.12}$$

where n can take nonnegative integer values. Rearranging this formula as

$$\boxed{\varepsilon=\left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}+\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\,,} \tag{22.1.13}$$

the electron energy is seen to be composed of two terms. The part of the kinetic energy coming from the motion parallel to the field is the same as in the zero-field case, whereas the contribution of the perpendicular motion is quantized in units of $\hbar\omega_\mathrm{c}$ and thus depends on the strength of the field.

When periodic boundary conditions are used, the z component of the wave vector, k_z, is quantized in units of $2\pi/L_z$. Therefore in macroscopic samples the energy of Landau states varies practically continuously with the quantum number k_z. However, the energy of the levels labeled by subsequent values of the quantum number n can differ greatly when the field is sufficiently strong. The states characterized by the same quantum number n then make up a continuum. They are said to belong to the nth Landau level or subband. However, the subbands can overlap on account of the dependence on k_z.

There are alternative ways to determine the energy spectrum. The Hamiltonian can be simplified by a suitably chosen canonical transformation, or ladder (creation and annihilation) operators can be used instead of the position and momentum operators, as was done for lattice vibrations in Chapter 12, but the description of the motion in the xy-plane requires two commuting sets of operators now. Since x and p_y appear together in the combination

$$x + \frac{1}{eB} p_y = x + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_y \,, \tag{22.1.14}$$

the appropriate choice in this case is

$$\begin{aligned} a &= \sqrt{\frac{m_\mathrm{e}\omega_\mathrm{c}}{2\hbar}} \left(x + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_y + \frac{\mathrm{i}}{m_\mathrm{e}\omega_\mathrm{c}} p_x \right) , \\ a^\dagger &= \sqrt{\frac{m_\mathrm{e}\omega_\mathrm{c}}{2\hbar}} \left(x + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_y - \frac{\mathrm{i}}{m_\mathrm{e}\omega_\mathrm{c}} p_x \right) , \\ b &= \sqrt{\frac{m_\mathrm{e}\omega_\mathrm{c}}{2\hbar}} \left(y + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_x + \frac{\mathrm{i}}{m_\mathrm{e}\omega_\mathrm{c}} p_y \right) , \\ b^\dagger &= \sqrt{\frac{m_\mathrm{e}\omega_\mathrm{c}}{2\hbar}} \left(y + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_x - \frac{\mathrm{i}}{m_\mathrm{e}\omega_\mathrm{c}} p_y \right) \end{aligned} \tag{22.1.15}$$

rather than (12.1.23). The inverse transformation is then

$$\begin{aligned} x + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_y &= \sqrt{\frac{\hbar}{2m_\mathrm{e}\omega_\mathrm{c}}} \left(a + a^\dagger \right) , \\ p_x &= \mathrm{i}\sqrt{\frac{\hbar m_\mathrm{e}\omega_\mathrm{c}}{2}} \left(a^\dagger - a \right) , \\ y + \frac{1}{m_\mathrm{e}\omega_\mathrm{c}} p_x &= \sqrt{\frac{\hbar}{2m_\mathrm{e}\omega_\mathrm{c}}} \left(b + b^\dagger \right) , \\ p_y &= \mathrm{i}\sqrt{\frac{\hbar m_\mathrm{e}\omega_\mathrm{c}}{2}} \left(b^\dagger - b \right) . \end{aligned} \tag{22.1.16}$$

It follows from the canonical commutation relations of the position and momentum operators that the ladder operators satisfy bosonic commutation relations:

$$[a, a^\dagger] = 1 \,, \qquad [b, b^\dagger] = 1 \,, \tag{22.1.17}$$

and

$$[a, a] = [a^\dagger, a^\dagger] = [b, b] = [b^\dagger, b^\dagger] = 0 \,, \tag{22.1.18}$$

moreover the operators a ($a^\dagger$) and b ($b^\dagger$) commute with each other, too. In terms of them the Hamiltonian (22.1.2) can be cast in diagonal form:

$$\mathcal{H} = \frac{\hbar\omega_{\mathrm{c}}}{2}\left(aa^\dagger + a^\dagger a\right) + \frac{1}{2m_{\mathrm{e}}}p_z^2 = \hbar\omega_{\mathrm{c}}\left(a^\dagger a + \tfrac{1}{2}\right) + \frac{1}{2m_{\mathrm{e}}}p_z^2 \,. \tag{22.1.19}$$

The ground state is the vacuum of the "particles" created by the operators $a^\dagger$ and $b^\dagger$,

$$a\psi_{0,0} = b\psi_{0,0} = 0\,, \tag{22.1.20}$$

while excited states are obtained from the ground state by acting on it with the creation operators $a^\dagger$ and $b^\dagger$:

$$\psi = \psi_{n,m}\mathrm{e}^{\mathrm{i}k_z z} = \frac{1}{\sqrt{n!\,m!}}(a^\dagger)^n(b^\dagger)^m\psi_{0,0}\mathrm{e}^{\mathrm{i}k_z z}\,. \tag{22.1.21}$$

However, the energy of the state depends only on the quantum number k_z and the number n of the oscillators created by $a^\dagger$; it is independent of the number of b-type oscillators.

22.1.2 Degree of Degeneracy of Landau Levels

As mentioned above, states can be characterized by three quantum numbers (n, k_y, k_z, or n, m, k_z), however only n and k_z appear in the energy expression. Since the energy does not depend on k_y (or m), the energy levels are highly degenerate. In the second-quantized form this manifests itself in the absence of the creation and annihilation operators of b-type oscillators in the Hamiltonian. As the operator $b^\dagger$ creates zero-energy oscillations, the energy does not depend on the occupation number of this oscillator state.

The number of possible states – that is, the degree of degeneracy – can be obtained most easily by counting the possible values for the quantum number k_y in the wavefunction (22.1.4). This quantum number is related to the center x_0 of the oscillator by (22.1.7), which can also be written as

$$x_0 = -l_0^2\,k_y\,. \tag{22.1.22}$$

The possible values for k_y and x_0 are limited by a geometric constraint. The formula (22.1.9) for $u_n(x)$ contains, in addition to the Hermite polynomials, an exponential function, whose rapid decay renders solutions physically meaningless unless x_0 is inside the sample, that is,

$$0 < x_0 < L_x\,. \tag{22.1.23}$$

This constraint and (22.1.7) imply that the quantum number k_y must be in the range

$$-\frac{m_{\mathrm{e}}\omega_{\mathrm{c}}}{\hbar}L_x < k_y < 0\,. \tag{22.1.24}$$

When periodic boundary conditions are imposed, k_y is quantized in units of $2\pi/L_y$, so the number of possible values for k_y is

$$N_{\mathrm{p}} = \frac{m_{\mathrm{e}}\omega_{\mathrm{c}}}{\hbar} L_x \left(\frac{2\pi}{L_y}\right)^{-1} = \frac{m_{\mathrm{e}}\omega_{\mathrm{c}}}{2\pi\hbar} L_x L_y = \frac{L_x L_y}{2\pi l_0^2} \,. \tag{22.1.25}$$

This formula gives the degree of degeneracy for Landau levels in each subband – that is, the number of states of the same energy when k_z is kept fixed. If ω_{c} is expressed in terms of the magnetic field, the equivalent formula

$$N_{\mathrm{p}} = \frac{eB}{2\pi\hbar} L_x L_y \tag{22.1.26}$$

clearly shows that the degree of degeneracy increases proportionally to the strength of the magnetic field.

By introducing the flux quantum[2] $\Phi_0^* = 2\pi\hbar/e = h/e$, the degree of degeneracy of Landau levels can be written as

$$N_{\mathrm{p}} = \frac{B}{\Phi_0^*} L_x L_y \,. \tag{22.1.27}$$

Since BL_xL_y is the total magnetic flux through the sample, N_{p} is determined by the ratio of the magnetic flux to the flux quantum. The result suggests that each state carries one flux quantum.

Up to this point we have completely ignored the spin degrees of freedom and their interaction with the magnetic field. When they are taken into account, the appropriate formula reads

$$\varepsilon_\sigma(n, k_z) = \left(n + \tfrac{1}{2}\right)\hbar\omega_{\mathrm{c}} + \frac{\hbar^2 k_z^2}{2m_{\mathrm{e}}} - \tfrac{1}{2} g_{\mathrm{e}} \mu_{\mathrm{B}} B \sigma \,, \tag{22.1.28}$$

where $\sigma = \pm 1$ for the two spin orientations. This leads to the spin splitting of Landau levels. For free electrons the spacing $\hbar\omega_{\mathrm{c}}$ of Landau levels is, to a good approximation, equal to the spin splitting, as $g_{\mathrm{e}} \approx -2$ and

$$\hbar\omega_{\mathrm{c}} = \hbar\frac{eB}{m_{\mathrm{e}}} = 2\mu_{\mathrm{B}} B \,. \tag{22.1.29}$$

Therefore the energy of spin-up electrons on the nth Landau level is practically the same as the energy of spin-down electrons on the $n+1$th level. Aside from the lowest level, the effects of spin can be taken into account by a factor 2 in the number of degenerate states. The situation will be more complicated for Bloch electrons, where the energy of Landau levels depends on the cyclotron mass through ω_{c}, and so the spin splitting will no longer be equal to the spacing of Landau levels. The effects of spin cannot then be lumped into a simple factor of two.

[2] This is twice the flux quantum $\Phi_0 = h/2e$ used in superconductivity; see page 454.

22.1.3 Density of States

When the energy spectrum and the degree of degeneracy for each level are known, we can proceed to determine the density of states, which plays a fundamental role in the calculation of thermodynamic quantities.

We shall first consider a two-dimensional electron gas in a perpendicular magnetic field. The electronic energy spectrum then consists only of the discrete values that correspond to the oscillator energies:

$$\varepsilon_n = \hbar\omega_{\mathrm{c}}\bigl(n + \tfrac{1}{2}\bigr) \,. \tag{22.1.30}$$

Ignoring spin, the number of states on each energy level is given by (22.1.27). The density of states thus contains regularly spaced Dirac delta peaks of equal amplitude N_{p}:

$$\rho_{2\mathrm{d}}(B,\varepsilon) = \sum_n N_{\mathrm{p}}\, \delta\bigl[\varepsilon - \hbar\omega_{\mathrm{c}}\bigl(n + \tfrac{1}{2}\bigr)\bigr] \,. \tag{22.1.31}$$

As the magnetic field is chosen weaker, the spacing of these peaks is reduced. In sufficiently weak fields a coarse-grained, continuous density of states can be defined, which is constant, and required to be equal to the density of states of the two-dimensional electron gas in the absence of a magnetic field.

To determine the latter, we shall follow the procedure introduced in Chapter 16, but now only the components k_x and k_y are taken into account in the kinetic energy, giving

$$\varepsilon_\perp(\boldsymbol{k}) = \frac{\hbar^2}{2m_{\mathrm{e}}}\left(k_x^2 + k_y^2\right) = \frac{\hbar^2 k_\perp^2}{2m_{\mathrm{e}}} \,. \tag{22.1.32}$$

Electrons with energies between ε and $\varepsilon + \mathrm{d}\varepsilon$ are located in an annulus of radius $k_\perp$ and thickness $\mathrm{d}k_\perp$, whose area is therefore $\mathrm{d}\mathcal{A} = 2\pi k_\perp\, \mathrm{d}k_\perp$. The relations between ε and $k_\perp$, and $\mathrm{d}\varepsilon$ and $\mathrm{d}k_\perp$ are

$$\varepsilon = \frac{\hbar^2 k_\perp^2}{2m_{\mathrm{e}}} \,, \qquad \mathrm{d}\varepsilon = \frac{\mathrm{d}\varepsilon}{\mathrm{d}k_\perp}\mathrm{d}k_\perp = \frac{\hbar^2 k_\perp}{m_{\mathrm{e}}}\mathrm{d}k_\perp \,. \tag{22.1.33}$$

By eliminating $\mathrm{d}k_\perp$ in favor of $\mathrm{d}\varepsilon$, we have

$$\mathrm{d}\mathcal{A} = \frac{2\pi m_{\mathrm{e}}}{\hbar^2}\,\mathrm{d}\varepsilon \,. \tag{22.1.34}$$

The number $\mathrm{d}N$ of states in the energy range of width $\mathrm{d}\varepsilon$ is obtained by dividing this area by the $\boldsymbol{k}$-space area $(2\pi/L_x)(2\pi/L_y)$ per allowed value of $\boldsymbol{k}$:

$$\mathrm{d}N = \mathrm{d}\mathcal{A}\left(\frac{2\pi}{L_x}\frac{2\pi}{L_y}\right)^{-1} = \frac{m_{\mathrm{e}}}{2\pi\hbar^2}L_x L_y\,\mathrm{d}\varepsilon \,. \tag{22.1.35}$$

The density of states per unit surface area and spin orientation in a two-dimensional electron gas is therefore

$$\rho_{2\mathrm{d},\sigma}(\varepsilon) = \frac{m_\mathrm{e}}{2\pi\hbar^2}\,. \tag{22.1.36}$$

For such a density of states, an energy range of width $\hbar\omega_\mathrm{c}$ in a sample of surface area L_xL_y contains the same number of states,

$$\rho_{2\mathrm{d},\sigma}(\varepsilon)\,\hbar\omega_\mathrm{c}\,L_xL_y = \frac{m_\mathrm{e}}{2\pi\hbar^2}\,\hbar\omega_\mathrm{c}\,L_xL_y = \frac{eB}{2\pi\hbar}L_xL_y\,, \tag{22.1.37}$$

as the degree of degeneracy of Landau levels according to (22.1.26). Therefore the lowest N_p states, which would fill an energy range of width $\hbar\omega_\mathrm{c}$ in the absence of a magnetic field, become degenerate at the lowest Landau level at $\hbar\omega_\mathrm{c}/2$ in the presence of a magnetic field. Similarly, the next N_p states in the range $\hbar\omega_\mathrm{c} < \varepsilon < 2\hbar\omega_\mathrm{c}$ are all pulled to the Landau level of energy $\frac{3}{2}\hbar\omega_\mathrm{c}$, etc.

Such a substantial rearrangement of the electron states is not surprising, since, according to (3.2.23), the leading term of the change in energy in an applied magnetic field is given by

$$\frac{e\hbar}{2m_\mathrm{e}}\boldsymbol{l}\cdot\boldsymbol{B}\,, \tag{22.1.38}$$

and so, assuming unit angular momentum, the shift of the energy of the states in a field of strength B is of order

$$\frac{e\hbar}{2m_\mathrm{e}}B = \tfrac{1}{2}\hbar\omega_\mathrm{c}\,. \tag{22.1.39}$$

As the magnetic field becomes stronger, the distance between adjacent Landau levels increases – and so does the degeneracy of states: more and more states condense into each Landau level. Figure 22.1 shows the energy spectrum for three different values of the magnetic field compared to the zero-field case.

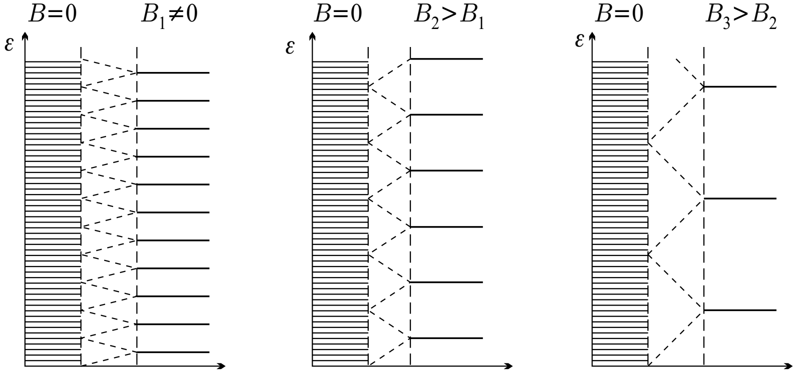


Fig. 22.1. Energy levels of a two-dimensional electron gas for three different values of the magnetic field

Now consider the three-dimensional case, where, in addition to the quantum number n and its degeneracy, k_z also needs to be taken into account. Since k_z is quantized in units of $2\pi/L_z$, there are $(L_z/2\pi)\,\mathrm{d}k_z$ possible values for k_z in the region of width $\mathrm{d}k_z$. For each value of k_z, k_y can take N_p different values. Therefore the total number of states in the nth Landau subband with wave numbers between k_z and $k_z + \mathrm{d}k_z$ is

$$\mathrm{d}N_n = N_\mathrm{p}\,\frac{L_z}{2\pi}\,\mathrm{d}k_z = \frac{eB}{2\pi\hbar}L_xL_y\frac{L_z}{2\pi}\,\mathrm{d}k_z\,. \tag{22.1.40}$$

When the factor of two coming from the two possible spin orientations is also included,

$$\mathrm{d}N_n = \frac{2eB}{(2\pi)^2\hbar}V\,\mathrm{d}k_z\,. \tag{22.1.41}$$

The relationship between the energy ε and k_z in the nth Landau subband is given by

$$\hbar k_z = \pm\sqrt{2m_\mathrm{e}\left[\varepsilon - \left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]} \tag{22.1.42}$$

and so

$$\mathrm{d}k_z = \frac{\mathrm{d}k_z}{\mathrm{d}\varepsilon}\mathrm{d}\varepsilon = \pm\frac{\sqrt{2m_\mathrm{e}}}{2\hbar}\left[\varepsilon - \left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{-1/2}\,\mathrm{d}\varepsilon\,. \tag{22.1.43}$$

Since the positive and negative values of k_z give the same contribution, the density of states per unit volume coming from the Landau subband of quantum number n is

$$\rho_n(\varepsilon) = \frac{1}{V}\frac{\mathrm{d}N_n}{\mathrm{d}\varepsilon} = 2eB\sqrt{2m_\mathrm{e}}\frac{1}{(2\pi\hbar)^2}\left[\varepsilon - \left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{-1/2}\,. \tag{22.1.44}$$

The total density of states is obtained by summing over Landau levels:

$$\begin{aligned}\rho(\varepsilon) &= 2eB\sqrt{2m_\mathrm{e}}\frac{1}{(2\pi\hbar)^2}\sum_{n=0}^{n_\mathrm{max}}\left[\varepsilon - \left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{-1/2}\\ &= \frac{1}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2}\frac{\hbar\omega_\mathrm{c}}{2}\sum_{n=0}^{n_\mathrm{max}}\left[\varepsilon - \left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{-1/2}\,,\end{aligned} \tag{22.1.45}$$

where the summation is up to the largest integer n_max that satisfies the condition $(n_\mathrm{max} + \frac{1}{2})\hbar\omega_\mathrm{c} \le \varepsilon$. As shown in Fig. 22.2, the density of states has a singularity at energies $\varepsilon = (n+\frac{1}{2})\hbar\omega_\mathrm{c}$. For weak fields, where the singularities are spaced rather densely, a coarse-grained, continuous density of states can be defined in which the singularities are smeared out. The $\sqrt{\varepsilon}$-type density of states derived in (16.2.54) for a free-electron gas is then recovered. This is indicated by a dashed line in the figure.

By increasing the magnetic field, more and more states are accommodated in each Landau subband, and so levels of higher quantum numbers become

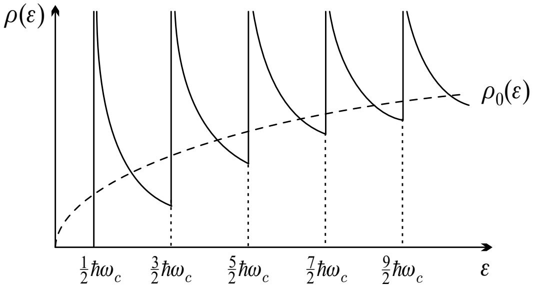


Fig. 22.2. The density of states of a three-dimensional electron gas in strong magnetic field. The density of states in the absence of the magnetic field is shown by the dashed line

successively empty as the electrons move to subbands of lower quantum numbers. When the density of states is considered at a fixed energy as a function of the magnetic field, singularities appear at those fields where new Landau levels become empty. The distance between such singularities increases with increasing B, whereas they are regularly spaced as a function of $1/B$. Figure 22.3 is a schematic plot of the density of states at the Fermi energy versus the magnetic induction B and its reciprocal at relatively strong fields.

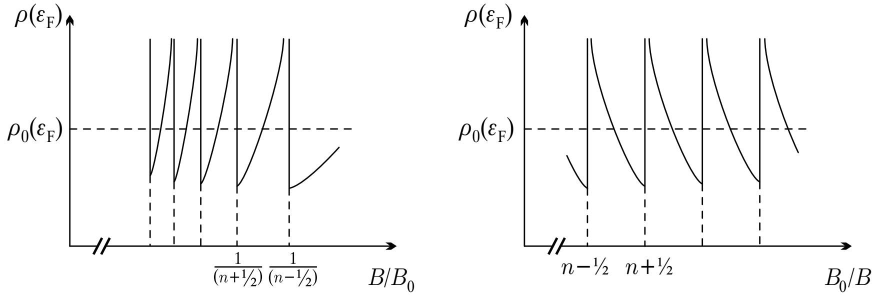


Fig. 22.3. The density of states at the Fermi energy as a function of the magnetic induction and its reciprocal

To be precise, the figure shows the density of states at the Fermi energy calculated for the zero-field case. We need to show now that, apart from very strong fields, the chemical potential depends weakly on the magnetic field.

Since fermions fill the states up to the Fermi energy at zero temperature, the integral of the density of states up to the chemical potential is just the number of electrons per unit volume:

$$n_e = \int_0^{\mu(B)} \rho(\varepsilon)\,d\varepsilon\,. \tag{22.1.46}$$

Comparison with the zero-field case gives

$$\int_0^{\mu(B)} \rho(\varepsilon)\,\mathrm{d}\varepsilon = \int_0^{\varepsilon_\mathrm{F}} \rho_0(\varepsilon)\,\mathrm{d}\varepsilon\,, \tag{22.1.47}$$

where ρ_0 is the density of states of the free-electron gas, which can also be written as

$$\rho_0(\varepsilon) = \frac{1}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon}\,. \tag{22.1.48}$$

Making use of (22.1.45), we have

$$\frac{\hbar\omega_\mathrm{c}}{2}\int_0^{\mu(B)}\sum_{n=0}^{n_\mathrm{max}}\left[\varepsilon-\left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{-1/2}\mathrm{d}\varepsilon = \int_0^{\varepsilon_\mathrm{F}}\sqrt{\varepsilon}\,\mathrm{d}\varepsilon\,. \tag{22.1.49}$$

By integrating this formula with respect to energy, and introducing the notations

$$\eta = \frac{\mu(B)}{\hbar\omega_\mathrm{c}}\,, \qquad \eta_0 = \frac{\varepsilon_\mathrm{F}}{\hbar\omega_\mathrm{c}}\,, \tag{22.1.50}$$

the equation

$$\sum_{n=0}^{n_\mathrm{max}}\left[\eta-\left(n+\tfrac{1}{2}\right)\right]^{1/2} = \tfrac{2}{3}\eta_0^{3/2} \tag{22.1.51}$$

is obtained, where n_max must satisfy the condition

$$n_\mathrm{max}+\tfrac{1}{2} < \eta < n_\mathrm{max}+\tfrac{3}{2}\,. \tag{22.1.52}$$

The simplest way to obtain the solution is to plot η_0 versus η. As shown in Fig. 22.4, apart from very small values of η the graph runs very close to the straight line $\eta_0 = \eta$. From the slope of the straight lines drawn from the origin

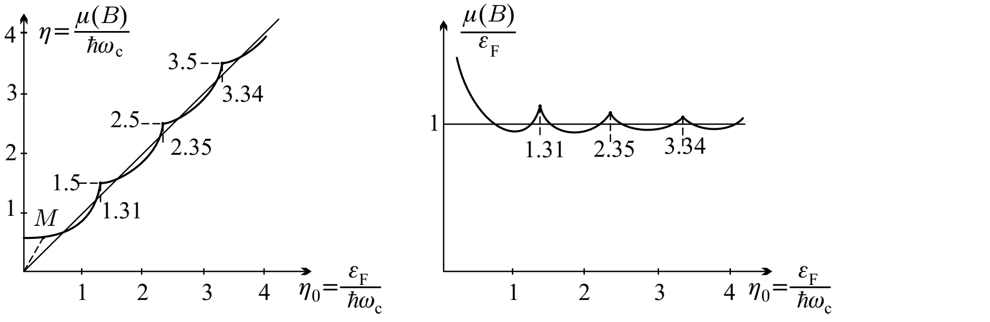


Fig. 22.4. The solution of the equation for the field dependence of the chemical potential, and the variation of the chemical potential with the inverse magnetic field

to different points of the curve the field dependence of $\eta/\eta_0 = \mu(B)/\varepsilon_{\mathrm{F}}$ (and through it that of the chemical potential) can be easily established. This is plotted in the right part of the figure.

Very strong fields aside, this ratio oscillates around unity with a small amplitude. Under these circumstances the analytical expression

$$\Delta\mu(B) = -\frac{\varepsilon_{\mathrm{F}}}{\pi}\left(\frac{\hbar\omega_{\mathrm{c}}}{2\varepsilon_{\mathrm{F}}}\right)^{3/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\sin\left(2\pi l\frac{\varepsilon_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}} - \frac{\pi}{4}\right) \tag{22.1.53}$$

is obtained for the oscillation of the chemical potential.

22.1.4 Visualization of the Landau States

When the magnetic field is turned on, the Landau subband of quantum number n becomes populated by those states for which the kinetic-energy contribution

$$\varepsilon_{\perp} = \frac{\hbar^2 k_{\perp}^2}{2m_{\mathrm{e}}} = \frac{\hbar^2}{2m_{\mathrm{e}}}(k_x^2 + k_y^2) \tag{22.1.54}$$

of the motion perpendicular to the field is in the range

$$n\hbar\omega_{\mathrm{c}} < \varepsilon_{\perp} < (n+1)\hbar\omega_{\mathrm{c}}\,. \tag{22.1.55}$$

As shown in Fig. 22.5, the quantum numbers k_x and k_y of these states fill an annulus in $\boldsymbol{k}$-space. Its inner radius $k_{\perp}$ can be determined from

$$n\hbar\omega_{\mathrm{c}} = \frac{\hbar^2 k_{\perp}^2}{2m_{\mathrm{e}}}\,, \tag{22.1.56}$$

and its outer radius $k_{\perp} + \mathrm{d}k_{\perp}$ from

$$(n+1)\hbar\omega_{\mathrm{c}} = \frac{\hbar^2 (k_{\perp} + \mathrm{d}k_{\perp})^2}{2m_{\mathrm{e}}}\,. \tag{22.1.57}$$

For large values of n, where $\mathrm{d}k_{\perp}$ is small compared to $k_{\perp}$,

$$\mathrm{d}k_{\perp} = \frac{m_{\mathrm{e}}}{\hbar^2 k_{\perp}}\hbar\omega_{\mathrm{c}}\,. \tag{22.1.58}$$

Since the area of the annulus is then

$$2\pi k_{\perp}\,\mathrm{d}k_{\perp} = \frac{2\pi m_{\mathrm{e}}}{\hbar^2}\hbar\omega_{\mathrm{c}}\,, \tag{22.1.59}$$

the number of allowed $\boldsymbol{k}$ vectors in the annulus is

$$2\pi k_{\perp}\,\mathrm{d}k_{\perp}\frac{L_x}{2\pi}\frac{L_y}{2\pi} = \frac{2\pi m_{\mathrm{e}}}{\hbar^2}\hbar\omega_{\mathrm{c}}\frac{L_x}{2\pi}\frac{L_y}{2\pi} = \frac{eB}{2\pi\hbar}L_x L_y\,. \tag{22.1.60}$$

This is the same as the degree of degeneracy of individual Landau levels, provided spin is neglected. Thus, when the magnetic field is turned on, states

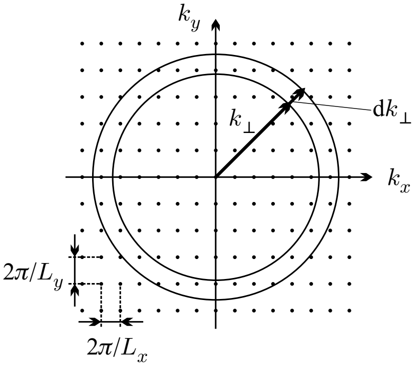


Fig. 22.5. Allowed values of the wave-vector components k_x and k_y for a free-electron gas. When a magnetic field is turned on, the corresponding states in the annulus are all condensed into the same Landau level

whose wave-vector component $k_\perp$ is located in an annulus of width $\mathrm{d}k_\perp$ in the zero-field case – that is, whose energy differs by less than $\hbar\omega_\mathrm{c}$ – become degenerate. They are pulled to the closest Landau level (subband) of quantum number n. Consequently, the Landau levels can also be visualized by the circles in the (k_x, k_y) plane shown in the left part of Fig. 22.6. The circles are drawn in such a way that the area of each annulus is the same:

$$\delta\mathcal{A} = 2\pi k_\perp\,\mathrm{d}k_\perp = \frac{2\pi m_\mathrm{e}}{\hbar}\omega_\mathrm{c} = \frac{2\pi eB}{\hbar} = \frac{2\pi}{l_0^2}\,, \tag{22.1.61}$$

and thus the same number of states condense into each circle. Note that the Landau states are not located at well-defined points on the circle but rotate with frequency ω_c.

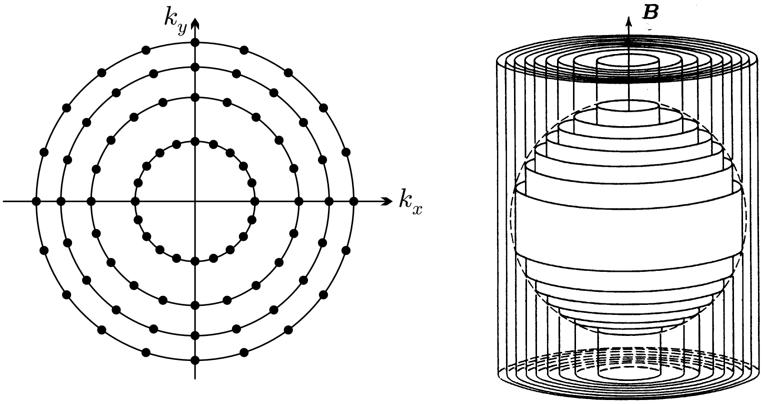


Fig. 22.6. Visualization of the Landau levels in the (k_x, k_y) plane by circles, and by cylinders drawn into the Fermi sphere

When the variable k_z is also taken into account, the states in the Landau subbands can be illustrated in $\boldsymbol{k}$-space by coaxial cylinders (tubes) drawn into the Fermi sphere. Their projection on the (k_x, k_y) plane is just the set of circles discussed earlier. Much like in the two-dimensional case, the states whose wave vectors in the zero-field case are close to the tube (either inside or outside) become degenerate and end up on the closest tube when the magnetic field is turned on. The tubes are characterized by the quantum number n. Points located in different heights on the cylinder correspond to states of different quantum numbers k_z. Because of the formula (22.1.61) for the area between the circles, the cross section of the tubes is said to be quantized. For the tube of quantum number n

$$\mathcal{A}_n = \left(n + \tfrac{1}{2}\right)\frac{2\pi eB}{\hbar} = \left(n + \tfrac{1}{2}\right)\frac{2\pi}{l_0^2}\,. \tag{22.1.62}$$

In the ground state the tubes are filled in the region between

$$\hbar k_z = \pm\sqrt{2m_\mathrm{e}\left[\mu - \left(n + \tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]}\,, \tag{22.1.63}$$

which depends on the quantum number n. Since the chemical potential can be identified with ε_F to a good approximation, the filled region can be obtained as the portions of the tubes inside the free-electron Fermi sphere.

22.1.5 Landau States in the Symmetric Gauge

To amend the picture given in the previous subsection, the spatial distribution of electrons in the Landau states needs to be specified. When calculated from the wavefunction given in (22.1.4) and (22.1.9), the electronic density varies only in the x-direction: it is nonvanishing around the center x_0 of the oscillator, over a width l_0 that is determined by (22.1.10). These regions become narrower as the field strength is increased, and electrons are then localized to planes $x = x_0 = l_0^2 k_y$ spaced at equal distances, where k_y is an integral multiple of $2\pi/L_y$. This configuration is very different from the set of circles in the xy-plane obtained in the semiclassical approximation. To recover the semiclassical result we need to take different linear combinations of the wavefunctions of degenerate states. This is achieved by using the symmetric gauge instead of the Landau gauge.

By choosing the vector potential as $\boldsymbol{A} = \frac{1}{2}\boldsymbol{B} \times \boldsymbol{r}$, the Hamiltonian whose eigenvalue problem has to be solved is now

$$\mathcal{H} = \frac{1}{2m_\mathrm{e}}\left[\left(p_x - \tfrac{1}{2}eBy\right)^2 + \left(p_y + \tfrac{1}{2}eBx\right)^2 + p_z^2\right]. \tag{22.1.64}$$

Since the z component is again separated from the x and y components, plane wave solutions are sought in the z-direction. We shall therefore discuss only the motion in the xy-plane, described by the Hamiltonian $\mathcal{H}_\perp$. By introducing the Larmor frequency

$$\omega_{\rm L} = \frac{eB}{2m_{\rm e}}, \tag{22.1.65}$$

which is just half of the cyclotron frequency, the transverse part of the Hamiltonian can be rewritten as

$$\mathcal{H}_\perp = \frac{p_x^2}{2m_{\rm e}} + \tfrac{1}{2}m_{\rm e}\omega_{\rm L}^2 x^2 + \frac{p_y^2}{2m_{\rm e}} + \tfrac{1}{2}m_{\rm e}\omega_{\rm L}^2 y^2 + \hbar\omega_{\rm L} L_z\,, \tag{22.1.66}$$

where $\hbar L_z = xp_y - yp_x$ is the z component of the angular momentum operator.[3] The previous expression is the sum of the Hamiltonians of two identical harmonic oscillators of frequency $\omega_{\rm L}$, however the two oscillators are not independent of one another: they are coupled by the last term of $\mathcal{H}_\perp$. Therefore the creation and annihilation operators used in the quantum mechanical treatment of oscillators, which are now written as

$$\begin{aligned} a &= \sqrt{\frac{m_{\rm e}\omega_{\rm L}}{2\hbar}}\left(x + \frac{\rm i}{m_{\rm e}\omega_{\rm L}}p_x\right), \quad a^\dagger = \sqrt{\frac{m_{\rm e}\omega_{\rm L}}{2\hbar}}\left(x - \frac{\rm i}{m_{\rm e}\omega_{\rm L}}p_x\right), \\ b &= \sqrt{\frac{m_{\rm e}\omega_{\rm L}}{2\hbar}}\left(y + \frac{\rm i}{m_{\rm e}\omega_{\rm L}}p_y\right), \quad b^\dagger = \sqrt{\frac{m_{\rm e}\omega_{\rm L}}{2\hbar}}\left(y - \frac{\rm i}{m_{\rm e}\omega_{\rm L}}p_y\right), \end{aligned} \tag{22.1.67}$$

rather than (22.1.15), do not diagonalize the Hamiltonian but yield

$$\mathcal{H}_\perp = \hbar\omega_{\rm L}\big(a^\dagger a + \tfrac{1}{2}\big) + \hbar\omega_{\rm L}\big(b^\dagger b + \tfrac{1}{2}\big) + {\rm i}\hbar\omega_{\rm L}\big(ab^\dagger - a^\dagger b\big). \tag{22.1.68}$$

The Hamiltonian can be diagonalized by a Bogoliubov-type transformation, by taking linear combinations of the two oscillators. To this end, we introduce the operators

$$\begin{aligned} \alpha &= \frac{1}{\sqrt{2}}\big(a - {\rm i}b\big), \qquad \alpha^\dagger = \frac{1}{\sqrt{2}}\big(a^\dagger + {\rm i}b^\dagger\big), \\ \beta &= \frac{1}{\sqrt{2}}\big(a + {\rm i}b\big), \qquad \beta^\dagger = \frac{1}{\sqrt{2}}\big(a^\dagger - {\rm i}b^\dagger\big). \end{aligned} \tag{22.1.69}$$

The inverse transformation is straightforward:

$$\begin{aligned} a &= \frac{1}{\sqrt{2}}\big(\alpha + \beta\big), \qquad a^\dagger = \frac{1}{\sqrt{2}}\big(\alpha^\dagger + \beta^\dagger\big), \\ b &= \frac{\rm i}{\sqrt{2}}\big(\alpha - \beta\big), \qquad b^\dagger = \frac{-{\rm i}}{\sqrt{2}}\big(\alpha^\dagger - \beta^\dagger\big). \end{aligned} \tag{22.1.70}$$

Substituting these into (22.1.68), and making use of the equality $\omega_{\rm c} = 2\omega_{\rm L}$,

$$\mathcal{H}_\perp = \hbar\omega_{\rm L}(2\alpha^\dagger\alpha + 1) = \hbar\omega_{\rm c}\big(\alpha^\dagger\alpha + \tfrac{1}{2}\big) \tag{22.1.71}$$

[3] The eigenvalues and eigenfunctions of this Hamiltonian were determined by V. Fock in 1928, before Landau, therefore the Landau spectrum is sometimes referred to as the *Fock–Landau spectrum*.

is obtained. This is just the Hamiltonian of a harmonic oscillator of angular frequency ω_{c}. The eigenvalues are thus the same as in (22.1.12). Note that just like in the Landau gauge, two creation and annihilation operators had to be introduced, however the oscillators created by $\beta^\dagger$ do not contribute to the energy. This gives rise to the degeneracy of the Landau levels.

The specific form of the wavefunctions can be obtained particularly simply by realizing that the operators α and β can be written in coordinate representation, provided the variables

$$w = \frac{1}{\sqrt{2}}(x + \mathrm{i}y) , \qquad w^* = \frac{1}{\sqrt{2}}(x - \mathrm{i}y) \tag{22.1.72}$$

and the corresponding canonical momenta,

$$p_w = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial w} = \frac{\hbar}{\mathrm{i}}\frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x} - \mathrm{i}\frac{\partial}{\partial y}\right) , \quad p_{w^*} = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial w^*} = \frac{\hbar}{\mathrm{i}}\frac{1}{\sqrt{2}}\left(\frac{\partial}{\partial x} + \mathrm{i}\frac{\partial}{\partial y}\right) \tag{22.1.73}$$

are used, since the substitution of (22.1.67) into (22.1.69) leads immediately to

$$\begin{aligned} \alpha &= \sqrt{\frac{m_{\mathrm{e}}\omega_{\mathrm{L}}}{2\hbar}}\left(w^* + \frac{\mathrm{i}}{m_{\mathrm{e}}\omega_{\mathrm{L}}}p_w\right) , \quad \alpha^\dagger = \sqrt{\frac{m_{\mathrm{e}}\omega_{\mathrm{L}}}{2\hbar}}\left(w - \frac{\mathrm{i}}{m_{\mathrm{e}}\omega_{\mathrm{L}}}p_{w^*}\right) , \\ \beta &= \sqrt{\frac{m_{\mathrm{e}}\omega_{\mathrm{L}}}{2\hbar}}\left(w + \frac{\mathrm{i}}{m_{\mathrm{e}}\omega_{\mathrm{L}}}p_{w^*}\right) , \quad \beta^\dagger = \sqrt{\frac{m_{\mathrm{e}}\omega_{\mathrm{L}}}{2\hbar}}\left(w^* - \frac{\mathrm{i}}{m_{\mathrm{e}}\omega_{\mathrm{L}}}p_w\right) . \end{aligned} \tag{22.1.74}$$

So w and its complex conjugate are the natural variables in the symmetric gauge. It seems therefore logical to replace the x and y coordinates by the polar coordinates defined by

$$x = r\cos\varphi , \qquad y = r\sin\varphi , \tag{22.1.75}$$

as

$$w = \frac{1}{\sqrt{2}}r\mathrm{e}^{\mathrm{i}\varphi} , \qquad w^* = \frac{1}{\sqrt{2}}r\mathrm{e}^{-\mathrm{i}\varphi} . \tag{22.1.76}$$

In terms of the polar coordinates, the Hamiltonian in (22.1.66) reads

$$\mathcal{H}_\perp = -\frac{\hbar^2}{2m_{\mathrm{e}}}\left[\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\varphi^2}\right] + \frac{1}{2}m_{\mathrm{e}}\omega_{\mathrm{L}}^2 r^2 + \omega_{\mathrm{L}}\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial\varphi} . \tag{22.1.77}$$

The eigenvalue problem can then be solved exactly. Eigenstates are characterized by two quantum numbers: l, which is related to the radial part, and m, which specifies the z component of the dimensionless angular momentum. The eigenfunctions are

$$\psi_{l,m}(r,\varphi) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathrm{i}m\varphi}\frac{1}{l_0}\left(\frac{l!}{(|m|+l)!}\right)^{1/2}\mathrm{e}^{-r^2/4l_0^2}\left(\frac{r^2}{2l_0^2}\right)^{|m|/2}L_l^{|m|}(r^2/2l_0^2) , \tag{22.1.78}$$

where $L_l^{|m|}$ is the generalized Laguerre polynomial defined in (C.4.10). The energy eigenvalues are given in terms of the quantum numbers l and m as

$$\varepsilon_\perp = \hbar\omega_\mathrm{c}\left[l + \frac{m+|m|}{2} + \frac{1}{2}\right], \tag{22.1.79}$$

in agreement with the result obtained in the Landau gauge.

By exploiting the properties of the generalized Laguerre polynomials it can be shown that the solutions give high electron densities in the xy-plane along the circles whose radii are related to the magnetic length l_0 in a simple way. Particularly simple are the wavefunctions for the lowest Landau level ($l = 0$):

$$\psi_{0,m}(r,\varphi) = \frac{1}{\sqrt{2\pi|m|!}\, l_0}\left(\frac{r^2}{2l_0^2}\right)^{|m|/2} \mathrm{e}^{\mathrm{i}m\varphi}\mathrm{e}^{-r^2/4l_0^2}\,, \tag{22.1.80}$$

which yield

$$\bigl\langle\psi_{0,m}(r,\varphi)\big|r^2\big|\psi_{0,m}(r,\varphi)\bigr\rangle = 2(m+1)l_0^2 \tag{22.1.81}$$

and

$$\bigl\langle\psi_{0,m}(r,\varphi)\big|L_z\big|\psi_{0,m}(r,\varphi)\bigr\rangle = m\,. \tag{22.1.82}$$

For increasing m the wavefunction is localized along circles of larger and larger radii. The degree of degeneracy can be determined from the requirement that the radius for the largest m should be inside the sample, a cylinder of radius R. This means

$$2l_0^2(m_\mathrm{max}+1) = R^2\,. \tag{22.1.83}$$

Expressed in terms of the cross-sectional area $F = R^2\pi$ of the sample,

$$m_\mathrm{max} + 1 = \frac{F}{2\pi l_0^2}\,, \tag{22.1.84}$$

which gives the same degree of degeneracy as (22.1.25).

For further reference, we shall write the wavefunction obtained in the symmetric gauge in another form:

$$\begin{aligned}\psi_{m,n}(x,y) = {} & \frac{1}{\bigl[2\pi l_0^2(2l_0^2)^{m+n}m!n!\bigr]^{1/2}}\mathrm{e}^{(x^2+y^2)/4l_0^2} \\ & \times\left(\frac{\partial}{\partial x} + \mathrm{i}\frac{\partial}{\partial y}\right)^m\left(\frac{\partial}{\partial x} - \mathrm{i}\frac{\partial}{\partial y}\right)^n \mathrm{e}^{-(x^2+y^2)/2l_0^2}\,.\end{aligned} \tag{22.1.85}$$

This can be written in a particularly simple form in terms of the operators $\alpha^\dagger$ and $\beta^\dagger$:

$$\psi_{m,n}(x,y) = \frac{1}{\sqrt{m!n!}}\left(\beta^\dagger\right)^m\left(\alpha^\dagger\right)^n\psi_{0,0}(x,y), \tag{22.1.86}$$

where

$$\psi_{0,0}(x,y) = \frac{1}{\sqrt{2\pi l_0^2}} \exp\left[-(x^2+y^2)/4l_0^2\right] . \tag{22.1.87}$$

The energy of these states is the same, $\varepsilon = \hbar\omega_\mathrm{c}(n+1/2)$, irrespective of m. The states that belong to the lowest Landau level will be particularly important. Their wavefunctions are

$$\psi_{m,0}(x,y) = \frac{1}{(2\pi l_0^2 2^m m!)^{1/2}} \left(\frac{x+\mathrm{i}y}{l_0}\right)^m \mathrm{e}^{-(x^2+y^2)/4l_0^2} . \tag{22.1.88}$$

22.1.6 Edge States

Even though the sample was assumed to be of finite extent in the directions perpendicular to the field, since this was necessary to determine the degree of degeneracy of the Landau levels, the finite height of the potential barrier at the edge of the sample was ignored. Close to the surface, the electronic wavefunctions may become distorted by this potential, and the energy of the state can change. This can be illustrated most simply in the semiclassical picture. As shown in Fig. 16.3, electrons placed in an applied magnetic field move in circular orbits that are perpendicular to the field direction. For electrons of energy ε in an applied field B, the radius of this classical circular orbit is the cyclotron radius $r_\mathrm{c} = (2m_\mathrm{e}\varepsilon)^{1/2}/eB$. This is true for electrons deep inside the sample. Electrons close to the boundaries move in circular arcs: they repeatedly hit the walls, and are reflected back from it, hence they move along the wall. However, the propagation direction is the opposite on the opposite face.

In the Landau gauge the degenerate oscillators are characterized by the coordinate x_0 of their center, which must be inside the sample. According to the above picture, one would expect that when the Schrödinger equation of electrons moving in a magnetic field is solved in a finite box bounded by potential walls using the Landau gauge, the previous result is recovered for the energy of those oscillators whose coordinate x_0 is more than r_c away from the boundary. The energy is independent of x_0, and discrete Landau levels appear. On the other hand, the energy of those states that are close to the edges of the sample are higher. This is shown schematically in Fig. 22.7.

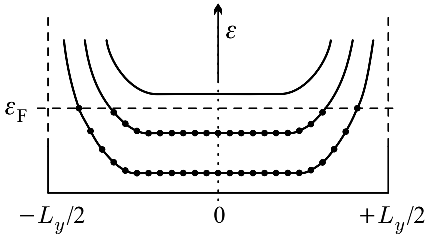


Fig. 22.7. Upward shift of the energy of the Landau levels near the surface and the appearance of edge states, shown in a section of the sample. Dots indicate occupied Landau states

The contribution of edge states can be neglected as far as the total energy of the system and the susceptibility derived from it are concerned. In the rest of this chapter we shall consistently ignore them. However, they need to be taken into account in the description of the quantum Hall effect.

22.2 Landau Diamagnetism

It was mentioned in Chapter 16 that the susceptibility due to the orbital motion of electrons vanishes in a classical electron gas. This is not the case in the quantum mechanical treatment. To determine the magnetization and the susceptibility, consider the ground-state energy of the electron system in the presence of a magnetic field:

$$E = V \int_0^{\varepsilon_\mathrm{F}} \varepsilon \rho(\varepsilon)\,\mathrm{d}\varepsilon\,. \tag{22.2.1}$$

Taking the density of states from (22.1.45),

$$E = \frac{V}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2} \frac{\hbar\omega_\mathrm{c}}{2} \sum_{n=0}^{n_\mathrm{max}} \int_{(n+\frac{1}{2})\hbar\omega_\mathrm{c}}^{\varepsilon_\mathrm{F}} \frac{\varepsilon}{\left[\varepsilon - \left(n+\frac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{1/2}} \mathrm{d}\varepsilon\,. \tag{22.2.2}$$

Since the number of electrons can also be expressed in terms of the density of states, we find

$$E = N_\mathrm{e}\varepsilon_\mathrm{F} + \frac{V}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2} \frac{\hbar\omega_\mathrm{c}}{2} \sum_{n=0}^{n_\mathrm{max}} \int_{(n+\frac{1}{2})\hbar\omega_\mathrm{c}}^{\varepsilon_\mathrm{F}} \frac{\varepsilon - \varepsilon_\mathrm{F}}{\left[\varepsilon - \left(n+\frac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{1/2}} \mathrm{d}\varepsilon\,. \tag{22.2.3}$$

After the change of variable $\varepsilon - (n+\frac{1}{2})\hbar\omega_\mathrm{c} \to \varepsilon$ the integral can be evaluated:

$$\begin{aligned} E &= N_\mathrm{e}\varepsilon_\mathrm{F} - \frac{V}{2\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2} \frac{\hbar\omega_\mathrm{c}}{2} \sum_{n=0}^{n_\mathrm{max}} \int_0^{\varepsilon_\mathrm{F}-(n+\frac{1}{2})\hbar\omega_\mathrm{c}} \frac{\varepsilon_\mathrm{F} - \left(n+\frac{1}{2}\right)\hbar\omega_\mathrm{c} - \varepsilon}{\sqrt{\varepsilon}} \mathrm{d}\varepsilon \\ &= N_\mathrm{e}\varepsilon_\mathrm{F} - \frac{V}{3\pi^2}\left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2} \hbar\omega_\mathrm{c} \sum_{n=0}^{n_\mathrm{max}} \left[\varepsilon_\mathrm{F} - \left(n+\tfrac{1}{2}\right)\hbar\omega_\mathrm{c}\right]^{3/2}. \end{aligned} \tag{22.2.4}$$

If the number of occupied Landau levels is sufficiently high, the sum can be approximated by an integral, and the Euler–Maclaurin formula

$$\sum_{n=0}^{n_0} f\left(n+\tfrac{1}{2}\right) = \int_0^{n_0+1} f(x)\,\mathrm{d}x - \frac{1}{24}\left[f'(n_0+1) - f'(0)\right] \tag{22.2.5}$$

can be used for the difference between the Riemann sum and the integral. The result is then

$$E_0 = N_e\varepsilon_F - \frac{V}{3\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\hbar\omega_c \int\limits_0^{n_{\max}+1} \left[\varepsilon_F - x\hbar\omega_c\right]^{3/2} \mathrm{d}x + \frac{V}{48\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(\hbar\omega_c)^2\varepsilon_F^{1/2} . \tag{22.2.6}$$

Since $n_{\max}$ is the largest integer for which $(n_{\max} + \frac{1}{2})\hbar\omega_c$ is less than the chemical potential, $(n_{\max} + 1)\hbar\omega_c$ is identified with ε_F. Making use of the relation

$$N_e = \frac{V}{3\pi^2}k_F^3 = \frac{V}{3\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\varepsilon_F^{3/2} \tag{22.2.7}$$

between the Fermi energy and the particle number, and expressing ω_c in terms of the magnetic induction B and the Fermi energy in terms of the Fermi momentum, we have

$$E_0 = \left(1 - \tfrac{2}{5}\right)N_e\varepsilon_F - \frac{V}{24\pi^2}(eB)^2\frac{k_F}{m_e} . \tag{22.2.8}$$

The first term is the ground-state energy of the free-electron gas in zero magnetic field; it is in agreement with (16.2.36). The susceptibility can be derived from the field-dependent second term using (3.2.40); this gives

$$\chi = -\frac{\mu_0 e^2 k_F}{3m_e}\frac{1}{(2\pi)^2} . \tag{22.2.9}$$

Introducing formally the Bohr magneton μ_B for $e\hbar/2m_e$,

$$\chi = -\frac{\mu_0\mu_B^2}{3}\frac{k_F m_e}{\pi^2\hbar^2} . \tag{22.2.10}$$

As comparison with (16.2.54) shows, the right-hand side contains the density of states at the Fermi energy, so the susceptibility due to the orbital motion in a free-electron gas is

$$\chi = -\tfrac{1}{3}\mu_0\mu_B^2\rho(\varepsilon_F) . \tag{22.2.11}$$

The negative sign indicates the diamagnetic behavior of the electron gas called *Landau diamagnetism.* The spin-related Pauli susceptibility also has to be included in the total susceptibility of the electron gas. Comparison with (16.2.113) reveals that the diamagnetic susceptibility is precisely one-third of the Pauli susceptibility for free electrons if $|g_e| = 2$ is taken. Therefore, despite the opposite signs, the combined orbital and spin contributions give an overall paramagnetic character to the electron gas. As we shall see in the next section, the situation may be different for Bloch electrons.

It is worth noting that the nonvanishing diamagnetic contribution comes formally from the correction term in the Euler–Maclaurin formula. If the density of states did not contain a sum over the discrete Landau levels, and the free energy could be written as an integral of a smooth function, then the second term in (22.2.8), which comes from the difference between the sum and the integral, would not appear, and there would be no Landau diamagnetism. That is why we found in Chapter 16 that the orbital motion gives no contribution to the susceptibility in the classical limit.

It should also be noted that the condition for the applicability of (22.2.5) is

$$|f(n) - f(n+1)| \ll f(n) \,. \tag{22.2.12}$$

Had the calculations been performed at a finite temperature, we would have seen that this condition is equivalent to the requirement $\mu_{\mathrm{B}} B \ll k_{\mathrm{B}} T$. Therefore the formula obtained for the susceptibility is valid only in the $B \to 0$ limit. For stronger fields, where $\mu_{\mathrm{B}} B \gg k_{\mathrm{B}} T$, the approximation used above cannot be applied. We shall discuss the details of more precise calculations in Section 22.4.

22.3 Bloch Electrons in Strong Magnetic Fields

The energy spectrum cannot be determined exactly for Bloch electrons moving in the periodic potential of a crystal. That is why we studied the dynamics of electrons in the semiclassical approximation in the previous chapter. We shall now demonstrate that the intuitive picture for the formation of Landau levels can be generalized to the case where the constant-energy surfaces are ellipsoids, or, if the dispersion relation is more general, to cases where the quantized nature of energy is important but the system is still far from the extreme quantum limit.

22.3.1 Electrons Characterized by an Effective-Mass Tensor

The Schrödinger equation (22.1.2) of free electrons contains the electron mass m_{e}. The calculation presented there can be straightforwardly generalized to the case where the energy spectrum of the Bloch electrons can be characterized by an effective-mass tensor, and the magnetic field is along a principal axis of the tensor – using *Wannier's theorem*[4] and the *Peierls substitution.*[5]

Wannier's theorem is based on the observation that if the energy of the Bloch state $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is $\varepsilon_n(\boldsymbol{k})$ in the presence of a periodic potential, then these Bloch functions are eigenfunctions of the operator $\varepsilon_n(-\mathrm{i}\boldsymbol{\nabla})$ obtained by replacing $\boldsymbol{k}$ in the dispersion relation by $-\mathrm{i}\boldsymbol{\nabla}$, with the same energy, that is,

[4] G. H. Wannier, 1937.

[5] R. E. Peierls, 1933.

$$\varepsilon_n(-\mathrm{i}\boldsymbol{\nabla})\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_n(\boldsymbol{k})\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,. \tag{22.3.1}$$

This can be most easily demonstrated by performing this substitution on the Fourier expansion[6]

$$\varepsilon_n(\boldsymbol{k}) = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_l} C_{nl}\mathrm{e}^{-\mathrm{i}\boldsymbol{R}_l\cdot\boldsymbol{k}}\,, \tag{22.3.2}$$

and then applying it to the wavefunction $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$. This gives

$$\begin{aligned}\varepsilon_n(-\mathrm{i}\boldsymbol{\nabla})\psi_{n\boldsymbol{k}}(\boldsymbol{r}) &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_l} C_{nl}\mathrm{e}^{-\boldsymbol{R}_l\cdot\boldsymbol{\nabla}}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\\ &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_l} C_{nl}\Big[1-\boldsymbol{R}_l\cdot\boldsymbol{\nabla}+\tfrac{1}{2}(\boldsymbol{R}_l\cdot\boldsymbol{\nabla})^2+\dots\Big]\psi_{n\boldsymbol{k}}(\boldsymbol{r})\\ &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_l} C_{nl}\psi_{n\boldsymbol{k}}(\boldsymbol{r}-\boldsymbol{R}_l)\,.\end{aligned} \tag{22.3.3}$$

Exploiting the translational properties of the wavefunction, this indeed leads to

$$\varepsilon_n(-\mathrm{i}\boldsymbol{\nabla})\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}_l} C_{nl}\mathrm{e}^{-\mathrm{i}\boldsymbol{R}_l\cdot\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \varepsilon_{n\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r})\,. \tag{22.3.4}$$

Therefore this operator can serve as an effective Hamiltonian:

$$\mathcal{H}_{\text{eff}} = \varepsilon_n(-\mathrm{i}\boldsymbol{\nabla})\,. \tag{22.3.5}$$

When the electrons are placed in a magnetic field described by a vector potential, the Peierls substitution is used. As demonstrated by PEIERLS, the wave vector characterizing the translational properties should be replaced by $-\mathrm{i}\boldsymbol{\nabla} + e\boldsymbol{A}/\hbar$ for Bloch electrons in a magnetic field. Thus the effects of a periodic potential and an electromagnetic field can be taken into account by an effective Hamiltonian that is obtained by using the above substitution in the dispersion relation of the Bloch electrons.

When the Fermi surface is ellipsoidal, the dispersion relation transformed to the principal axes can be characterized by the diagonal elements m_1^*, m_2^*, and m_3^* of the effective-mass tensor according to (21.2.33). By choosing the coordinate axes along the principal axes, the effective Hamiltonian reads

$$\mathcal{H} = \frac{1}{2m_1^*}(p_x+eA_x)^2 + \frac{1}{2m_2^*}\left(p_y+eA_y\right)^2 + \frac{1}{2m_3^*}(p_z+eA_z)^2\,. \tag{22.3.6}$$

If the magnetic field is along the z-direction, and the Landau gauge is used, we have

[6] Note that the expansion contains a sum over the translation vectors of the direct lattice, since $\varepsilon_n(\boldsymbol{k})$ is periodic in $\boldsymbol{k}$ in the reciprocal lattice.

$$\mathcal{H} = -\frac{\hbar^2}{2m_1^*}\frac{\partial^2}{\partial x^2} - \frac{\hbar^2}{2m_2^*}\left(\frac{\partial}{\partial y} + \mathrm{i}\frac{eB}{\hbar}x\right)^2 - \frac{\hbar^2}{2m_3^*}\frac{\partial^2}{\partial z^2}\,. \tag{22.3.7}$$

The eigenvalue problem can then be solved exactly by repeating the steps of the free-electron case. The eigenenergies can be written as

$$\varepsilon = \left(n + \tfrac{1}{2}\right)\hbar\omega_{\mathrm{c}} + \frac{\hbar^2 k_z^2}{2m_{||}^*}\,, \tag{22.3.8}$$

where the cyclotron frequency formula $\omega_{\mathrm{c}} = eB/m_{\mathrm{c}}$ contains the cyclotron mass $m_{\mathrm{c}} = (m_1^* m_2^*)^{1/2}$, and $m_{||}^* = m_3^*$. The calculation also shows that the degree of degeneracy is once again given by (22.1.26).

Using the Peierls substitution, the energy of Landau levels can also be calculated after a great deal of tedious algebra in the case where the dispersion relation of the Bloch electrons is still quadratic but the orientation of the magnetic field with respect to the principal axes is arbitrary. Specifying this arbitrary direction by the direction cosines α_1, α_2, α_3, as in (21.2.34), the energy spectrum can again be written as

$$\varepsilon = \left(n + \tfrac{1}{2}\right)\hbar\omega_{\mathrm{c}} + \frac{\hbar^2 k_{||}^2}{2m_{||}^*}\,, \tag{22.3.9}$$

where $k_{||}$ is the wave number of the motion along the direction of the magnetic field, and the corresponding effective mass is

$$m_{||}^* = m_1^*\alpha_1^2 + m_2^*\alpha_2^2 + m_3^*\alpha_3^2\,, \tag{22.3.10}$$

whereas the formula for the cyclotron frequency of the perpendicular motion contains the cyclotron mass (21.2.26) obtained in the semiclassical approximation. This is quite natural, since the semiclassical and quantum mechanical approaches should lead to the same cyclotron frequency. Figure 22.8 shows the oblique Landau tubes formed by the corresponding states of the same quantum number n.

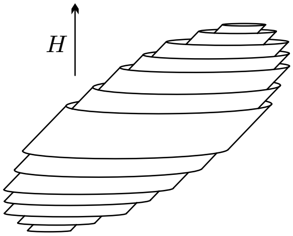


Fig. 22.8. Landau tubes associated with the Landau states of the same quantum number n for an ellipsoidal Fermi surface and general orientation of the magnetic field with respect to the principal axes

22.3.2 Semiclassical Quantization

Quantum effects become important when the distance of energy levels becomes comparable to the thermal energy. In magnetic fields of $B \sim 1\,\mathrm{T}$ the energy difference $\hbar\omega_\mathrm{c}$ is on the order of $10^{-4}\,\mathrm{eV}$, which corresponds to the thermal energy at 1 kelvin. Therefore the quantized nature of energy needs to be taken into account at low temperatures for such fields. Since the spectrum in magnetic field cannot be determined exactly for a general dispersion relation, further approximations are required. It is useful to bear in mind that in metals the Fermi energy is of order 1 eV, $\hbar\omega_\mathrm{c}/\varepsilon_\mathrm{F} \sim 10^{-4}$, and the number of interesting Landau levels is high, about 10^4. According to L. ONSAGER's proposition (1952), under such circumstances BOHR's semiclassical quantization can be used to calculate the Landau levels of Bloch electrons.

According to the correspondence principle, the energy difference between two neighboring levels can be related to the frequency ν of the motion in the classical orbit:

$$\varepsilon(n+1) - \varepsilon(n) = h\nu\,. \tag{22.3.11}$$

Identifying this frequency with

$$\nu_\mathrm{c} = T_\mathrm{c}^{-1} = \frac{eB}{\hbar^2}\left(\frac{\partial\mathcal{A}}{\partial\varepsilon}\right)^{-1}, \tag{22.3.12}$$

the frequency determined in the semiclassical approximation and given in (21.2.19), the energy difference of adjacent Landau levels with k_z fixed should be

$$\varepsilon(n+1,k_z) - \varepsilon(n,k_z) = \frac{2\pi eB}{\hbar}\left(\frac{\partial\mathcal{A}}{\partial\varepsilon}\right)^{-1}. \tag{22.3.13}$$

For large quantum numbers

$$\frac{\partial\mathcal{A}}{\partial\varepsilon} = \frac{\mathcal{A}[\varepsilon(n+1,k_z)] - \mathcal{A}[\varepsilon(n,k_z)]}{\varepsilon(n+1,k_z) - \varepsilon(n,k_z)} \tag{22.3.14}$$

to a good approximation, therefore

$$\mathcal{A}[\varepsilon(n+1,k_z)] - \mathcal{A}[\varepsilon(n,k_z)] = \frac{2\pi eB}{\hbar}\,, \tag{22.3.15}$$

and hence

$$\mathcal{A}[\varepsilon(n,k_z)] = (n+\gamma)\frac{2\pi eB}{\hbar}\,, \tag{22.3.16}$$

where γ is a fractional number that cannot be determined exactly because of the approximation.

Since $\mathcal{A}$ is the area of the semiclassical orbit in the (k_x, k_y) plane, this result has another intuitive interpretation: for electrons moving in periodic potentials, too, only those $\boldsymbol{k}$-space orbits are allowed whose area in the plane

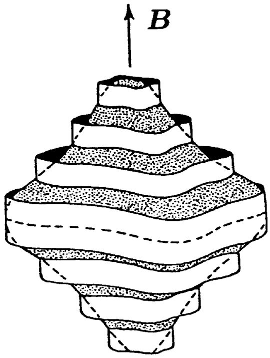


Fig. 22.9. Visualization of the Landau states for a general Fermi surface

perpendicular to the magnetic field is quantized in units of $2\pi eB/\hbar$. Since the semiclassical motion is on constant-energy surfaces, the lines of constant energy in the (k_x, k_y) plane are drawn in such a way that the area between neighboring contours should be $2\pi eB/\hbar$. This can be considered as a generalization of Fig. 22.6. By drawing the closed contours for all values of k_z, the Landau tubes shown in Fig. 22.9 are obtained. In contrast to Fig. 22.6, their cross sections are not circular but of the shape of the lines of constant energy.

22.3.3 Quantization of the Orbit in Real Space

As illustrated in Fig. 21.5, electrons trace out similar orbits in real space and $\boldsymbol{k}$-space in the semiclassical approximation. Their dimensions are related by a scaling factor $\hbar/eB$, as given in (21.2.6). For large values of the quantum number n the same result holds for electrons on quantized Landau levels. Since the $\boldsymbol{k}$-space area of the orbits is quantized, so is the area enclosed by the orbit in real space: in the plane that is perpendicular to the magnetic field the electron can only trace out orbits whose area is given by

$$F_n = \left(\frac{\hbar}{eB}\right)^2 \frac{2\pi eB}{\hbar}(n+\gamma) = \frac{2\pi\hbar}{eB}(n+\gamma) = 2\pi l_0^2(n+\gamma)\,. \qquad (22.3.17)$$

The magnetic flux through this area is

$$\varPhi_n = (n+\gamma)\frac{2\pi\hbar}{e} = (n+\gamma)\varPhi_0^*\,, \qquad (22.3.18)$$

where $\varPhi_0^*$ is the flux quantum. Electrons in a strong magnetic field are observed to move in orbits for which the flux enclosed by the area is an integral multiple of the flux quantum – provided the constant γ is neglected.

22.3.4 Energy Spectrum in the Tight-Binding Approximation

Though semiclassical quantization gives a good approximation in most cases, it is worth examining what happens to the electrons moving in a periodic potential in a very strong magnetic field using a different approach. In zero magnetic field the dispersion relation in the tight-binding approximation for electrons moving in a square lattice is

$$\varepsilon_{\boldsymbol{k}} = -2t\left[\cos(k_x a) + \cos(k_y a)\right]. \tag{22.3.19}$$

As has been mentioned, the effects of the magnetic field can be taken into account by means of the Peierls substitution, so the expression obtained from the dispersion relation via

$$\boldsymbol{k} \to \frac{1}{\hbar}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e\boldsymbol{A}\right) \tag{22.3.20}$$

will be considered as the effective Hamiltonian. Using the Landau gauge,

$$\mathcal{H} = -2t\left\{\cos\left[\frac{1}{\mathrm{i}}\frac{\partial}{\partial x}a\right] + \cos\left[\left(\frac{1}{\mathrm{i}}\frac{\partial}{\partial y} + \frac{e}{\hbar}Bx\right)a\right]\right\}. \tag{22.3.21}$$

Note that the Hamiltonian now contains the shift operators $\exp(a\partial/\partial x)$ and $\exp(a\partial/\partial y)$, which shift the wavefunction by a lattice constant in the x- and y-directions, respectively:

$$\mathrm{e}^{a\partial/\partial x}\psi(\boldsymbol{r}) = \psi(\boldsymbol{r} + a\hat{\boldsymbol{x}}), \qquad \mathrm{e}^{a\partial/\partial y}\psi(\boldsymbol{r}) = \psi(\boldsymbol{r} + a\hat{\boldsymbol{y}}). \tag{22.3.22}$$

Applying them to the eigenvalue problem of the Hamiltonian (22.3.21),

$$\begin{aligned} -t\Big[\psi(\boldsymbol{r} + a\hat{\boldsymbol{x}}) + \psi(\boldsymbol{r} - a\hat{\boldsymbol{x}}) + \mathrm{e}^{\mathrm{i}eaBx/\hbar}\psi(\boldsymbol{r} + a\hat{\boldsymbol{y}}) \\ + \mathrm{e}^{-\mathrm{i}eaBx/\hbar}\psi(\boldsymbol{r} - a\hat{\boldsymbol{y}})\Big] = \varepsilon\psi(\boldsymbol{r}). \end{aligned} \tag{22.3.23}$$

Since the Wannier functions are better adapted to the tight-binding approximation, we shall now expand the wavefunction of one-particle states in terms of the Wannier functions associated with the lattice points, with coefficients $g(\boldsymbol{R}_i)$:

$$\psi(\boldsymbol{r}) = \sum_i g(\boldsymbol{R}_i)\phi(\boldsymbol{r} - \boldsymbol{R}_i) = \sum_i g(\boldsymbol{R}_i)c_i^\dagger|0\rangle. \tag{22.3.24}$$

Substituting this expression into (22.3.23), the coefficients $g(\boldsymbol{R}_i)$ are found to satisfy a similar equation:

$$\begin{aligned} -t\Big[g(\boldsymbol{R}_i + a\hat{\boldsymbol{x}}) + g(\boldsymbol{R}_i - a\hat{\boldsymbol{x}}) + \mathrm{e}^{\mathrm{i}eaBx/\hbar}g(\boldsymbol{R}_i + a\hat{\boldsymbol{y}}) \\ + \mathrm{e}^{-\mathrm{i}eaBx/\hbar}g(\boldsymbol{R}_i - a\hat{\boldsymbol{y}})\Big] = \varepsilon g(\boldsymbol{R}_i). \end{aligned} \tag{22.3.25}$$

Specifying the lattice points $\boldsymbol{R}_i = (ma, na)$ by the coordinates (m, n), the equation

$$\begin{aligned}\Big[g(m+1,n) + g(m-1,n) + \mathrm{e}^{\mathrm{i}ea^2Bm/\hbar}g(m,n+1)& \\ + \mathrm{e}^{-\mathrm{i}ea^2Bm/\hbar}g(m,n-1)\Big] &= \varepsilon g(m,n)\end{aligned} \tag{22.3.26}$$

is obtained, where ε now denotes the dimensionless energy (i.e., the energy divided by $-t$).

Note that the same result could have been obtained by exploiting the property that in the presence of a magnetic field the amplitude t of hopping between lattice points contains a field-dependent phase factor. Neglecting the spin variable, the Hamiltonian is

$$\mathcal{H} = \sum_{ij} t_{ij} c_i^\dagger c_j \mathrm{e}^{\mathrm{i}\phi_{ij}}, \tag{22.3.27}$$

where the phase factor is given by the line integral of the vector potential along a path joining two neighboring lattice points:

$$\phi_{ij} = -\frac{2\pi}{\Phi_0^*} \int_j^i \boldsymbol{A}(\boldsymbol{r}) \cdot \mathrm{d}\boldsymbol{l}\,. \tag{22.3.28}$$

Because of the choice of the Landau gauge, only the x coordinate appears in the phase factors, and the variation in the y-direction can be chosen as a plane wave:

$$g(m,n) = \mathrm{e}^{\mathrm{i}kna} g(m)\,. \tag{22.3.29}$$

This leads to the Harper equation for the variations in the x-direction:

$$g(m+1) + g(m-1) + 2\cos(2\pi m\alpha + ka)g(m) = \varepsilon g(m)\,, \tag{22.3.30}$$

where

$$\alpha = \frac{ea^2B}{h} = \frac{a^2B}{\Phi_0^*} \tag{22.3.31}$$

is the number of flux quanta through the primitive cell. Writing the recursive equation as

$$\begin{pmatrix} g(m+1) \\ g(m) \end{pmatrix} = \begin{pmatrix} \varepsilon - 2\cos(2\pi m\alpha + ka) & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} g(m) \\ g(m-1) \end{pmatrix}, \tag{22.3.32}$$

and making use of the property that under periodic boundary conditions the starting point is reached after N steps, the equation for the energy eigenvalues is

$$\prod_{m=1}^{N} \begin{pmatrix} \varepsilon - 2\cos(2\pi m\alpha + ka) & -1 \\ 1 & 0 \end{pmatrix} = 1\,. \tag{22.3.33}$$

In the absence of a magnetic field, the energy eigenvalues make up a band of width $4t$. This is also the case when the strength of the magnetic field is such that the flux through each primitive cell is an integral multiple of the flux quantum Φ_0^*. This is because when a closed path is traversed, the phase factors give $2\pi/\Phi_0^*$ times the enclosed magnetic flux on account of Gauss's law, and these phase factors can be transformed away by a suitable gauge transformation in the above-mentioned cases. For intermediate values of the field strength the band is split into subbands. Based on the Harper equation, the energy spectrum was first determined by D. R. HOFSTADTER in 1976. The spectrum, known as the *Hofstadter butterfly*, is shown in Fig. 22.10.

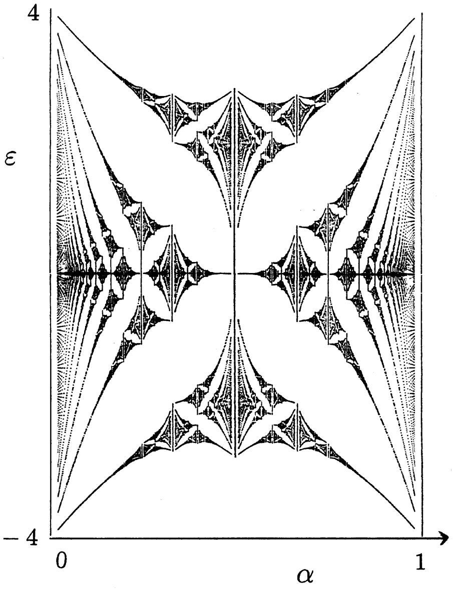


Fig. 22.10. The Hofstadter spectrum of electrons in magnetic field in the tight-binding approximation. The dimensionless energy is plotted against the dimensionless flux α [Reprinted with permission from D. R. Hofstadter, *Phys. Rev. B* **14**, 2239 (1976). ©1976 by the American Physical Society]

For relatively weak fields, the spectrum at the bottom and top of the band is similar to the spectrum of Landau levels for free electrons; the regularly spaced discrete energies vary linearly with the field. In stronger fields, these Landau levels are split further, which is most conspicuous for the lowest level, while Landau levels of higher quantum numbers disappear gradually. The number of subbands depends on the fraction of the flux quantum per primitive cell. If the area a^2 encloses one-third of the flux quantum then there are three subbands, separated by gaps. If it encloses one-fifth or two-fifth then there are five subbands. In general, if α is a rational number that can be written as p/q, where q is odd, then the band is split into q subbands (minibands). When q is even, the subbands may touch at the center.

We shall see in Chapter 24 on transport phenomena that the rearrangement of electron states in two-dimensional electron systems placed in a strong magnetic field – that is, the appearance of relatively well separated Landau levels – leads to an interesting phenomenon: the quantization of the Hall resistance. Plateaux appear in the Hall resistance vs. magnetic field plot, where the inverse Hall resistance is quantized in units of e^2/h, and the longitudinal resistance vanishes. As the field strength is increased, the integer ν that labels the plateau decreases in steps of unity. This is illustrated in Fig. 24.13. It can be shown that further rearrangement of the Landau levels due to the periodic potential of the lattice does not change the quantized character, however, the plateaux no longer make up a monotonically decreasing set of steps on account of the "minigaps" separating the subbands. The observation of this pattern is difficult, since for the customary, atomic-size lattice constants of crystals the parameter $\alpha = Ba^2/\Phi_0^*$ is too small even for the strongest attainable fields. However, in superlattices fabricated in semiconductor heterostructures α can be sufficiently large for that the minigaps produce observable effects. As the experimental results in Fig. 22.11 show, the quantum number of the plateaux does not change monotonically, and the longitudinal resistivity has peaks even in individual Hall plateau regions.

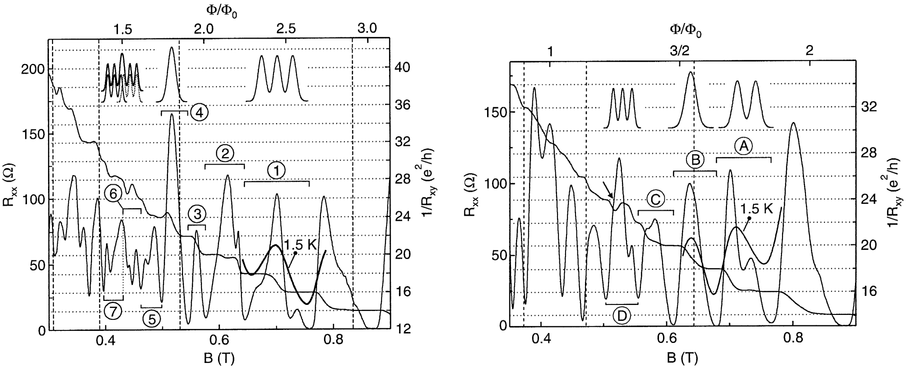


Fig. 22.11. The longitudinal resistance R_{xx} and inverse Hall resistance $1/R_{xy}$ at 50 mK for two superlattices of lattice constants 120 and 100 nm, respectively, fabricated in a semiconductor GaAs/AlGaAs heterostructure. Letters and numbers label the resistance peaks [Reprinted with permission from C. Albrecht et al., *Phys. Rev. Lett.* **86**, 147 (2001). ©2001 by the American Physical Society]

22.3.5 Diamagnetic Susceptibility of Bloch Electrons

The diamagnetic susceptibility of electrons moving in a periodic potential can be calculated in the same way as for free electrons, provided the conditions of semiclassical quantization are met. If the energy spectrum can be characterized by a scalar effective mass m^*, the result can be expressed in the same form as for free electrons, however, the electron mass needs to be replaced by the effective mass. Using the form (22.2.9), the diamagnetic susceptibility is

$$\chi_{\mathrm{dia}} = -\mu_0 \frac{e^2 k_{\mathrm{F}}}{3m^*} \frac{1}{(2\pi)^2} \,. \tag{22.3.34}$$

The formal introduction of the Bohr magneton then gives

$$\chi_{\mathrm{dia}} = -\tfrac{1}{3}\mu_0 \mu_{\mathrm{B}}^2 \frac{k_{\mathrm{F}} m^*}{\pi^2 \hbar^2} \left(\frac{m_{\mathrm{e}}}{m^*}\right)^2 \,. \tag{22.3.35}$$

Since $k_{\mathrm{F}} m^* / \pi^2 \hbar^2$ is the density of states of electrons of effective mass m^* at the Fermi energy,

$$\chi_{\mathrm{dia}} = -\tfrac{1}{3}\mu_0 \mu_{\mathrm{B}}^2 \rho(\varepsilon_{\mathrm{F}}) \left(\frac{m_{\mathrm{e}}}{m^*}\right)^2 \,. \tag{22.3.36}$$

The Bohr magneton appeared naturally in the analogous expression (17.4.43) for the Pauli paramagnetic susceptibility of Bloch electrons, since that was due to the spins, but it showed up in the previous formula only through a formal substitution. Therefore the relative magnitude of the paramagnetic and diamagnetic contributions in the total susceptibility

$$\chi_{\mathrm{m}} = \mu_0 \mu_{\mathrm{B}}^2 \rho(\varepsilon_{\mathrm{F}}) \left[1 - \tfrac{1}{3}\left(\frac{m_{\mathrm{e}}}{m^*}\right)^2\right] \tag{22.3.37}$$

depends on the effective mass. When this is sufficiently low, the diamagnetic contribution can exceed the paramagnetic one, as in bismuth. Note that in addition to the total susceptibility, the contribution of the electron spins can also be measured in experiments using ESR techniques, and so the two contributions may be separated.

22.4 Quantum Oscillations in Magnetic Fields

It was established in the previous section that the separation of Landau levels and the number of degenerate states on each level is proportional to the magnetic field. Therefore the density of states at the Fermi energy has a singularity at each value of the magnetic field where a Landau level becomes completely empty on account of the rearrangement of states. At low temperatures, where the thermal energy $k_{\mathrm{B}}T$ is lower than the magnetic energy $\hbar\omega_{\mathrm{c}}$ separating the Landau levels, this gives rise to jumps and oscillations in other macroscopic properties of the system, too. Such oscillations did not appear in the above expression for the diamagnetic susceptibility because the Euler–Maclaurin formula is a too simple approximation for the sum over the quantum number n of the Landau levels. In fact it can be justified only in magnetic fields for which $\mu_{\mathrm{B}}B$ is smaller than the thermal energy $k_{\mathrm{B}}T$. Below we shall present a more rigorous treatment.

22.4.1 Oscillations in a Two-Dimensional Electron Gas

We shall first examine a two-dimensional electron gas, and choose the direction of the magnetic field to be perpendicular to the plane. This choice is not made solely for the simplicity of the treatment and the possibility of illustrating oscillatory phenomena by an easily tractable example: it is equally motivated by the experimental realizability of the two-dimensional electron gas in which neat oscillations can be measured.

First consider such values of the magnetic field for which the Landau levels are completely filled up to and including the level of quantum number n, while all levels above it are completely empty. Ignoring spins, the $n+1$ filled levels can accommodate

$$N_{\mathrm{e}} = (n+1)N_{\mathrm{p}} = (n+1)\frac{B}{\Phi_0^*}L_xL_y \tag{22.4.1}$$

electrons, provided (22.1.27) is used for N_{p}. Using the converse of this relationship, if the number N_{e} of electrons is given, the level of quantum number n is completely filled while the next one is completely empty in a magnetic field B_n satisfying

$$B_n = \frac{1}{n+1}B_0\,, \qquad \text{where} \qquad B_0 = N_{\mathrm{e}}\frac{\Phi_0^*}{L_xL_y}\,. \tag{22.4.2}$$

B_0 is the field at which the lowest Landau level ($n=0$) is completely filled and all others are empty.

Since the energy of the Landau level is in the middle of the range from which states condense into the Landau level in question upon the application of the magnetic field, and the density of states of a two-dimensional electron gas is independent of the energy, the same number of electrons gain and lose energy in the magnetic field when a Landau level is completely filled, as illustrated in Fig. 22.1. Thus, the ground-state energy at the magnetic fields that satisfy the above condition is obviously the same as in the zero-field case,

$$E_0(B_n) = E_0(B=0)\,. \tag{22.4.3}$$

When the particle number is kept fixed and the magnetic field is increased from B_n to $B > B_n$, the number of allowed states on each level increases. As long as the condition $B < B_{n-1}$ is met, the lowest n Landau levels remain completely filled and the $(n+1)$th (of quantum number n) only partially filled. The total energy of the system is therefore

$$E_0(B) = N_{\mathrm{p}}\sum_{l=0}^{n-1}\hbar\omega_{\mathrm{c}}\bigl(l+\tfrac{1}{2}\bigr) + (N_{\mathrm{e}} - N_{\mathrm{p}}n)\hbar\omega_{\mathrm{c}}\bigl(n+\tfrac{1}{2}\bigr)\,. \tag{22.4.4}$$

The first term is the energy of the electrons on the completely filled levels of quantum numbers $l = 0, 1, \ldots, n-1$, while the second term is the energy of the remaining $N_\mathrm{e} - nN_\mathrm{p}$ electrons on the level of quantum number n. Summation then gives

$$\begin{aligned} E_0(B) &= N_\mathrm{p}\hbar\omega_\mathrm{c}\left[\tfrac{1}{2}n(n-1) + \tfrac{1}{2}n\right] + (N_\mathrm{e} - N_\mathrm{p}n)\hbar\omega_\mathrm{c}\bigl(n+\tfrac{1}{2}\bigr) \\ &= N_\mathrm{e}\hbar\omega_\mathrm{c}\bigl(n+\tfrac{1}{2}\bigr) - \tfrac{1}{2}N_\mathrm{p}\,\hbar\omega_\mathrm{c}n(n+1)\,. \end{aligned} \tag{22.4.5}$$

Making use of (22.1.27), (22.4.2), and the well-known form of the cyclotron frequency,

$$\begin{aligned} E_0(B) &= N_\mathrm{e}\hbar\omega_\mathrm{c}\left[n + \tfrac{1}{2} - \tfrac{1}{2}n(n+1)\frac{B}{B_0}\right] \\ &= N_\mathrm{e}\hbar\frac{eB}{m_\mathrm{e}}\left[n + \tfrac{1}{2} - \tfrac{1}{2}n(n+1)\frac{B}{B_0}\right]. \end{aligned} \tag{22.4.6}$$

It should be stressed that this formula is valid in the region $B_0/(n+1) \le B \le B_0/n$. The piecewise parabolic pattern of the energy vs. magnetic field plot is illustrated in Fig. 22.12. The figure also shows the field dependence of the magnetization, which is just the negative partial derivative of the ground-state energy with respect to B.

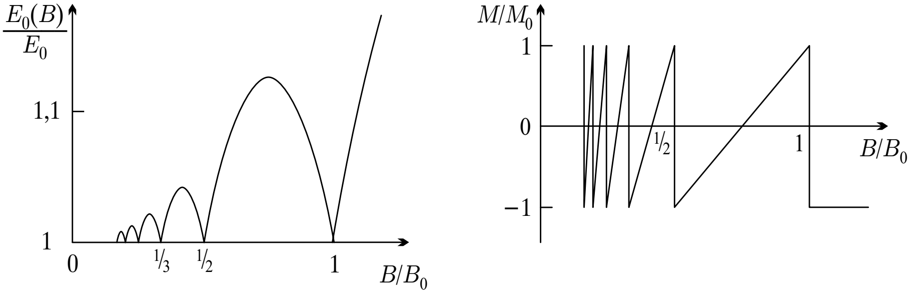


Fig. 22.12. Energy and magnetization of a two-dimensional electron gas as functions of the magnetic field

If the rapid variations of the energy for weak fields were approximated by a smooth curve (with the same area beneath), a quadratically increasing function would be obtained. This indicates that in an average sense the susceptibility is negative – that is, the spinless electron system is diamagnetic. The same conclusion can be drawn from the magnetization vs. magnetic field graph, if proper account is taken of the singularly large negative values of the susceptibility in those points where the magnetization is discontinuous. By taking an average of these and the intermediate regions that give positive contributions, the overall susceptibility is found to be diamagnetic.

As a function of the magnetic field, the energy exhibits kinks and the magnetization has jumps at those values B_n where a Landau level becomes completely empty. It follows from (22.4.2) for B_n that when these quantities are plotted against $1/B$, as in Fig. 22.13, kinks and jumps are spaced at regular distances. In fields where the quantum number of the highest completely filled Landau level is sufficiently large (on the order of hundreds or more), the magnetization shows regular sawtooth oscillations as a function of $1/B$.

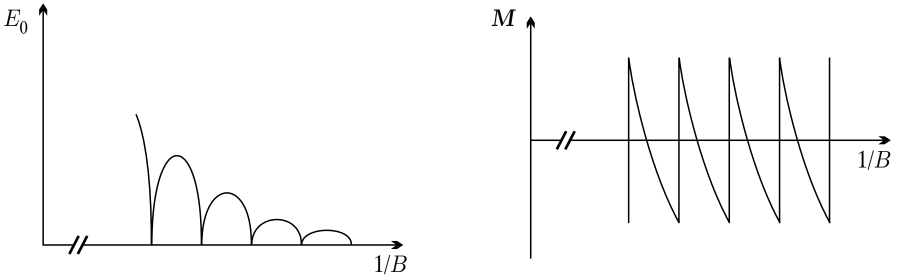


Fig. 22.13. Energy and magnetization of a two-dimensional electron gas as functions of $1/B$

There are several naturally occurring materials in which the motion of electrons can be considered two-dimensional. A prime example is Θ–(BEDT-TTF)$_2$I$_3$,[7] an organic conductor. Figure 22.14 shows the measured magnetization against the magnetic field, as well as against its inverse over a small region. The sawtooth-like pattern is in good agreement with the theoretical predictions of Fig. 22.13.

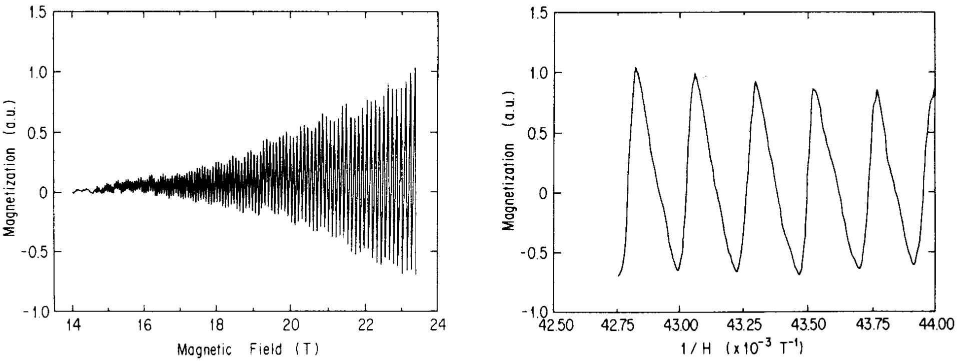


Fig. 22.14. Magnetization of the quasi-two-dimensional Θ–(BEDT-TTF)$_2$I$_3$ in a strong magnetic field as a function of the magnetic field and its inverse [M. Tokumoto et al., *Solid State Commun.* **75**, 439 (1990)]

[7] BEDT-TTF stands for *bis(ethylenedithio)tetrathiafulvalene*.

22.4.2 Energy of a Three-Dimensional Electron Gas in a Magnetic Field

The contribution of the motion in the z-direction also needs to be included in the description of the oscillations in a three-dimensional electron gas. In the presence of a magnetic field the density of states is no longer a set of sharp Dirac delta-like peaks but rather a set of peaks smeared out to one side, as shown in Fig. 22.2. However, the energy and possibly other physical quantities are expected to show oscillations because of the singularities in the density of states.

To determine the ground-state energy we shall first consider those electrons for which the z component of the wave vector is between k_z and $k_z + \mathrm{d}k_z$. Such electrons constitute a quasi-two-dimensional electron gas, whose effective Fermi energy is

$$\varepsilon_{\mathrm{F}}'(k_z) = \varepsilon_{\mathrm{F}} - \frac{\hbar^2 k_z^2}{2m_{\mathrm{e}}} . \tag{22.4.7}$$

At this value of k_z those Landau levels are filled for which

$$n \le \frac{\varepsilon_{\mathrm{F}}'(k_z)}{\hbar\omega_{\mathrm{c}}} - \tfrac{1}{2} , \tag{22.4.8}$$

that is, the quantum number of the highest completely filled level satisfies the condition

$$\frac{\varepsilon_{\mathrm{F}}'(k_z)}{\hbar\omega_{\mathrm{c}}} - \tfrac{3}{2} < n_{\max} \le \frac{\varepsilon_{\mathrm{F}}'(k_z)}{\hbar\omega_{\mathrm{c}}} - \tfrac{1}{2} . \tag{22.4.9}$$

As the magnetic field is increased, this number $n_{\max}$ decreases in unit steps.

There is an important difference compared to the two-dimensional case. When a slice of width $\mathrm{d}k_z$ is considered at a fixed k_z, the states are now either completely filled or completely empty in a Landau subband because when the field is changed, and electrons rearrange themselves, they can end up on another subband with a different k_z. Consequently, the electron number oscillates in the slice in question. The number of occupied states and the energy of the slice exhibit kinks at those values of the field $B_n'(k_z)$ where the energy of the Landau level is the same as the effective Fermi energy:

$$\varepsilon_{\mathrm{F}}'(k_z) = \bigl(n + \tfrac{1}{2}\bigr)\hbar\omega_{\mathrm{c}} = \bigl(n + \tfrac{1}{2}\bigr)\hbar \frac{eB_n'(k_z)}{m_{\mathrm{e}}} , \tag{22.4.10}$$

since for stronger fields the Landau tube of quantum number n moves just outside the Fermi sphere in this height, and so becomes empty. The location of the kinks is given by

$$\frac{1}{B_n'(k_z)} = \bigl(n + \tfrac{1}{2}\bigr)\frac{e\hbar}{m_{\mathrm{e}}}\frac{1}{\varepsilon_{\mathrm{F}}'(k_z)} , \tag{22.4.11}$$

indicating that they are regularly spaced in $1/B$, and their separation is

$$\Delta\left(\frac{1}{B}\right) = \frac{e\hbar}{m_{\mathrm{e}}}\frac{1}{\varepsilon'_{\mathrm{F}}(k_z)}\,. \tag{22.4.12}$$

Instead of the effective Fermi energy of the slice at k_z, the previous formulas can also be expressed in terms of the cross-sectional area

$$\mathcal{A}(k_z) = k_\perp^2\pi = \frac{2m_{\mathrm{e}}}{\hbar^2}\left(\varepsilon_{\mathrm{F}} - \frac{\hbar^2k_z^2}{2m_{\mathrm{e}}}\right)\pi = \frac{2\pi m_{\mathrm{e}}}{\hbar^2}\varepsilon'_{\mathrm{F}}(k_z) \tag{22.4.13}$$

of the Fermi sphere in height k_z. In a given height the Landau tube of quantum number n becomes empty when its quantized cross-sectional area

$$\mathcal{A} = \frac{2\pi eB}{\hbar}\left(n+\tfrac{1}{2}\right) \tag{22.4.14}$$

becomes larger than the cross-sectional area of the Fermi sphere at that height. The period of oscillations is given by

$$\Delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar}\frac{1}{\mathcal{A}(k_z)}\,. \tag{22.4.15}$$

Now consider a magnetic field such that

$$B'_{n+1}(k_z) < B < B'_n(k_z)\,, \tag{22.4.16}$$

that is,

$$\left(n+\tfrac{1}{2}\right)\frac{e\hbar}{m_{\mathrm{e}}}\frac{1}{\varepsilon'_{\mathrm{F}}(k_z)} < \frac{1}{B} < \left(n+1+\tfrac{1}{2}\right)\frac{e\hbar}{m_{\mathrm{e}}}\frac{1}{\varepsilon'_{\mathrm{F}}(k_z)}\,. \tag{22.4.17}$$

According to the foregoing, there are $n+1$ Landau tubes inside the Fermi sphere in height k_z, and they are all filled in this height. To determine the energy for such an intermediate value of B, we have to make use of the formula for the number of electrons on the $n+1$ Landau levels in a slice of thickness at k_z,

$$N_{\mathrm{e}}(B,k_z)\,\mathrm{d}k_z = N_{\mathrm{p}}(n+1)\frac{L_z}{2\pi}\mathrm{d}k_z = (n+1)\frac{eBV}{(2\pi)^2\hbar}\mathrm{d}k_z\,, \tag{22.4.18}$$

which implies that the density of electrons per unit thickness of the slice is

$$\rho(B) = (n+1)\frac{eB}{(2\pi)^2\hbar} = \frac{m_{\mathrm{e}}}{(2\pi\hbar)^2}(n+1)\hbar\omega_{\mathrm{c}}\,. \tag{22.4.19}$$

Considering, in addition to the energy of the oscillators on the completely filled Landau levels, the kinetic energy of the motion in the z-direction (which is the same for each electron of the slice), the energy of the slice is

$$E_0(B,k_z)\,\mathrm{d}k_z = N_{\mathrm{p}}\,\frac{L_z}{2\pi}\,\mathrm{d}k_z\sum_{l=0}^{n}\hbar\omega_{\mathrm{c}}\left(l+\tfrac{1}{2}\right)+N_{\mathrm{p}}\left(n+1\right)\frac{L_z}{2\pi}\,\mathrm{d}k_z\,\frac{\hbar^2k_z^2}{2m_{\mathrm{e}}}\,. \tag{22.4.20}$$

Evaluating the sum in the first term gives

$$E_0(B,k_z)\,\mathrm{d}k_z = N_\mathrm{p}\,\frac{L_z}{2\pi}\,\mathrm{d}k_z\,\hbar\omega_\mathrm{c}\frac{(n+1)^2}{2} + N_\mathrm{p}(n+1)\,\frac{L_z}{2\pi}\,\mathrm{d}k_z\,\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\,. \quad (22.4.21)$$

Using the value of N_p, the energy of the slice can be expressed in terms of the density $\rho(B)$ as

$$E_0(B,k_z)\,\mathrm{d}k_z = \left[\frac{(2\pi\hbar)^2}{2m_\mathrm{e}}\rho^2(B) + \rho(B)\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\right] V\,\mathrm{d}k_z\,. \quad (22.4.22)$$

This shows that between two kinks the energy varies quadratically with the magnetic field.

Let us introduce a second characteristic magnetic field, $B_n(k_z)$, which is defined by the requirement that the number of electron states in the slice of thickness $\mathrm{d}k_z$ at k_z in the presence of the magnetic field be equal to the same in the absence of the field. From our previous results the latter can be easily established by means of the cross-sectional area $\mathcal{A}(k_z)$ of the Fermi sphere in height k_z. This section contains

$$\mathcal{A}(k_z)\left(\frac{2\pi}{L_x}\,\frac{2\pi}{L_y}\right)^{-1} = \frac{m_\mathrm{e}}{2\pi\hbar^2}\varepsilon'_\mathrm{F}(k_z)L_xL_y \quad (22.4.23)$$

allowed vectors $\boldsymbol{k}_\perp$. Since the number of allowed k_z values in a region of thickness $\mathrm{d}k_z$ is $L_z\mathrm{d}k_z/2\pi$, the total number of allowed electron states, neglecting spins, is

$$N_\mathrm{e}(k_z)\,\mathrm{d}k_z = \frac{m_\mathrm{e}}{2\pi\hbar^2}\,\varepsilon'_\mathrm{F}(k_z)L_xL_y\,\frac{L_z}{2\pi}\,\mathrm{d}k_z = m_\mathrm{e}\frac{V}{(2\pi\hbar)^2}\,\varepsilon'_\mathrm{F}(k_z)\,\mathrm{d}k_z\,. \quad (22.4.24)$$

On the other hand, using (22.1.26) for the degree of degeneracy of the Landau levels, the condition for having exactly $n+1$ filled Landau levels in the magnetic field $B_n(k_z)$ is that

$$N_\mathrm{e}(k_z)\,\mathrm{d}k_z = N_\mathrm{p}\,(n+1)\frac{L_z}{2\pi}\,\mathrm{d}k_z = (n+1)\frac{eB_n(k_z)}{(2\pi)^2\hbar}\,V\,\mathrm{d}k_z\,. \quad (22.4.25)$$

Comparison of the two formulas for $N_\mathrm{e}(k_z)$ gives

$$B_n(k_z) = \frac{1}{n+1}\,\frac{m_\mathrm{e}\varepsilon'_\mathrm{F}(k_z)}{\hbar e}\,. \quad (22.4.26)$$

It can be shown that for such magnetic fields $B_n(k_z)$ the total energy of the electrons in the slice is independent of the number of filled Landau levels – just like in the two-dimensional case. Following the arguments that led to (22.4.20) and (22.4.21), the energy contribution of the slice is

$$E_0(B_n,k_z)\,\mathrm{d}k_z = N_\mathrm{p}\,\frac{L_z}{2\pi}\,\mathrm{d}k_z\,\hbar\omega_\mathrm{c}\frac{(n+1)^2}{2} + N_\mathrm{p}(n+1)\,\frac{L_z}{2\pi}\,\mathrm{d}k_z\,\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\,. \quad (22.4.27)$$

To cast it in a more practical form, we introduce the notation $\rho_0 = N_\mathrm{e}(k_z)/V$ for the density of electrons in a region of unit thickness at k_z:

$$\rho_0 = \frac{m_\mathrm{e}}{(2\pi\hbar)^2}\,\varepsilon'_\mathrm{F}(k_z)\,. \tag{22.4.28}$$

Using (22.4.25), this can be written in the equivalent forms

$$\rho_0 = \frac{1}{V}N_\mathrm{p}(n+1)\frac{L_z}{2\pi} = (n+1)\frac{eB_n(k_z)}{(2\pi)^2\hbar}\,. \tag{22.4.29}$$

The energy of the slice is then

$$E_0(B_n,k_z)\,\mathrm{d}k_z = \left[\frac{(2\pi\hbar)^2}{2m_\mathrm{e}}\rho_0^2 + \rho_0\frac{\hbar^2k_z^2}{2m_\mathrm{e}}\right]V\,\mathrm{d}k_z\,. \tag{22.4.30}$$

The last formula directly shows that for such fields $B_n(k_z)$ the energy is always the same, independently of the number of Landau tubes in the given cross section.

It follows directly from the comparison of (22.4.10) and (22.4.26) that

$$B'_{n+1}(k_z) < B_n(k_z) < B'_n(k_z)\,. \tag{22.4.31}$$

Another particularity of the field $B_n(k_z)$ and the associated density ρ_0, which can be seen immediately from the expression (22.4.22) for the quadratically changing energy between $B'_{n+1}(k_z)$ and $B'_n(k_z)$, is that the energy has its minimum at the density given by (22.4.28) – that is, at the field $B_n(k_z)$.

It is useful to add a further term to the formula (22.4.22) of the energy of the slice at k_z that vanishes upon integration with respect to k_z but nonetheless simplifies the energy expression of individual slices. Starting from the magnetic field associated with the energy minimum, more and more states appear in the Landau tubes as the field strength is increased. Until the magnetic field reaches the value B'_n satisfying (22.4.10), more and more electrons arrive in the slice at k_z from slices with different k_z values at which the inflating Landau tubes intersect the Fermi sphere. The number of occupied electron states in the slice at k_z changes by

$$\delta N_\mathrm{e} = (\rho-\rho_0)V\,\mathrm{d}k_z\,. \tag{22.4.32}$$

Since these electrons arrive in this slice from regions where their energy is equal to the Fermi energy, the energy of the other slices is reduced by

$$-\,\varepsilon_\mathrm{F}\,\delta N_\mathrm{e} = -(\rho-\rho_0)\varepsilon_\mathrm{F}\,V\,\mathrm{d}k_z\,. \tag{22.4.33}$$

Adding this term to (22.4.22), the contribution of this slice to the ground-state energy is

$$E_0(B,k_z)\,\mathrm{d}k_z = \left[\frac{(2\pi\hbar)^2}{2m_\mathrm{e}}\rho^2 + \rho\frac{\hbar^2k_z^2}{2m_\mathrm{e}} - (\rho-\rho_0)\varepsilon_\mathrm{F}\right]V\,\mathrm{d}k_z\,. \tag{22.4.34}$$

With respect to its value in the fields B_n, the energy changes by

$$\begin{aligned}\Delta E_0(B,k_z) &= [E_0(B,k_z) - E_0(B_n,k_z)]\,\mathrm{d}k_z \\ &= \left[\frac{(2\pi\hbar)^2}{2m_\mathrm{e}}(\rho^2-\rho_0^2) - (\rho-\rho_0)\left(\varepsilon_\mathrm{F} - \frac{\hbar^2k_z^2}{2m_\mathrm{e}}\right)\right] V\,\mathrm{d}k_z\,.\end{aligned} \tag{22.4.35}$$

Using (22.4.28), the formula for the effective Fermi energy, we have

$$\Delta E_0(B,k_z) = \frac{(2\pi\hbar)^2}{2m_\mathrm{e}}(\rho-\rho_0)^2 V\,\mathrm{d}k_z\,. \tag{22.4.36}$$

For magnetic fields in the range (22.4.16), which is close to B_n, the density of electrons in the slice varies linearly with the magnetic field,

$$\rho - \rho_0 = (n+1)\frac{e}{(2\pi)^2\hbar}\left[B - B_n(k_z)\right], \tag{22.4.37}$$

while the energy of the slice shows a quadratic field dependence. At the boundary of the region, determined by (22.4.10), when the outermost Landau tube at the given height just crosses the Fermi sphere, a jump appears in the occupation number and a kink in the energy. This is illustrated in Fig. 22.15, where the variations are plotted against $1/B$, since the jumps and kinks are spaced at regular distances in $1/B$. When the quantum number of the Landau level is sufficiently high ($n \gg 1$), the jump in the occupation number is

$$\Delta(\rho-\rho_0) = \frac{1}{n}\frac{m_\mathrm{e}\varepsilon_\mathrm{F}'}{(2\pi\hbar)^2}\,. \tag{22.4.38}$$

22.4.3 De Haas–van Alphen Effect

It was demonstrated above that in strong magnetic fields the ground-state energy of the electron gas oscillates as the magnetic field varies. Naturally, such oscillations do not appear in the energy alone but also in other physical quantities that can be derived from the energy, e.g., the magnetization or the susceptibility. This was first observed by L. V. Shubnikov and W. J. de Haas in 1930 in the low-temperature resistivity of bismuth, and later the same year by W. J. de Haas and P. M. van Alphen in the magnetization. The latter phenomenon, called the de Haas–van Alphen effect became particulary important when it was established that the shape of the Fermi surface can be inferred from the frequency of the oscillations and the temperature dependence of the amplitude. Since the calculation is rather tedious for a general Fermi surface, we shall just outline the most important results below.

To determine the magnetization of an electron system in a strong magnetic field, we shall follow the method used for evaluating the energy, and calculate first the contribution of a slice of thickness $\mathrm{d}k_z$ to the magnetization from the field dependence (22.4.36) of the energy:

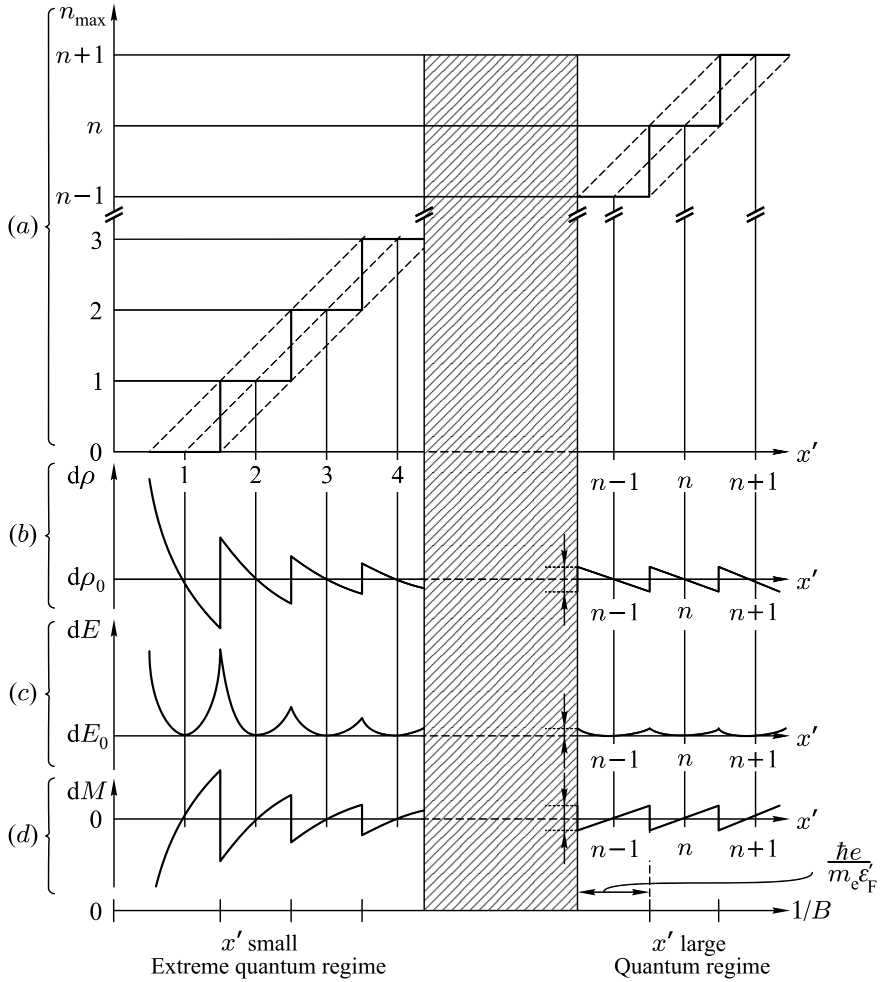


Fig. 22.15. (a) The quantum number of the highest occupied Landau states in the slice of width $\mathrm{d}k_z$ at k_z, as a function of $x' = m_e \varepsilon_F'(k_z)/e\hbar B$. Parts (b), (c), and (d) show the variations of the electron density, energy, and magnetization with x' (that is, practically the inverse magnetic field)

$$\delta M(B, k_z) = -\frac{1}{V}\frac{\partial E_0(B, k_z)}{\partial B} = -\frac{(2\pi\hbar)^2}{m_e}(\rho - \rho_0)\frac{\mathrm{d}\rho}{\mathrm{d}B}\,\mathrm{d}k_z\,. \tag{22.4.39}$$

Making use of (22.4.19) and (22.4.37),

$$\begin{aligned}\delta M(B, k_z) &= -(\rho - \rho_0)(n+1)\frac{e\hbar}{m_e}\,\mathrm{d}k_z \\ &= -(n+1)^2\frac{e^2}{(2\pi)^2 m_e}\big[B - B_n(k_z)\big]\,\mathrm{d}k_z\end{aligned} \tag{22.4.40}$$

that is, the magnetization of the slice is proportional to the field in the region specified by (22.4.17). For Landau levels of sufficiently large quantum numbers, for which

$$B'_n(k_z) - B_n(k_z) = \frac{m_e \varepsilon'_F(k_z)}{e\hbar}\left(\frac{1}{n+\frac{1}{2}} - \frac{1}{n+1}\right) \approx \frac{1}{2}\frac{m_e \varepsilon'_F(k_z)}{e\hbar}\frac{1}{(n+1)^2}\,, \tag{22.4.41}$$

the magnetization varies between $\pm\delta M_{\max}$, where

$$\delta M_{\max} = \frac{1}{2}\frac{e\varepsilon'_F}{(2\pi)^2\hbar}\mathrm{d}k_z\,. \tag{22.4.42}$$

As illustrated in Fig. (22.15), the magnetization is a sawtooth-like periodic function of $1/B$ to a good approximation. It follows from (22.4.12) that when the variable

$$x = \frac{2\pi}{B}\frac{m_e}{e\hbar}\varepsilon'_F = 2\pi\frac{\varepsilon'_F}{\hbar\omega_c} \tag{22.4.43}$$

is used, $\delta M(x)$ is periodic with a period of 2π. Expanding the magnetization into a Fourier series as

$$\delta M(x, k_z) = \mathrm{d}k_z \sum_{l=1}^{\infty} A_l \sin lx\,, \tag{22.4.44}$$

and making use of (C.1.52), the formula for the Fourier series of the sawtooth wave,

$$A_l = \frac{e}{4\pi^3\hbar}\varepsilon'_F\frac{(-1)^l}{l} \tag{22.4.45}$$

is obtained for the Fourier coefficients, and hence

$$\delta M(x, k_z) = \frac{e}{4\pi^3\hbar}\varepsilon'_F\sum_{l=1}^{\infty}(-1)^l\frac{\sin lx}{l}\,\mathrm{d}k_z\,. \tag{22.4.46}$$

The oscillatory part of the total magnetization is obtained by integration with respect to k_z over the Fermi sphere:

$$M_{\text{osc}} = \frac{e}{4\pi^3\hbar}\sum_{l=1}^{\infty}\frac{(-1)^l}{l}\int_{-k_F}^{k_F}\left(\varepsilon_F - \frac{\hbar^2k_z^2}{2m_e}\right)\sin\left[2\pi l\frac{m_e}{e\hbar B}\left(\varepsilon_F - \frac{\hbar^2k_z^2}{2m_e}\right)\right]\mathrm{d}k_z\,. \tag{22.4.47}$$

Owing to the rapid oscillations in the integrand, only the contribution of the $k_z \sim 0$ region is important. When the prefactor ε'_F of the integrand is approximated by ε_F, a Fresnel integral arises. Extending the limits of integration from $\pm k_F$ to $\pm\infty$, and exploiting that

$$\int_0^{\infty}\sin\frac{\pi}{2}x^2\,\mathrm{d}x = \int_0^{\infty}\cos\frac{\pi}{2}x^2\,\mathrm{d}x = \frac{1}{2}\,, \tag{22.4.48}$$

the integral in (22.4.47) can be evaluated and yields

$$M_{\text{osc}} = \frac{e}{4\pi^3\hbar}\varepsilon_{\text{F}}\left(\frac{eB}{\hbar}\right)^{1/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\sin\left(2\pi l\frac{m_{\text{e}}}{e\hbar B}\varepsilon_{\text{F}} - \frac{\pi}{4}\right). \tag{22.4.49}$$

Expressing the coefficient formally in terms of the Bohr magneton $\mu_{\text{B}} = e\hbar/2m_{\text{e}}$, and making use of the relationship between the electron density and the Fermi energy, the oscillatory part of the magnetization can be rewritten as

$$M_{\text{osc}} = \frac{3n_{\text{e}}\mu_{\text{B}}^2 B}{4\pi\varepsilon_{\text{F}}}\left(\frac{2\varepsilon_{\text{F}}}{\hbar\omega_{\text{c}}}\right)^{1/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\sin\left(2\pi l\frac{\varepsilon_{\text{F}}}{\hbar\omega_{\text{c}}} - \frac{\pi}{4}\right). \tag{22.4.50}$$

Comparing the amplitude of the oscillatory term to the nonoscillatory term obtained in the calculation of the Landau diamagnetism,

$$M_{\text{osc}}/M_0 \sim (\varepsilon_{\text{F}}/\hbar\omega_{\text{c}})^{1/2}. \tag{22.4.51}$$

In the magnetic fields customarily used in experiments, this ratio is much larger than one, and so the oscillations are easy to observe. The magnetic susceptibility features very similar oscillations.

22.4.4 Role of Spin in Oscillatory Phenomena

The electron spin was neglected in the foregoing calculations, even though its role is much more than just giving a contribution to the magnetization which is proportional to the Pauli susceptibility: it also gives rise to a spin-dependent energy shift of the Landau levels. This means that the Landau tubes do not move outside the Fermi sphere at the previously determined magnetic fields but at somewhat weaker or stronger fields, depending on the spin quantum number. The magnetization due to the orbital motion shows sawtooth-like oscillations for spin-up and spin-down electrons alike, however the loci of the discontinuities are shifted with respect to the spinless case. The easiest way to incorporate this shift into the calculations is to replace the effective Fermi energy (22.4.7) of the electrons at k_z by the spin-dependent expression

$$\varepsilon'_{\text{F}\sigma}(k_z) = \varepsilon_{\text{F}} - \frac{\hbar^2 k_z^2}{2m_{\text{e}}} + \tfrac{1}{2}g_{\text{e}}\mu_{\text{B}}B\sigma. \tag{22.4.52}$$

Repeating the steps of the previous calculation, the oscillatory part of the magnetization is found to be

$$\begin{aligned} M_{\text{osc}} &= \frac{e}{4\pi^3\hbar}\varepsilon_{\text{F}}\left(\frac{eB}{\hbar}\right)^{1/2}\sum_{\sigma}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\sin\left[2\pi l\frac{m_{\text{e}}}{e\hbar B}\left(\varepsilon_{\text{F}} + \tfrac{1}{2}g_{\text{e}}\mu_{\text{B}}B\sigma\right) - \frac{\pi}{4}\right] \\ &= \frac{e}{2\pi^3\hbar}\varepsilon_{\text{F}}\left(\frac{eB}{\hbar}\right)^{1/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\cos\left(\tfrac{1}{2}\pi l g_{\text{e}}\right)\sin\left(2\pi l\frac{m_{\text{e}}}{e\hbar B}\varepsilon_{\text{F}} - \frac{\pi}{4}\right). \end{aligned} \tag{22.4.53}$$

In terms of the Bohr magneton this reads

$$M_{\mathrm{osc}} = \frac{3n_{\mathrm{e}}\mu_{\mathrm{B}}^2 B}{2\pi\varepsilon_{\mathrm{F}}}\left(\frac{2\varepsilon_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}}\right)^{1/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\cos\left(\tfrac{1}{2}\pi l g_{\mathrm{e}}\right)\sin\left(2\pi l\frac{m_{\mathrm{e}}}{e\hbar B}\varepsilon_{\mathrm{F}} - \frac{\pi}{4}\right). \tag{22.4.54}$$

It should be noted that if the effective mass m^* of the Bloch electrons differs from the electron mass, then m_{e} needs to be replaced by m^* in (22.4.49), since the cyclotron frequency and the energy of Landau levels are determined by the latter. Naturally, the Bohr magneton contains the electron mass, so

$$M_{\mathrm{osc}} = \frac{3n_{\mathrm{e}}\mu_{\mathrm{B}}^2 B}{2\pi\varepsilon_{\mathrm{F}}}\left(\frac{2\varepsilon_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}}\right)^{1/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}}\cos\left(\frac{1}{2}\pi l g_{\mathrm{e}}\frac{m^*}{m_{\mathrm{e}}}\right)\sin\left(2\pi l\frac{\varepsilon_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}} - \frac{\pi}{4}\right). \tag{22.4.55}$$

22.4.5 Oscillations in the Magnetization at Finite Temperatures

The previous calculation can be performed at finite temperatures as well; in this case the free energy has to be determined instead of the energy. If the chemical potential rather than the particle number is fixed, a grand canonical ensemble has to be considered. According to the results of statistical mechanics, the grand canonical potential for a noninteracting fermion gas is

$$\Omega = -k_{\mathrm{B}}T\sum_i \ln\left(1 + \mathrm{e}^{-\beta(\varepsilon_i - \mu)}\right), \tag{22.4.56}$$

where the sum is over all one-particle states of energy ε_i in the system. The free energy is then obtained from the thermodynamic relation

$$F = \Omega + \mu N. \tag{22.4.57}$$

Characterizing the Landau levels by the quantum number n and the wave number k_z, we have to sum over both of them, taking care of the N_{p}-fold degeneracy of the states and the additional double degeneracy due to the electron spin. (This is not true for the level $n = 0$ but the difference is immaterial if $N_{\mathrm{e}} \gg N_{\mathrm{p}}$.) Using (22.1.26), the free energy is

$$F = N_{\mathrm{e}}\mu - 2k_{\mathrm{B}}T\frac{eB}{2\pi\hbar}L_x L_y\sum_{n=0}^{\infty}\sum_{k_z}\ln\left(1 + \mathrm{e}^{-\beta[\varepsilon(n,k_z)-\mu]}\right). \tag{22.4.58}$$

For macroscopic samples the spacing $2\pi/L_z$ of the k_z values is sufficiently small for that the sum can be replaced by an integral:

$$\begin{aligned} F &= N_{\mathrm{e}}\mu - 2k_{\mathrm{B}}T\frac{eB}{2\pi\hbar}L_x L_y\sum_{n=0}^{\infty}\frac{L_z}{2\pi}\int_{-\infty}^{\infty}\ln\left(1 + \mathrm{e}^{-\beta[\varepsilon(n,k_z)-\mu]}\right)\mathrm{d}k_z \\ &= N_{\mathrm{e}}\mu - k_{\mathrm{B}}T\frac{2eB}{\hbar}\frac{V}{(2\pi)^2}\sum_{n=0}^{\infty}\int_{-\infty}^{\infty}\ln\left(1 + \mathrm{e}^{-\beta[\varepsilon(n,k_z)-\mu]}\right)\mathrm{d}k_z\,. \end{aligned} \tag{22.4.59}$$

Since the integrand depends on k_z only through the energy, the k_z-integral can be replaced by an energy integral, making use of the relationship

$$k_z = \pm\sqrt{\frac{2m_\mathrm{e}}{\hbar^2}\left[\varepsilon(n,k_z) - (n+\tfrac{1}{2})\hbar\omega_\mathrm{c}\right]}\,. \tag{22.4.60}$$

The two values k_z that belong to a given energy $\varepsilon(n,k_z)$ contribute equally to the free energy, so we can consider only the branch with the positive sign, and multiply its contribution by two. Since the lower limit of the energy integral for the Landau level of quantum number n is $(n+\frac{1}{2})\hbar\omega_\mathrm{c}$, we find

$$F = N_\mathrm{e}\mu - k_\mathrm{B}T\frac{4eB}{\hbar}\frac{V}{(2\pi)^2}\sum_{n=0}^{\infty}\int\limits_{(n+\frac{1}{2})\hbar\omega_\mathrm{c}}^{\infty}\ln\left(1+\mathrm{e}^{-\beta(\varepsilon-\mu)}\right)\frac{\mathrm{d}k_z}{\mathrm{d}\varepsilon}\,\mathrm{d}\varepsilon\,. \tag{22.4.61}$$

Upon integration by parts, the integrated part vanishes, so

$$F = N_\mathrm{e}\mu - 4eB\frac{V}{(2\pi\hbar)^2}\sum_{n=0}^{\infty}\int\limits_{(n+\frac{1}{2})\hbar\omega_\mathrm{c}}^{\infty}\frac{\sqrt{2m_\mathrm{e}\left[\varepsilon-\left(n+\frac{1}{2}\right)\hbar\omega_\mathrm{c}\right]}}{\mathrm{e}^{\beta(\varepsilon-\mu)}+1}\,\mathrm{d}\varepsilon\,. \tag{22.4.62}$$

The previously derived formula for the ground-state energy, (22.2.4), is recovered in the $T \to 0$ limit. The sum over the quantum number n was approximated using the Euler–Maclaurin formula; that is how we arrived at the Landau diamagnetic susceptibility. In retrospect we can see that that procedure was not sufficiently precise, since it did not account for the oscillatory correction. We shall therefore return to (22.4.59) and try to evaluate it more accurately.

To this end, we shall use the Poisson summation formula, which asserts that

$$\begin{aligned}\sum_{n=0}^{\infty} f(n+\tfrac{1}{2}) &= \int\limits_0^{\infty} f(x)\,\mathrm{d}x + 2\sum_{l=1}^{\infty}\int\limits_0^{\infty} f(x)\cos\left[2\pi l\left(x-\tfrac{1}{2}\right)\right]\mathrm{d}x \\ &= \int\limits_0^{\infty} f(x)\,\mathrm{d}x + 2\sum_{l=1}^{\infty}(-1)^l\int\limits_0^{\infty} f(x)\cos(2\pi l x)\,\mathrm{d}x\,.\end{aligned} \tag{22.4.63}$$

This can be easily proved using the relation

$$\begin{aligned}\sum_{n=-\infty}^{\infty}\delta\left[x-\left(n+\tfrac{1}{2}\right)\right] &= \sum_{l=-\infty}^{\infty}\mathrm{e}^{2\pi i l(x-\frac{1}{2})} = 1 + 2\sum_{l=1}^{\infty}\cos\left[2\pi l\left(x-\tfrac{1}{2}\right)\right] \\ &= 1 + 2\sum_{l=1}^{\infty}(-1)^l\cos(2\pi l x)\,,\end{aligned} \tag{22.4.64}$$

which is in fact the Fourier representation of an infinite set of periodically spaced Dirac delta peaks. By multiplying both sides by a function $f(x)$ and integrating over the interval $(0,\infty)$, the above result is indeed recovered.

We shall now apply (22.4.63) to the thermodynamic potential (22.4.59) of the grand canonical ensemble. Considering the spinless case first,

$$\begin{aligned}\Omega = &- 2k_{\mathrm{B}}T\frac{eB}{\hbar}\frac{V}{(2\pi)^2}\int\limits_0^\infty\int\limits_{-\infty}^\infty \ln\left(1+\mathrm{e}^{-\beta[\varepsilon(x,k_z)-\mu]}\right)\mathrm{d}k_z\,\mathrm{d}x\\ &- 4k_{\mathrm{B}}T\frac{eB}{\hbar}\frac{V}{(2\pi)^2}\sum_{l=1}^\infty(-1)^l\int\limits_0^\infty\int\limits_{-\infty}^\infty \ln\left(1+\mathrm{e}^{-\beta[\varepsilon(x,k_z)-\mu]}\right)\\ &\qquad\times\cos(2\pi lx)\,\mathrm{d}k_z\,\mathrm{d}x\,,\end{aligned}\tag{22.4.65}$$

where

$$\varepsilon(x,k_z) = \hbar\omega_{\mathrm{c}}x + \frac{\hbar^2k_z^2}{2m_{\mathrm{e}}}\,.\tag{22.4.66}$$

Taking a term of the l-sum, we shall consider the expression

$$I_l(k_z) = \int\limits_0^\infty \ln\left(1+\mathrm{e}^{-\beta[\varepsilon(x,k_z)-\mu]}\right)\cos(2\pi lx)\,\mathrm{d}x\tag{22.4.67}$$

before integration with respect to k_z. Making use of the result

$$-k_{\mathrm{B}}T\frac{\partial}{\partial\varepsilon}\ln\left[1+\mathrm{e}^{-\beta[\varepsilon(x,k_z)-\mu]}\right] = f_0(\varepsilon)\,,\tag{22.4.68}$$

and integrating by parts twice, we have

$$\begin{aligned}I_l(k_z) = &\frac{1}{4\pi^2l^2k_{\mathrm{B}}T}\left[f_0(\varepsilon)\frac{\partial\varepsilon}{\partial x}\right]_{x=0}\\ &+\int\limits_0^\infty\frac{\cos(2\pi lx)}{4\pi^2l^2k_{\mathrm{B}}T}\left[f_0(\varepsilon)\frac{\partial^2\varepsilon}{\partial x^2}+\frac{\partial f_0(\varepsilon)}{\partial x}\frac{\partial\varepsilon}{\partial x}\right]\mathrm{d}x\,.\end{aligned}\tag{22.4.69}$$

The first term can be ignored as it does not give an oscillatory contribution. Since ε is linear in x, and $\partial f_0/\partial x$ is nonvanishing only near the Fermi energy,

$$I_l(k_z) = \frac{\hbar\omega_{\mathrm{c}}}{4\pi^2l^2k_{\mathrm{B}}T}\int\limits_{-\infty}^\infty\cos(2\pi lx)\frac{\partial f_0(\varepsilon(x))}{\partial x}\,\mathrm{d}x\,.\tag{22.4.70}$$

The Sommerfeld expansion cannot be used now because the cosine function oscillates too rapidly. However, as the negative derivative of the Fermi function has its maximum at that particular value of x where the energy is equal to the chemical potential, for each value of k_z we shall seek the value $x = X$ for which

$$\varepsilon(X, k_z) = \mu\,. \tag{22.4.71}$$

Exploiting the property that the chemical potential is practically independent of the magnetic field,

$$\hbar\omega_{\rm c} X + \frac{\hbar^2 k_z^2}{2m_{\rm e}} = \varepsilon_{\rm F}\,. \tag{22.4.72}$$

Using (22.4.13), X can be related to the cross-sectional area $\mathcal{A}(k_z)$ of the Fermi sphere in height k_z:

$$X = \frac{\varepsilon_{\rm F}'}{\hbar\omega_{\rm c}} = \frac{\hbar}{2\pi eB}\mathcal{A}(k_z)\,. \tag{22.4.73}$$

Expanding the integration variable about $x = X$, and changing the variable from x to $\eta = \hbar\omega_{\rm c}(x - X)/k_{\rm B}T$,

$$I_l(k_z) = \frac{\hbar\omega_{\rm c}}{4\pi^2 l^2 k_{\rm B}T} \int\limits_{-\infty}^{\infty} \cos\left[2\pi l \left(X + \frac{k_{\rm B}T}{\hbar\omega_{\rm c}}\eta\right)\right] \frac{{\rm d}f_0}{{\rm d}\eta}\,{\rm d}\eta \tag{22.4.74}$$

is obtained. Since

$$\frac{{\rm d}f_0(x)}{{\rm d}x} = -\frac{{\rm e}^x}{({\rm e}^x + 1)^2} = -\frac{1}{4\cosh^2(x/2)} \tag{22.4.75}$$

is an even function of x, we have

$$I_l(k_z) = \frac{\hbar\omega_{\rm c}}{4\pi^2 l^2 k_{\rm B}T}\cos(2\pi l X) \int\limits_{-\infty}^{\infty} \cos\left(\frac{2\pi l k_{\rm B}T}{\hbar\omega_{\rm c}}\eta\right) \frac{{\rm d}f_0}{{\rm d}\eta}\,{\rm d}\eta\,. \tag{22.4.76}$$

The integral can be evaluated exactly, using

$$\int\limits_0^{\infty} \frac{\cos ax}{\cosh^2 \beta x}{\rm d}x = \frac{a\pi}{2\beta^2 \sinh(a\pi/2\beta)}\,, \tag{22.4.77}$$

which leads to

$$I_l(k_z) = -\frac{1}{2l}\frac{1}{\sinh(2\pi^2 l k_{\rm B}T/\hbar\omega_{\rm c})}\cos\left(\frac{l\hbar\mathcal{A}(\mu, k_z)}{eB}\right). \tag{22.4.78}$$

The periodicity in $1/B$ comes from the cosine function.

Writing this expression back into the thermodynamic potential, the oscillatory part that we are interested in reads

$$\varOmega_{\rm osc} = k_{\rm B}T \sum_{l=1}^{\infty} (-1)^l \int\limits_{-\infty}^{\infty} \frac{eB}{2\pi^2\hbar} g(l) \cos\left(\frac{l\hbar\mathcal{A}(\mu, k_z)}{eB}\right) {\rm d}k_z\,, \tag{22.4.79}$$

where

$$g(l) = \frac{1}{l} \frac{1}{\sinh(2\pi^2 l k_\mathrm{B} T/\hbar\omega_\mathrm{c})} . \tag{22.4.80}$$

The most important contribution to the integral comes from that region of k_z where the cosine function varies slowly, that is, the cross-sectional area is stationary,

$$\frac{\partial \mathcal{A}(\mu, k_z)}{\partial k_z} = 0 . \tag{22.4.81}$$

The series expansion about such a point k_0 leads to a quadratic variation of the cross-sectional area:

$$\mathcal{A} = \mathcal{A}_0 - \tfrac{1}{2} k'^2 \mathcal{A}_0'' + \dots , \tag{22.4.82}$$

where $k' = k_z - k_0$, and $\mathcal{A}_0''$ is negative if the cross-sectional area of the Fermi surface has a local minimum. Obviously, for the spherical Fermi surface of free electrons the region $k_z \approx 0$ gives the largest contribution, and $\mathcal{A}_0'' = 2\pi$. Inserting this series expansion into the formula for the thermodynamic potential, we have

$$\begin{aligned}
\Omega_\mathrm{osc} &= k_\mathrm{B} T \sum_{l=1}^{\infty} (-1)^l \frac{eB}{2\pi^2 \hbar} g(l) \int\limits_{-\infty}^{\infty} \cos\left(\frac{l\hbar}{eB} (\mathcal{A}_0 - \tfrac{1}{2} k'^2 \mathcal{A}_0'') \right) \mathrm{d}k' \\
&= k_\mathrm{B} T \sum_{l=1}^{\infty} (-1)^l \frac{eB}{2\pi^2 \hbar} g(l) \left(\frac{2\pi eB}{l\hbar |\mathcal{A}_0''|} \right)^{1/2} \cos\left(\frac{l\hbar \mathcal{A}_0}{eB} - \frac{\pi}{4} \right) \\
&= 2 k_\mathrm{B} T |\mathcal{A}_0''|^{-1/2} \sum_{l=1}^{\infty} (-1)^l \left(\frac{eB}{2\pi l \hbar} \right)^{3/2} \frac{1}{\sinh(2\pi^2 l k_\mathrm{B} T/\hbar\omega_\mathrm{c})} \\
&\quad \times \cos\left(\frac{l\hbar \mathcal{A}_0}{eB} - \frac{\pi}{4} \right) .
\end{aligned} \tag{22.4.83}$$

The amplitude of oscillations is small in the thermodynamic potential – and thus in the free energy, too. In the $T \to 0$ limit the oscillatory part of Ω is proportional to the 5/2th power of $\hbar\omega_\mathrm{c}$. Comparison with the formula for the ground-state energy at $B = 0$ gives

$$\frac{\Omega_\mathrm{osc}}{E_0(B = 0)} \sim (\hbar\omega_\mathrm{c}/\varepsilon_\mathrm{F})^{5/2} . \tag{22.4.84}$$

The oscillatory term is found to be even smaller than the field-dependent but not oscillatory term that gives rise to Landau diamagnetism – which is smaller than $E_0(B = 0)$ by a factor of order $(\hbar\omega_\mathrm{c}/\varepsilon_\mathrm{F})^2$.

However, this is not the case for the magnetization. The dominant contribution comes from the derivative of the argument of the cosine function with respect to the magnetic field:

$$
\begin{aligned}
M_{\mathrm{osc}} = k_{\mathrm{B}}T|\mathcal{A}_0''|^{-1/2}\frac{e\mathcal{A}_0}{2\pi^2\hbar}\left(\frac{eB}{2\pi\hbar}\right)^{-1/2}\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{3/2}} \\
\times\frac{1}{\sinh(2\pi^2 lk_{\mathrm{B}}T/\hbar\omega_{\mathrm{c}})}\sin\left(\frac{l\hbar\mathcal{A}_0}{eB}-\frac{\pi}{4}\right).
\end{aligned}
\tag{22.4.85}
$$

This expression is valid for a general Fermi surface, and is known as the *Lifshitz–Kosevich formula.*[8] At $T = 0$ (22.4.49) is recovered, and the assertion made on page 317 that the oscillatory part has a larger amplitude than the nonoscillatory part is also established. Since at finite temperatures an additional multiplicative factor

$$
\frac{2\pi^2 lk_{\mathrm{B}}T/\hbar\omega_{\mathrm{c}}}{\sinh(2\pi^2 lk_{\mathrm{B}}T/\hbar\omega_{\mathrm{c}})}
\tag{22.4.86}
$$

appears, the amplitude of the oscillation decreases exponentially with increasing temperature when $k_{\mathrm{B}}T \gg \hbar\omega_{\mathrm{c}}$. Therefore the experimental observation of the de Haas–van Alphen effect is possible only at very low temperatures.

When the contribution of the coupling between the electron spin and the magnetic field is also taken into account in the energy, an additional factor

$$
\cos\left(\tfrac{1}{2}\pi lg_{\mathrm{e}}\right)
\tag{22.4.87}
$$

appears in the oscillatory part of the magnetization (just like in the zero-temperature case). If, moreover, proper care is taken of the subtlety that for Bloch electrons the cyclotron frequency – which determines the energy of the Landau levels – contains the cyclotron mass rather than the electron mass, and therefore the Zeeman splitting due to spins is not the same as the separation of the Landau levels, the above factor is replaced by

$$
\cos\left(\tfrac{1}{2}\pi lg_{\mathrm{e}}\frac{m^*}{m_{\mathrm{e}}}\right).
\tag{22.4.88}
$$

Keeping the nonoscillatory parts, too, for an electron system that can be characterized by a scalar effective mass, and for which the area $\mathcal{A}_0$ of the cross section of maximum diameter is related to the Fermi energy by

$$
\mathcal{A}_0 = \frac{2\pi m^*}{\hbar^2}\varepsilon_{\mathrm{F}}\,,
\tag{22.4.89}
$$

the final result for the magnetization is

$$
\begin{aligned}
M = \mu_{\mathrm{B}}^2\rho(\varepsilon_{\mathrm{F}})B\Biggl[\left(\frac{g_{\mathrm{e}}}{2}\right)^2-\frac{1}{3}\left(\frac{m_{\mathrm{e}}}{m^*}\right)^2+\frac{2\pi k_{\mathrm{B}}T}{\hbar\omega_{\mathrm{c}}}\left(\frac{2\varepsilon_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}}\right)^{1/2} \\
\times\sum_{l=1}^{\infty}\frac{(-1)^l}{l^{1/2}}\cos\left(\tfrac{1}{2}\pi lg_{\mathrm{e}}\frac{m^*}{m_{\mathrm{e}}}\right)\frac{\sin\left(2\pi l\varepsilon_{\mathrm{F}}/\hbar\omega_{\mathrm{c}}-\pi/4\right)}{\sinh(2\pi^2 lk_{\mathrm{B}}T/\hbar\omega_{\mathrm{c}})}\Biggr].
\end{aligned}
\tag{22.4.90}
$$

[8] I. M. Lifshitz and A. M. Kosevich, 1955. The result for a spherical Fermi surface was derived by L. D. Landau in 1939.

At $T = 0$ the oscillatory term is the same as in (22.4.55).

Deriving the susceptibility from the magnetization,

$$\begin{aligned}\chi = \mu_0 \mu_{\mathrm{B}}^2 \rho(\varepsilon_{\mathrm{F}}) \Bigg[\left(\frac{g_{\mathrm{e}}}{2}\right)^2 - \frac{1}{3}\left(\frac{m_{\mathrm{e}}}{m^*}\right)^2 + \frac{\pi k_{\mathrm{B}} T}{\varepsilon_{\mathrm{F}}} \left(\frac{2\varepsilon_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}}\right)^{3/2} \\ \times \sum_{l=1}^{\infty} \frac{(-1)^l}{l^{1/2}} \cos\left(\tfrac{1}{2}\pi l g_{\mathrm{e}} \frac{m^*}{m_{\mathrm{e}}}\right) \frac{\sin\left(2\pi l \varepsilon_{\mathrm{F}}/\hbar\omega_{\mathrm{c}} - \pi/4\right)}{\sinh(2\pi^2 l k_{\mathrm{B}} T/\hbar\omega_{\mathrm{c}})} \Bigg] ,\end{aligned} \tag{22.4.91}$$

which contains both the Pauli susceptibility due to spins and the Landau diamagnetic susceptibility.

22.4.6 Oscillations for General Fermi Surfaces

The previous procedure can be applied not only to electron systems with a spherical Fermi surface but also to arbitrarily shaped Fermi surfaces. The calculations showed that the slices at different k_z give oscillations of different frequencies, but these interfere and cancel out, and only the oscillation due to electrons in stationary cross sections survives. As (22.4.85) indicates, the sine function that characterizes the oscillations in the magnetization is of the form

$$\sin\left(l \frac{\hbar}{eB} \mathcal{A}_0 - \frac{\pi}{4}\right) , \tag{22.4.92}$$

where $\mathcal{A}_0$ is the area of the stationary cross section of the Fermi surface perpendicular to the magnetic field, and so the oscillations are, once again, regularly spaced in $1/B$, with a spacing of

$$\Delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar} \frac{1}{\mathcal{A}_0} . \tag{22.4.93}$$

If the Fermi surface has two extremal cross sections in a given direction then oscillations appear at both frequencies. Figure 22.16 shows such an example, the experimental results for zinc when the applied magnetic field is along a characteristic crystallographic direction of the sample. A higher-frequency oscillation is superposed on the slower variation, indicating the presence of two extremal cross sections.

Thus, when the magnetization is plotted as a function of $1/B$, the period of oscillation gives directly the maximal and minimal cross-sectional areas of the Fermi surface perpendicular to the field. By measuring the oscillation period in different directions, information can be obtained about the shape of the Fermi surface, while the effective mass can be inferred from the temperature dependence of the amplitude.

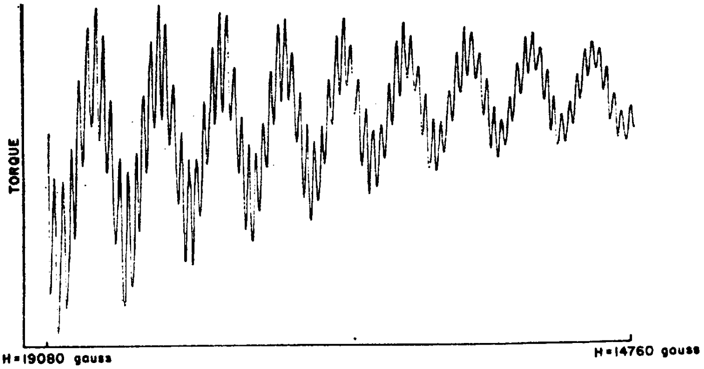


Fig. 22.16. De Haas–van Alphen oscillations in zinc [Reprinted with permission from A. S. Joseph and W. L. Gordon, *Phys. Rev.* **126**, 489 (1962). ©1962 by the American Physical Society]

22.4.7 Experimental Results

One possibility for studying the de Haas–van Alphen effect is to measure the frequency of the oscillations of the magnetization. Denoting the frequency in magnetic field by ν – which is defined by the requirement that the argument of the sine function characterizing the oscillations should contain integral multiples of $2\pi\nu/B$ –, the extremal cross section of the Fermi surface and the frequency are related by

$$\mathcal{A} = \frac{2\pi e}{\hbar}\nu \,. \tag{22.4.94}$$

Figure 22.17 shows the oscillations obtained for copper. By applying a magnetic field in the [111] direction, the superposition of a slow and a rapid oscillation is observed. They correspond to the two stationary cross sections of the Fermi surface perpendicular to the [111] direction. To see this better, the Fermi surface of copper shown in Fig. 19.4(*b*) is also presented, but this time in the repeated-zone scheme.

The spherical regions of the Fermi surface are connected by "necks" in the [111] direction. In the perpendicular direction the smallest-area section is at the neck, while the largest-area section at the great circle of the sphere. This is the "belly" of the Fermi surface. Measurements in magnetic fields applied in other directions would show "dog bone" shaped sections.

It should be emphasized that oscillations can be observed only in sufficiently pure materials and at low enough temperatures. As discussed earlier, the large-amplitude oscillations are smeared out at finite temperatures. Scattering by impurities, which gives rise to a finite relaxation time τ, has a similar effect. The quantized Landau levels are broadened, and an extra exponential factor,

$$\exp(-2\pi/(\omega_{\mathrm{c}}\tau)) = \exp(-2\pi m_{\mathrm{e}}/eB\tau) \,, \tag{22.4.95}$$

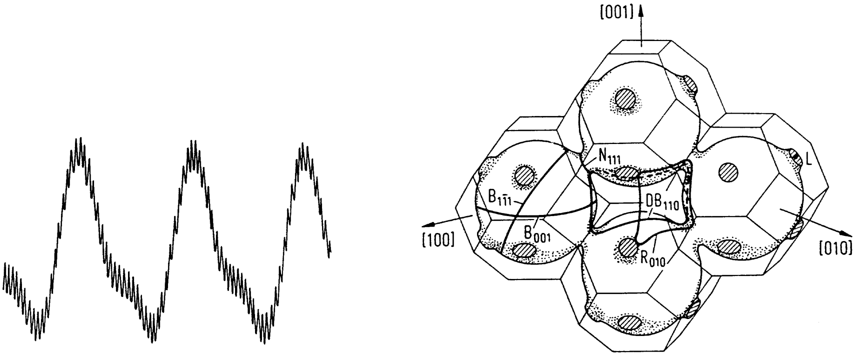


Fig. 22.17. Characteristic de Haas–van Alphen oscillations in copper, as a function of the strength of the magnetic field applied in the [111] direction [Reprinted with permission from A. S. Joseph et al., *Phys. Rev.* **148**, 569 (1966). ©1966 by the American Physical Society], and the Fermi surface of copper in the repeated-zone scheme

called the *Dingle factor*,[9] appears in the amplitude of the oscillations. The amplitude is considerably reduced when the relaxation time becomes comparable to or smaller than the reciprocal of ω_c. Writing the Dingle factor as

$$\exp(-2\pi^2 k_B T_D/\hbar\omega_c)\,, \tag{22.4.96}$$

the thermal energy that corresponds to the Dingle temperature T_D needs to be small compared to the magnetic energy for that the de Haas–van Alphen effect could be observed.

It was mentioned in connection with the data for the coefficient of the linear term in the low-temperature specific heat of metals listed in Table 16.7 that in experiments carried out in the 1980s γ was found to be several orders of magnitude larger in a family of materials (certain cerium and uranium compounds) than its usual value, indicating a large density of states and a high effective mass. Examining the Fermi surface using de Haas–van Alphen techniques, the cyclotron mass can be determined from the temperature-dependent coefficient

$$\frac{1}{\sinh(2\pi^2 k_B T/\hbar\omega_c)} \tag{22.4.97}$$

in the amplitude of the oscillatory terms. By fitting the amplitude of the oscillations, the value obtained for m_c/m_e was between 11 and 40 for CeCu$_6$[10] and between 25 and 90 for UPt$_3$.[11]

[9] R. B. Dingle, 1952.

[10] P. H. P. Reinders et al., *Phys. Rev. Lett.* **57**, 1631 (1986).

[11] L. Taillefer and G. G. Lonzarich, *Phys. Rev. Lett.* **60**, 1570 (1988).

22.4.8 Further Oscillation Phenomena

In addition to the magnetization, oscillations may also occur in other physical quantities. As mentioned on page 314, L. V. Shubnikov and W. J. de Haas observed such behavior in the dependence of resistivity on the applied magnetic field. This is the *Shubnikov–de Haas effect.* The oscillations of the low-temperature resistivity of the organic conductor β–(BEDT-TTF)$_2$I$_3$ in strong magnetic fields is plotted against the field strength in Fig. 22.18.

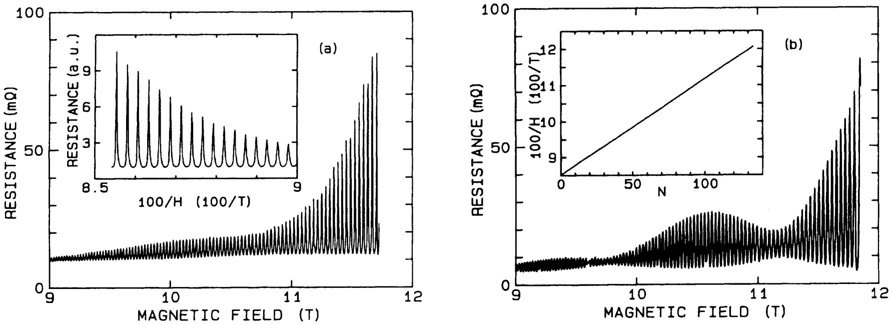


Fig. 22.18. Shubnikov–de Haas oscillations in two samples of β–(BEDT-TTF)$_2$I$_3$, at low temperature ($T = 380\,\text{mK}$) [Reprinted with permission from W. Kang et al., *Phys. Rev. Lett.* **62**, 2559 (1989). ©1989 by the American Physical Society]

Similar oscillations have been observed in the thermoelectric power as well as the thermal conductivity. The anomalous behavior of the Hall effect in strong magnetic fields will be discussed separately, in Chapter 24.

Further Reading

1. A. A. Abrikosov, *Fundamentals of the Theory of Metals*, North Holland, Amsterdam (1988).
2. I. M. Lifshitz, M. Ya. Azbel' and M. I. Kaganov, *Electron Theory of Metals*, Consultants Bureau, New York (1973).
3. W. Mercouroff, *La surface de Fermi des mètaux*, Collection de Monographies de Physique, Masson et C$^{\text{ie}}$, Éditeurs, Paris (1967).
4. D. Shoenberg, *Magnetic Oscillation in Metals*, Cambridge University Press, Cambridge (1984).

23

Electrons in Thermally Vibrating Lattices

Up to now we have either neglected the potential due to ion cores or considered it strictly periodic in our calculations of the electron states. The thermal motion of ions has been completely ignored. Similarly, the effect of electrons on the dynamics of ions was taken into account only by a mean static potential, paying no attention to the possibility that the motion of the electrons may influence that of the ions. Instead of this approach, the coupled system of electrons and ions should have been studied, and the problem of electron states and lattice vibrations should have been solved simultaneously. More precisely, instead of assuming a lattice-periodic potential, the eigenstates of electrons should have been calculated in the field of ions oscillating about their equilibrium positions, while the vibration of ions should have been studied in a fluctuating sea of electrons. In this chapter we shall first demonstrate that the two kinds of degrees of freedom can be treated separately in general – that is, the Born–Oppenheimer approximation,[1] also called the *adiabatic decoupling*, is justified. This also means that the effects of lattice vibrations on the electron states can be studied in perturbation theory. For this reason, it is practical to write the Hamiltonian of the interaction between the vibrating lattice and the electrons in second-quantized form. In this representation the Hamiltonian appears as if the electrons absorbed and emitted phonons due to the interactions. Thus, even though the periodicity of the potential is broken by the motion of the ions, the electron states can still be characterized by a wave vector $\boldsymbol{k}$ – but it is no longer conserved. The electrons can be scattered into a state with a different wave vector, transferring energy and momentum to the phonon subsystem, and so both the electron states and phonons have finite lifetimes. After specifying the interaction Hamiltonian, we shall explore what other consequences these scattering processes have on electrons and lattice vibrations.

[1] M. Born and J. R. Oppenheimer, 1927.

23.1 Adiabatic Decoupling

As mentioned in Chapter 3, the full Hamiltonian of the system of electrons and ions, including their interactions, can be written as a sum of three terms, just like in (3.1.8):

$$\mathcal{H} = \mathcal{H}_{\text{el}} + \mathcal{H}_{\text{ion}} + \mathcal{H}_{\text{el–ion}} \,. \tag{23.1.1}$$

The first term, $\mathcal{H}_{\text{el}}$, which depends only on the position coordinates $\boldsymbol{r}_i$ of the electrons, contains the kinetic energy and the mutual Coulomb interactions of the electrons, as in (3.1.4). Instead of applying the notation used in Chapter 11, we shall denote the instantaneous position vector of the ions by $\boldsymbol{R}_l$ for clarity. The label l now contains both the index m of the primitive cell and the index μ that distinguishes the atoms of the basis from each other. Assuming pair potentials among the ions, the Hamiltonian $\mathcal{H}_{\text{ion}}$, which consists of the kinetic energy and the mutual interactions of the ions, reads

$$\mathcal{H}_{\text{ion}} = \mathcal{H}_{\text{ion}}^{\text{kin}} + \mathcal{H}_{\text{ion–ion}} = -\sum_l \frac{\hbar^2}{2M_l}\frac{\partial^2}{\partial \boldsymbol{R}_l^2} + \tfrac{1}{2}\sum_{l,l'} U_{\text{ion}}(\boldsymbol{R}_l - \boldsymbol{R}_{l'}) \,. \tag{23.1.2}$$

Finally, the part $\mathcal{H}_{\text{el–ion}}$, which describes the interactions between electrons and ions, depends on the coordinates of both electrons and ions:

$$\mathcal{H}_{\text{el–ion}} \equiv U_{\text{el–ion}}(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_i, \dots, \boldsymbol{r}_{N_\text{e}}, \boldsymbol{R}_1, \boldsymbol{R}_2, \dots, \boldsymbol{R}_l, \dots, \boldsymbol{R}_N) \,. \tag{23.1.3}$$

Most of our considerations are valid without any restrictions on the form of the electron–ion potential. Nonetheless, for simplicity, we shall sometimes assume that the potential is essentially the sum of electron–ion pair potentials, that is,

$$\mathcal{H}_{\text{el–ion}} = \sum_{i,l} U'_{\text{el–ion}}(\boldsymbol{r}_i - \boldsymbol{R}_l) \,. \tag{23.1.4}$$

Below we shall use the notation $\{\boldsymbol{r}_i\}$ for the position vectors of all electrons $(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_{N_\text{e}})$ and $\{\boldsymbol{R}_l\}$ for those of all ions $(\boldsymbol{R}_1, \boldsymbol{R}_2, \dots, \boldsymbol{R}_N)$. In terms of these,

$$\mathcal{H}_{\text{el–ion}} \equiv \mathcal{H}_{\text{el–ion}}(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_l\}) \,, \tag{23.1.5}$$

and the total wavefunction of the system is written concisely as

$$\Psi(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_l\}) \,. \tag{23.1.6}$$

To calculate this wavefunction and the total energy of the system, the Schrödinger equation for the full Hamiltonian would need to be solved, without the simplifications arising from the invariance under discrete translations. In this general case it is impossible to find the complete solution of the problem. However, as Born and Oppenheimer pointed out, the characteristic velocity of electrons in solids, the Fermi velocity ($v_\text{F} \sim 10^6$ m/s) is much larger than the sound velocity ($c_\text{s} \sim 10^3$ m/s), which is the characteristic velocity of the vibrating ion system. Hence ions are affected by an averaged electron distribution, whereas electrons feel the instantaneous positions of the ions, following their motion adiabatically. With this assumption, the wavefunction of the electrons can be determined by first considering ions to be stationary at their instantaneous nonequilibrium positions $\boldsymbol{R}_1$, $\boldsymbol{R}_2, \dots \boldsymbol{R}_l, \dots, \boldsymbol{R}_N$, and solving the Schrödinger equation

$$(\mathcal{H}_{\text{el}} + \mathcal{H}_{\text{el–ion}})\, \phi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) = E_{\text{el}}(\{\boldsymbol{R}_l\})\phi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) \tag{23.1.7}$$

in this nonperiodic potential, where $\phi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})$ is the wavefunction of the electron states with the positions of the ions "fixed".

The solutions of this Schrödinger equation, indexed by the label n, constitute a complete orthonormal set. The wavefunction $\Psi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})$ of the complete system of electrons and ions can be expanded in terms of them, allowing the coefficients $\Phi_n(\{\boldsymbol{R}_l\})$ to depend on the coordinates of the ions:

$$\Psi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) = \sum_n \Phi_n(\{\boldsymbol{R}_l\})\, \phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\,. \tag{23.1.8}$$

Acting on this wavefunction by the full Hamiltonian, and exploiting the relation (23.1.7) for $\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})$ as well as the orthonormality of the wavefunctions,

$$\begin{aligned}\Big\{ -\sum_l \frac{\hbar^2}{2M_l}\frac{\partial^2}{\partial \boldsymbol{R}_l^2} + \tfrac{1}{2}\sum_{l,l'} U(\boldsymbol{R}_l - \boldsymbol{R}_{l'}) + E_n^{\text{el}}(\{\boldsymbol{R}_l\})\Big\}\Phi_n(\{\boldsymbol{R}_l\}) \quad (23.1.9)\\ -\sum_l \sum_{n'} \Big\{ M_{n,n'} \frac{\partial \Phi_{n'}(\{\boldsymbol{R}_l\})}{\partial \boldsymbol{R}_l} + N_{n,n'}\, \Phi_{n'}(\{\boldsymbol{R}_l\})\Big\} = E\Phi_n(\{\boldsymbol{R}_l\})\,,\end{aligned}$$

where

$$M_{n,n'} = \frac{\hbar^2}{M_l}\int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) \frac{\partial \phi_{n'}(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{R}_l} \prod_i \mathrm{d}\boldsymbol{r}_i \tag{23.1.10}$$

and

$$N_{n,n'} = \frac{\hbar^2}{2M_l}\int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) \frac{\partial^2 \phi_{n'}(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{R}_l^2} \prod_i \mathrm{d}\boldsymbol{r}_i\,. \tag{23.1.11}$$

By neglecting the terms that contain the matrix elements $M_{n,n'}$ and $N_{n,n'}$ in (23.1.9), we obtain a Schrödinger equation for the ions in which the effects of the electrons are taken into account by an effective potential $E_n^{\text{el}}(\{\boldsymbol{R}_l\})$ that depends on the ionic coordinates $\boldsymbol{R}_l$ – in the same way as in the analysis of lattice vibrations. Since the states of different labels n are not mixed, this approximation is obviously equivalent to replacing the full wavefunction (23.1.8) of electrons and ions by a single term,

$$\Psi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) = \Phi(\{\boldsymbol{R}_l\})\phi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\,. \tag{23.1.12}$$

Owing to the product form of the wavefunction, the two kinds of degrees of freedom can be separated to a certain degree. It is quite plausible to assume that the electrons are always in their ground state that corresponds to the instantaneous configuration of the slowly moving ions. Therefore, the electronic wavefunction $\phi(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})$ is the ground-state wavefunction in the particular ionic configuration, and $E^{\text{el}}(\{\boldsymbol{R}_l\})$ is the ground-state energy. This approximation is called the *adiabatic approximation.*

To justify this approximation, and to understand the role of the neglected terms, we shall first show that their contribution to the full energy of the system is indeed negligible. To this end, the diagonal matrix elements have to be considered. Since in the absence of a magnetic field the wavefunction can always be chosen real, the relationship

$$\begin{aligned}\int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\frac{\partial}{\partial \boldsymbol{R}_l}\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\prod_i \mathrm{d}\boldsymbol{r}_i \\ = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{R}_l}\int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\prod_i \mathrm{d}\boldsymbol{r}_i\end{aligned} \tag{23.1.13}$$

can be used for the integral in the diagonal matrix element $M_{n,n}$. The integral on the right-hand side is just the total number of electrons in the system. The derivative vanishes on account of the conservation of the particle number, that is, the diagonal elements of $M_{n,n'}$ are zero.

The diagonal elements of $N_{n,n'}$ give a finite but small contribution to the energy. To evaluate it, we shall assume that the wavefunction $\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})$ depends only on the relative position vectors, that is, the variables $\boldsymbol{r}_i - \boldsymbol{R}_l$. In terms of them, the formula for $N_{n,n}$ reads

$$\begin{aligned}&\int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\frac{\hbar^2}{2M_l}\frac{\partial^2\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{R}_l^2}\prod_i \mathrm{d}\boldsymbol{r}_i \\ &= \int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\frac{\hbar^2}{2M_l}\frac{\partial^2\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{r}_i^2}\prod_i \mathrm{d}\boldsymbol{r}_i \\ &= \frac{m_{\text{e}}}{M_l}\int \phi_n^*(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\frac{\hbar^2}{2m_{\text{e}}}\frac{\partial^2}{\partial \boldsymbol{r}_i^2}\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})\prod_i \mathrm{d}\boldsymbol{r}_i\,.\end{aligned} \tag{23.1.14}$$

The reader may recognize the kinetic energy of the electrons, multiplied by a factor of m_{e}/M_l. Because of the great disparity between the mass of electrons and ions, the contribution of this term is 10^{-3} to 10^{-4} times smaller than the usual electron energies. By neglecting this tiny correction, we may say that the electronic contribution to the lattice energy is $E^{\text{el}}(\{\boldsymbol{R}_l\})$ in the adiabatic approximation.

Even if the diagonal elements can be neglected, the nonvanishing off-diagonal elements indicate the possibility of transitions between different electron states due to the motion of the ions. Their accurate evaluation would require knowing the wavefunctions $\phi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\})$ precisely. Because of the presence of the electron–electron interaction term in $\mathcal{H}_{\text{el}}$, the solution of this problem would require the full apparatus of the many-body problem. Even when electron–electron interactions are treated in the simplest approximation, the Hartree–Fock approximation (to be discussed in Chapter 28), and are absorbed in an effective one-particle potential, we still have to overcome another difficulty: the states of electrons need to be calculated in the field of displaced electrons, that is, in a nonperiodic potential. By taking the displacement of ions from their equilibrium positions to be small, and expanding the electronic wavefunction about the equilibrium position of the ions, the leading term, which is proportional to the displacement of ions, comes from the matrix element $M_{n,n'}$, as the terms that are linear in the displacement vanish in $N_{n,n'}$. The off-diagonal matrix elements $M_{n,n'}$ correspond to the interaction of the electrons and the vibrating lattice. When the lattice vibrations are described in terms of phonons, this interaction can be viewed as an interaction between electrons and phonons. It can be shown that it is weak, and the smallness of the coupling constant can be traced back to the smallness of the electron-to-ion mass ratio, thus the electron–phonon interaction can usually be considered as a perturbation, and can be treated in the framework of perturbation theory.

23.2 Hamiltonian of the Electron–Phonon Interaction

The leading-order contribution to the interaction between the system of electrons and phonons thus appears in the off-diagonal elements of the matrix $M_{n,n'}$. However, it is rather difficult to study them in the present form. The Bloch form[2] of the interaction turns out to be much more practical as well as intuitive. Suppose that the potential $U_{\text{el–ion}}$ of the interaction between electrons and ions can be written as a sum of one-particle potentials for the individual electrons:

$$U_{\text{el–ion}}(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_l\}) = \sum_i U_{\text{el–ion}}(\boldsymbol{r}_i;\{\boldsymbol{R}_l\})\,. \tag{23.2.1}$$

Because of the displacement of the ions, the potential felt by the ith electron is modified. For small displacements this modification can be taken to be proportional to the displacement. Denoting the equilibrium position of the ions by $\boldsymbol{R}_l^0$ and the corresponding displacement by $\boldsymbol{u}(\boldsymbol{R}_l^0)$,

$$\begin{aligned} U_{\text{el–ion}}(\boldsymbol{r}_i;\{\boldsymbol{R}_l\}) &= U_{\text{el–ion}}(\boldsymbol{r}_i;\{\boldsymbol{R}_l^0+\boldsymbol{u}(\boldsymbol{R}_l^0)\}) \\ &= U_{\text{el–ion}}(\boldsymbol{r}_i;\{\boldsymbol{R}_l^0\}) + \delta U_{\text{el–ion}}(\boldsymbol{r}_i) \end{aligned} \tag{23.2.2}$$

in leading order, where

[2] F. Bloch, 1928.

$$\delta U_{\text{el–ion}} = \sum_l \boldsymbol{u}(\boldsymbol{R}_l^0) \frac{\partial U_{\text{el–ion}}(\boldsymbol{r}_i; \{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)} \bigg|_{\boldsymbol{u}(\boldsymbol{R}_l^0)=0} . \tag{23.2.3}$$

Below we shall not indicate that the derivative has to be taken at the equilibrium position. The first term gives the equilibrium value of the electron–ion interaction. This lattice-periodic potential has to be included in (17.1.1) to calculate the energy spectrum of the electrons in the static lattice. The second term contains the interaction of the electrons with lattice vibrations. Since the latter is described by phonons in the quantum picture, this term represents the *electron–phonon interaction.* When the contribution of all electrons is taken into account:

$$\mathcal{H}_{\text{el–ph}} = \sum_i \sum_l \boldsymbol{u}(\boldsymbol{R}_l^0) \frac{\partial U_{\text{el–ion}}(\boldsymbol{r}_i; \{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)} . \tag{23.2.4}$$

In terms of the electron density

$$n(\boldsymbol{r}) = \sum_i \delta(\boldsymbol{r} - \boldsymbol{r}_i) , \tag{23.2.5}$$

it can also be written as

$$\mathcal{H}_{\text{el–ph}} = \sum_l \boldsymbol{u}(\boldsymbol{R}_l^0) \int n(\boldsymbol{r}) \frac{\partial U_{\text{el–ion}}(\boldsymbol{r}; \{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)} \, \mathrm{d}\boldsymbol{r} . \tag{23.2.6}$$

We now have to cast it in another, simpler form, and then examine the consequences of the interaction.

This simpler form is based on the second-quantized representation of wavefunctions and operators. Therefore we shall repeatedly refer to the formulas of Appendix H, which presents second quantization. In this formulation the interaction between electrons and the vibrating lattice appears as scattering of particle-like elementary excitations – that is, scattering of electrons by phonons. However, unlike electrons, for which charge conservation implies the conservation of the electron number, the number of phonons may change: they can be created and annihilated in scattering processes.

23.2.1 Second-Quantized Form of the Hamiltonian

For simplicity, we shall consider a crystal with a monatomic basis below, keeping in mind that it is straightforward to generalize the expressions to crystals with a polyatomic basis. In the latter case both m and μ – which label the primitive cells and the atoms inside a cell – are included in the index l, however, the sum is taken only over the vectors of the primitive cells in the Fourier transform.

Expressing the operators of the displacement of ions in terms of the phonon creation and annihilation operators, (12.1.39) is simplified to

$$\boldsymbol{u}(\boldsymbol{R}_l^0) = \sum_{\boldsymbol{q},\lambda} \sqrt{\frac{\hbar}{2MN\omega_\lambda(\boldsymbol{q})}} \boldsymbol{e}^{(\lambda)}(\boldsymbol{q}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{R}_l^0} \left(a_\lambda(\boldsymbol{q}) + a_\lambda^\dagger(-\boldsymbol{q}) \right) \tag{23.2.7}$$

for crystals with a monatomic basis. Substituting this into (23.2.4), the Hamiltonian is found to be linear in the $a_\lambda^\dagger$ and a_λ.

In conjunction with the phonon creation and annihilation operators, it is practical to write the electron states in second-quantized form, too. Therefore we express the wavefunction of the many-electron system in terms of the creation operators $c_{n\boldsymbol{k}\sigma}^\dagger$ of the one-particle states characterized by the Bloch functions $\psi_{n\boldsymbol{k}\sigma}(\boldsymbol{r})$. For the sake of simplicity, we shall consider electron states within a single band, and suppress the band index.[3] According to (H.2.9) and (H.2.10), a one-particle operator $\sum_i f(\boldsymbol{r}_i)$ that acts on the electron states can be written as

$$\sum_i f(\boldsymbol{r}_i) = \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma} \left\{ \int \psi_{\boldsymbol{k}'\sigma}^*(\boldsymbol{r}) f(\boldsymbol{r}) \psi_{\boldsymbol{k}\sigma}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} \right\} c_{\boldsymbol{k}'\sigma}^\dagger c_{\boldsymbol{k}\sigma} \tag{23.2.8}$$

in terms of the electron creation and annihilation operators. Applying this to the derivative term in the electron–phonon interaction formula (23.2.4), its second-quantized form reads

$$\begin{aligned} &\sum_i \frac{\partial U_{\text{el–ion}}(\boldsymbol{r}_i; \{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)} \\ &\quad = \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma} \left\{ \int \psi_{\boldsymbol{k}'\sigma}^*(\boldsymbol{r}) \frac{\partial U_{\text{el–ion}}(\boldsymbol{r}; \{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)} \psi_{\boldsymbol{k}\sigma}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} \right\} c_{\boldsymbol{k}'\sigma}^\dagger c_{\boldsymbol{k}\sigma} \,. \end{aligned} \tag{23.2.9}$$

Inserting this formula and (23.2.7) into (23.2.4), the following general form is obtained for the Hamiltonian of the electron–phonon interaction:

$$\mathcal{H}_{\text{el–ph}} = \sum_{\boldsymbol{q}\lambda} \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma} D_{\lambda\sigma}(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{q}) c_{\boldsymbol{k}'\sigma}^\dagger c_{\boldsymbol{k}\sigma} \left(a_\lambda(\boldsymbol{q}) + a_\lambda^\dagger(-\boldsymbol{q}) \right), \tag{23.2.10}$$

where

$$\begin{aligned} D_{\lambda\sigma}(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{q}) = &\sum_l \sqrt{\frac{\hbar}{2MN\omega_\lambda(\boldsymbol{q})}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{R}_l^0} \\ &\times \int \psi_{\boldsymbol{k}'\sigma}^*(\boldsymbol{r}) \boldsymbol{e}^{(\lambda)}(\boldsymbol{q}) \cdot \frac{\partial U_{\text{el–ion}}(\boldsymbol{r}; \{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)} \psi_{\boldsymbol{k}\sigma}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} \,. \end{aligned} \tag{23.2.11}$$

The electron–ion potential $U_{\text{el–ion}}(\boldsymbol{r}; \{\boldsymbol{R}_l^0\})$ is obviously lattice periodic in $\boldsymbol{r}$ as it contains the equilibrium positions of the ions. On the other hand, the derivative with respect to the position vector of the lth ion is a function of

[3] Physically speaking, this means that processes in which an electron is scattered into another band upon the absorption or emission of a photon are neglected.

$\boldsymbol{r}-\boldsymbol{R}_l^0$. By exploiting the simple translational properties of the Bloch functions in the integral of the matrix element, we have

$$\begin{aligned}\int \psi^*_{\boldsymbol{k}'\sigma}(\boldsymbol{r})\boldsymbol{e}^{(\lambda)}(\boldsymbol{q})\cdot\frac{\partial U_{\text{el-ion}}(\boldsymbol{r};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)}\psi_{\boldsymbol{k}\sigma}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} &= \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{R}_l^0} \\ \times \int \psi^*_{\boldsymbol{k}'\sigma}(\boldsymbol{r}-\boldsymbol{R}_l^0)\boldsymbol{e}^{(\lambda)}(\boldsymbol{q})\cdot\frac{\partial U_{\text{el-ion}}(\boldsymbol{r};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)}&\psi_{\boldsymbol{k}\sigma}(\boldsymbol{r}-\boldsymbol{R}_l^0)\,\mathrm{d}\boldsymbol{r}\,.\end{aligned} \tag{23.2.12}$$

After the separation of the phase factors, the remaining integral is independent of the position $\boldsymbol{R}_l^0$ of the lattice point. Therefore when the sum over $\boldsymbol{R}_l^0$ is calculated in (23.2.11), the exponential factors cancel the contributions unless

$$\boldsymbol{q}+\boldsymbol{k}-\boldsymbol{k}'=\boldsymbol{G}\,. \tag{23.2.13}$$

The Hamiltonian of the electron–phonon interaction can then be written as

$$\mathcal{H}_{\text{el-ph}}=\sum_{\boldsymbol{q}\lambda}\sum_{\boldsymbol{k}\boldsymbol{G}\sigma} D_{\lambda\sigma}(\boldsymbol{k},\boldsymbol{q},\boldsymbol{G})c^\dagger_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}\sigma}c_{\boldsymbol{k}\sigma}\left(a_\lambda(\boldsymbol{q})+a^\dagger_\lambda(-\boldsymbol{q})\right) \tag{23.2.14}$$

where

$$\begin{aligned}D_{\lambda\sigma}(\boldsymbol{k},\boldsymbol{q},\boldsymbol{G})=\sqrt{\frac{\hbar N}{2M\omega_\lambda(\boldsymbol{q})}}\int &\psi^*_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}\sigma}(\boldsymbol{r}-\boldsymbol{R}_l^0) \\ &\times \boldsymbol{e}^{(\lambda)}(\boldsymbol{q})\cdot\frac{\partial U_{\text{el-ion}}(\boldsymbol{r};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)}\psi_{\boldsymbol{k}\sigma}(\boldsymbol{r}-\boldsymbol{R}_l^0)\,\mathrm{d}\boldsymbol{r}\,.\end{aligned} \tag{23.2.15}$$

The interaction can be described by words like this: Because of the vibrational motion of the atoms, regular periodicity is broken in the crystal, so the Bloch states are no longer eigenstates. Since charge conservation implies that the number of electrons is conserved, whereas the number of phonons may change, the interaction with lattice vibrations may scatter electrons from their initial state to some other state, and additional phonons may appear in the system or existing ones may disappear from it. As the discrete translational symmetry is broken, the crystal momentum of electrons is not conserved any more. When, however, vibrations are specified in terms of phonons, and the Bloch electrons as well as phonons are characterized by the wave vectors as quantum numbers defined in the regular lattice, then the conservation of quasimomentum to within an additive reciprocal-lattice vector is restored when the combined system of electrons and phonons is considered. The general conservation laws presented in Chapter 6 as the consequences of translational symmetry can be applied to the interacting system of electrons and phonons by stating that an electron of wave number $\boldsymbol{k}$ can be scattered to the state of wave number $\boldsymbol{k}'=\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}$ upon the absorption of a phonon of wave number $\boldsymbol{q}$ or the emission of a phonon of wave number $-\boldsymbol{q}$. These processes are shown in Fig. 23.1.

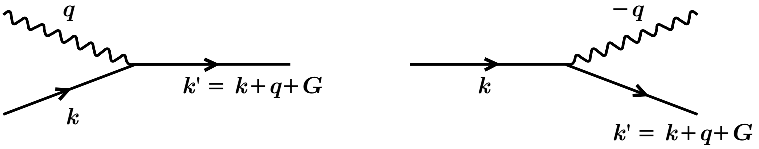


Fig. 23.1. Electron–phonon interactions, in which a phonon is emitted and absorbed. Electrons are represented by straight lines, phonons by wavy ones

Wave vectors are usually reduced to the first Brillouin zone, that is why the reciprocal-lattice vector $\boldsymbol{G}$ appears in (23.2.13) and the Hamiltonian (23.2.14). As mentioned in Chapter 6, those processes that do not require such a reduction – because $\boldsymbol{k}+\boldsymbol{q}$ is already in the first Brillouin zone, so $\boldsymbol{G}=0$ – are called normal processes. When $\boldsymbol{k}+\boldsymbol{q}$ is outside the first Brillouin zone, and so a reciprocal-lattice vector $\boldsymbol{G}\neq 0$ is needed to reduce it to the first Brillouin zone, the process is called *umklapp*. Both types are shown in Fig. 23.2.

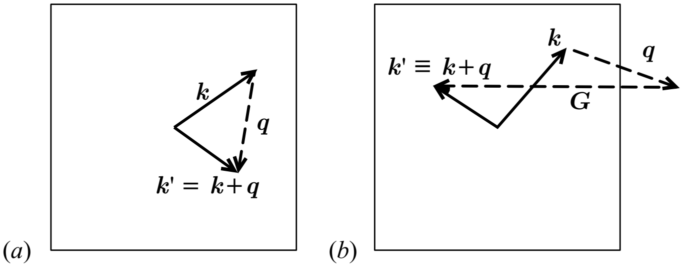


Fig. 23.2. Electron–photon interaction through (a) a normal and (b) an umklapp process

23.2.2 Electron–Phonon Matrix Element

To determine the strength of the electron–phonon interaction, further assumptions have to be made about the potential. In the simplest approximation, proposed by L. Nordheim in 1931, the full potential is the sum of individual atomic potentials U_a that ions carry with themselves rigidly as they move around:

$$U_{\text{el–ion}}(\boldsymbol{r};\{\boldsymbol{R}_l\}) = \sum_l U_\mathrm{a}(\boldsymbol{r}-\boldsymbol{R}_l)\,. \tag{23.2.16}$$

That is why this approximation is called the *rigid-ion approximation.* Figure 23.3 shows the rigid atomic potentials displaced from the equilibrium positions.

Even when the atomic potential is assumed to be nonvanishing only within the Wigner–Seitz cell, once the atoms are displaced, the rigidly comoving

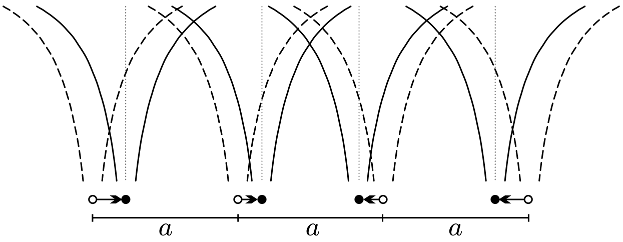


Fig. 23.3. The potential felt by the electrons in the rigid-ion approximation. Dashed lines show the potential around the equilibrium position of the ions, and solid lines around the displaced ions

potentials do not match properly at the cell boundaries. The rigid-ion approximation must certainly be improved in this respect. However, as we are interested only in the qualitative description of the electron–phonon interaction here, we shall not be concerned with this problem.

Using the relation

$$\left.\frac{\partial U_{\text{el–ion}}(\boldsymbol{r};\{\boldsymbol{R}_l\})}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)}\right|_{\boldsymbol{u}(\boldsymbol{R}_l^0)=0} = \left.\frac{\partial U_{\text{a}}(\boldsymbol{r}-\boldsymbol{R}_l)}{\partial \boldsymbol{u}(\boldsymbol{R}_l^0)}\right|_{\boldsymbol{u}(\boldsymbol{R}_l^0)=0} = -\frac{\partial U_{\text{a}}(\boldsymbol{r}-\boldsymbol{R}_l^0)}{\partial \boldsymbol{r}}\,, \tag{23.2.17}$$

the Fourier representation

$$U_{\text{a}}(\boldsymbol{r}-\boldsymbol{R}_l^0) = \sum_{\boldsymbol{q}'} U_{\text{a}}(\boldsymbol{q}')\text{e}^{\text{i}\boldsymbol{q}'\cdot(\boldsymbol{r}-\boldsymbol{R}_l^0)} \tag{23.2.18}$$

of the one-particle atomic potential, and the customary form of the Bloch functions, the coupling constant that determines the strength of the electron–phonon interaction is

$$\begin{aligned} D_{\lambda\sigma}(\boldsymbol{k},\boldsymbol{q},\boldsymbol{G}) = {}& -\text{i}\sqrt{\frac{\hbar N}{2M\omega_\lambda(\boldsymbol{q})}} \sum_{\boldsymbol{q}'} \left(\boldsymbol{e}^{(\lambda)}(\boldsymbol{q})\cdot\boldsymbol{q}'\right) U_{\text{a}}(\boldsymbol{q}') \\ & \times \frac{1}{V}\int \text{e}^{\text{i}\boldsymbol{q}'\cdot(\boldsymbol{r}-\boldsymbol{R}_l^0)}\text{e}^{-\text{i}(\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G})\cdot(\boldsymbol{r}-\boldsymbol{R}_l^0)} \\ & \times u^*_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}\sigma}(\boldsymbol{r}-\boldsymbol{R}_l^0)\text{e}^{\text{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{R}_l^0)}u_{\boldsymbol{k}\sigma}(\boldsymbol{r}-\boldsymbol{R}_l^0)\,\text{d}\boldsymbol{r}\,. \end{aligned} \tag{23.2.19}$$

Since the functions $u_{\boldsymbol{k}}(\boldsymbol{r})$ are lattice periodic, the previous integral vanishes unless $\boldsymbol{q}' = \boldsymbol{q}+\boldsymbol{G}$. Separating the volume integral into an integral over a primitive cell and a sum over the cells, we have

$$\begin{aligned} D_{\lambda\sigma}(\boldsymbol{k},\boldsymbol{q},\boldsymbol{G}) = {}& -\text{i}\frac{N}{V}\sqrt{\frac{\hbar N}{2M\omega_\lambda(\boldsymbol{q})}}\boldsymbol{e}^{(\lambda)}(\boldsymbol{q})\cdot(\boldsymbol{q}+\boldsymbol{G})U_{\text{a}}(\boldsymbol{q}+\boldsymbol{G}) \\ & \int_v u^*_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}\sigma}(\boldsymbol{r})u_{\boldsymbol{k}\sigma}(\boldsymbol{r})\,\text{d}\boldsymbol{r}\,. \end{aligned} \tag{23.2.20}$$

The factor $\boldsymbol{e}^{(\lambda)}(\boldsymbol{q})\cdot(\boldsymbol{q}+\boldsymbol{G})$ in the interaction strength is called the *polarization factor*. Because of its presence, only longitudinal phonons may participate in normal processes. While this is not true for umklapp processes, it is often justified to assume that scattering by longitudinal phonons gives the most important contribution to the electron–phonon interaction.

Starting with the Schrödinger equation of the electronic wavefunction, the quantities in the interaction matrix element can be estimated after some tedious algebra. The result is

$$U(\boldsymbol{q}) \int\limits_{v} u^{*}_{\boldsymbol{k}+\boldsymbol{q}\sigma}(\boldsymbol{r}) u_{\boldsymbol{k}\sigma}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} = \frac{2}{5}\varepsilon_{\mathrm{F}} G(|\boldsymbol{q}|r_{\mathrm{WS}})\,, \tag{23.2.21}$$

where r_{WS} is the Wigner–Seitz radius and

$$G(x) = 3\left(\frac{x\cos x - \sin x}{x^3}\right). \tag{23.2.22}$$

$2\varepsilon_{\mathrm{F}}/5$ is the *energy factor*, while $G(|\boldsymbol{q}|r_{\mathrm{WS}})$, which depends on the momentum transfer in the process, is the *interference factor*. For increasing $|\boldsymbol{q}|$ the interaction gets weaker. In the rigid-ion approximation forward scattering processes are much more probable than backscattering processes.

23.2.3 Deformation Potential

The possibility that the displacement of ions can modify the electronic charge distribution, leading to an additional potential felt by the electrons, was ignored in the rigid-ion approximation. This may be particularly important for long-wavelength longitudinal acoustic phonons, since the propagation of such vibrations deforms the primitive cells, and the lattice constant and the ion density are locally modified. The breaking of the local charge neutrality leads to local variations in the electron density, and this results in a modification of the potential. Below we shall give a simple estimate for this additional potential due to the deformation of the lattice, called the *deformation potential*, and also determine the strength of the electron–phonon interaction caused by the potential.

In the long-wavelength limit longitudinal acoustic vibrations can be considered as compressional waves propagating in an elastic continuum. The displacements $\boldsymbol{u}(\boldsymbol{R}_l^0)$ defined only in discrete lattice points can then be replaced by a continuous displacement field $\boldsymbol{u}(\boldsymbol{r})$ via the generalization of (23.2.7):

$$\boldsymbol{u}(\boldsymbol{r}) = \sum_{\boldsymbol{q}\lambda} \sqrt{\frac{\hbar}{2MN\omega_\lambda(\boldsymbol{q})}}\, \boldsymbol{e}^{(\lambda)}(\boldsymbol{q}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \left(a_\lambda(\boldsymbol{q}) + a^{\dagger}_{\lambda}(-\boldsymbol{q})\right). \tag{23.2.23}$$

The relative change of the lattice constant in the propagation direction is proportional to the gradient of the ionic displacement. The local relative change in the volume is then

$$\varDelta(\boldsymbol{r}) \equiv \frac{\delta V}{V} = \frac{\partial \boldsymbol{u}(\boldsymbol{r})}{\partial \boldsymbol{r}} . \tag{23.2.24}$$

The deformation potential arising from the breaking of local charge neutrality because of the long-wavelength variations of the lattice constant is proportional to $\varDelta(\boldsymbol{r})$:

$$U_{\text{def}}(\boldsymbol{r}) = C\varDelta(\boldsymbol{r}) . \tag{23.2.25}$$

The Hamiltonian contains the product of the potential and the local electron density:

$$\mathcal{H}_{\text{def}} = C \int n_{\text{e}}(\boldsymbol{r})\varDelta(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} . \tag{23.2.26}$$

The coefficient C can be estimated in the free-electron model. The deformation $\varDelta(\boldsymbol{r})$ modifies the local electron density, and therefore the local Fermi energy as well. Making use of the relation

$$\varepsilon_{\text{F}} = \frac{\hbar^2}{2m_{\text{e}}} \left(\frac{3\pi^2 N_{\text{e}}}{V} \right)^{2/3} \tag{23.2.27}$$

for free electrons, where k_{F} can be expressed from (16.2.24), the change δV in the volume modifies the Fermi energy locally by

$$\delta\varepsilon_{\text{F}}(\boldsymbol{r}) = -\tfrac{2}{3}\varepsilon_{\text{F}} \frac{\delta V}{V} = -\tfrac{2}{3}\varepsilon_{\text{F}}\varDelta(\boldsymbol{r}) . \tag{23.2.28}$$

In the presence of the deformation potential the electrons are rearranged in such a way that the chemical potential should be the same. Consequently,

$$U_{\text{def}}(\boldsymbol{r}) = \tfrac{2}{3}\varepsilon_{\text{F}}\varDelta(\boldsymbol{r}) , \tag{23.2.29}$$

that is, $C = \frac{2}{3}\varepsilon_{\text{F}}$ and

$$\mathcal{H}_{\text{def}} = \tfrac{2}{3}\varepsilon_{\text{F}} \int n_{\text{e}}(\boldsymbol{r})\varDelta(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} . \tag{23.2.30}$$

To find the second-quantized form of the Hamiltonian, we shall express the electron density and the deformation in terms of creation and annihilation operators. Writing the field operators in (H.2.46) in terms of the Bloch states,

$$n_{\text{e}}(\boldsymbol{r}) = \frac{1}{V} \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\boldsymbol{r}} u^*_{\boldsymbol{k}'}(\boldsymbol{r}) u_{\boldsymbol{k}}(\boldsymbol{r}) c^\dagger_{\boldsymbol{k}'\sigma} c_{\boldsymbol{k}\sigma} , \tag{23.2.31}$$

while for $\varDelta(\boldsymbol{r})$ we shall use

$$\varDelta(\boldsymbol{r}) = \mathrm{i} \sum_{\boldsymbol{q}\lambda} \sqrt{\frac{\hbar}{2MN\omega_\lambda(\boldsymbol{q})}} (\boldsymbol{e}^{(\lambda)}(\boldsymbol{q}) \cdot \boldsymbol{q}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \left(a_\lambda(\boldsymbol{q}) + a^\dagger_\lambda(-\boldsymbol{q}) \right) , \tag{23.2.32}$$

which can be obtained by differentiating (23.2.23) with respect to $\boldsymbol{q}$. Substituting these into (23.2.26), and separating the volume integral into an integral over a primitive cell and a sum over the cells, as before, the Hamiltonian is similar to the one derived in the rigid-ion approximation. Neglecting umklapp processes for simplicity,

$$\mathcal{H}_{\text{el–ph}} = \sum_{\boldsymbol{q}\lambda} \sum_{\boldsymbol{k}\sigma} D_\lambda(\boldsymbol{q}) c^\dagger_{\boldsymbol{k}+\boldsymbol{q}\sigma} c_{\boldsymbol{k}\sigma} \Big(a_\lambda(\boldsymbol{q}) + a^\dagger_\lambda(-\boldsymbol{q}) \Big) , \tag{23.2.33}$$

where

$$D_\lambda(\boldsymbol{q}) = \frac{\mathrm{i}C}{V} \sqrt{\frac{\hbar N}{2M\omega_\lambda(\boldsymbol{q})}} \Big(\boldsymbol{e}^{(\lambda)}(\boldsymbol{q}) \cdot \boldsymbol{q} \Big) \int\limits_v u^*_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}\sigma}(\boldsymbol{r}) u_{\boldsymbol{k}\sigma}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} . \tag{23.2.34}$$

The argument that led to this form of the electron–phonon interaction is justified only for the interaction with long-wavelength longitudinal acoustic phonons, where $\omega_\lambda(\boldsymbol{q})$ is proportional to q, and thus the interaction strength is proportional to $q^{1/2}$. The strength C of the deformation potential can be treated as a free parameter and fitted to measured data.

If more accurate calculations are required for the strength of the electron–phonon scattering, then the real potential cannot be approximated by the sum of some plausible atomic potentials. The ion cores may carry their potentials rigidly with themselves, so the rigid-ion approximation can be used as a starting point, however this potential is screened by the redistribution of the electronic charge. This leads to a weakening of the Coulomb potential of the ion core, which can be taken into account by the wavelength-dependent dielectric function. We shall discuss this screening mechanism in Chapter 29 on electron–electron interactions.

23.2.4 Interaction of Electrons with Optical Phonons

In the previous subsection we dealt with the modifications of the electron states due to the long-wavelength deformations of the crystal lattice, that is, the interaction of Bloch electrons with acoustic phonons. It is plausible to expect that in ionic crystals, where optical phonons correspond to the vibrations of oppositely charged ions (i.e., polarization waves), the scattering by optical phonons is even more important in the interaction between the Bloch electrons and the vibrating lattice. Since transverse optical phonons give rise to a much smaller polarization than longitudinal optical phonons, we shall focus on longitudinal optical (LO) phonons. We shall denote their creation and annihilation operators by $b^\dagger_{\boldsymbol{q}}$ and $b_{\boldsymbol{q}}$, suppressing the index λ of polarization.

To determine the Hamiltonian of the interaction with LO phonons, we shall start with the electric polarization vector $\boldsymbol{P}$, which can be considered proportional to the amplitude of optical phonons:

$$\boldsymbol{P} = F \sum_{\boldsymbol{q}} \boldsymbol{e}(\boldsymbol{q}) \left(b_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} + b^\dagger_{\boldsymbol{q}} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \right) , \tag{23.2.35}$$

where the polarization vector $\boldsymbol{e}(\boldsymbol{q})$ is the unit vector in the direction of $\boldsymbol{q}$, thus the polarization can be written as

$$\boldsymbol{P} = F \sum_{\boldsymbol{q}} \frac{\boldsymbol{q}}{q} \left(b_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} + b_{\boldsymbol{q}}^{\dagger} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \right) . \tag{23.2.36}$$

Since there are no free charges in the crystal, $\operatorname{div} \boldsymbol{D} = 0$. As its Fourier transform satisfies $\boldsymbol{q} \cdot \boldsymbol{D} = 0$, we have $\epsilon_0 \boldsymbol{E} + \boldsymbol{P} = 0$ in the longitudinal case, and so

$$\boldsymbol{E} = -\frac{F}{\epsilon_0} \sum_{\boldsymbol{q}} \frac{\boldsymbol{q}}{q} \left(b_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} + b_{\boldsymbol{q}}^{\dagger} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \right) . \tag{23.2.37}$$

When the electric field $\boldsymbol{E}$ is derived from a scalar potential $\varphi(\boldsymbol{r})$ through $\boldsymbol{E} = -\operatorname{grad} \varphi(\boldsymbol{r})$, the potential that leads to the field in the previous formula is

$$\varphi(\boldsymbol{r}) = -\mathrm{i} \frac{F}{\epsilon_0} \sum_{\boldsymbol{q}} \frac{1}{q} \left(b_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} - b_{\boldsymbol{q}}^{\dagger} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \right) . \tag{23.2.38}$$

Writing the Hamiltonian of the electron–phonon interaction as the energy of an electron system of density $n(\boldsymbol{r})$ moving in such a potential, we have

$$\mathcal{H}_{\text{el–ph}} = -e \int \varphi(\boldsymbol{r}) n(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} . \tag{23.2.39}$$

In terms of the electron field operators this can be rewritten as

$$\mathcal{H}_{\text{el–ph}} = -e \sum_{\sigma} \int \varphi(\boldsymbol{r}) \psi_{\sigma}^{\dagger}(\boldsymbol{r}) \psi_{\sigma}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} . \tag{23.2.40}$$

Approximating the Bloch states of electrons by plane waves, and using the corresponding creation and annihilation operators,

$$\mathcal{H}_{\text{el–ph}} = -\mathrm{i} \frac{Fe}{\epsilon_0} \sum_{\boldsymbol{q}\sigma} \frac{1}{q} \left(b_{\boldsymbol{q}} c_{\boldsymbol{k}+\boldsymbol{q},\sigma}^{\dagger} c_{\boldsymbol{k}\sigma} - b_{\boldsymbol{q}}^{\dagger} c_{\boldsymbol{k}-\boldsymbol{q},\sigma}^{\dagger} c_{\boldsymbol{k}\sigma} \right) . \tag{23.2.41}$$

The coefficient F can be expressed in terms of the frequency ω_{LO} of LO phonons and the low- and high-frequency values of the dielectric constant:

$$F^2 = \frac{\hbar \omega_{\mathrm{LO}} \epsilon_0}{2V} \left[\frac{1}{\epsilon_{\mathrm{r}}(\infty)} - \frac{1}{\epsilon_{\mathrm{r}}(0)} \right] . \tag{23.2.42}$$

This formula is very similar to its counterpart for acoustic phonons. This is because the structure of the interaction Hamiltonian is determined by the conservation laws for the number of electrons and the crystal momentum. The only important difference is that the interaction is singularly strong for long-wavelength LO phonons, whereas the corresponding matrix element is proportional to $\sqrt{q}$ for acoustic phonons.

23.3 Consequences of the Electron–Phonon Interaction

The Hamiltonian of the electron–phonon interaction thus describes two distinct processes: the creation and annihilation of a phonon, depicted in Fig. 23.1. An electron cannot remain in a Bloch state of wave vector $\boldsymbol{k}$ indefinitely in the vibrating lattice: it is scattered into another state of different wave vector by the emission or absorption of a phonon. Because of this finite lifetime, the current carried by the electrons would dissipate in a finite amount of time in the absence of an electric field that accelerates the electrons. To maintain a constant current, a constant electric field has to be applied. As we shall see it in detail in the next chapter on transport phenomena, electron–phonon scattering gives one of the most important contributions to the resistivity of metals.

The electron–phonon interaction processes shown in Fig. 23.1 are redrawn differently in Fig. 23.4. The left-hand process can be interpreted as the annihilation of a phonon into an electron–hole pair, with the energy and crystal momentum conserved. This process contributes to the attenuation of sound waves in metals. As is well known, phonons can also decay because of the slight anharmonicity of the potential felt by the atoms. The possibility that the number of phonons may also be reduced by the electron–phonon interaction gives a further contribution to the decay rate (inverse lifetime) of phonons. Of course, the inverse process, the annihilation of an electron–hole pair into a phonon, is also possible.

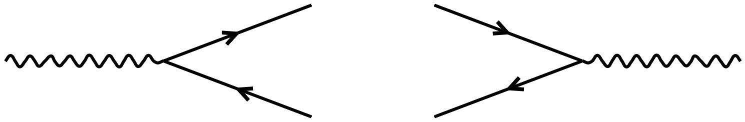


Fig. 23.4. First-order processes of the electron–phonon interaction

Figure 23.5 shows three second-order processes of the electron–phonon interaction.

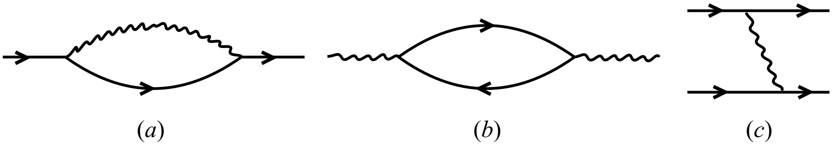


Fig. 23.5. Second-order processes of the electron–phonon interaction

The first process corresponds to the emission and subsequent absorption of the same phonon by an electron.[4] If this process is repeated several times, it is as if the electron were surrounded by a comoving phonon cloud, and so the measure of its inertia, its effective mass, is also changed. In an analogous, two-step process showed in Fig. 23.5(*b*), the phonon is transformed into an electron–hole pair for a short time, and when the pair recombines, the phonon emerges again. Such processes modify the energy of the phonon. However, the correction of these second-order processes to the particle energy contains an imaginary part as well. While the real part gives the energy shift of the particles (electrons and phonons) due to the interaction, the imaginary part can be interpreted as their decay rate. Thus, the electron–phonon interaction is seen to render the lifetime of electrons and phonons finite.

The third possibility is shown in Fig. 23.5(*c*): an electron propagating in the crystal emits a phonon, and thus modifies the vibrational state of the crystal; then the phonon is absorbed by another electron, and the initial vibrational state is restored. In this process the state of two electrons is changed, so the exchange of the phonon gives rise to an effective interaction between the two electrons. As will be discussed in detail in Chapter 34 on the microscopic theory of superconductors, the appearance of Cooper pairs, which are responsible for superconductivity, is due precisely to such processes in the majority of superconductors.

23.3.1 Finite Lifetime of Electron States

As mentioned in the previous section, owing to the electron–phonon interaction, the Bloch states determined in an ideal crystal are not eigenstates of the full Hamiltonian, so electrons are scattered from the Bloch state of wave vector $\boldsymbol{k}$ in a finite amount of time, and thus their lifetime and mean free path becomes finite. They can be estimated by analyzing the scattering by acoustic phonons.

In terms of the creation operator the state containing a Bloch electron of wave vector $\boldsymbol{k}$ can be written as

$$|\boldsymbol{k}\rangle = c^{\dagger}_{\boldsymbol{k}\sigma}|0\rangle\,. \tag{23.3.1}$$

Because of the electron–phonon interaction, this can be scattered into a state of wave vector $\boldsymbol{k}'$, while a phonon of wave vector $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}'$ is simultaneously created. For simplicity, we shall neglect umklapp processes. In second quantization this state is given by

$$|\boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}\rangle = c^{\dagger}_{\boldsymbol{k}-\boldsymbol{q},\sigma} a^{\dagger}_{\lambda}(\boldsymbol{q})|0\rangle\,. \tag{23.3.2}$$

The transition probability for this quantum mechanical process is

[4] While momentum conservation must apply to each elementary process, energy conservation applies only to real processes, between an initial and a finial state, but not to the intermediate virtual states.

$$W_{\boldsymbol{k}\boldsymbol{k}'} = \frac{2\pi}{\hbar} \left|\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}\rangle\right|^2 \delta(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_\lambda(\boldsymbol{q}))\delta_{\boldsymbol{k}',\boldsymbol{k}-\boldsymbol{q}} \,. \tag{23.3.3}$$

The probability of the very similar process in which the electron of wave vector $\boldsymbol{k}$ is scattered into another state by the absorption of a phonon can be written analogously.

The inverse lifetime is obtained by summing over all possible scattering processes:

$$\begin{aligned} \frac{1}{\tau} = \frac{2\pi}{\hbar} \sum_{\boldsymbol{q}\lambda} &\left|\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}\rangle\right|^2 \delta(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_\lambda(\boldsymbol{q})) \\ + \frac{2\pi}{\hbar} \sum_{\boldsymbol{q}\lambda} &\left|\langle \boldsymbol{k}+\boldsymbol{q}|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}, 1_{\boldsymbol{q}\lambda}\rangle\right|^2 \delta(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}+\boldsymbol{q}} + \hbar\omega_\lambda(\boldsymbol{q})) \,. \end{aligned} \tag{23.3.4}$$

It should be noted that the same result can be simply obtained from the energy correction

$$\tilde{\varepsilon}_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}} + \sum_{\boldsymbol{q}\lambda} \left[\frac{|\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}, 0\rangle|^2}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_\lambda(\boldsymbol{q})} + \frac{|\langle \boldsymbol{k}+\boldsymbol{q}|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}, 1_{\boldsymbol{q}\lambda}\rangle|^2}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}+\boldsymbol{q}} + \hbar\omega_\lambda(\boldsymbol{q})} \right] \tag{23.3.5}$$

in the second order of perturbation theory.[5] By adding an infinitesimal imaginary part iα to the energy denominator, the lifetime of the particle is determined by the imaginary part of the particle energy:

$$-\operatorname{Im}\tilde{\varepsilon}_{\boldsymbol{k}} = \varGamma_{\boldsymbol{k}} = \frac{\hbar}{2\tau} \,. \tag{23.3.6}$$

When the dispersion relation of the electrons is characterized by an effective mass m^* and that of the acoustic phonons by the sound velocity c_{s}, energy and momentum conservation jointly lead to the requirement

$$\frac{\hbar^2(\boldsymbol{k} \pm \boldsymbol{q})^2}{2m^*} = \frac{\hbar^2\boldsymbol{k}^2}{2m^*} \pm \hbar c_{\text{s}} q \tag{23.3.7}$$

for normal processes involving phonon absorption and emission, respectively. Rewritten in terms of the angle θ between $\boldsymbol{k}$ and $\boldsymbol{q}$,

$$q = \mp 2k\cos\theta \pm \frac{2m^* c_{\text{s}}}{\hbar} \,. \tag{23.3.8}$$

In metals, where the sound velocity is several orders of magnitude smaller than the Fermi velocity, the second term on the right-hand side is much smaller than k_{F}, thus

[5] The energy of the quasiparticle can be determined from the location of the pole of the Green function $G(\boldsymbol{k},\omega)$ in the complex ω plane. For the retarded Green function the pole is in the lower half-plane, at $\tilde{\varepsilon}_{\boldsymbol{k}} - \mathrm{i}\varGamma_{\boldsymbol{k}}$, at a finite distance from the real axis. When the Green function is transformed back to real time, the probability of finding the particle of energy $\tilde{\varepsilon}_{\boldsymbol{k}}$ decays exponentially with the time constant $\hbar/(2\varGamma_{\boldsymbol{k}})$ on account of the imaginary part of $\varGamma_{\boldsymbol{k}}$.

$$q = \mp 2k\cos\theta \tag{23.3.9}$$

to a good approximation. In spite of the possibly large momentum transfer, the scattering can be considered practically elastic, and only electrons close to the Fermi energy participate in the processes. The lifetime due to these processes is found to be

$$\tau \propto T^{-3} \tag{23.3.10}$$

at very low temperatures (which will be demonstrated more accurately in the next chapter), while at higher temperatures

$$\tau \propto T^{-1}\,. \tag{23.3.11}$$

This estimate cannot be applied to semiconductors, where the characteristic value of k is related to the thermal de Broglie wavelength. Using the formula $\hbar k = \sqrt{m^* k_{\mathrm{B}} T}$, the change in the energy of electrons is found to be negligible for temperatures above 1 K, and the scattering is practically elastic again, however the lifetime depends strongly on the energy of the electrons:

$$\tau \propto T^{-1}\varepsilon^{-1/2}\,. \tag{23.3.12}$$

23.3.2 Polarons

As has been mentioned, electrons propagating in a crystal emit or absorb phonons when scattered by the vibrating lattice. We may say that the electrons are surrounded by a cloud of phonons – in other words the "bare" Bloch electrons of effective mass m^* in the rigid lattice become "dressed" in the presence of phonons, and so their effective mass is modified.

To study this mass enhancement, we shall consider a Bloch state of the electron system, and assume that there are no phonons in the unperturbed ground state. The interaction admixes to the state $|\boldsymbol{k}\rangle$ other states in which a phonon of wave vector $\boldsymbol{q}$ and polarization λ is present, and the wave vector of the electron is $\boldsymbol{k}-\boldsymbol{q}$. According to the formulas of perturbation theory, in the first order of the electron–phonon interaction the electron wavefunction is

$$|\boldsymbol{k}\rangle^{(1)} = |\boldsymbol{k}\rangle + \sum_{\boldsymbol{q}\lambda} |\boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}\rangle \frac{\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}|\mathcal{H}_{\text{el–ph}}|\boldsymbol{k}\rangle}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_\lambda(\boldsymbol{q})}\,. \tag{23.3.13}$$

The electron propagating in the system is accompanied by phonons in this perturbed state, therefore the state can be characterized by the mean number of phonons it contains. This is the expectation value of the phonon-number operator in the perturbed state:

$$\begin{aligned}\langle n_{\text{ph}}\rangle &= {}^{(1)}\langle \boldsymbol{k}| \sum_{\boldsymbol{q}\lambda} a^\dagger_\lambda(\boldsymbol{q}) a_\lambda(\boldsymbol{q}) |\boldsymbol{k}\rangle^{(1)} \\ &= \sum_{\boldsymbol{q}\lambda} \frac{|\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}\lambda}|\mathcal{H}_{\text{el–ph}}|\boldsymbol{k}\rangle|^2}{\big(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_\lambda(\boldsymbol{q})\big)^2}\,.\end{aligned} \tag{23.3.14}$$

Assuming that only the interaction with long-wavelength acoustic phonons gives an important contribution, we shall drop the polarization index λ, and write

$$\mathcal{H}_{\text{el–ph}} = \mathrm{i}C \sum_{\boldsymbol{k},\boldsymbol{q}\sigma} \sqrt{\frac{\hbar N}{2M\omega_{\boldsymbol{q}}}} q c^{\dagger}_{\boldsymbol{k}+\boldsymbol{q}\sigma} c_{\boldsymbol{k}\sigma} (a_{\boldsymbol{q}} + a^{\dagger}_{-\boldsymbol{q}}) , \tag{23.3.15}$$

where only the $\boldsymbol{q}$-dependent factors are written out explicitly, and the square of the matrix element is thus

$$\left| \langle \boldsymbol{k} - \boldsymbol{q}, 1_{\boldsymbol{q}} | \mathcal{H}_{\text{el–ph}} | \boldsymbol{k}, 0 \rangle \right|^2 = \frac{C^2 \hbar N |q|}{2Mc_{\text{s}}} . \tag{23.3.16}$$

Expressing the electron spectrum in terms of an effective mass m^*, and assuming a linear dispersion relation for the phonons, the energy denominator is

$$\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_{\boldsymbol{q}} = \frac{\hbar^2}{2m^*}(2\boldsymbol{k} \cdot \boldsymbol{q} - q^2) - \hbar c_{\text{s}} q . \tag{23.3.17}$$

Let us first examine what happens to the electrons at the bottom of the conduction band in semiconductors. For slow electrons $\boldsymbol{k} \cdot \boldsymbol{q}$ can be neglected compared to q^2. Taking the wave numbers of phonons, according to the Debye approximation, inside a sphere of radius q_{D}, and making use of the integral

$$\int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3} \frac{q}{\left(q^2 + \dfrac{2m^* c_{\text{s}}}{\hbar} q \right)^2} = \frac{1}{2\pi^2} \int\limits_0^{q_{\text{D}}} \frac{q}{(q + q_{\text{c}})^2} \, \mathrm{d}q , \tag{23.3.18}$$

where $q_{\text{c}} = 2m^* c_{\text{s}}/\hbar$,

$$\langle n_{\text{ph}} \rangle = \frac{1}{\pi^2} \frac{{m^*}^2 N C^2}{\hbar^3 c_{\text{s}} M} \ln \left(\frac{q_{\text{D}}}{q_{\text{c}}} \right) \tag{23.3.19}$$

to a good approximation. When the characteristic values of covalent semiconductors are inserted, the result is $\langle n_{\text{ph}} \rangle \sim 0.01$ to 0.02, that is, phonons mix very weakly with electrons. For metals, where k is on the order of the Fermi wave number, the mixing can be more important, and acoustic phonons may lead to a perceivable but still rather small mass enhancement.

The situation is radically different for the optical phonons in ionic crystals, where the polarizability is high and the number of phonons in the phonon cloud is not small. Then the elementary excitations are, strictly speaking, mixtures of electron and phonon states. These composite particles are called *polarons*. For simplicity, we shall examine the formation of polarons in the framework of perturbation theory, and determine the change in the wavefunction and energy of the Bloch electron of wave vector $\boldsymbol{k}$. Calculating the matrix element from the Hamiltonian (23.2.41) of the interaction between electrons and LO phonons, and using the value of the coupling constant given in (23.2.42), we have

$$\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\mathrm{LO}\boldsymbol{q}}|\mathcal{H}_{\text{el–ph}}|\boldsymbol{k}, 0\rangle = -\frac{\mathrm{i}}{q}\left\{\frac{e^2\hbar\omega_{\mathrm{LO}}}{2\epsilon_0 V}\left[\frac{1}{\epsilon_{\mathrm{r}}(\infty)} - \frac{1}{\epsilon_{\mathrm{r}}(0)}\right]\right\}^{1/2} . \tag{23.3.20}$$

The energy denominator of $\langle n_{\mathrm{ph}}\rangle$ in (23.3.14) is now the square of

$$\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_{\mathrm{LO}\boldsymbol{q}} = \frac{\hbar^2}{2m^*}(2\boldsymbol{k}\cdot\boldsymbol{q} - q^2) - \hbar\omega_{\mathrm{LO}} . \tag{23.3.21}$$

Neglecting, once again, the term proportional to $\boldsymbol{k}$ for slow electrons, and including the factor $1/q^2$ in the square of the matrix element, the following q-integral arises:

$$\int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\frac{1}{q^2}\frac{1}{(q^2+q_{\mathrm{LO}}^2)^2} , \tag{23.3.22}$$

where $q_{\mathrm{LO}}^2 = (2m^*/\hbar)\omega_{\mathrm{LO}}$. Instead of integrating over the Brillouin zone, the limits of integration can be extended to infinity. Since

$$\int_0^\infty \frac{\mathrm{d}x}{(x^2+a^2)^2} = \frac{\pi}{4a^3} , \tag{23.3.23}$$

the mean number of phonons in the cloud around the electron is

$$\langle n_{\mathrm{ph}}\rangle = \frac{e^2}{16\pi\epsilon_0\hbar\omega_{\mathrm{LO}}}\left(\frac{2m^*\omega_{\mathrm{LO}}}{\hbar}\right)^{1/2}\left[\frac{1}{\epsilon_{\mathrm{r}}(\infty)} - \frac{1}{\epsilon_{\mathrm{r}}(0)}\right] . \tag{23.3.24}$$

By introducing the dimensionless coupling constant

$$\alpha = \frac{\tilde{e}^2}{2\hbar\omega_{\mathrm{LO}}}\left(\frac{2m^*\omega_{\mathrm{LO}}}{\hbar}\right)^{1/2}\left[\frac{1}{\epsilon_{\mathrm{r}}(\infty)} - \frac{1}{\epsilon_{\mathrm{r}}(0)}\right] \tag{23.3.25}$$

we find

$$\langle n_{\mathrm{ph}}\rangle = \tfrac{1}{2}\alpha . \tag{23.3.26}$$

If the electron–phonon interaction is sufficiently weak, and perturbation theory may be applied, then the effective mass of the polaron can also be determined. The second-order term is the lowest nonvanishing correction to the energy of an electron in a phonon cloud:

$$\widetilde{\varepsilon}_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}} + \sum_{\boldsymbol{q}}\frac{|\langle \boldsymbol{k}-\boldsymbol{q}, 1_{\boldsymbol{q}}|\mathcal{H}_{\text{el–ph}}|\boldsymbol{k}, 0\rangle|^2}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_{\mathrm{LO}}} . \tag{23.3.27}$$

Expanding the energy denominator for small values of $\boldsymbol{k}$, the correction is proportional to k^2. Integration then leads to

$$\tilde{\varepsilon}_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}} - \alpha\left(\hbar\omega_{\mathrm{LO}} + \frac{\hbar^2k^2}{12m^*}\right) . \tag{23.3.28}$$

Neglecting the correction to the ground-state energy, the kinetic energy of the polaron is

$$\varepsilon_{\text{kin}} = \frac{\hbar^2}{2m^*}\big(1 - \tfrac{1}{6}\alpha\big)k^2 \,, \tag{23.3.29}$$

and so its effective mass is

$$m_{\text{pol}} = \frac{m^*}{1 - \frac{1}{6}\alpha} \approx m^* \left(1 + \tfrac{1}{6}\alpha\right) . \tag{23.3.30}$$

When the coupling is weak, the size of the polaron can also be estimated. Since the electron absorbs and emits phonons of energy $\hbar\omega_{\text{LO}}$, the energy uncertainty is $\Delta\varepsilon = \hbar\omega_{\text{LO}}$. The related uncertainty of the wave number can be determined from

$$\frac{\hbar^2(\Delta k)^2}{2m^*} = \hbar\omega_{\text{LO}} \,, \tag{23.3.31}$$

leading to

$$\Delta k = \sqrt{\frac{2m^*\omega_{\text{LO}}}{\hbar}} \,. \tag{23.3.32}$$

Identifying the size of the polaron with the position uncertainty obtained from the Heisenberg uncertainty principle,

$$r_0 = \frac{1}{\Delta k} = \sqrt{\frac{\hbar}{2m^*\omega_{\text{LO}}}} \,. \tag{23.3.33}$$

Since this usually means a distance of 10 to 100 lattice constants, they are called *large polarons.*

If the mean number of phonons around the electron is estimated from (23.3.24), and the measured low- and high-frequency values are used for the dielectric constant, $\langle n_{\text{ph}} \rangle$ is found to be between 1 and 3 for alkali halides. Since the dimensionless coupling constant is of order unity now, keeping the first few terms in the perturbation series is obviously not sufficient. Other methods need to be applied that treat electron–phonon interactions nonperturbatively. Such calculations show that the size of the polaron is then comparable to the lattice constant, hence the name *small polaron.* As the size decreases and the mass increases, these excitations are more and more localized. This change in the character of the electron states – the transition from states extending over the entire lattice to localized states due to the interaction with the vibrations of the lattice – can be studied in the *Holstein model.*[6] In this model electrons are described in the tight-binding approximation, with their spins neglected, in the basis of the Wannier states, and the vibrations of the lattice are specified in terms of dispersionless optical vibrations, as in the Einstein model (that is, atoms in each lattice point vibrate independently of one another with the same frequency ω_0). The interaction between electrons and lattice vibrations is assumed to be local, that is, the vibrational state of an atom may change only when an electron is located at the lattice point in question. Writing the Hamiltonian in second-quantized representation,

[6] T. Holstein, 1959.

$$\mathcal{H} = -t \sum_{<ij>} \left(c_i^\dagger c_j + c_j^\dagger c_i \right) + \hbar\omega_0 \sum_i b_i^\dagger b_i - g \sum_i c_i^\dagger c_i \left(b_i^\dagger + b_i \right) , \qquad (23.3.34)$$

where the operators $c_i^\dagger$ and $b_i^\dagger$ create and annihilate a lattice vibration on the ith lattice point, and $\langle ij \rangle$ denotes nearest-neighbor lattice points. Calculations show that the localization of the electrons occurs gradually as the coupling constant g increases. The study of transport properties requires very different methods in this localized state than in the general case.

23.3.3 Kohn Anomaly and Peierls Instability

Just like for electrons, the energy of phonons is also shifted (renormalized) by the electron–phonon interaction. The two calculations differ in the choice of the unperturbed wavefunction. It now corresponds to a state in which a single phonon of wave number $\boldsymbol{q}$ is present, and the electron system is in its ground state denoted by $|\mathrm{FS}\rangle$ – that is, the Bloch states are filled up to the Fermi energy. The unperturbed wavefunction is then

$$|1_{\boldsymbol{q}\lambda}\rangle = a_\lambda^\dagger(\boldsymbol{q})|\mathrm{FS}\rangle . \qquad (23.3.35)$$

Because of the interaction, this phonon may be annihilated while an electron–hole pair is created. The reverse process may also occur, and the phonon may reappear after a while. Denoting the wave vector of the hole by $\boldsymbol{k}$ and that of the electron by $\boldsymbol{k}+\boldsymbol{q}$, the wavefunction in this intermediate state is

$$|\boldsymbol{k}+\boldsymbol{q}, \boldsymbol{k}\rangle = c_{\boldsymbol{k}+\boldsymbol{q}\sigma}^\dagger c_{\boldsymbol{k}\sigma}|\mathrm{FS}\rangle . \qquad (23.3.36)$$

The phonon energy in the first nonvanishing order of perturbation theory is

$$\hbar\omega_\lambda^{(1)}(\boldsymbol{q}) = \hbar\omega_\lambda(\boldsymbol{q}) + \sum_{\boldsymbol{k}\sigma} \frac{|\langle \boldsymbol{k}, \boldsymbol{k}+\boldsymbol{q}|\mathcal{H}_{\text{el-ph}}|1_{\lambda\boldsymbol{q}}\rangle|^2}{\hbar\omega_\lambda(\boldsymbol{q}) - (\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{\boldsymbol{k}})} . \qquad (23.3.37)$$

When evaluating the sum, some care must be exercised, as the electron–hole pair was created in a ground-state Fermi sea, thus the hole is below and the electron is above the Fermi level. Consequently, the sum is over the region for which $|\boldsymbol{k}| < k_{\mathrm{F}}$ and $|\boldsymbol{k}+\boldsymbol{q}| > k_{\mathrm{F}}$. This region is shown in Fig. 23.6.

Since the most interesting effects appear in the vicinity of $2k_{\mathrm{F}}$, we shall approximate the strength of the electron–phonon interaction by a constant, and denote it by g. Specifying the range of summation in terms of the Fermi function,

$$\hbar\omega_\lambda^{(1)}(\boldsymbol{q}) = \hbar\omega_\lambda(\boldsymbol{q}) + \sum_{\boldsymbol{k}\sigma} g^2 \frac{f_0(\boldsymbol{k})[1 - f_0(\boldsymbol{k}+\boldsymbol{q})]}{\hbar\omega_\lambda(\boldsymbol{q}) - (\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{\boldsymbol{k}})} . \qquad (23.3.38)$$

Neglecting the phonon energy compared to the electron energies in the energy denominator,

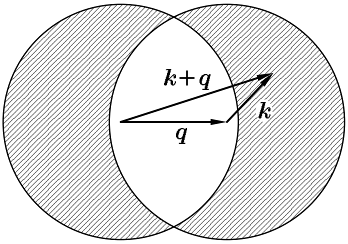


Fig. 23.6. The possible momenta of the electron and hole illustrated by two Fermi spheres displaced by $\boldsymbol{q}$

$$\hbar\omega_\lambda^{(1)}(\boldsymbol{q}) = \hbar\omega_\lambda(\boldsymbol{q}) - g^2 \sum_{\boldsymbol{k}\sigma} \frac{f_0(\boldsymbol{k})[1 - f_0(\boldsymbol{k}+\boldsymbol{q})]}{\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{\boldsymbol{k}}} . \tag{23.3.39}$$

Now the sum can be approximated by an integral, and the latter can be evaluated. We shall delve into the details in Chapter 33. Using the results to be derived there, it can be shown that the derivative of the energy correction for general three-dimensional systems exhibits a singularity at $q = 2k_F$ at zero temperature. Therefore anomalous dispersion appears in the phonon energy at this wave number. This is called *Kohn anomaly.*[7]

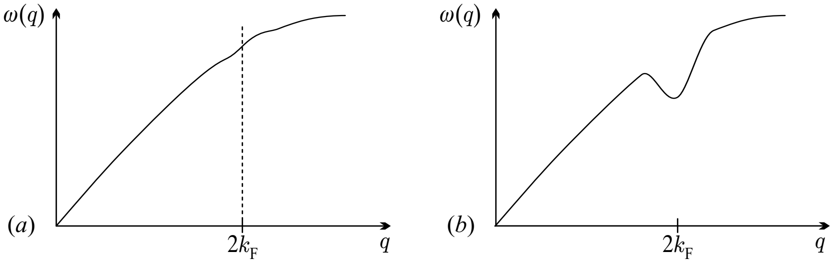


Fig. 23.7. The Kohn anomaly in the phonon spectrum due to the electron–phonon interaction at $q = 2k_F$, in a (a) three-dimensional and (b) one-dimensional system

As illustrated in Fig. 23.7(a), this anomaly is usually hardly observable. This is not the case in strongly anisotropic systems, in which the propagation of electrons is practically limited to a single direction. We shall demonstrate in Chapter 33 that the value of the integral in (23.3.39) depends greatly on the shape of the Fermi surface. In quasi-one-dimensional systems the energy correction itself shows a logarithmic singularity at $q = 2k_F$ at low temperatures. The conditions of adiabatic decoupling are no longer met then, and so it is not sufficient to keep the lowest-order corrections in perturbation theory.

[7] W. Kohn, 1959. Walter Kohn (1923–) was awarded the 1998 Nobel Prize in Chemistry "for his development of the density-functional theory".

More precise calculations based on the methods of the many-body problem show that the equation that determines the renormalized phonon frequency $\widetilde{\omega}_\lambda$ is

$$(\hbar\widetilde{\omega}_\lambda(\boldsymbol{q}))^2 = (\hbar\omega_\lambda(\boldsymbol{q}))^2 - 2\hbar\omega_\lambda(\boldsymbol{q})g^2 \sum_{\boldsymbol{k}\sigma} \frac{f_0(\boldsymbol{k})[1 - f_0(\boldsymbol{k}+\boldsymbol{q})]}{\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{\boldsymbol{k}}} \,. \tag{23.3.40}$$

Naturally, the result obtained in perturbation theory is recovered in leading order. In one-dimensional systems the energy of the phonon with wave number $q = 2k_{\mathrm{F}}$ shows strong temperature dependence, and even vanishes at a finite temperature:

$$k_{\mathrm{B}}T_{\mathrm{c}} = 2.28\,\varepsilon_{\mathrm{F}}\,\mathrm{e}^{-\hbar\omega_\lambda(2k_{\mathrm{F}})/g^2\rho(\varepsilon_{\mathrm{F}})} \,. \tag{23.3.41}$$

Slightly above this critical point, the energy of the phonon is

$$(\hbar\widetilde{\omega}_\lambda(2k_{\mathrm{F}}))^2 = \hbar\omega_\lambda(2k_{\mathrm{F}})g^2\rho(\varepsilon_{\mathrm{F}}) \ln\frac{T}{T_{\mathrm{c}}} \,. \tag{23.3.42}$$

The $2k_{\mathrm{F}}$ phonon is said to become *soft* at T_{c}, and the phonon state of vanishing energy may become macroscopically populated (just like in the Bose condensation). The system then becomes unstable against the static distortions of the lattice. Since the periodicity of the distorted lattice is determined by $2k_{\mathrm{F}}$, this quantity will be the primitive reciprocal-lattice vector of the new lattice. Figure 23.8 shows the distorted lattice and its new energy spectrum in the special case where the band was initially half filled ($2k_{\mathrm{F}} = \pi/a$), and so the distortion corresponds to the dimerization of the lattice, and hence a doubling of the unit cell.

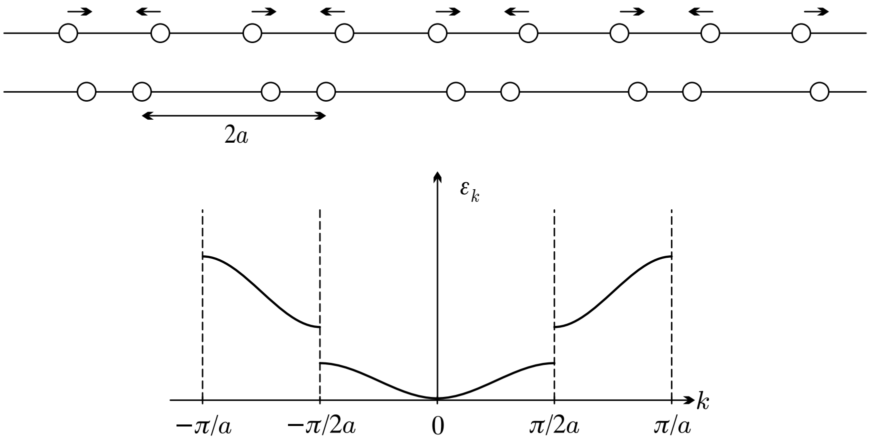


Fig. 23.8. The dimerization of the lattice and the formation of gaps in a one-dimensional system due to the Peierls instability

Even though below T_{c} the energy of the ion system increases because of the distortion, this is compensated for by the decrease in the energy of the electron system as the boundary of the Brillouin zone moves to k_{F}, and thus the energy of each occupied electron state is lowered. In the one-dimensional case the decrease is always larger than the increase in the elastic energy, that is why a $2k_{\mathrm{F}}$ distortion occurs in the position of the ions. It is accompanied by a density wave of the same periodicity in the electron system. This will be discussed in Chapter 33.

Note that the initially partially (half) filled band becomes completely filled now. Because of the gap at the zone boundary the material that shows metallic behavior at high temperatures becomes insulator at T_{c}. This is the *Peierls instability.*[8]

It is known that in strictly one-dimensional systems no phase transition is possible at finite temperatures: they are washed out by fluctuations. However, phase transition may occur in quasi-one-dimensional systems made up of weakly coupled chains, where the Fermi surface has a particular nesting property: two pieces on opposite sides of the Fermi surface can be matched when one is translated through a vector $\boldsymbol{q}$. In this case the integral for the shift of the phonon frequency shows similar logarithmic singularity to the one-dimensional case. However, a true phase transition may now take place, since the fluctuations are limited by the three-dimensional ordering of the density waves on individual chains. The more accurate description must take account of the electrostatic energy between chains.

23.3.4 Jahn–Teller Distortion

As discussed above, in strongly anisotropic, quasi-one-dimensional systems, where the motion of the electrons in the conduction band is essentially limited to a single direction, the interaction with the lattice makes the lattice distorted – in particular, it becomes dimerized when the band is half filled. However, the distortion of the lattice may also occur when the deformation lowers the energy of the electrons of the ion core (rather than the conduction electrons) by an amount that exceeds the increment in the elastic energy caused by the distortion. This phenomenon is called the Jahn–Teller effect.[9]

To understand this phenomenon, consider a crystal in which Cu^{2+} ions sit in an environment of cubic symmetry. The outermost open 3d shell accommodates nine electrons. In a free atom the state ${}^2\mathrm{D}_{5/2}$ of quantum numbers $S = 1/2$, $L = 2$, $J = 5/2$ would be the ground state according to Hund's rules. When the copper atoms sit in a crystal, the ground-state configuration can be determined by the method worked out in Chapter 6: the S and L values are specified by Hund's first and second rules, and then the crystal field, rather than the spin–orbit coupling, is considered as the essential perturbation. This splits the tenfold degenerate energy level further.

The $L = 2$ energy level is known to split into two in a cubic environment. The d_{xy}, d_{yz}, and d_{zx} states remain degenerate, and transform according

[8] R. E. Peierls, 1955.

[9] H. Jahn and E. Teller, 1936.

to the representation $\varGamma'_{25}$ ($T_{2\mathrm{g}}$), while the d_{z^2} and $d_{x^2-y^2}$ states transform according to $\varGamma_{12}$ (E_{g}). For more than half filled bands the $t_{2\mathrm{g}}$ states are known to be lower. Therefore they accommodate six electrons, and the e_{g} levels three.

When the environment of the ions does not exhibit a cubic symmetry but a lower one, the splitting pattern is somewhat different. Suppose that the crystal is slightly elongated or compressed along the z-axis, and so the cubic symmetry is reduced to tetragonal. By considering the distortion along the z-axis as a perturbation, the splitting pattern of the $t_{2\mathrm{g}}$ and e_{g} levels of the cubic structure can be determined. This is shown in Fig. 23.9.

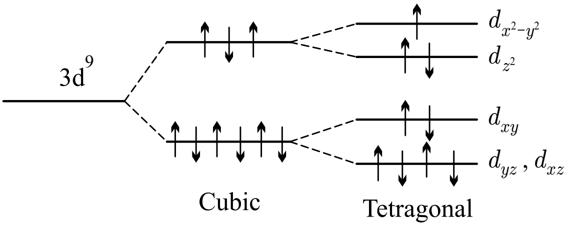


Fig. 23.9. The splitting of the ground-state energy level $S = 1/2$, $L = 2$ of the Cu^{2+} ion in a cubic environment, and in the presence of an additional weak tetragonal component

The total energy of the electrons on the $t_{2\mathrm{g}}$ level is not changed by the splitting caused by the distortion. Of the three electrons on the e_g level two will have lower and one higher energy, so the total energy of the electron system is lowered in the transition to the distorted state by an amount that is expected to be proportional to the distortion.

On the other hand, the elastic energy of the lattice increases, as for the Peierls distortion – and this increment is proportional to the square of the distortion parameter. When the two contributions are examined together, the linear term is found to be dominant for small distortions, and so the cubic lattice becomes spontaneously distorted to reach a lower-energy state.

This is generalized by the Jahn–Teller theorem, which states that *if the symmetry of the crystal field is so high that the highest-energy electrons would occupy degenerate orbitals in the ground state, then it is energetically more favorable for the crystal to undergo a distortion so that this degeneracy is lifted.*

At high temperature, where many different configurations are mixed, this distortion is no longer favorable. Just like the Peierls distortion, the Jahn–Teller distortion also occurs at a finite temperature, as a structural transformation.

23.3.5 Effective Electron–Electron Interaction

The second-order process of the electron–phonon interaction shown in Fig. 23.5(c) was interpreted as an effective interaction between two electrons through the emission of a phonon by one and its subsequent absorption by the other. The most convenient way to describe this effective interaction is to use the canonical transformation presented in detail in Appendix I, but the same result can also be obtained in the second order of perturbation theory.

In the latter approach we start with FRÖHLICH's[10] second-quantized Hamiltonian of the joint system of electrons and phonons, which contains the kinetic energy of the electrons, the energy of the free phonons, and the interaction between the electrons and the long-wavelength longitudinal acoustic phonons. Neglecting the electron spin,

$$\mathcal{H} = \sum_{\boldsymbol{k}} \varepsilon_{\boldsymbol{k}} c^{\dagger}_{\boldsymbol{k}} c_{\boldsymbol{k}} + \sum_{\boldsymbol{q}} \hbar\omega_{\boldsymbol{q}} a^{\dagger}_{\boldsymbol{q}} a_{\boldsymbol{q}} + \mathrm{i} \sum_{\boldsymbol{k},\boldsymbol{q}} D_{\boldsymbol{q}} c^{\dagger}_{\boldsymbol{k}+\boldsymbol{q}} c_{\boldsymbol{k}} \left(a_{\boldsymbol{q}} + a^{\dagger}_{-\boldsymbol{q}} \right) . \tag{23.3.43}$$

Writing the Fröhlich Hamiltonian as

$$\mathcal{H}_0 + \lambda \mathcal{H}_1 , \tag{23.3.44}$$

the electron–phonon interaction is chosen as the perturbation. Then a canonical transformation

$$\tilde{\mathcal{H}} = \mathrm{e}^{S} \mathcal{H} \mathrm{e}^{-S} \tag{23.3.45}$$

is performed, which transforms phonons out in the lowest order, leaving back only electrons. This can be achieved by expanding the transformed Hamiltonian in powers of S, and requiring that

$$\lambda \mathcal{H}_1 + [S, \mathcal{H}_0] = 0 , \tag{23.3.46}$$

thereby eliminating the direct electron–phonon interaction. The new Hamiltonian is then

$$\tilde{\mathcal{H}} = \mathcal{H}_0 + \tfrac{1}{2} \lambda \, [S, \mathcal{H}_1] + \mathcal{O}(\lambda^3) . \tag{23.3.47}$$

In our particular case, where the interaction term describes the creation or annihilation of a phonon, S has nonvanishing matrix elements between states whose phonon numbers differ by one. By choosing a state in which the number of phonons with wave vector $\boldsymbol{q}$ is $n_{\boldsymbol{q}}$, and keeping in mind that, according to (23.2.34), $D_{\boldsymbol{q}}$ is an odd function of $\boldsymbol{q}$, the matrix elements of interest are

$$\begin{aligned} \big\langle (n+1)_{\boldsymbol{q}} \big| S \big| n_{\boldsymbol{q}} \big\rangle &= \mathrm{i} D_{\boldsymbol{q}} \sum_{\boldsymbol{k}} c^{\dagger}_{\boldsymbol{k}-\boldsymbol{q}} c_{\boldsymbol{k}} \frac{1}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_{\boldsymbol{q}}} , \\ \big\langle n_{\boldsymbol{q}} \big| S \big| (n+1)_{\boldsymbol{q}} \big\rangle &= -\mathrm{i} D_{\boldsymbol{q}} \sum_{\boldsymbol{k}'} c^{\dagger}_{\boldsymbol{k}'+\boldsymbol{q}} c_{\boldsymbol{k}'} \frac{1}{\varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}'+\boldsymbol{q}} + \hbar\omega_{\boldsymbol{q}}} . \end{aligned} \tag{23.3.48}$$

[10] H. FRÖHLICH, 1954.

The term $\frac{1}{2}\lambda\,[S,\mathcal{H}_1]$ in the transformed Hamiltonian has nonvanishing matrix elements between those many-electron states that differ in the occupation of four one-particle states: relative to the initial state, two states become empty, and two others occupied in the final state. By taking those matrix elements for which the phonon state is the same in the initial and final states, this corresponds to an effective electron–electron scattering. Other matrix elements of $\widetilde{\mathcal{H}}$ describe the absorption or emission of two phonons, however we shall not be concerned with these processes now.

The Hamiltonian of the scattering of electrons – that is, of the effective electron–electron interaction – reads

$$\begin{aligned}\mathcal{H}_{\text{eff}} = \frac{1}{2}\sum_{\boldsymbol{q}} D_{\boldsymbol{q}}^2 \sum_{\boldsymbol{k},\boldsymbol{k}'} & c_{\boldsymbol{k}'+\boldsymbol{q}}^{\dagger} c_{\boldsymbol{k}'} c_{\boldsymbol{k}-\boldsymbol{q}}^{\dagger} c_{\boldsymbol{k}} \\ & \times \left(\frac{1}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}} - \hbar\omega_{\boldsymbol{q}}} - \frac{1}{\varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}'+\boldsymbol{q}} + \hbar\omega_{\boldsymbol{q}}} \right).\end{aligned} \tag{23.3.49}$$

Changing the variables as $\boldsymbol{k} \leftrightarrow \boldsymbol{k}'$ and $\boldsymbol{q} \to -\boldsymbol{q}$ in the second term, and making use of $\omega_{\boldsymbol{q}} = \omega_{-\boldsymbol{q}}$,

$$\mathcal{H}_{\text{eff}} = \sum_{\boldsymbol{q}} D_{\boldsymbol{q}}^2 \sum_{\boldsymbol{k},\boldsymbol{k}'} c_{\boldsymbol{k}'+\boldsymbol{q}}^{\dagger} c_{\boldsymbol{k}'} c_{\boldsymbol{k}-\boldsymbol{q}}^{\dagger} c_{\boldsymbol{k}} \frac{\hbar\omega_{\boldsymbol{q}}}{(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}})^2 - (\hbar\omega_{\boldsymbol{q}})^2}. \tag{23.3.50}$$

Up to now, the electron spin has been ignored. Since it is not changed by the emission or absorption of a phonon, the Hamiltonian of the effective interaction including spin is

$$\mathcal{H}_{\text{eff}} = \sum_{\boldsymbol{q}} D_{\boldsymbol{q}}^2 \sum_{\substack{\boldsymbol{k},\boldsymbol{k}' \\ \sigma\sigma'}} c_{\boldsymbol{k}'+\boldsymbol{q}\sigma'}^{\dagger} c_{\boldsymbol{k}'\sigma'} c_{\boldsymbol{k}-\boldsymbol{q}\sigma}^{\dagger} c_{\boldsymbol{k}\sigma} \frac{\hbar\omega_{\boldsymbol{q}}}{(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}})^2 - (\hbar\omega_{\boldsymbol{q}})^2}. \tag{23.3.51}$$

For electrons close to the Fermi surface, $|\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}-\boldsymbol{q}}|$ may be smaller than $\hbar\omega_{\boldsymbol{q}}$, therefore the emission of a phonon by such an electron and its absorption by another leads to an attractive interaction between the electrons. If this attraction overcomes the Coulomb repulsion between them, the pair of electrons forms a bound state. Such pairs are at the origin of superconductivity.

Further Reading

1. G. Grimvall, *The Electron-Phonon Interaction in Metals*, Selected Topics in Solid State Physics, Editor E. P. Wohlfarth, Volume XVI. North-Holland Publishing Company, Amsterdam (1981).
2. J. M. Ziman, *Electrons and Phonons, The Theory of Transport Phenomena in Solids*, Oxford Classic Texts in the Physical Sciences, Oxford University Press, Oxford (2001).

24

Transport Phenomena

In Chapter 16 we already discussed the propagation of electric and heat currents in the simplest model of metals, the classical Drude model, and then in the quantum mechanical Sommerfeld model, however the treatment was not consistently quantum mechanical in the latter. Since the complete solution of the quantum mechanical problem of electrons in an applied electric field or under other external forces (for example, a temperature gradient) is not known, we tacitly opted for the application of the semiclassical treatment by assuming that the occupation of the electron states can be characterized by a position- and wave-vector-dependent nonequilibrium stationary distribution function $f(\boldsymbol{r}, \boldsymbol{k})$. Having given a more precise meaning to the semiclassical approximation in Chapter 21, we are now in the position to apply it to the description of conduction phenomena, and examine the current carried by the electrons moving in the periodic potential of the lattice.

In addition to the semiclassical approximation, our previous considerations were also based on the assumption that, just like in the Drude model, the effects of collisions can be taken into account by a relaxation time in the Sommerfeld model, too. Since this quantity is introduced phenomenologically, nothing can be said about its temperature dependence within these models – and consequently, the correct results for the temperature dependence of the electrical resistivity and thermal conductivity cannot be derived, either.

Moreover, these models may, at best, be applied to the description of the properties of simple metals, in which the conduction electrons behave like free electrons; the transport properties of semiconductors or transition metals with a narrow d-band are obviously outside their realm. Therefore another approach is required for the description of the transport properties of Bloch electrons and phonons interacting with them.

Before starting the discussion of the characteristic phenomena in solids, we briefly summarize the most important results of the irreversible thermodynamical treatment of transport phenomena. Then we introduce the Boltzmann equation, whose solution gives the nonequilibrium distribution of electrons and phonons in the presence of an external perturbation. To determine the distribution function, we shall first study the Boltzmann equation in the relaxation-time approximation, and then take into account the scattering of electrons more accurately than before. Finally, we shall use the results to calculate the temperature dependence of the electrical resistivity and thermal conductivity of metals and semiconductors.

24.1 General Formulation of Transport Phenomena

In a system of charged particles an applied electric field induces an electric current, while a temperature gradient generates a heat current. However, the heat current may also be carried by uncharged particles, for example, phonons. The behavior of solids is primarily determined by the propagation of these currents inside them. In multicomponent materials the particle currents of individual components are also studied, while in semiconductors the currents carried by electrons and holes are customarily separated.

By writing the electric field as the gradient of an electrostatic potential, currents are seen to be induced by the gradient of this potential and the temperature gradient. The diffusion current of particles appears when the concentration has a nonvanishing gradient. In general, such gradients are the driving forces of currents.

In solids the electric field is usually sufficiently weak to justify dealing with linear phenomena alone. Then Ohm's law implies

$$\boldsymbol{j} = \boldsymbol{\sigma} \boldsymbol{E} = -\boldsymbol{\sigma} \boldsymbol{\nabla} \varphi \,, \tag{24.1.1}$$

where $\boldsymbol{\sigma}$ is the conductivity tensor. For a temperature gradient, too, the response to the perturbation – the heat current – is proportional to the driving force (unless its spatial variations are excessively rapid):

$$\boldsymbol{j}_Q = -\lambda \boldsymbol{\nabla} T \,, \tag{24.1.2}$$

where λ is the thermal or heat conductivity. If the particle density n is not uniform, diffusion leads to a nonvanishing particle-current density:

$$\boldsymbol{j}_n = -D \boldsymbol{\nabla} n \,, \tag{24.1.3}$$

where D is the diffusion coefficient.

It is well known from classical physics that an electric field can also induce a heat current, and a temperature gradient can generate an electric current. Likewise, electric and heat currents are induced when the chemical potential shows spatial variations. The corresponding component of the current is proportional to the gradient of the chemical potential. It was demonstrated in Chapter 16 that the coefficients of these cross effects are related to each other in a simple way. We shall give a more general formulation of this statement below.

24.1.1 Currents and Driving Forces

In the analysis of currents, we need to distinguish particle currents, electric currents, energy currents, heat currents, and entropy currents. Their densities are denoted by $\boldsymbol{j}_n$, $\boldsymbol{j}$, $\boldsymbol{j}_E$, $\boldsymbol{j}_Q$, and $\boldsymbol{j}_S$, respectively. The relationships between $\boldsymbol{j}_n$ and $\boldsymbol{j}$, and $\boldsymbol{j}_Q$ and $\boldsymbol{j}_S$ are easily established:

$$\boldsymbol{j}_Q = T\boldsymbol{j}_S \tag{24.1.4}$$

is generally valid, while for the density of the electric current carried by electrons

$$\boldsymbol{j} = -e\boldsymbol{j}_n \,. \tag{24.1.5}$$

One more relation can be written down if changes in the volume can be neglected ($\mathrm{d}V = 0$), which is a fairly good approximation for solids. From the thermodynamic relation

$$T\mathrm{d}S = \mathrm{d}E - \mu \mathrm{d}N \tag{24.1.6}$$

we then have

$$T\boldsymbol{j}_S = \boldsymbol{j}_E - \mu\boldsymbol{j}_n \,. \tag{24.1.7}$$

The same thermodynamic relation also implies a formula between the time rates of change of the entropy, energy, and particle number:

$$T\frac{\partial S}{\partial t} = \frac{\partial E}{\partial t} - \mu\frac{\partial N}{\partial t} \,. \tag{24.1.8}$$

When each quantity is referred to unit volume, and therefore denoted by the corresponding lowercase letter (except for the energy density, which is denoted by w to distinguish it from the elementary charge e), we have

$$T\frac{\partial s}{\partial t} = \frac{\partial w}{\partial t} - \mu\frac{\partial n}{\partial t} \,. \tag{24.1.9}$$

The particle density and the particle-current density have to satisfy the continuity equation

$$\frac{\partial n}{\partial t} + \operatorname{div} \boldsymbol{j}_n = 0 \,. \tag{24.1.10}$$

A similar equation is valid for the energy density and energy-current density; however, the generation of Joule heat is taken into account by an additional term in the energy balance:

$$\frac{\partial w}{\partial t} + \operatorname{div} \boldsymbol{j}_E = \boldsymbol{E} \cdot \boldsymbol{j} \,. \tag{24.1.11}$$

The analogous equation for the entropy density also contains an additional term on the right-hand side that corresponds to the local entropy production:

$$\frac{\partial s}{\partial t} + \operatorname{div} \boldsymbol{j}_S = \dot{s} \,. \tag{24.1.12}$$

Using the previous equations, $\dot{s}$ can be written as

$$\begin{aligned}\dot{s} &= \frac{1}{T}\left(\frac{\partial w}{\partial t} - \mu\frac{\partial n}{\partial t}\right) + \operatorname{div}\frac{\boldsymbol{j}_Q}{T} \\ &= \frac{1}{T}\left(\boldsymbol{E}\cdot\boldsymbol{j} - \operatorname{div}\boldsymbol{j}_E + \mu\operatorname{div}\boldsymbol{j}_n + \operatorname{div}\boldsymbol{j}_Q - \boldsymbol{j}_Q\cdot\frac{\boldsymbol{\nabla}T}{T}\right).\end{aligned} \tag{24.1.13}$$

From (24.1.4) and (24.1.7) we have

$$\operatorname{div}\boldsymbol{j}_Q = \operatorname{div}\boldsymbol{j}_E - \boldsymbol{j}_n\cdot\boldsymbol{\nabla}\mu - \mu\operatorname{div}\boldsymbol{j}_n\,, \tag{24.1.14}$$

and so the entropy-production formula can be simplified to

$$\begin{aligned}\dot{s} &= \frac{1}{T}\left\{\boldsymbol{E}\cdot\boldsymbol{j} - \boldsymbol{j}_n\cdot\boldsymbol{\nabla}\mu - \boldsymbol{j}_Q\cdot\frac{\boldsymbol{\nabla}T}{T}\right\} \\ &= \frac{1}{T}\left\{\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right)\cdot\boldsymbol{j} - \frac{\boldsymbol{\nabla}T}{T}\cdot\boldsymbol{j}_Q\right\}.\end{aligned} \tag{24.1.15}$$

Rewriting this as

$$\dot{s} = \frac{1}{T}\left(\boldsymbol{X}_e\cdot\boldsymbol{j} + \boldsymbol{X}_Q\cdot\boldsymbol{j}_Q\right), \tag{24.1.16}$$

the multiplying factor of the current densities is called the driving force of the current. If, in addition to the electric and heat currents, other currents also flow in the system, the corresponding driving force is defined through the relation

$$\dot{s} = \frac{1}{T}\sum_i \boldsymbol{X}_i\cdot\boldsymbol{j}_i\,. \tag{24.1.17}$$

The previous formulas confirm that the driving force of the electric current is not simply the electric field $\boldsymbol{E}$ but the combination $\boldsymbol{E} + \boldsymbol{\nabla}\mu/e$. When the electric field is derived from a scalar potential,

$$\boldsymbol{X}_e = \boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e} = -\boldsymbol{\nabla}\left(\varphi - \frac{\mu}{e}\right), \tag{24.1.18}$$

that is, the driving force is not the gradient of the electrostatic potential φ but that of the electrochemical potential $\varphi - \mu/e$, as was demonstrated for free electrons on page 53. Similarly, the driving force of the heat current is not simply $-\boldsymbol{\nabla}T$ but

$$\boldsymbol{X}_Q = -\frac{\boldsymbol{\nabla}T}{T}\,. \tag{24.1.19}$$

24.1.2 Onsager Relations

As mentioned earlier, in the overwhelming majority of transport phenomena in solids the currents can be taken to be proportional to the driving forces:

$$\boldsymbol{j}_i = \sum_j L_{ij} \boldsymbol{X}_j \,. \tag{24.1.20}$$

According to irreversible thermodynamics, the cross effects satisfy the Onsager relations in this linear approximation:

$$L_{ij}^{\alpha\beta} = L_{ji}^{\beta\alpha} \,, \tag{24.1.21}$$

while in a magnetic field

$$L_{ij}^{\alpha\beta}(B) = L_{ji}^{\beta\alpha}(-B) \,. \tag{24.1.22}$$

When only an electric field and a temperature gradient are applied, and only the electric and heat currents are considered,

$$\begin{aligned} \boldsymbol{j} &= L_{11}\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) + L_{12}\left(-\frac{\boldsymbol{\nabla}T}{T}\right), \\ \boldsymbol{j}_Q &= L_{21}\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) + L_{22}\left(-\frac{\boldsymbol{\nabla}T}{T}\right). \end{aligned} \tag{24.1.23}$$

In the Sommerfeld model these coefficients are scalars on account of the isotropy of the electron gas. Comparison with (16.3.32) shows that L_{12} and L_{21} are equal, and their value is the negative of the integral K_1 defined in (16.3.33). The equality of L_{12} and L_{21} is now seen not to be accidental but the consequence of the Onsager relations of nonequilibrium thermodynamics, and is thus generally valid.

24.2 Boltzmann Equation

An exact quantum mechanical treatment of conduction phenomena requires the apparatus of the many-body problem. As we shall briefly discuss in Chapter 36, this is indeed necessary in disordered systems. However, in solid-state physics we usually deal with systems in which the mean free path of electrons is much larger than their de Broglie wavelength, and therefore the conduction electrons can be considered to make up a semiclassical electron gas as far as the conduction properties are concerned – so their equation of motion is known. The kinetic theory of gases can then be applied to this electron gas, with the difference that the most important collision mechanism is not the scattering of particles by one another but the interaction with the crystal lattice, its vibrations or defects. The reason why transport phenomena were not analyzed in depth in Chapter 16 on the Sommerfeld model is that the reader may not have been familiar with the details of the interaction between the electrons and the vibrating lattice (the electron–phonon interaction).

In thermal equilibrium the occupation of electron states is characterized by the Fermi–Dirac distribution, while phonon states are described by the Bose–Einstein distribution. In the semiclassical approximation the nonequilibrium state can be specified by a nonequilibrium distribution function. We shall first derive the Boltzmann equation that determines this distribution function, and then present the approximations used for solving it.

24.2.1 Nonequilibrium Distribution Function

Consider a point $\boldsymbol{r}, \boldsymbol{k}$ in the phase space of the position and wave vector of electrons, and denote the number of electrons that are inside the phase-space volume element $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}$ around the point $\boldsymbol{r}, \boldsymbol{k}$ at time t by $\mathrm{d}N(\boldsymbol{r}, \boldsymbol{k}, t)$. The nonequilibrium distribution function $f(\boldsymbol{r}, \boldsymbol{k}, t)$ is then defined as

$$\mathrm{d}N(\boldsymbol{r}, \boldsymbol{k}, t) = f(\boldsymbol{r}, \boldsymbol{k}, t)\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}}{4\pi^3}\,. \tag{24.2.1}$$

This formula contains the information that the $\boldsymbol{k}$-space density of the allowed wave vectors is $V/(2\pi)^3$ in a sample of volume V, and the factor 2 due to the two possible spin orientations is also included. The electric current density at point $\boldsymbol{r}$ at time t is then

$$\boldsymbol{j}(\boldsymbol{r}, t) = -e\int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{\boldsymbol{k}} f(\boldsymbol{r}, \boldsymbol{k}, t)\,, \tag{24.2.2}$$

while the heat-current density is

$$\boldsymbol{j}_Q(\boldsymbol{r}, t) = \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}(\varepsilon_{\boldsymbol{k}} - \mu)\boldsymbol{v}_{\boldsymbol{k}} f(\boldsymbol{r}, \boldsymbol{k}, t)\,. \tag{24.2.3}$$

If electrons populate several bands, then the number of electrons can be considered separately for each band. When $\mathrm{d}N_n$ of the $\mathrm{d}N$ electrons in the phase-space volume element occupy states of the band of index n, the distribution function for that band is defined by

$$\mathrm{d}N_n(\boldsymbol{r}, \boldsymbol{k}, t) = f_n(\boldsymbol{r}, \boldsymbol{k}, t)\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}}{4\pi^3}\,. \tag{24.2.4}$$

Since electrons in completely filled bands do not carry a net current, we shall be concerned with the contribution of incomplete shells in our treatment of transport phenomena. In most of our calculations we shall assume that there is a single band of interest, and thus suppress the band index. As individual bands contribute additively to the currents, the generalization to the case of multiple partially filled bands is straightforward.

In magnetic systems, the number of electrons in the volume element $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}$ may be different for the two spin orientations ($\sigma = \uparrow, \downarrow$). We shall also see examples where the strength of the interaction that governs the scattering of electrons depends on the orientation of the electron spin. Their proper treatment requires the introduction of a spin-dependent distribution function, which can be defined in terms of the number of spin-σ particles as

$$\mathrm{d}N_\sigma(\boldsymbol{r},\boldsymbol{k},t) = f_\sigma(\boldsymbol{r},\boldsymbol{k},t)\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}}{8\pi^3}\,. \tag{24.2.5}$$

Below we shall suppress the spin index, unless it is expressly needed.

In thermal equilibrium the distribution of the electrons is specified by the well-known Fermi function

$$f_0(\boldsymbol{k}) = \frac{1}{\exp\left[\left(\varepsilon_{\boldsymbol{k}} - \mu\right)/k_\mathrm{B}T\right] + 1}\,. \tag{24.2.6}$$

If the temperature or the chemical potential is nonuniform, the distribution function contains their local values:

$$f_0(\boldsymbol{r},\boldsymbol{k}) = \frac{1}{\exp\left[\left(\varepsilon_{\boldsymbol{k}} - \mu(\boldsymbol{r})\right)/k_\mathrm{B}T(\boldsymbol{r})\right] + 1}\,. \tag{24.2.7}$$

Phonons are treated on the same footing. A temperature gradient modifies the phonon distribution with respect to the thermal equilibrium. Denoting the number of phonons of polarization λ in the region $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{q}$ around the point $\boldsymbol{r},\boldsymbol{q}$ of the phase space by $\mathrm{d}N_\lambda$, the phonon distribution function $g_\lambda(\boldsymbol{r},\boldsymbol{q},t)$ is defined as

$$\mathrm{d}N_\lambda(\boldsymbol{r},\boldsymbol{q},t) = g_\lambda(\boldsymbol{r},\boldsymbol{q},t)\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{q}}{8\pi^3}\,. \tag{24.2.8}$$

Since phonons obey the Bose–Einstein statistics, and their chemical potential vanishes in thermal equilibrium, we have

$$g_\lambda^0(\boldsymbol{q}) = \frac{1}{\exp\left[\hbar\omega_\lambda(\boldsymbol{q})/k_\mathrm{B}T\right] - 1}\,, \tag{24.2.9}$$

while if the system is in local equilibrium in a nonuniform temperature distribution,

$$g_\lambda^0(\boldsymbol{r},\boldsymbol{q}) = \frac{1}{\exp\left[\hbar\omega_\lambda(\boldsymbol{q})/k_\mathrm{B}T(\boldsymbol{r})\right] - 1}\,. \tag{24.2.10}$$

24.2.2 Boltzmann Equation for Electrons

To determine the nonequilibrium distribution function, the phase-space motion of the particles inside the volume element $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}$ around the point $\boldsymbol{r},\boldsymbol{k}$ at time t needs to be studied. If no collisions occur, then $\mathrm{d}t$ time later they are in the region $\mathrm{d}\boldsymbol{r}'\mathrm{d}\boldsymbol{k}'$ around $\boldsymbol{r}',\boldsymbol{k}'$, where the primed and unprimed quantities are related by $\boldsymbol{r}' = \boldsymbol{r} + \dot{\boldsymbol{r}}\,\mathrm{d}t$ and $\boldsymbol{k}' = \boldsymbol{k} + \dot{\boldsymbol{k}}\,\mathrm{d}t$. Because of the conservation of the particle number,

$$f(\boldsymbol{r},\boldsymbol{k},t)\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k} = f(\boldsymbol{r} + \dot{\boldsymbol{r}}\,\mathrm{d}t, \boldsymbol{k} + \dot{\boldsymbol{k}}\,\mathrm{d}t, t + \mathrm{d}t)\,\mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{k}'\,. \tag{24.2.11}$$

According to Liouville's theorem, the phase-space volume remains constant during the motion, $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k} = \mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{k}'$, and so

$$f(\boldsymbol{r} + \dot{\boldsymbol{r}}\,\mathrm{d}t, \boldsymbol{k} + \dot{\boldsymbol{k}}\,\mathrm{d}t, t + \mathrm{d}t) = f(\boldsymbol{r},\boldsymbol{k},t)\,. \tag{24.2.12}$$

For small time differences the linear-order expansion of the left-hand side gives

$$\frac{\mathrm{d}f}{\mathrm{d}t} \equiv \frac{\partial f}{\partial t} + \dot{\boldsymbol{r}} \cdot \frac{\partial f}{\partial \boldsymbol{r}} + \dot{\boldsymbol{k}} \cdot \frac{\partial f}{\partial \boldsymbol{k}} = 0\,, \tag{24.2.13}$$

which is in fact the equation of continuity in phase space.

On the other hand, when collisions occur during the infinitesimal time interval $\mathrm{d}t$, by which certain particles are scattered out of the phase-space trajectory (outscattering) – or other particles are scattered into the vicinity of the phase-space point $\boldsymbol{r}', \boldsymbol{k}'$ (inscattering) – then a collision term appears on the right-hand side:

$$f(\boldsymbol{r} + \dot{\boldsymbol{r}}\,\mathrm{d}t, \boldsymbol{k} + \dot{\boldsymbol{k}}\,\mathrm{d}t, t + \mathrm{d}t) = f(\boldsymbol{r}, \boldsymbol{k}, t) + \left(\frac{\partial f(\boldsymbol{r}, \boldsymbol{k}, t)}{\partial t}\right)_{\text{coll}} \mathrm{d}t\,, \tag{24.2.14}$$

where

$$\left(\frac{\partial f(\boldsymbol{r}, \boldsymbol{k}, t)}{\partial t}\right)_{\text{coll}} \mathrm{d}t = \left(\frac{\partial f(\boldsymbol{r}, \boldsymbol{k}, t)}{\partial t}\right)_{\text{in}} \mathrm{d}t - \left(\frac{\partial f(\boldsymbol{r}, \boldsymbol{k}, t)}{\partial t}\right)_{\text{out}} \mathrm{d}t \tag{24.2.15}$$

is the change in particle number due to the difference of inscattering and outscattering. Expanding this formula to linear order in $\mathrm{d}t$,

$$\frac{\partial f}{\partial t} + \dot{\boldsymbol{r}} \cdot \frac{\partial f}{\partial \boldsymbol{r}} + \dot{\boldsymbol{k}} \cdot \frac{\partial f}{\partial \boldsymbol{k}} = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}}\,. \tag{24.2.16}$$

This is the *Boltzmann equation.*[1]

In the semiclassical approximation the phase-space motion of electrons is given by the equations

$$\dot{\boldsymbol{r}} = \boldsymbol{v}_{\boldsymbol{k}} = \frac{1}{\hbar}\frac{\partial \varepsilon_{\boldsymbol{k}}}{\partial \boldsymbol{k}}\,, \qquad \hbar\dot{\boldsymbol{k}} = -e(\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B})\,. \tag{24.2.17}$$

Thus in the stationary case the distribution function can be determined by solving

$$\boldsymbol{v}_{\boldsymbol{k}} \cdot \frac{\partial f}{\partial \boldsymbol{r}} - \frac{e}{\hbar}(\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}) \cdot \frac{\partial f}{\partial \boldsymbol{k}} = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}}\,. \tag{24.2.18}$$

Assuming that the distribution function is just slightly different from the thermal-equilibrium function f_0, using the notation $f = f_0 + f_1$ we may transform the Boltzmann equation into an equation for the departure f_1 from the equilibrium distribution:

$$\begin{aligned} &\boldsymbol{v}_{\boldsymbol{k}} \cdot \frac{\partial f_0}{\partial \boldsymbol{r}} - \frac{e}{\hbar}(\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}) \cdot \frac{\partial f_0}{\partial \boldsymbol{k}} \\ &\qquad = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} - \boldsymbol{v}_{\boldsymbol{k}} \cdot \frac{\partial f_1}{\partial \boldsymbol{r}} + \frac{e}{\hbar}(\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}) \cdot \frac{\partial f_1}{\partial \boldsymbol{k}}\,. \end{aligned} \tag{24.2.19}$$

[1] L. Boltzmann, 1872.

Since f_0 depends on the wave and position vectors only through the combination $(\varepsilon_{\boldsymbol{k}} - \mu(\boldsymbol{r}))/k_{\mathrm{B}}T(\boldsymbol{r})$, the left-hand side can be rewritten as

$$\begin{aligned}\boldsymbol{v}_{\boldsymbol{k}} \cdot \frac{\partial f_0}{\partial \boldsymbol{r}} - \frac{e}{\hbar}\left(\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}\right) \cdot \frac{\partial f_0}{\partial \boldsymbol{k}} \\ = -\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right) \boldsymbol{v}_{\boldsymbol{k}} \cdot \left[-e\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) - \frac{\varepsilon_{\boldsymbol{k}} - \mu}{T}\boldsymbol{\nabla}T\right].\end{aligned} \tag{24.2.20}$$

The explicit magnetic-field dependence of the left-hand side has thus been eliminated. Nonetheless an implicit dependence remains, as the trajectory of electrons – and, consequently, their velocity, too – still depends on the applied field. This clearly shows the special role of the magnetic field: when the motion of electrons is examined, it cannot be considered as a weak perturbation to which electrons react linearly. The situation is different for the electric field. Customarily, only the linear response to the electric field is studied. As f_1 itself is proportional to the applied field, the term that is proportional to $\boldsymbol{E}$ can be neglected on the right-hand side of (24.2.19), as it would lead to a quadratic correction. The Boltzmann equation for electrons is therefore

$$\begin{aligned}-\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right)\boldsymbol{v}_{\boldsymbol{k}} \cdot \left[-e\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) - \frac{\varepsilon_{\boldsymbol{k}} - \mu}{T}\boldsymbol{\nabla}T\right] \\ = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} - \boldsymbol{v}_{\boldsymbol{k}} \cdot \frac{\partial f_1}{\partial \boldsymbol{r}} + \frac{e}{\hbar}(\boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}) \cdot \frac{\partial f_1}{\partial \boldsymbol{k}}.\end{aligned} \tag{24.2.21}$$

We shall usually study conduction phenomena in uniform fields and in the presence of a constant temperature gradient, and be concerned with the dependence of the distribution function only on the wave vector and energy.

Up to now nothing has been said about how the collision term in the Boltzmann equation should be chosen. Therefore we shall first determine the collision term due to the scattering of electrons by the defects of a rigid lattice, and then turn to the collision terms in a vibrating lattice due to the electron–phonon interaction (coupled system of electrons and phonons). We shall also see what collision terms arise when phonons interact only among themselves.

24.2.3 Collision Term for Scattering by Lattice Defects

In a rigid lattice, the collision term comes from the interaction of electrons with other electrons and their scattering by lattice defects. The collision integral can be written down simply for the latter provided the matrix elements of the scattering by lattice defects are known.

Let us denote the probability for an electron in state $\boldsymbol{k}$ to be scattered in time $\mathrm{d}t$ into an empty state in the volume element $\mathrm{d}\boldsymbol{k}'$ around $\boldsymbol{k}'$ by $W_{\boldsymbol{k}\boldsymbol{k}'}\dfrac{\mathrm{d}\boldsymbol{k}'\,\mathrm{d}t}{(2\pi)^3}$. Since the probability for a given state to be empty is $1 - f(\boldsymbol{k}')$, and the state $\boldsymbol{k}$ is present with a weight $f(\boldsymbol{k})$, the probability of outscattering is

$$\left(\frac{\partial f}{\partial t}\right)_{\text{out}} = f(\boldsymbol{k}) \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} \left[1 - f(\boldsymbol{k}')\right] . \tag{24.2.22}$$

In the inscattering process electrons are scattered from $\boldsymbol{k}'$ to $\boldsymbol{k}$; this requires that $\boldsymbol{k}'$ be occupied and $\boldsymbol{k}$ empty initially. The probability of this process is therefore

$$\left(\frac{\partial f}{\partial t}\right)_{\text{in}} = [1 - f(\boldsymbol{k})] \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}'\boldsymbol{k}}\, f(\boldsymbol{k}') . \tag{24.2.23}$$

Since the spin is conserved in the scattering, there is no summation over the spin variable. The full collision integral of the two processes is

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} \left\{ W_{\boldsymbol{k}'\boldsymbol{k}} f(\boldsymbol{k}') \left[1 - f(\boldsymbol{k})\right] - W_{\boldsymbol{k}\boldsymbol{k}'} f(\boldsymbol{k}) \left[1 - f(\boldsymbol{k}')\right] \right\} . \tag{24.2.24}$$

When this is substituted into the Boltzmann equation, a nonlinear integrodifferential equation is obtained for the electronic distribution function even in the simplest case. Its solution obviously requires further approximations. Before turning to their discussion, let us write down the Boltzmann equation for the nonequilibrium distribution of phonons, which sometimes also play an important role in the determination of transport properties of solids.

24.2.4 Boltzmann Equation for Phonons

When the perturbations vary slowly both in space and time, the properties determined by the lattice vibrations (phonons) can also be treated in the semiclassical approximation – that is, a position- and time-dependent semiclassical distribution function can be introduced for phonons much in the same way as was done for electrons. Apart from piezoelectric effects, the electromagnetic field does not act on phonons; however, when a temperature gradient is present, the equilibrium distribution of phonons is disturbed, and they can also carry a heat current. Consequently, the equation governing the phase-space variations of the distribution function $g_\lambda(\boldsymbol{r}, \boldsymbol{q}, t)$ for phonons of polarization λ differs formally from the Boltzmann equation for electrons in the absence of the term arising from the variation of the wave vector:

$$\frac{\partial g_\lambda(\boldsymbol{r}, \boldsymbol{q}, t)}{\partial t} + \boldsymbol{c}_\lambda(\boldsymbol{q}) \cdot \frac{\partial g_\lambda(\boldsymbol{r}, \boldsymbol{q}, t)}{\partial \boldsymbol{r}} = \left(\frac{\partial g_\lambda}{\partial t}\right)_{\text{coll}} , \tag{24.2.25}$$

where $\boldsymbol{c}_\lambda(\boldsymbol{q})$ is the group velocity of phonons of polarization λ, which can be derived from the dispersion relation as

$$\boldsymbol{c}_\lambda(\boldsymbol{q}) = \frac{\partial \omega_\lambda(\boldsymbol{q})}{\partial \boldsymbol{q}} . \tag{24.2.26}$$

In the stationary case

$$\boldsymbol{c}_\lambda \cdot \frac{\partial g_\lambda(\boldsymbol{r},\boldsymbol{q},t)}{\partial \boldsymbol{r}} = \left(\frac{\partial g_\lambda}{\partial t}\right)_{\text{coll}} . \tag{24.2.27}$$

By relating the spatial variations to the temperature gradient,

$$\boldsymbol{c}_\lambda \cdot \boldsymbol{\nabla} T \frac{\partial g_\lambda(\boldsymbol{r},\boldsymbol{q},t)}{\partial T} = \left(\frac{\partial g_\lambda}{\partial t}\right)_{\text{coll}} . \tag{24.2.28}$$

When the system of phonons is studied in itself, the decay and merger processes of phonons due to anharmonicity need to be taken into account to determine the collision integral. For simplicity, we shall consider only three-phonon processes, and neglect umklapp processes. The processes in which a phonon of wave vector $\boldsymbol{q}$ can participate are shown in Fig. 24.1.

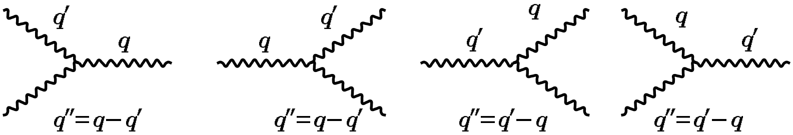


Fig. 24.1. Decay and merger processes with three phonons involved, which modify the distribution function of the phonons of wave vector $\boldsymbol{q}$

We shall denote by $W_{\boldsymbol{q}\boldsymbol{q}'}$ the transition probability of the process in which a phonon of wave vector $\boldsymbol{q}$ decays into two phonons, $\boldsymbol{q}'$ and $\boldsymbol{q}'' = \boldsymbol{q} - \boldsymbol{q}'$. Obviously, the inverse process – in which two phonons, $\boldsymbol{q}'$ and $\boldsymbol{q}'' = \boldsymbol{q} - \boldsymbol{q}'$, merge into one, of wave vector $\boldsymbol{q}$ – has the same transition probability. The number of phonons of wave vector $\boldsymbol{q}$ is reduced by the first and increased by the second process, so they contribute to outscattering and inscattering, respectively. To determine the actual contribution, one has to sum over the possible values of $\boldsymbol{q}'$. It must, however, be borne in mind that for inscattering processes the transition probability $W_{\boldsymbol{q}\boldsymbol{q}'}$ has to be multiplied by the probability that the system initially contains the two phonons, $\boldsymbol{q}'$ and $\boldsymbol{q}''$ (that is, by the product of the corresponding distribution functions).

Owing to its bosonic character, the creation of a phonon $\boldsymbol{q}$ does not require that the state should be empty – moreover, the creation probability is even increased by the presence of existing phonons because of stimulated emission, yielding a factor $1 + g_\lambda(\boldsymbol{q})$. By taking into account the contributions of the other processes in Fig. 24.1,

$$\begin{aligned}\left(\frac{\partial g_\lambda}{\partial t}\right)_{\text{coll}} = \sum_{\lambda'\lambda''} \int \frac{\mathrm{d}\boldsymbol{q}'}{(2\pi)^3} \Big\{ & \tfrac{1}{2} W_{\boldsymbol{q}\boldsymbol{q}'} \left[g_{\lambda'}(\boldsymbol{q}') g_{\lambda''}(\boldsymbol{q}-\boldsymbol{q}')(1+g_\lambda(\boldsymbol{q})) \right. \\ & \left. - g_\lambda(\boldsymbol{q})(1+g_{\lambda'}(\boldsymbol{q}'))(1+g_{\lambda''}(\boldsymbol{q}-\boldsymbol{q}'))\right] \\ & + W_{\boldsymbol{q}'\boldsymbol{q}} \left[g_{\lambda'}(\boldsymbol{q}')(1+g_\lambda(\boldsymbol{q}))(1+g_{\lambda''}(\boldsymbol{q}'-\boldsymbol{q})) \right. \\ & \left. - g_\lambda(\boldsymbol{q}) g_{\lambda''}(\boldsymbol{q}'-\boldsymbol{q})(1+g_{\lambda'}(\boldsymbol{q}'))\right] \Big\} . \end{aligned} \tag{24.2.29}$$

The factor $\frac{1}{2}$ in the first term is due to the indistinguishability of phonons: since the processes related by $\boldsymbol{q}' \leftrightarrow \boldsymbol{q}-\boldsymbol{q}'$ are identical, a twofold overcounting occurs in the $\boldsymbol{q}'$ integral.

24.2.5 Coupled Electron–Phonon Systems

Because of the interaction between electrons and phonons, the two distribution functions are coupled. The collision term in the Boltzmann equation for the electron distribution function $f(\boldsymbol{r},\boldsymbol{k},t)$ contains the contributions of four kinds of process. The number of electrons with wave vector $\boldsymbol{k}$ increases when an electron of wave vector $\boldsymbol{k}'=\boldsymbol{k}+\boldsymbol{q}$ emits a phonon of wave vector $\boldsymbol{q}$ or absorbs one of wave vector $-\boldsymbol{q}$, and is thus scattered into the state of quantum number $\boldsymbol{k}$. On the other hand, the number of such electrons is reduced by the processes in which an electron of wave vector $\boldsymbol{k}$ absorbs or emits a phonon. These possibilities are shown in Fig. 24.2.

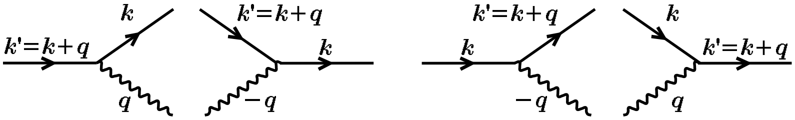


Fig. 24.2. Electron scattering processes with phonon emission and absorption that increase and decrease the number of electrons with wave vector $\boldsymbol{k}$

Consequently, the collision integral for electrons contains four terms. The collision term is proportional to the transition probabilities of the phonon-emission and -absorption processes. For the process in which an electron of wave vector $\boldsymbol{k}$ absorbs one of the $n_\lambda(\boldsymbol{q})$ phonons of polarization λ and wave vector $\boldsymbol{q}$,

$$W_{\boldsymbol{k},\boldsymbol{q},\lambda;\boldsymbol{k}'} = \frac{2\pi}{\hbar}\left|\langle \boldsymbol{k}', n_\lambda(\boldsymbol{q})-1|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}, n_\lambda(\boldsymbol{q})\rangle\right|^2 \delta(\varepsilon_{\boldsymbol{k}}+\hbar\omega_\lambda(\boldsymbol{q})-\varepsilon_{\boldsymbol{k}'})\delta_{\boldsymbol{k}+\boldsymbol{q},\boldsymbol{k}'}\,. \tag{24.2.30}$$

The transition probabilities of other processes are similar in form. As the matrix element of the electron–phonon interaction is the same whether the electron emits or absorbs the phonon, after the separation of the factors granting energy and momentum conservation each process is proportional to

$$\begin{aligned} I_{\boldsymbol{k},\boldsymbol{q},\lambda} &= \frac{2\pi}{\hbar}\left|\langle \boldsymbol{k}'=\boldsymbol{k}+\boldsymbol{q}, n_\lambda(\boldsymbol{q})-1|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}, n_\lambda(\boldsymbol{q})\rangle\right|^2 \\ &= \frac{2\pi}{\hbar}\left|\langle \boldsymbol{k}'=\boldsymbol{k}-\boldsymbol{q}, n_\lambda(\boldsymbol{q})+1|\mathcal{H}_{\text{el-ph}}|\boldsymbol{k}, n_\lambda(\boldsymbol{q})\rangle\right|^2 . \end{aligned} \tag{24.2.31}$$

Of course, in phonon absorption and emission alike, the initial electron state must be occupied, while the final state that becomes occupied in the scattering must be initially empty according to the Pauli exclusion principle.

Similarly, the initial phonon state must be occupied in phonon-absorption processes, while in phonon-emission processes a factor $1+g$ appears because of stimulated emission. Adding the scattering processes in their order in Fig. 24.2,

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \frac{1}{V}\sum_{\boldsymbol{k}'\boldsymbol{q}\lambda} \Big\{ & W_{\boldsymbol{k}';\boldsymbol{k},\boldsymbol{q},\lambda} f(\boldsymbol{k}')[1-f(\boldsymbol{k})][1+g_\lambda(\boldsymbol{q})] \\ & + W_{\boldsymbol{k}',-\boldsymbol{q},\lambda;\boldsymbol{k}} f(\boldsymbol{k}') g_\lambda(-\boldsymbol{q})[1-f(\boldsymbol{k})] \\ & - W_{\boldsymbol{k};\boldsymbol{k}',-\boldsymbol{q},\lambda} f(\boldsymbol{k})[1-f(\boldsymbol{k}')][1+g_\lambda(-\boldsymbol{q})] \\ & - W_{\boldsymbol{k},\boldsymbol{q},\lambda;\boldsymbol{k}'} f(\boldsymbol{k}) g_\lambda(\boldsymbol{q})[1-f(\boldsymbol{k}')]\Big\}.\end{aligned} \tag{24.2.32}$$

The same scattering processes also contribute to the collision integral in the Boltzmann equation for phonons, since the phonon number is changed by phonon emission and absorption:

$$\begin{aligned}\left(\frac{\partial g_\lambda}{\partial t}\right)_{\text{coll}} = \frac{1}{V}\sum_{\boldsymbol{k}\boldsymbol{k}'} \Big\{ & W_{\boldsymbol{k}';\boldsymbol{k},\boldsymbol{q},\lambda} f(\boldsymbol{k}')[1-f(\boldsymbol{k})][1+g_\lambda(\boldsymbol{q})] \\ & - W_{\boldsymbol{k},\boldsymbol{q},\lambda;\boldsymbol{k}'} f(\boldsymbol{k}) g_\lambda(\boldsymbol{q})[1-f(\boldsymbol{k}')]\Big\}.\end{aligned} \tag{24.2.33}$$

To quantify the role of electron–phonon scattering, the coupled system of equations for the electron and phonon distribution functions needs to be solved. In thermal equilibrium the contributions of inscattering and outscattering processes cancel out, because when energy conservation is taken into account, $\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} = \varepsilon_{\boldsymbol{k}} + \hbar\omega_\lambda(\boldsymbol{q})$ implies

$$f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})[1-f_0(\varepsilon_{\boldsymbol{k}})][1+g_0(\omega_\lambda(\boldsymbol{q}))] = f_0(\varepsilon_{\boldsymbol{k}})g_0(\omega_\lambda(\boldsymbol{q}))[1-f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})]\,, \tag{24.2.34}$$

and similarly, $\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} + \hbar\omega_\lambda(-\boldsymbol{q}) = \varepsilon_{\boldsymbol{k}}$ implies

$$f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})g_0(\omega_\lambda(-\boldsymbol{q}))[1-f_0(\varepsilon_{\boldsymbol{k}})] = f_0(\varepsilon_{\boldsymbol{k}})[1-f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})][1+g_0(\omega_\lambda(-\boldsymbol{q}))]\,. \tag{24.2.35}$$

This is the *principle of detailed balance* for individual processes. The collision integral vanishes unless an external perturbation drives the system out of thermal equilibrium. If the departure from equilibrium is small, and the occupation of electron states is expected to change only around the Fermi energy, the nonequilibrium distribution functions can be written as

$$\begin{aligned} f(\boldsymbol{k}) &= f_0(\boldsymbol{k}) - k_{\mathrm{B}}T\frac{\partial f_0(\boldsymbol{k})}{\partial\varepsilon_{\boldsymbol{k}}}\chi(\boldsymbol{k})\,, \\ g_\lambda(\boldsymbol{q}) &= g_\lambda^0(\boldsymbol{q}) - \frac{k_{\mathrm{B}}T}{\hbar}\frac{\partial g_0(\boldsymbol{q})}{\partial\omega_\lambda(\boldsymbol{q})}\phi_\lambda(\boldsymbol{q})\,.\end{aligned} \tag{24.2.36}$$

Writing out explicitly the derivative of the equilibrium distribution function,

$$\begin{aligned} f(\boldsymbol{k}) &= f_0(\boldsymbol{k}) + f_0(\boldsymbol{k})(1-f_0(\boldsymbol{k}))\chi(\boldsymbol{k})\,, \\ g_\lambda(\boldsymbol{q}) &= g_\lambda^0(\boldsymbol{q}) + g_\lambda^0(\boldsymbol{q})[1+g_\lambda^0(\boldsymbol{q})]\phi_\lambda(\boldsymbol{q})\,.\end{aligned} \tag{24.2.37}$$

By linearizing the collision integral in the small dimensionless quantities χ and ϕ, and making use of the property that the equilibrium phonon distribution function is even in $\boldsymbol{q}$, we find

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = &-\sum_{\boldsymbol{q}\lambda} I_{\boldsymbol{k},\boldsymbol{q},\lambda} f_0(\boldsymbol{k})[1-f_0(\boldsymbol{k}+\boldsymbol{q})] g_\lambda^0(\boldsymbol{q}) \\ &\quad\times \bigl\{\chi(\boldsymbol{k})-\chi(\boldsymbol{k}+\boldsymbol{q})+\phi_\lambda(\boldsymbol{q})\bigr\}\delta(\varepsilon_{\boldsymbol{k}}+\hbar\omega_\lambda(\boldsymbol{q})-\varepsilon_{\boldsymbol{k}+\boldsymbol{q}}) \\ &-\sum_{\boldsymbol{q}\lambda} I_{\boldsymbol{k},\boldsymbol{q},\lambda} f_0(\boldsymbol{k})[1-f_0(\boldsymbol{k}+\boldsymbol{q})][1+g_\lambda^0(\boldsymbol{q})] \\ &\quad\times \bigl\{\chi(\boldsymbol{k})-\chi(\boldsymbol{k}+\boldsymbol{q})-\phi_\lambda(-\boldsymbol{q})\bigr\}\delta(\varepsilon_{\boldsymbol{k}}-\hbar\omega_\lambda(\boldsymbol{q})-\varepsilon_{\boldsymbol{k}+\boldsymbol{q}}),\end{aligned} \tag{24.2.38}$$

and

$$\begin{aligned}\left(\frac{\partial g_\lambda}{\partial t}\right)_{\text{coll}} = &-\sum_{\boldsymbol{q}\lambda} I_{\boldsymbol{k},\boldsymbol{q},\lambda} f_0(\boldsymbol{k})[1-f_0(\boldsymbol{k}+\boldsymbol{q})] g_\lambda^0(\boldsymbol{q}) \\ &\quad\times \bigl\{\chi(\boldsymbol{k})-\chi(\boldsymbol{k}+\boldsymbol{q})+\phi_\lambda(\boldsymbol{q})\bigr\}\delta(\varepsilon_{\boldsymbol{k}}+\hbar\omega_\lambda(\boldsymbol{q})-\varepsilon_{\boldsymbol{k}+\boldsymbol{q}}).\end{aligned} \tag{24.2.39}$$

The system of equations is quite complicated even after this linearization. Instead of solving it, we usually study either the electron or phonon distribution function, and assume that the other subsystem is in thermal equilibrium.

24.3 Relaxation-Time Approximation

In the previous section we saw that the difficulty in solving the Boltzmann equation is rooted in the fact that the distribution function to be determined appears in the integrand of the collision term as well. Before turning to the general treatment of transport phenomena based on the Boltzmann equation, we shall first examine under what conditions the collision term can be interpreted in terms of a relaxation time, and where this treatment leads to. We shall then compare the results with those derived in the Drude and Sommerfeld models, where the finite relaxation time was assumed to be the same for each electron.

24.3.1 Relaxation Time

To introduce the relaxation time, we shall start with the impurity scattering formula (24.2.24). The collision term vanishes in thermal equilibrium, since the inscattering and outscattering processes compensate each other, that is,

$$W_{\boldsymbol{k}\boldsymbol{k}'} f_0(\boldsymbol{k})\bigl[1-f_0(\boldsymbol{k}')\bigr] = W_{\boldsymbol{k}'\boldsymbol{k}} f_0(\boldsymbol{k}')\bigl[1-f_0(\boldsymbol{k})\bigr]. \tag{24.3.1}$$

This requirement, which formulates the condition of detailed balance for these processes, is customarily written as

$$W_{\boldsymbol{k}\boldsymbol{k}'}\exp[-\varepsilon_{\boldsymbol{k}}/k_{\mathrm{B}}T] = W_{\boldsymbol{k}'\boldsymbol{k}}\exp[-\varepsilon_{\boldsymbol{k}'}/k_{\mathrm{B}}T]\,, \tag{24.3.2}$$

too. Naturally, $W_{\boldsymbol{k}\boldsymbol{k}'} = W_{\boldsymbol{k}'\boldsymbol{k}}$ for elastic scattering. Then

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} &= \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} f_0(\boldsymbol{k})\left[1 - f_0(\boldsymbol{k}')\right] \\ &\quad\times\left[\frac{f(\boldsymbol{k}')\left[1 - f(\boldsymbol{k})\right]}{f_0(\boldsymbol{k}')\left[1 - f_0(\boldsymbol{k})\right]} - \frac{f(\boldsymbol{k})\left[1 - f(\boldsymbol{k}')\right]}{f_0(\boldsymbol{k})\left[1 - f_0(\boldsymbol{k}')\right]}\right].\end{aligned} \tag{24.3.3}$$

Assuming that the departure $f_1 \equiv f - f_0$ from equilibrium is small,

$$\begin{aligned}&\frac{f(\boldsymbol{k}')\left[1 - f(\boldsymbol{k})\right]}{f_0(\boldsymbol{k}')\left[1 - f_0(\boldsymbol{k})\right]} - \frac{f(\boldsymbol{k})\left[1 - f(\boldsymbol{k}')\right]}{f_0(\boldsymbol{k})\left[1 - f_0(\boldsymbol{k}')\right]} \\ &\qquad = \frac{f_1(\boldsymbol{k}')}{f_0(\boldsymbol{k}')\left[1 - f_0(\boldsymbol{k}')\right]} - \frac{f_1(\boldsymbol{k})}{f_0(\boldsymbol{k})\left[1 - f_0(\boldsymbol{k})\right]}\,.\end{aligned} \tag{24.3.4}$$

Making use of the formula

$$f_0(\boldsymbol{k})\left[1 - f_0(\boldsymbol{k})\right] = -k_{\mathrm{B}}T\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \tag{24.3.5}$$

for the Fermi function, and rewriting f_1 in the previously used form

$$f_1(\boldsymbol{k}) = -k_{\mathrm{B}}T\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\chi(\boldsymbol{k})\,, \tag{24.3.6}$$

we have

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} f_0(\boldsymbol{k})\left[1 - f_0(\boldsymbol{k}')\right]\left[\chi(\boldsymbol{k}') - \chi(\boldsymbol{k})\right]. \tag{24.3.7}$$

For elastic scattering, where $\varepsilon_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}'}$,

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} &= -k_{\mathrm{B}}T\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'}\left[\chi(\boldsymbol{k}') - \chi(\boldsymbol{k})\right] \\ &= k_{\mathrm{B}}T\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\chi(\boldsymbol{k})\int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'}\left[1 - \frac{\chi(\boldsymbol{k}')}{\chi(\boldsymbol{k})}\right] \\ &= -(f(\boldsymbol{k}) - f_0(\boldsymbol{k}))\int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'}\left[1 - \frac{\chi(\boldsymbol{k}')}{\chi(\boldsymbol{k})}\right].\end{aligned} \tag{24.3.8}$$

The distribution function in a selected point $\boldsymbol{r}, \boldsymbol{k}$ of the phase space usually depends on its value in other points; it is precisely for this reason that the collision term contains the integral of the distribution function. For elastic scattering, the collision integral could be cast in a simple form in which the departure from the equilibrium distribution is multiplied by a $\boldsymbol{k}$-dependent factor. To understand the physical meaning of the proportionality factor we shall determine the variation of the number of particles due to collisions by another method.

Let $\tau(\boldsymbol{r},\boldsymbol{k})$ be the mean time between collisions for a particle in the vicinity of point $\boldsymbol{r},\boldsymbol{k}$ of the phase space; in other words, let $1/\tau(\boldsymbol{r},\boldsymbol{k})$ be the probability for the same particle to be scattered in unit time. Out of the $\mathrm{d}N_{\mathrm{e}}$ particles in volume $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}/4\pi^3$ of the phase space

$$\frac{\mathrm{d}t}{\tau(\boldsymbol{r},\boldsymbol{k})}\mathrm{d}N_{\mathrm{e}}=\frac{\mathrm{d}t}{\tau(\boldsymbol{r},\boldsymbol{k})}f(\boldsymbol{r},\boldsymbol{k},t)\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}}{4\pi^3} \tag{24.3.9}$$

are then scattered in time $\mathrm{d}t$. These outscattering processes modify the distribution function by

$$\mathrm{d}f(\boldsymbol{r},\boldsymbol{k},t)_{\mathrm{out}}=-\frac{\mathrm{d}t}{\tau(\boldsymbol{r},\boldsymbol{k})}f(\boldsymbol{r},\boldsymbol{k},t)\,. \tag{24.3.10}$$

The analogous results for inscattering are just as straightforward to derive, only the two previously discussed fundamental conditions are required. The first condition is very natural: the collisions should not modify the distribution function in thermal equilibrium. The second is much less intuitive. It requires that the distribution after collisions should be independent of the state before the collisions – in other words, the collisions should erase the memory of the system. This implies that the strength of inscattering is independent of the instantaneous state of the local environment in phase space – that is, its departure from equilibrium is immaterial. Since inscattering and outscattering compensate each other in thermal equilibrium, we shall assume that the distribution function changes by

$$\mathrm{d}f(\boldsymbol{r},\boldsymbol{k},t)_{\mathrm{in}}=\frac{\mathrm{d}t}{\tau(\boldsymbol{r},\boldsymbol{k})}f_0(\boldsymbol{r},\boldsymbol{k})\,. \tag{24.3.11}$$

The total change due to collisions is then

$$\left(\frac{\partial f(\boldsymbol{r},\boldsymbol{k},t)}{\partial t}\right)_{\mathrm{coll}}=-\frac{1}{\tau(\boldsymbol{r},\boldsymbol{k})}\left[f(\boldsymbol{r},\boldsymbol{k},t)-f_0(\boldsymbol{r},\boldsymbol{k})\right]. \tag{24.3.12}$$

If collisions were the only processes, the system would relax toward equilibrium with a characteristic time $\tau(\boldsymbol{r},\boldsymbol{k})$ called the *relaxation time.*

Comparing this general expression with (24.3.8) leads to the following formula for the reciprocal of the relaxation time:

$$\frac{1}{\tau(\boldsymbol{k})}=\int\frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3}W_{\boldsymbol{k}\boldsymbol{k}'}\left[1-\frac{\chi(\boldsymbol{k}')}{\chi(\boldsymbol{k})}\right]. \tag{24.3.13}$$

Attention should be paid to a subtlety that turns out to be crucial in our later considerations: in the relaxation time (24.3.13), the transition probability $W_{\boldsymbol{k}\boldsymbol{k}'}$ is not simply integrated over each process (as would be if the inverse lifetime of a particle were to be calculated), but scattering processes are taken into account by a weight factor that depends on the wave vectors $\boldsymbol{k}$ and $\boldsymbol{k}'$. Therefore, the relaxation time τ is often called transport lifetime and sometimes denoted by τ_{tr} in order to distinguish it from the ordinary lifetime obtained from

$$\frac{1}{\tau(\boldsymbol{k})} = \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} \,. \tag{24.3.14}$$

Following the same steps, the collision term for phonons is often well approximated by expressing it in terms of the phonon relaxation time τ_{ph} as

$$\left(\frac{\partial g_\lambda(\boldsymbol{q})}{\partial t} \right)_{\text{coll}} = -\frac{g_\lambda(\boldsymbol{q}) - g_\lambda^0(\boldsymbol{q})}{\tau_{\text{ph}}(\boldsymbol{q})} \,. \tag{24.3.15}$$

The introduction of the relaxation time was based on the assumption of elastic scattering. It can be demonstrated that the relaxation-time approximation can also be used for quasielastic scattering, where the energy transfer is smaller than the thermal energy $k_{\text{B}}T$; otherwise the whole integral has to be considered. Looking back at (24.3.13), it should be noted that, through χ, the relaxation time also depends on the nonequilibrium distribution, and so it is not just a characteristic parameter of scattering. This observation also indicates the limitations of the relaxation-time approximation, even for elastic scattering.

24.3.2 Distribution Function in the Relaxation-Time Approximation

The most straightforward method to determine the distribution function is the direct integration of the Boltzmann equation. If no magnetic field is present and the driving forces are uniform – i.e., T is not constant in space but $\boldsymbol{\nabla} T$ is –, this can be carried out without difficulty. When the collision term is written in the relaxation-time approximation, and the other terms on the right-hand side of the Boltzmann equation (24.2.21) are neglected – since, according to our assumptions, $\boldsymbol{B} = 0$, and the distribution function is spatially uniform –, the solution of the equation

$$-\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) \boldsymbol{v}_{\boldsymbol{k}} \cdot \left[-e \left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e} \right) - \frac{\varepsilon_{\boldsymbol{k}} - \mu}{T} \boldsymbol{\nabla} T \right] = -\frac{f(\boldsymbol{k}) - f_0(\boldsymbol{k})}{\tau(\varepsilon_{\boldsymbol{k}})} \tag{24.3.16}$$

is

$$f(\boldsymbol{k}) = f_0(\boldsymbol{k}) + \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) \tau(\varepsilon_{\boldsymbol{k}}) \boldsymbol{v}_{\boldsymbol{k}} \cdot \left[-e \left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e} \right) - \frac{\varepsilon_{\boldsymbol{k}} - \mu}{T} \boldsymbol{\nabla} T \right] . \tag{24.3.17}$$

Thus (16.3.27), which was derived using much simpler considerations, is recovered. This straightforward approach can no longer be used in the presence of a magnetic field. By applying CHAMBERS' method,[2] intuitive formulas are obtained that can be used in a wide range of applications.

[2] R. G. CHAMBERS, 1963.

Consider the phase-space volume element $\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}$ around a point $\boldsymbol{r},\boldsymbol{k}$ at time t, and determine the electron trajectories $\boldsymbol{r}(t'),\boldsymbol{k}(t')$ for anterior times $t' < t$ by integrating the equation of motion backward in time from $\boldsymbol{r}(t),\boldsymbol{k}(t)$. These trajectories are shown in Fig. 24.3.

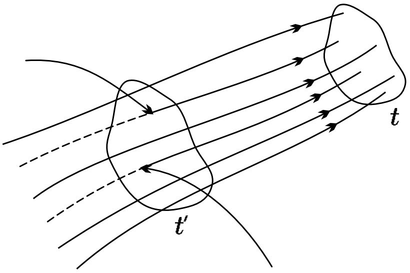


Fig. 24.3. Semiclassical trajectories of electrons running into a small neighborhood of a selected point of the phase space at time t

If there were no collisions, each electron that traverses one of these semiclassical trajectories would be in the selected region at time t. However, electrons trace out these trajectories only in the intervals between collisions. Some of the electrons are kicked off from them upon scattering, while others emerge on them after a collision. Eventually, those electrons reach the neighborhood of point $\boldsymbol{r},\boldsymbol{k}$ at time t that were scattered onto one of the trajectories at time $t' < t$ and did not undergo subsequent scattering before t.

The number of electrons that emerge on these trajectories after a collision during the time interval $\mathrm{d}t'$ around t' is given by

$$\frac{\mathrm{d}t'}{\tau(\boldsymbol{r}(t'),\boldsymbol{k}(t'))} f_0(\boldsymbol{r}(t'),\boldsymbol{k}(t'))\frac{\mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{k}'}{4\pi^3} \tag{24.3.18}$$

in the relaxation-time approximation. According to Liouville's theorem, $\mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{k}' = \mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}$, thus the number of inscattered electrons is

$$\frac{\mathrm{d}t'}{\tau(\boldsymbol{r}(t'),\boldsymbol{k}(t'))} f_0(\boldsymbol{r}(t'),\boldsymbol{k}(t'))\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}}{4\pi^3}\,. \tag{24.3.19}$$

Denoting the probability that the particle does not undergo any further scattering by $P(\boldsymbol{r},\boldsymbol{k},t,t')$, the total number of electrons that arrive in the selected phase-space volume element at time t is given by the integral

$$\mathrm{d}N = \int\limits_{-\infty}^{t} \frac{\mathrm{d}t'}{\tau(\boldsymbol{r}(t'),\boldsymbol{k}(t'))} f_0(\boldsymbol{r}(t'),\boldsymbol{k}(t'))P(\boldsymbol{r},\boldsymbol{k},t,t')\,\frac{\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{k}}{4\pi^3}\,. \tag{24.3.20}$$

Comparison with (24.2.1), the defining equation of the distribution function, gives

$$f(\boldsymbol{r},\boldsymbol{k},t)=\int\limits_{-\infty}^{t}\mathrm{d}t'\frac{1}{\tau(\boldsymbol{r}(t'),\boldsymbol{k}(t'))}f_0(\boldsymbol{r}(t'),\boldsymbol{k}(t'))P(\boldsymbol{r},\boldsymbol{k},t,t')\,. \tag{24.3.21}$$

For simplicity, we keep only the time arguments:

$$f(t)=\int\limits_{-\infty}^{t}\mathrm{d}t'\frac{1}{\tau(t')}f_0(t')P(t,t')\,. \tag{24.3.22}$$

Since the collision probability in time $\mathrm{d}t'$ is $\mathrm{d}t'/\tau$, and that of collisionless propagation is $1-\mathrm{d}t'/\tau$, the probability $P(t,t')$ that no collision occurs from t' to t is

$$P(t,t')=P(t,t'+\mathrm{d}t')\left[1-\frac{\mathrm{d}t'}{\tau(t')}\right], \tag{24.3.23}$$

and hence

$$\frac{\partial}{\partial t'}P(t,t')=\frac{P(t,t')}{\tau(t')}\,. \tag{24.3.24}$$

The solution that satisfies the initial condition $P(t,t)=1$ is

$$P(t,t')=\exp\left(-\int\limits_{t'}^{t}\frac{\mathrm{d}t''}{\tau(t'')}\right). \tag{24.3.25}$$

Instead of using the explicit form of $P(t,t')$ in (24.3.22), it is more convenient to substitute (24.3.24) and perform an integration by parts. Since the probability that a particle never underwent collisions is zero $[P(t,-\infty)=0]$, we have

$$f(t)=\int\limits_{-\infty}^{t}\mathrm{d}t' f_0(t')\frac{\partial}{\partial t'}P(t,t')=f_0(t)-\int\limits_{-\infty}^{t}\mathrm{d}t' P(t,t')\frac{\mathrm{d}}{\mathrm{d}t'}f_0(t')\,. \tag{24.3.26}$$

As the t'-dependence appears through the arguments $\boldsymbol{r}(t')$ and $\boldsymbol{k}(t')$, the rules of implicit differentiation give

$$\frac{\mathrm{d}f_0(t')}{\mathrm{d}t'}=\frac{\partial f_0(t')}{\partial\varepsilon_{\boldsymbol{k}}}\frac{\partial\varepsilon_{\boldsymbol{k}}}{\partial\boldsymbol{k}}\cdot\frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t'}+\frac{\partial f_0(t')}{\partial T}\frac{\partial T}{\partial\boldsymbol{r}}\cdot\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t'}+\frac{\partial f_0(t')}{\partial\mu}\frac{\partial\mu}{\partial\boldsymbol{r}}\cdot\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t'}\,. \tag{24.3.27}$$

Since f_0 depends on the combination $[\varepsilon_{\boldsymbol{k}}-\mu(\boldsymbol{r})]/T(\boldsymbol{r})$, its derivatives can be expressed in terms of the derivatives with respect to the energy. Making use of the semiclassical equation of motion, we find

$$\begin{aligned}f(t)=f_0(t)+&\int\limits_{-\infty}^{t}\mathrm{d}t' P(t,t')\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\\&\times\boldsymbol{v}_{\boldsymbol{k}}(t')\cdot\left[-e\boldsymbol{E}(\boldsymbol{r}(t'))-\boldsymbol{\nabla}\mu(\boldsymbol{r}(t'))-\frac{\varepsilon_{\boldsymbol{k}}-\mu}{T}\boldsymbol{\nabla}T(\boldsymbol{r}(t'))\right].\end{aligned} \tag{24.3.28}$$

This is the general form of the distribution function in the relaxation-time approximation. The magnetic field does not appear explicitly, but it does implicitly, since it has strong influence on the semiclassical electron trajectories.

If the electric field and the temperature gradient are weak, and linear effects are considered, $P(t,t')$, which depends on the collisions, can often be approximated by the formulas obtained for $\boldsymbol{E}=0$ and $\boldsymbol{\nabla}T=0$. In a uniform field nothing depends explicitly on $\boldsymbol{r}(t')$, only implicitly, through $\boldsymbol{k}(t')$. It is customary to assume that the relaxation time depends on $\boldsymbol{k}$ only through $\varepsilon_{\boldsymbol{k}}$. Since $\varepsilon_{\boldsymbol{k}}$ is constant in a uniform magnetic field, the integral in $P(t,t')$ can be evaluated:

$$P(t,t') = \mathrm{e}^{-(t-t')/\tau(\varepsilon_{\boldsymbol{k}})}\,, \tag{24.3.29}$$

and

$$\begin{aligned} f(t) = f_0(t) + \int\limits_{-\infty}^{t} \mathrm{d}t' \mathrm{e}^{-(t-t')/\tau(\varepsilon_{\boldsymbol{k}})} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right) \\ \times\, \boldsymbol{v}_{\boldsymbol{k}}(t') \cdot \left[-e\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) - \frac{\varepsilon_{\boldsymbol{k}}-\mu}{T}\boldsymbol{\nabla}T\right]. \end{aligned} \tag{24.3.30}$$

In terms of the quantity

$$\boldsymbol{w}_{\boldsymbol{k}}(t) = \frac{1}{\tau(\varepsilon_{\boldsymbol{k}})} \int\limits_{-\infty}^{t} \mathrm{d}t' \mathrm{e}^{-(t-t')/\tau(\varepsilon_{\boldsymbol{k}})} \boldsymbol{v}_{\boldsymbol{k}}(t')\,, \tag{24.3.31}$$

the distribution function can be written as

$$f(t) = f_0(t) + \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right) \tau(\varepsilon_{\boldsymbol{k}}) \boldsymbol{w}_{\boldsymbol{k}}(t) \cdot \left[-e\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) - \frac{\varepsilon_{\boldsymbol{k}}-\mu}{T}\boldsymbol{\nabla}T\right]. \tag{24.3.32}$$

When no external magnetic field is present, and the driving forces are uniform, the drift velocity is constant in time and, according to (24.3.31), $\boldsymbol{w}_{\boldsymbol{k}} = \boldsymbol{v}_{\boldsymbol{k}}$. The stationary distribution function is therefore

$$f(\boldsymbol{k}) = f_0(\boldsymbol{k}) + \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right) \tau(\varepsilon_{\boldsymbol{k}}) \boldsymbol{v}_{\boldsymbol{k}} \cdot \left[-e\left(\boldsymbol{E} + \frac{\boldsymbol{\nabla}\mu}{e}\right) - \frac{\varepsilon_{\boldsymbol{k}}-\mu}{T}\boldsymbol{\nabla}T\right]. \tag{24.3.33}$$

Note that this is the same as (24.3.17), which was obtained for the same situation. In the presence of a magnetic field, where the velocity changes on account of moving in a circular orbit, the relationship between the results of the two approaches is less conspicuous. Depending on the character of the particular situation, we shall use either form below.

24.3.3 DC Conductivity

According to the foregoing, in the absence of a magnetic field and a temperature gradient, when $\boldsymbol{E}$ is constant, the distribution function is given by

$$f(\boldsymbol{k}) = f_0(\boldsymbol{k}) - e\,(\boldsymbol{E}\cdot\boldsymbol{v}_{\boldsymbol{k}})\,\tau(\varepsilon_{\boldsymbol{k}})\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right). \tag{24.3.34}$$

This is the same as our previous result (16.3.17) obtained by assuming that the stationary distribution function can be derived from the Fermi function by shifting the energy in its argument by $e\tau\boldsymbol{E}/\hbar$, the energy gained from the electric field. As (24.3.34) shows, the approach based on the Boltzmann equation asserts that the same formula is applicable to Bloch electrons, too.

Writing the current density carried by electrons in its customary form, and noting that the current vanishes in thermal equilibrium,

$$\boldsymbol{j} = -e\frac{1}{V}\sum_{\boldsymbol{k},\sigma}\boldsymbol{v}_{\boldsymbol{k}}\big[f(\boldsymbol{k}) - f_0(\boldsymbol{k})\big] = -e\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{\boldsymbol{k}}\big[f(\boldsymbol{k}) - f_0(\boldsymbol{k})\big]. \tag{24.3.35}$$

By separating the part of the distribution function that is proportional to the electric field, the conductivity tensor is found to be

$$\boldsymbol{\sigma} = e^2\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{\boldsymbol{k}}\circ\boldsymbol{v}_{\boldsymbol{k}}\tau(\varepsilon_{\boldsymbol{k}})\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right), \tag{24.3.36}$$

where $\boldsymbol{v}_{\boldsymbol{k}}\circ\boldsymbol{v}_{\boldsymbol{k}}$ denotes the diadic product of the two vectors. This tensor exhibits the symmetries of the crystal lattice. In cubic crystals $\sigma_{\alpha\beta} = \delta_{\alpha\beta}\sigma$, and thus the conductivity can be specified by a scalar. For lower symmetries, $\boldsymbol{j}$ and $\boldsymbol{E}$ are not necessarily parallel, and, depending on the symmetry, off-diagonal tensor elements may also appear.

Because of the derivative of the Fermi function only those electrons contribute that are close to ε_{F}, within a region of width $k_{\mathrm{B}}T$. When the conductivity at $T = 0$ is considered, the factor $\tau(\varepsilon_{\mathrm{F}})$ can be taken outside the integral. Making use of

$$\boldsymbol{v}_{\boldsymbol{k}}\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right) = -\frac{1}{\hbar}\frac{\partial f_0(\varepsilon_{\boldsymbol{k}})}{\partial \boldsymbol{k}}, \tag{24.3.37}$$

the formula for the conductivity is

$$\boldsymbol{\sigma} = -\frac{1}{\hbar}e^2\tau(\varepsilon_{\mathrm{F}})\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\boldsymbol{v}_{\boldsymbol{k}}\circ\frac{\partial f_0(\varepsilon_{\boldsymbol{k}})}{\partial \boldsymbol{k}}. \tag{24.3.38}$$

Integrating by parts, and assuming that the dispersion relation can be specified in terms of a scalar effective mass, at zero temperature we have

$$\sigma_0 = e^2\tau(\varepsilon_{\mathrm{F}})\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\frac{1}{\hbar}\frac{\partial \boldsymbol{v}_{\boldsymbol{k}}}{\partial \boldsymbol{k}}f_0(\varepsilon_{\boldsymbol{k}}) = e^2\tau(\varepsilon_{\mathrm{F}})\int\limits_{\text{occupied}}\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\frac{1}{m^*_{\boldsymbol{k}}}. \tag{24.3.39}$$

If the effective mass is independent of $\boldsymbol{k}$, the Drude formula

$$\sigma_0 = \frac{n_{\mathrm{e}}e^2\tau}{m^*} \tag{24.3.40}$$

is recovered, with the sole difference that the electron mass is replaced by the effective mass. A perfectly intuitive meaning can be given to n_{e}: it is the number (density) of electrons in the partially filled conduction band, while τ is the transport relaxation time of electrons on the Fermi surface.

If the integral in the conductivity is performed over the entire Brillouin zone rather than the occupied states, it vanishes, as completely filled bands do not contribute to the current. Therefore, a negative sign aside, the integration could be done over the empty states as well, leading to

$$\sigma_0 = e^2 \tau(\varepsilon_{\mathrm{F}}) \int\limits_{\text{empty}} \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3} \left(-\frac{1}{m^*_{\boldsymbol{k}}} \right) . \tag{24.3.41}$$

Since the electron and hole masses are defined to be opposite in sign, the same form is obtained for the current and conductivity as above, provided the latter is expressed in terms of the effective hole mass m^*_{h}:

$$\sigma_0 = \frac{n_{\mathrm{h}} e^2 \tau}{m^*_{\mathrm{h}}} , \tag{24.3.42}$$

where n_{h} is the number of holes per unit volume.

24.3.4 AC and Optical Conductivity

To determine the AC conductivity, we start with the distribution function formula (24.3.30) obtained using Chambers' method. In an electric field $\boldsymbol{E}(t) = \boldsymbol{E}(\omega)\mathrm{e}^{-\mathrm{i}\omega t}$ of frequency ω, the integral over the semiclassical trajectory of electrons can be evaluated:

$$\begin{aligned} f(\boldsymbol{k},t) &= f_0(\boldsymbol{k}) + \int\limits_{-\infty}^{t} \mathrm{d}t' \mathrm{e}^{-(t-t')/\tau(\varepsilon_{\boldsymbol{k}})} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) (-e) \boldsymbol{v}_{\boldsymbol{k}}(t') \cdot \boldsymbol{E}(\omega) \mathrm{e}^{-\mathrm{i}\omega t'} \\ &= f_0(\boldsymbol{k}) - \mathrm{e}^{-\mathrm{i}\omega t} \int\limits_{-\infty}^{t} \mathrm{d}t' \mathrm{e}^{-(t-t')[1/\tau(\varepsilon_{\boldsymbol{k}}) - \mathrm{i}\omega]} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) e \boldsymbol{v}_{\boldsymbol{k}}(t') \cdot \boldsymbol{E}(\omega) \\ &= f_0(\boldsymbol{k}) - \mathrm{e}^{-\mathrm{i}\omega t} \frac{1}{1/\tau(\varepsilon_{\boldsymbol{k}}) - \mathrm{i}\omega} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) e \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{E}(\omega) . \end{aligned} \tag{24.3.43}$$

Expressing the current density through this formula, the frequency-dependent (AC) conductivity is

$$\sigma_{\alpha\beta}(\omega) = e^2 \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3} v^{\alpha}_{\boldsymbol{k}} v^{\beta}_{\boldsymbol{k}} \frac{1}{1/\tau(\varepsilon_{\boldsymbol{k}}) - \mathrm{i}\omega} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) . \tag{24.3.44}$$

If the relaxation time is independent of the energy, and the system is isotropic or exhibits cubic symmetry, we have

$$\sigma(\omega) = \frac{\sigma_0}{1 - \mathrm{i}\omega\tau}\,, \tag{24.3.45}$$

just like in the Drude model, where

$$\sigma_0 = e^2 \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3} \tfrac{1}{3} v_{\boldsymbol{k}}^2 \tau \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right). \tag{24.3.46}$$

This means that when the periodic potential of the lattice is taken into account, then the same results are obtained for the optical properties of metals and the role of the plasma frequency in the relaxation-time approximation as in the classical model of free electrons.

When σ_0 is written in the Drude form, the mass that appears in the denominator is neither the electron mass nor the dynamical effective mass of Bloch electrons but a new parameter called the *optical mass*:

$$\sigma_0 = n_{\mathrm{e}} e^2 \tau / m_{\mathrm{opt}}\,, \tag{24.3.47}$$

where the inverse of the optical mass is given by

$$m_{\mathrm{opt}}^{-1} = \frac{1}{3n_{\mathrm{e}}} \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3} v_{\boldsymbol{k}}^2 \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right). \tag{24.3.48}$$

Following the same steps as for the density of states, we now decompose the $\boldsymbol{k}$-space integral into integrals over the constant-energy surfaces and the component of $\boldsymbol{k}$ that is perpendicular to them,

$$\int \mathrm{d}\boldsymbol{k} = \int \mathrm{d}\varepsilon \int \frac{\mathrm{d}S}{\hbar v_{\boldsymbol{k}}}\,, \tag{24.3.49}$$

which gives

$$m_{\mathrm{opt}}^{-1} = \frac{1}{12\pi^3 \hbar n_{\mathrm{e}}} \int\limits_{S(\varepsilon_{\mathrm{F}})} v_{\boldsymbol{k}}\, \mathrm{d}S\,, \tag{24.3.50}$$

where the integral is over the Fermi surface. For free electrons m_{e} is recovered, but in general m_{opt} is different from the dynamical effective mass.

24.3.5 General Form of Transport Coefficients

In the foregoing the distribution function was determined in two different ways for uniform driving forces and zero magnetic field. It was also mentioned that the results given in (24.3.17) and (24.3.33) are formally identical to the free-electron formula (16.3.27). The only differences are that $\varepsilon_{\boldsymbol{k}}$ is now taken from band-structure calculations (instead of the simple isotropic and quadratic dispersion relation for free electrons), and the electron velocity is the group velocity derived from the dispersion relation. Moreover, the relaxation time may depend on $\varepsilon_{\boldsymbol{k}}$.

Using this distribution function in the formulas for the electric and heat currents, the kinetic coefficients defined by the relations

$$\begin{aligned} \boldsymbol{j} &= L_{11}\left(\boldsymbol{E}+\frac{\boldsymbol{\nabla}\mu}{e}\right)+L_{12}\left(-\frac{\boldsymbol{\nabla}T}{T}\right), \\ \boldsymbol{j}_Q &= L_{21}\left(\boldsymbol{E}+\frac{\boldsymbol{\nabla}\mu}{e}\right)+L_{22}\left(-\frac{\boldsymbol{\nabla}T}{T}\right) \end{aligned} \tag{24.3.51}$$

are tensor quantities that exhibit the symmetries of the crystal itself:

$$\begin{aligned} L_{11}^{\alpha\beta} &= e^2\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\tau(\varepsilon_{\boldsymbol{k}})v_\alpha(\boldsymbol{k})v_\beta(\boldsymbol{k}), \\ L_{12}^{\alpha\beta} = L_{21}^{\alpha\beta} &= -e\int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\tau(\varepsilon_{\boldsymbol{k}})v_\alpha(\boldsymbol{k})v_\beta(\boldsymbol{k})(\varepsilon_{\boldsymbol{k}}-\mu), \\ L_{22}^{\alpha\beta} &= \int\frac{\mathrm{d}\boldsymbol{k}}{4\pi^3}\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\tau(\varepsilon_{\boldsymbol{k}})v_\alpha(\boldsymbol{k})v_\beta(\boldsymbol{k})(\varepsilon_{\boldsymbol{k}}-\mu)^2. \end{aligned} \tag{24.3.52}$$

The $\boldsymbol{k}$-space integral can be decomposed into an integral over the constant-energy surface and an energy integral. By introducing

$$\sigma_{\alpha\beta}(\varepsilon)=\frac{e^2\tau(\varepsilon)}{4\pi^3}\int\limits_{\varepsilon_{\boldsymbol{k}}=\varepsilon}\frac{\mathrm{d}S}{|\boldsymbol{\nabla}\varepsilon_{\boldsymbol{k}}|}v_\alpha(\boldsymbol{k})v_\beta(\boldsymbol{k}), \tag{24.3.53}$$

the kinetic coefficients can be written as

$$\begin{aligned} L_{11}^{\alpha\beta} &= \int\mathrm{d}\varepsilon\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)\sigma_{\alpha\beta}(\varepsilon), \\ L_{12}^{\alpha\beta} = L_{21}^{\alpha\beta} &= -\frac{1}{e}\int\mathrm{d}\varepsilon\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)(\varepsilon-\mu)\sigma_{\alpha\beta}(\varepsilon), \\ L_{22}^{\alpha\beta} &= \frac{1}{e^2}\int\mathrm{d}\varepsilon\left(-\frac{\partial f_0}{\partial\varepsilon_{\boldsymbol{k}}}\right)(\varepsilon-\mu)^2\sigma_{\alpha\beta}(\varepsilon). \end{aligned} \tag{24.3.54}$$

Assuming that close to the Fermi surface $\sigma_{\alpha\beta}(\varepsilon)$ varies slowly, the Sommerfeld expansion can be applied. Keeping only the leading term, the value of $\sigma_{\alpha\beta}(\varepsilon)$ at the Fermi energy gives directly the conductivity tensor.

Applying the Sommerfeld expansion to L_{22}, too, we have

$$L_{22}^{\alpha\beta}=\frac{\pi^2}{3e^2}(k_{\mathrm{B}}T)^2\sigma_{\alpha\beta}(\varepsilon_{\mathrm{F}}). \tag{24.3.55}$$

Since the leading term of the thermal conductivity is $\lambda = L_{22}/T$, the Wiedemann–Franz law, which was derived for free electrons in Section 16.1.3, applies also to Bloch electrons in the semiclassical approximation.

24.3.6 Hall Effect

If the effect of the magnetic field on electrons were taken into account in (24.3.34) by replacing $\boldsymbol{E}$ by $\boldsymbol{E} + \boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B}$, as in the Lorentz force, nothing would change, since this combination is multiplied by $\boldsymbol{v}_{\boldsymbol{k}}$. However, it is well known from classical physics that crossed electric and magnetic fields give rise to the Hall effect. To derive it in our present framework, we must either solve the Boltzmann equation more accurately, or start with Chambers' formula of the distribution function, exploiting that electrons in a magnetic field move in cyclotron orbits semiclassically. We shall see both approaches below.

We start with the Boltzmann equation (24.2.21), and apply it to the case of crossed uniform electric and magnetic fields. If the collision term is treated in the relaxation-time approximation, we have

$$e\boldsymbol{E} \cdot \boldsymbol{v}_{\boldsymbol{k}} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) = -\frac{f_1}{\tau} + \frac{e}{\hbar} \left(\boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B} \right) \cdot \frac{\partial f_1}{\partial \boldsymbol{k}} \,. \tag{24.3.56}$$

The solution can be sought in the form

$$f_1 \equiv f - f_0 = -e \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) \tau \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{A} \,, \tag{24.3.57}$$

where $\boldsymbol{A}$ is yet to be determined. Then, following the steps of Section 24.3.3, the particularly simple relation

$$\boldsymbol{j} = \sigma_0 \boldsymbol{A} \tag{24.3.58}$$

is obtained in the isotropic case, where σ_0 is the usual Drude conductivity.

Substituting (24.3.57) into the right-hand side of (24.3.56), the derivative of f_1 with respect to $\boldsymbol{k}$ can be calculated using the relation $\boldsymbol{v}_{\boldsymbol{k}} = \hbar \boldsymbol{k}/m^*$, which is valid in the effective-mass approximation. We then have

$$e\boldsymbol{E} \cdot \boldsymbol{v}_{\boldsymbol{k}} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) = e \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{A} - \frac{e^2 \tau}{m^*} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) \left(\boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B} \right) \cdot \boldsymbol{A} \,, \tag{24.3.59}$$

which implies

$$\boldsymbol{E} \cdot \boldsymbol{v}_{\boldsymbol{k}} = \boldsymbol{v}_{\boldsymbol{k}} \cdot \boldsymbol{A} - \frac{e\tau}{m^*} \left(\boldsymbol{v}_{\boldsymbol{k}} \times \boldsymbol{B} \right) \cdot \boldsymbol{A} \,. \tag{24.3.60}$$

By rearranging the scalar triple product in the second term of the right-hand side, it is readily seen that the equation can be satisfied for any $\boldsymbol{v}_{\boldsymbol{k}}$ only when

$$\boldsymbol{E} = \boldsymbol{A} - \frac{e\tau}{m^*} \left(\boldsymbol{B} \times \boldsymbol{A} \right) . \tag{24.3.61}$$

Whatever the relative orientation of the electric and magnetic fields, this equation can be solved. The components of $\boldsymbol{E}$, $\boldsymbol{A}$, and $\boldsymbol{j}$ that are parallel to the magnetic field satisfy

$$\boldsymbol{E}_{\parallel} = \boldsymbol{A}_{\parallel} \,, \tag{24.3.62}$$

and thus

$$\boldsymbol{j}_\parallel = \sigma_0 \boldsymbol{E}_\parallel \,. \tag{24.3.63}$$

The magnetic field does not affect the parallel component of the current. The resistivity is the same as in the absence of the magnetic field. There is no longitudinal magnetoresistance in isotropic metals with a spherical Fermi surface.

As for the components that are perpendicular to the magnetic field: according to (24.3.61), $\boldsymbol{A}_\perp$ is also perpendicular to the magnetic field, and is thus in the plane spanned by the vectors $\boldsymbol{E}_\perp$ and $\boldsymbol{B} \times \boldsymbol{E}_\perp$. By seeking it in the form $\boldsymbol{A}_\perp = a\boldsymbol{E}_\perp + b(\boldsymbol{B} \times \boldsymbol{E}_\perp)$, and substituting that into (24.3.61), it follows immediately that

$$\boldsymbol{A}_\perp = \frac{\boldsymbol{E}_\perp + \dfrac{e\tau}{m^*}\boldsymbol{B} \times \boldsymbol{E}_\perp}{1 + \left(\dfrac{e\tau}{m^*}\right)^2 \boldsymbol{B}^2} \,. \tag{24.3.64}$$

Substituting this into (24.3.58),

$$\boldsymbol{j}_\perp = \sigma_0 \frac{\boldsymbol{E}_\perp + \dfrac{e\tau}{m^*}\boldsymbol{B} \times \boldsymbol{E}_\perp}{1 + \left(\dfrac{e\tau}{m^*}\right)^2 \boldsymbol{B}^2} \tag{24.3.65}$$

is obtained. Thus in crossed electric and magnetic fields the current flow is not parallel to the electric field: a component that is perpendicular to both the electric and magnetic fields is also present.

This can be cast in another form by eliminating $\boldsymbol{A}$ from (24.3.61):

$$\boldsymbol{E} = \frac{1}{\sigma_0}\boldsymbol{j} - \frac{e\tau}{m^*}\boldsymbol{B} \times \frac{1}{\sigma_0}\boldsymbol{j} = \varrho_0 \boldsymbol{j} - \frac{e\tau}{m^*}\varrho_0 \boldsymbol{B} \times \boldsymbol{j} \,. \tag{24.3.66}$$

When the applied magnetic field is perpendicular to the current, the electric field acquires a new component that is perpendicular both to the current and the magnetic field:

$$E_{\mathrm{H}} = -\frac{e\tau}{m^*}\varrho_0 B j \,. \tag{24.3.67}$$

Choosing the z-axis along the magnetic, and the y-axis along the electric field, the Hall coefficient is defined by

$$R_{\mathrm{H}} = \frac{E_y}{j_x B} = \frac{\varrho_{yx}(B)}{B} \,. \tag{24.3.68}$$

Note that for nonspherical Fermi surfaces $\varrho_{yx}(B = 0)$ can be finite. The proper definition of the Hall coefficient is then

$$R_{\mathrm{H}} = \frac{1}{2B}\big[\varrho_{yx}(B) - \varrho_{yx}(-B)\big]. \tag{24.3.69}$$

In our case the Hall coefficient is

$$R_{\mathrm{H}} = -\frac{e\tau}{m^*}\varrho_0 = -\frac{e\tau}{m^*}\frac{m^*}{n_{\mathrm{e}}e^2\tau} = -\frac{1}{n_{\mathrm{e}}e}\,. \tag{24.3.70}$$

The calculation is similar for hole conduction. In line with our previous observation that holes behave as positively charged particles,

$$R_{\mathrm{H}} = \frac{1}{n_{\mathrm{h}}e} \tag{24.3.71}$$

is obtained for the Hall coefficient. This explains why the Hall coefficient measured in experiments is not always negative – even though the Sommerfeld model would imply that.

24.3.7 Alternative Treatment of Transport in Magnetic Fields

In the previous subsection we examined how an applied magnetic field affected transport properties for systems with a spherical Fermi surface. In the more general case the transport coefficients can be determined using Chambers' method. Before turning to this discussion, we shall demonstrate how the previously derived results can be recovered in Chambers' approach.

According to (24.3.32), when both electric and magnetic fields are present, and the temperature is uniform (thus $\boldsymbol{\nabla}\mu = 0$), we have

$$f(t) = f_0(t) - e\left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}}\right)\tau(\varepsilon_{\boldsymbol{k}})\boldsymbol{w}_{\boldsymbol{k}}(t)\cdot\boldsymbol{E}\,, \tag{24.3.72}$$

where $\boldsymbol{w}_{\boldsymbol{k}}(t)$ can be calculated from the velocity of electrons by integrating over the trajectory:

$$\boldsymbol{w}_{\boldsymbol{k}}(t) = \frac{1}{\tau(\varepsilon_{\boldsymbol{k}})}\int\limits_{-\infty}^{t} \mathrm{d}t'\mathrm{e}^{-(t-t')/\tau(\varepsilon_{\boldsymbol{k}})}\boldsymbol{v}_{\boldsymbol{k}}(t')\,. \tag{24.3.73}$$

When the Fermi surface is spherical, and a z-directed magnetic field is applied, the electron moves in a helical path, and its projection on the xy-plane is a circular motion of angular frequency ω_{c}. If the velocity is given by $v_{\boldsymbol{k}}^x = v_{\boldsymbol{k}\perp}\cos\phi$, $v_{\boldsymbol{k}}^y = v_{\boldsymbol{k}\perp}\sin\phi$, $v_{\boldsymbol{k}}^z = v_{\boldsymbol{k}\|}$ at $t=0$, then at an arbitrary time t it is

$$\begin{aligned}\boldsymbol{v}_{\boldsymbol{k}}(t) &= \left[v_{\boldsymbol{k}\perp}\cos(\omega_{\mathrm{c}}t+\phi),\, v_{\boldsymbol{k}\perp}\sin(\omega_{\mathrm{c}}t+\phi),\, v_{\boldsymbol{k}\|}\right] \qquad (24.3.74)\\ &= \left[v_{\boldsymbol{k}}^x\cos\omega_{\mathrm{c}}t - v_{\boldsymbol{k}}^y\sin\omega_{\mathrm{c}}t,\, v_{\boldsymbol{k}}^x\sin\omega_{\mathrm{c}}t + v_{\boldsymbol{k}}^y\cos\omega_{\mathrm{c}}t,\, v_{\boldsymbol{k}\|}\right].\end{aligned}$$

Substituting this back into the above formula for $\boldsymbol{w}_{\boldsymbol{k}}$, and making use of the relations

$$\frac{1}{\tau}\int\limits_{-\infty}^{0} \mathrm{d}t'\, \mathrm{e}^{t'/\tau} \left\{ \begin{matrix} \cos\omega_\mathrm{c} t' \\ \sin\omega_\mathrm{c} t' \end{matrix} \right\} = \frac{1}{1+(\omega_\mathrm{c}\tau)^2} \left\{ \begin{matrix} 1 \\ -\omega_\mathrm{c}\tau \end{matrix} \right\}, \tag{24.3.75}$$

we have

$$\boldsymbol{w}_{\boldsymbol{k}} = \left\{ \frac{v_{\boldsymbol{k}}^x}{1+(\omega_\mathrm{c}\tau)^2} + \frac{v_{\boldsymbol{k}}^y \omega_\mathrm{c}\tau}{1+(\omega_\mathrm{c}\tau)^2},\ \frac{v_{\boldsymbol{k}}^y}{1+(\omega_\mathrm{c}\tau)^2} - \frac{v_{\boldsymbol{k}}^x \omega_\mathrm{c}\tau}{1+(\omega_\mathrm{c}\tau)^2},\ v_{\boldsymbol{k}\|}(\boldsymbol{k}) \right\}. \tag{24.3.76}$$

Using this formula in the distribution function to determine the current, the conductivity tensor is

$$\sigma_{\alpha\beta}(B) = \sigma_0 \begin{pmatrix} \dfrac{1}{1+(\omega_\mathrm{c}\tau)^2} & \dfrac{-\omega_\mathrm{c}\tau}{1+(\omega_\mathrm{c}\tau)^2} & 0 \\ \dfrac{\omega_\mathrm{c}\tau}{1+(\omega_\mathrm{c}\tau)^2} & \dfrac{1}{1+(\omega_\mathrm{c}\tau)^2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{24.3.77}$$

where $\sigma_0 = n_\mathrm{e} e^2 \tau / m^*$. By inverting it,

$$\varrho_{\alpha\beta}(B) = \varrho_0 \begin{pmatrix} 1 & \omega_\mathrm{c}\tau & 0 \\ -\omega_\mathrm{c}\tau & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{24.3.78}$$

is obtained for the resistivity tensor. The off-diagonal elements give the Hall coefficient, and the previous result is duly recovered. The diagonal terms are, on the other hand, independent of the magnetic field. This means that the current along the electric field is the same whether or not a magnetic field is present.

24.3.8 Magnetoresistance

The previous results are valid only for spherical Fermi surfaces. For nonspherical Fermi surfaces the conductivity can substantially change on the application of a magnetic field, since electrons no longer traverse circular orbits in $\boldsymbol{k}$-space – and, consequently, in real space, either. In many cases their orbits are not even closed.

When the inverse of the conductivity tensor, the resistivity tensor

$$\varrho_{\alpha\beta}(\boldsymbol{B}) = \left[\sigma^{-1}(\boldsymbol{B})\right]_{\alpha\beta} \tag{24.3.79}$$

is examined in the presence of a z-directed magnetic field, $\varrho_{xx}(B)$ and $\varrho_{yy}(B)$ specify the transverse, and $\varrho_{zz}(B)$ the longitudinal magnetoresistance.

Suppose that electrons move in a closed orbit in a section of a nonspherical Fermi surface. To obtain generally valid results in strong fields, symmetry considerations need to be applied. When the field is so strong that the period of the cyclotron orbit is much shorter than the mean time between collisions, the exponential factor in the integral for $\boldsymbol{w_k}$ varies little over a period T_c, and so

$$\boldsymbol{w_k} \propto \int\limits_0^{T_\mathrm{c}} \mathrm{d}t\, \boldsymbol{v_k}(t)\,. \tag{24.3.80}$$

The integral over the closed orbit vanishes for the velocity components perpendicular to the magnetic field; only the integral of the field-directed z component survives. Thus in the $B \to \infty$ limit σ_{zz} is the single nonvanishing component of the conductivity tensor.

In large but finite magnetic fields the magnetic-field-dependent corrections can be expanded into a power series of $1/B$,

$$\sigma_{\alpha\beta}(B) = A_{\alpha\beta} + \frac{1}{B}B_{\alpha\beta} + \frac{1}{B^2}C_{\alpha\beta} + \dots\,. \tag{24.3.81}$$

Because of the Onsager relation, the condition

$$\sigma_{\alpha\beta}(B) = \sigma_{\beta\alpha}(-B) \tag{24.3.82}$$

must be satisfied. Applying it to the series above, we have

$$A_{\alpha\beta} + \frac{1}{B}B_{\alpha\beta} + \frac{1}{B^2}C_{\alpha\beta} + \dots = A_{\beta\alpha} - \frac{1}{B}B_{\beta\alpha} + \frac{1}{B^2}C_{\beta\alpha} + \dots\,. \tag{24.3.83}$$

Comparison of the terms of the same order shows that each diagonal element must be either a constant or proportional to $1/B^2$. If the Fermi surface is closed, then, according to our previous considerations, only the zz component can be a constant. On the other hand, the off-diagonal elements can be of order $1/B$ or smaller. Therefore,

$$\sigma_{\alpha\beta}(B) = \begin{pmatrix} \dfrac{C_{xx}}{B^2} & \dfrac{B_{xy}}{B} & \dfrac{B_{xz}}{B} \\ -\dfrac{B_{xy}}{B} & \dfrac{C_{yy}}{B^2} & \dfrac{B_{yz}}{B} \\ -\dfrac{B_{xz}}{B} & -\dfrac{B_{yz}}{B} & A_{zz} \end{pmatrix}. \tag{24.3.84}$$

By inverting this matrix,

$$\varrho_{xx} \sim \frac{C_{yy}A_{zz} + B_{yz}^2}{A_{zz}B_{xy}^2} \tag{24.3.85}$$

in leading order. A similar expression holds for ϱ_{yy}. They show that the transverse magnetoresistance tends to a finite value.

The situation is radically different when the electrons move in open trajectories. If the $\boldsymbol{k}$-space orbit does not close in the k_y-direction, then the x

component of the velocity does not average out to zero in the high-field limit, and only those components of $\sigma_{\alpha\beta}(B)$ vanish in this limit for which either α or β is equal to y. Thus we have

$$\sigma_{\alpha\beta}(B) = \begin{pmatrix} A_{xx} & \dfrac{B_{xy}}{B} & A_{xz} \\ -\dfrac{B_{xy}}{B} & \dfrac{C_{yy}}{B^2} & \dfrac{B_{yz}}{B} \\ -A_{xz} & -\dfrac{B_{yz}}{B} & A_{zz} \end{pmatrix} . \tag{24.3.86}$$

Once again, ϱ_{xx} becomes saturated – but ϱ_{yy} does not: $\varrho_{yy} \sim B^2$.

By rotating the sample relative to the magnetic field direction, and measuring the variation of the resistivity, one can infer the directions along which the electron orbits are closed and open. This provides information about the topology of the cross sections of the Fermi surface. Figure 24.4(a) shows the angular dependence of the resistivity of β-(BEDT-TTF)$_2$I$_3$ when the magnetic field direction is rotated in the plane perpendicular to a plane of high conductivity. Part (b) shows the calculated angular dependence of ρ_{zz} for the highly anisotropic Fermi surface illustrated in the top left part.

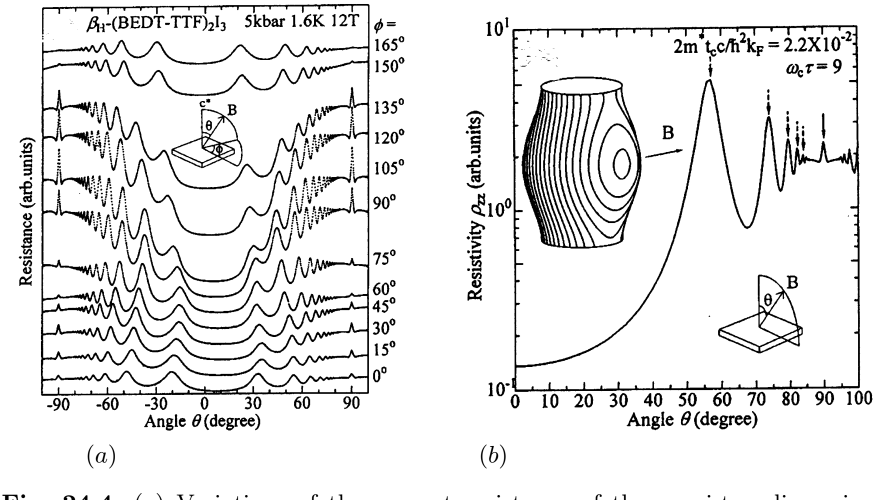


Fig. 24.4. (a) Variations of the magnetoresistance of the quasi-two-dimensional β-(BEDT-TTF)$_2$I$_3$ as the magnetic field is rotated in a plane perpendicular to the conducting plane. (b) Electron orbits on the quasi-two-dimensional Fermi surface in a magnetic field and the calculated angular dependence of the resistance ϱ_{zz} [Reprinted with permission from N. Hanasaki et al., *Phys. Rev. B* **57**, 1336 (1998). ©1998 by the American Physical Society]

24.4 Transport Coefficients in Metals and Semiconductors

In the previous sections we ignored the microscopic origin of scattering processes, and studied transport phenomena in the relaxation-time approximation. We shall now examine how individual scattering processes – such as scattering by impurities and crystal defects, interaction with lattice vibrations (phonons), and electron–electron interactions – modify the transport coefficients. We shall also discuss the applicability of the relaxation-time approximation.

Studying individual scattering processes separately is justified if they do not interfere. *Matthiessen's rule*[3] is the formulation of the empirical observation that the resistivities due to different scattering processes – such as ϱ_{imp}, due to impurity scattering, and $\varrho_{\text{el-ph}}$, due to the electron–phonon interaction – add up:

$$\varrho = \varrho_{\text{imp}} + \varrho_{\text{el-ph}} \,. \tag{24.4.1}$$

For the relaxation times this implies reciprocal additivity:

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{imp}}} + \frac{1}{\tau_{\text{el-ph}}} \,. \tag{24.4.2}$$

Even though deviations from Matthiessen's rule are not rare, assuming its validity allows us to separate experimental results into the contributions of individual scattering processes.

24.4.1 Scattering of Electrons by Impurities

Assuming that the sample contains a small number (n_{i}) of randomly distributed immobile impurities, the quantum mechanical transition probability from state $\boldsymbol{k}$ to state $\boldsymbol{k}'$ is given by the Fermi golden rule:

$$W_{\boldsymbol{k}\boldsymbol{k}'} = \frac{2\pi}{\hbar} n_{\text{i}} \delta(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}'}) \left| \langle \boldsymbol{k}' | U | \boldsymbol{k} \rangle \right|^2 , \tag{24.4.3}$$

where U is not simply the bare Coulomb potential of impurity ion: the screening effect of the other electrons are also incorporated in it. We shall discuss this in more detail at the end of the section. The matrix element is calculated between Bloch states:

$$\langle \boldsymbol{k}' | U | \boldsymbol{k} \rangle = \int \mathrm{d}\boldsymbol{r} \psi^*_{n\boldsymbol{k}'}(\boldsymbol{r}) U(\boldsymbol{r}) \psi_{n\boldsymbol{k}}(\boldsymbol{r}) \,. \tag{24.4.4}$$

Because of the random distribution of the impurities, the electrons scattered incoherently by different impurities do not interfere. That is why the transition probability is proportional to the number of scatterers.

[3] A. Matthiessen, 1864.

Since scattering by rigid impurities is elastic, the approach of Section 24.2.1 for the treatment of elastic scattering can be applied, and a relaxation time can be introduced through (24.3.13). This expression is further simplified for isotropic Fermi surfaces. Since the velocity $\boldsymbol{v}_{\boldsymbol{k}}$ is then parallel to $\boldsymbol{k}$, it is natural to assume that $\chi(\boldsymbol{k})$, the quantity characterizing the departure from equilibrium – which must be proportional to the electric field – can be written as

$$\chi(\boldsymbol{k}) = a(\varepsilon_{\boldsymbol{k}})\boldsymbol{k} \cdot \boldsymbol{E} . \tag{24.4.5}$$

Substituting this into the formula for the relaxation time, and making use of the elastic character of the scattering, we have

$$\frac{1}{\tau(\boldsymbol{k})} = \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} \left(1 - \frac{\boldsymbol{E} \cdot \boldsymbol{k}'}{\boldsymbol{E} \cdot \boldsymbol{k}}\right) . \tag{24.4.6}$$

Let us now separate $\boldsymbol{k}'$ into components that are parallel and perpendicular to $\boldsymbol{k}$. As the two vectors are of the same length,

$$\boldsymbol{k}' = \boldsymbol{k}'_{\perp} + (\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{k}}')\boldsymbol{k} = \boldsymbol{k}'_{\perp} + \cos\theta\, \boldsymbol{k} , \tag{24.4.7}$$

where $\hat{\boldsymbol{k}}$ ($\hat{\boldsymbol{k}}'$) is the unit vector in the direction $\boldsymbol{k}$ ($\boldsymbol{k}'$), and θ is the angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$. Since only those scattering processes for which both $\boldsymbol{k}$ and $\boldsymbol{k}'$ are essentially on the Fermi surface give important contributions, $W_{\boldsymbol{k}\boldsymbol{k}'}$ depends only on θ, and is independent of the perpendicular component. The part of $(1 - \boldsymbol{E} \cdot \boldsymbol{k}'/\boldsymbol{E} \cdot \boldsymbol{k})$ that contains $\boldsymbol{k}'_{\perp}$ gives vanishing contribution after integration over the azimuthal angle φ, and so the relaxation time is given by

$$\frac{1}{\tau(\boldsymbol{k})} = \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} (1 - \cos\theta) . \tag{24.4.8}$$

As has already been mentioned, the relaxation time used for the calculation of the electrical conductivity is not the same as the lifetime of an electron of wave vector $\boldsymbol{k}$ obtained using simple quantum mechanical considerations, which is

$$\frac{1}{\tau(\boldsymbol{k})} = \int \frac{\mathrm{d}\boldsymbol{k}'}{(2\pi)^3} W_{\boldsymbol{k}\boldsymbol{k}'} . \tag{24.4.9}$$

In this formula each process that scatters the electron of wave vector $\boldsymbol{k}$ into another state is weighted by its proper transition probability. However, when the resistivity – i.e., the decay of the current – is considered, other factors have to be taken into account as well. Those processes in which the wave vector changes little hardly reduce the current. The effect is the strongest when the electron of wave vector $\boldsymbol{k}$ is scattered into the state $-\boldsymbol{k}$, that is, in backscattering. The factor $1 - \cos\theta$ in the transport relaxation time comes precisely from the different weights of forward and backward scattering.

Since the relaxation time was found to be temperature independent, impurity scattering gives a temperature-independent contribution to the resistivity.

Figure 24.5 shows the low-temperature resistivity for two copper samples of different degrees of purity. At very low temperatures all other scattering processes are frozen out, and impurities cause a finite *residual resistivity*, which depends sensitively on the purity of the sample. The ratio of this quantity to the room-temperature resistivity can be used to characterize the purity of the sample.

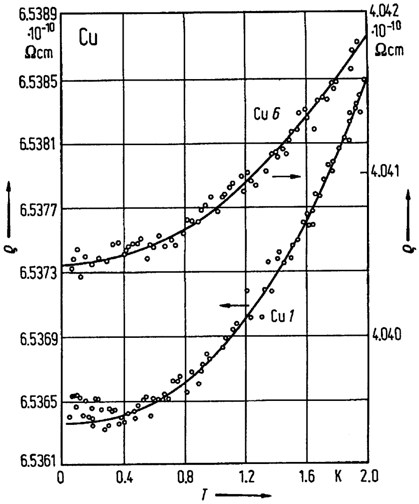


Fig. 24.5. Low-temperature resistivity in two copper samples of different purity [Reprinted with permission from M. Khoshenevisan et al., *Phys. Rev. B* **19**, 3873 (1979). ©1979 by the American Physical Society]

To evaluate the order of magnitude of this resistivity contribution, assumptions have to be made about the strength of the interaction. The simplest option is to assume that the impurity ion has a certain charge, and conduction electrons are scattered by this long-range potential via the Coulomb interaction. We shall see later that far from the impurity conduction electrons screen the charge of the impurity ion, therefore instead of the long-range Coulomb interaction electrons feel a short-range scattering potential. The results obtained in this modified picture for the transport relaxation time and resistivity due to impurities are in agreement with measurement data.

24.4.2 Contribution of Electron–Phonon Scattering to the Resistivity

To determine the contribution of the electron–phonon interaction to the resistivity, the transport equation needs to be solved simultaneously for the electron and phonon systems. In the equation for the electrons the form (24.2.38) of the collision integral has to be used. By making the approximation that, for the purpose of determining the electron distribution, the system of phonons is in thermal equilibrium ($\phi_\lambda(\boldsymbol{q})=0$), the collision term takes the form

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = -\frac{1}{V}\sum_{\boldsymbol{q}\lambda} & I_{\boldsymbol{k},\boldsymbol{q},\lambda} f_0(\boldsymbol{k})\bigl[1-f_0(\boldsymbol{k}+\boldsymbol{q})\bigr]\bigl[\chi(\boldsymbol{k})-\chi(\boldsymbol{k}+\boldsymbol{q})\bigr] \\ & \times\bigl\{g_\lambda^0(\boldsymbol{q})\delta(\varepsilon_{\boldsymbol{k}}+\hbar\omega_\lambda(\boldsymbol{q})-\varepsilon_{\boldsymbol{k}+\boldsymbol{q}}) \\ & \quad +[1+g_\lambda^0(\boldsymbol{q})]\delta(\varepsilon_{\boldsymbol{k}}-\hbar\omega_\lambda(\boldsymbol{q})-\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})\bigr\}\,.\end{aligned} \tag{24.4.10}$$

Before deriving an approximately valid formula, we shall analyze two limiting cases, in which the temperature dependence of the resistivity can be simply determined from the number of phonons participating in the collisions and the conservation of energy and momentum.

At temperature T, the electrons in the region of width $k_{\text{B}}T$ around the Fermi energy can participate in collision. At temperatures higher than the characteristic Debye temperature of the phonon spectrum the energy of the phonons is smaller than the thermal energy, therefore the phonon-absorption or -emission processes are quasielastic from the viewpoint of electrons. If the phonon energy may be neglected in the energy-conservation delta function then the two terms in the previous collision integral can be combined. The arising formula is similar to the result obtained in the relaxation-time approximation, however, an additional factor $1+2g_\lambda^0(\boldsymbol{q})$ appears, indicating the presence of phonons. As mentioned before, if it were not for this factor, the relaxation time would be temperature independent. However, the occupation function of phonons is proportional to the temperature in the $T \gg \Theta_{\text{D}}$ region,

$$g_\lambda^0(\boldsymbol{q}) = \frac{1}{\exp(\beta\hbar\omega_\lambda(\boldsymbol{q}))-1} \approx \frac{k_{\text{B}}T}{\hbar\omega_\lambda(\boldsymbol{q})}\,, \tag{24.4.11}$$

and thus the inverse relaxation time and the resistivity are both proportional to it, too:

$$\varrho \sim T\,, \qquad \text{if} \qquad T \gg \Theta_{\text{D}}\,. \tag{24.4.12}$$

This result can be interpreted alternatively. The relaxation time contains a certain average of the square of the matrix element of the electron–phonon interaction. Since the matrix element is proportional to the displacement of the ions, $1/\tau$ contains the mean square displacement. It was shown in (12.3.12) that at high temperatures $\langle \boldsymbol{u}^2\rangle \sim T$, so the resistivity increases in direct proportion to the temperature.

The situation is more complicated at low temperatures. Even though scattering is quasielastic, the number of electrons and phonons participating in the scattering need to be estimated more accurately: in addition to energy and momentum conservation, the wave-number dependence of the matrix element as well as the asymmetry between the contributions of forward- and backward-scattering processes also have to be taken into account.

Assuming that a single phonon is created or annihilated in the collision,

$$\varepsilon_{\boldsymbol{k}\pm\boldsymbol{q}} = \varepsilon_{\boldsymbol{k}} \pm \hbar\omega_\lambda(\boldsymbol{q}). \tag{24.4.13}$$

If the conservation of energy were considered alone, upon the absorption of a sufficiently energetic phonon an electron sitting relatively deep below the Fermi energy could also be scattered to an empty state above the Fermi level. In reality, however, only energy transfers of order $k_\mathrm{B}T$ occur: only those phonons are likely to participate in absorption for which $\hbar\omega_\lambda(\boldsymbol{q}) \le k_\mathrm{B}T$ because only these phonons are excited thermally in sufficient numbers. This implies that only acoustic phonons are relevant, and their wave numbers must satisfy the condition $q \le k_\mathrm{B}T/\hbar c_\mathrm{s}$.

The same bound applies to phonon emission. Once a thermally excited electron has emitted a phonon and transferred a part of its energy to the lattice, it has to occupy an initially empty state. Thus both the initial and final states of the electron must be close to the Fermi energy, inside a region that is a few times $k_\mathrm{B}T$ wide. Hence the energy loss must be of the same order of magnitude or smaller. This change in the energy is small compared to the energy of the electrons, which is on the order of the Fermi energy, therefore these processes may be called quasielastic.

Now consider the collision integral for wave vectors $\boldsymbol{k}$ that are close to the Fermi surface. It follows from our previous considerations that substantial contributions come only from those scattering processes for which $\boldsymbol{k}+\boldsymbol{q}$ is inside the sphere of radius $k_\mathrm{B}T/\hbar c_\mathrm{s}$ centered at $\boldsymbol{k}$. Energy conservation imposes another constraint, so the wave vectors of phonons participating in scattering lie on a two-dimensional surface inside the sphere. Thus in the sum over the wave vector $\boldsymbol{q}$ the phase-space restriction gives a factor that is proportional to T^2.

Another factor comes from the strength of the scattering. According to (24.2.30), the transition probability $W_{\boldsymbol{k};\boldsymbol{k}',\boldsymbol{q},\lambda}$ in the collision integral contains the square of the matrix element of the electron–phonon interaction. As mentioned above, only acoustic phonons need to be taken into account at low temperatures. For acoustic phonons the matrix element of the electron–phonon interaction is proportional to $\sqrt{q}$. According to the previous dimensional consideration, the factor q coming from the square of the matrix element brings in an additional factor of T.

Altogether, this would mean that the inverse relaxation time due to the electron–phonon interaction is proportional to T^3 – which would lead to $\varrho \sim T^3$ in the $T \ll \Theta_\mathrm{D}$ region. However, just like for impurity scattering, a factor $1-\chi(\boldsymbol{k}')/\chi(\boldsymbol{k})$ appears in the collision integral for electron–phonon scattering. At low temperatures, where electron–phonon scattering is quasielastic, this factor can be rewritten as $1-\cos\theta$. Owing to this factor, forward and small-angle scattering contribute much less to the effective collision rate than large-angle and backward scattering. Since the electron has to be close to the Fermi surface both in the initial and final states, and at low temperatures the wavelength of phonons participating in these scattering processes is much smaller than the Fermi wave number, small-angle scattering dominates, for which

$$1 - \cos\theta = 2\sin^2(\theta/2) \approx \tfrac{1}{2}\big(q/k_F\big)^2 . \qquad (24.4.14)$$

This is illustrated in Fig. 24.6.

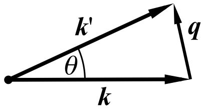


Fig. 24.6. Electron and phonon wave vectors in small-angle quasielastic scattering

According to our previous dimensional considerations, this q^2 term gives an additional factor T^2 to the collision integral, and thus

$$\varrho \sim T^5 . \qquad (24.4.15)$$

This T^5 dependence becomes modified if the Fermi surface approaches the zone boundary, and thus umklapp scattering become important.

Calculations are very tedious in the region between the two limiting cases. However, a good interpolation formula is offered by the partly empirical *Bloch–Grüneisen relation*:[4]

$$\varrho(T) = K(T/\Theta_{\mathrm{D}})^5 J_5(\Theta_{\mathrm{D}}/T), \qquad (24.4.16)$$

where K is a constant, which will not be specified here, and

$$J_5(x) = \int_0^x \frac{\mathrm{e}^\xi \xi^5 \,\mathrm{d}\xi}{(\mathrm{e}^\xi - 1)^2} = 5\int_0^x \frac{\xi^4\,\mathrm{d}\xi}{\mathrm{e}^\xi - 1} - \frac{x^5}{\mathrm{e}^x - 1} . \qquad (24.4.17)$$

This integral lends itself to simple evaluation in two limits:

$$J_5(x) = \begin{cases} \dfrac{1}{4}x^4 - \dfrac{1}{72}x^6 + \ldots & x \ll 1 , \\ 5!\displaystyle\sum_{k=1}^{\infty} \frac{1}{k^5} = 5!\zeta(5) = 124.43 & x \gg 1 . \end{cases} \qquad (24.4.18)$$

At low temperatures ($\Theta_{\mathrm{D}}/T \gg 1$), $J_5(\Theta_{\mathrm{D}}/T)$ is constant, so the resistivity is proportional to the fifth power of the temperature. Likewise, at high temperatures ($\Theta_{\mathrm{D}}/T \ll 1$), $J_5(\theta_{\mathrm{D}}/T)$ is proportional to $(1/T)^4$, so the resistivity is proportional to the temperature. In both limits the previous results are recovered, and the interpolation formula provides a good approximation for the resistivity of simple metals in the intermediate region, too. This is shown in Fig. 24.7.

At very low temperatures electron–electron scattering has to be taken into account, too. It can be demonstrated that the requirements of energy

[4] F. Bloch, 1930; E. Grüneisen, 1933.

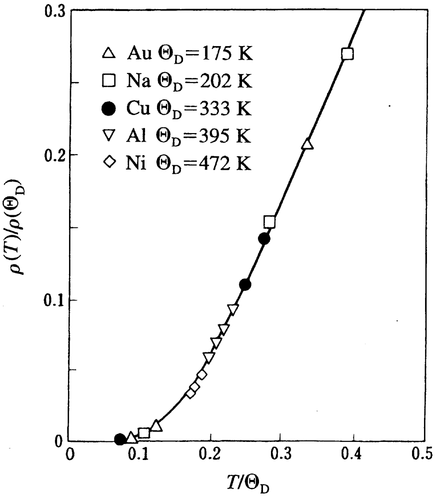


Fig. 24.7. The resistivity of some simple metals scaled by the resistivity at the Debye temperature, as a function of T/Θ_D. The solid line is the prediction of the Bloch–Grüneisen formula [F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill Book Co., New York (1968)]

and momentum conservation can be satisfied simultaneously only by umklapp processes. After some algebra, the resistivity is found to be proportional to the square of the temperature:

$$\varrho_{\text{el–el}} \sim \left(\frac{k_B T}{\varepsilon_F}\right)^2 . \tag{24.4.19}$$

In alkali metals $\tau_{\text{el–ph}}$ and $\tau_{\text{el–el}}$ are comparable at a few kelvins (around 4 K for sodium), however a relatively low concentration of impurities gives rise to a similar relaxation time. The contributions of individual processes are difficult to separate.

The situation is different in transition metals. The electric current is dominantly carried by s-electrons, nevertheless the interaction with d-electrons, whose density of states is high, can give an important contribution to the resistivity. In some transition metals (Mn, Fe, Co, Ni, Pd, Pt, W, and Nb) the temperature dependence of the resistivity below 10 K is fairly well approximated by a quadratic fit, as shown in Fig. 24.8. It should be noted that other scattering processes can also lead to such a temperature dependence in magnetic materials – but the majority of the materials listed above are not magnetic.

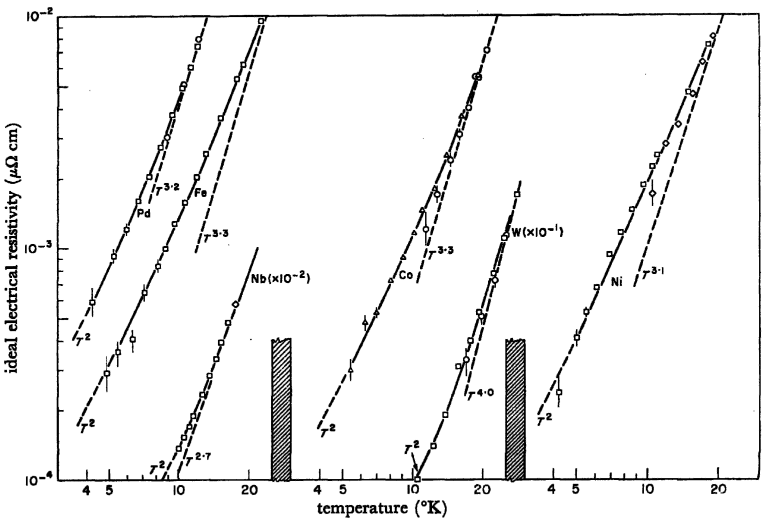


Fig. 24.8. Temperature dependence of resistivity for some transition metals (Pd, Fe, Nb, Co, W, Ni) at temperatures below 20 K [G. K. White and S. B. Woods, *Philosophical Transactions of the Royal Society*, **251** A, 273 (1959)]

24.4.3 Scattering by Magnetic Impurities and the Kondo Effect

The ubiquitous impurities in metals give a temperature-independent contribution to the resistivity, the so-called residual resistivity (page 389). However, strong temperature dependence was observed in several low-temperature experiments: the resistivity attained a minimum at a few kelvins, and then started to grow with decreasing temperature as $\log T$, see Fig. 24.9.

Since measurements indicated the possibility that the effect might be due to the presence of magnetic impurities, J. Kondo (1964) justified by theoretical calculations that scattering by such impurities indeed gives rise to an increase in the resistivity with decreasing temperature. Earlier calculations of impurity scattering were based on a spin-independent scattering potential. However, when magnetic impurities (Mn, Cr, Fe, etc.) are introduced into a nonmagnetic metal (Cu, Ag, Au, Al, etc.), spin-flip processes become possible: the spin of the scattered electron and the spin of the magnetic impurity are flipped simultaneously, while the component of the total spin along the quantization axis is conserved. This opens a new scattering channel. The interaction of the impurity with conduction electrons is described in terms of the so-called s–d interaction, in which the spin $\boldsymbol{S}_i$ of the impurity located at $\boldsymbol{R}_i$ – which is an internal degree of freedom of the impurity – interacts with the local spin density of the conduction electrons. When electrons are represented by field operators or by creation and annihilation operators of free-electron states (as the periodic potential of the lattice is neglected in the present calculation), the interaction Hamiltonian takes the form

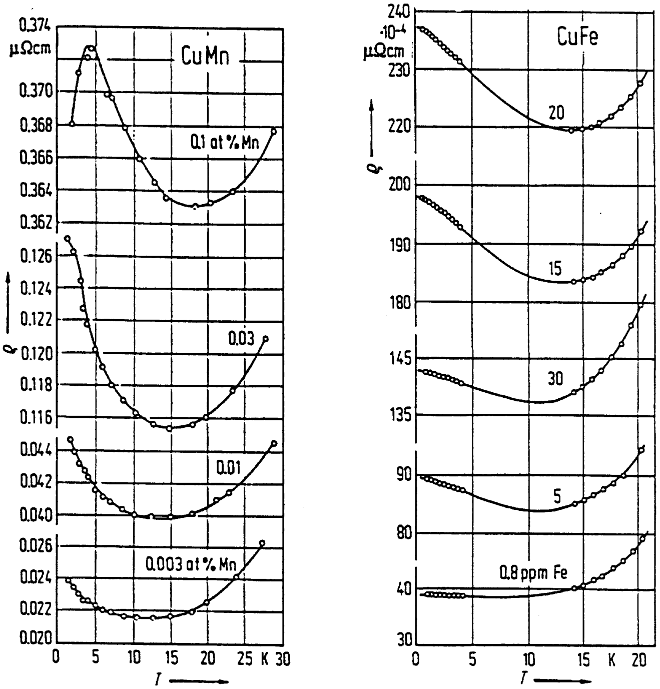


Fig. 24.9. Temperature dependence of resistivity in copper containing a small concentration of magnetic impurities (manganese and iron) [G. Grüner, *Advances in Physics* **23**, 941 (1974), and B. Knook, *Thesis*, Leiden (1962)]

$$\begin{aligned}\mathcal{H}_{\text{s-d}} &= -J\sum_{i\alpha\beta}\int \hat{\psi}^\dagger_\alpha(\boldsymbol{r})\boldsymbol{\sigma}_{\alpha\beta}\hat{\psi}_\beta(\boldsymbol{r})\cdot \boldsymbol{S}_i\delta(\boldsymbol{r}-\boldsymbol{R}_i)\,\mathrm{d}\boldsymbol{r} \\ &= -\frac{J}{V}\sum_i\sum_{\substack{\boldsymbol{k},\boldsymbol{k}'\\ \alpha\beta}} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{R}_i} c^\dagger_{\boldsymbol{k}'\alpha}\boldsymbol{\sigma}_{\alpha\beta}c_{\boldsymbol{k}\beta}\cdot\boldsymbol{S}_i\,.\end{aligned} \qquad (24.4.20)$$

Since the flip of the impurity spin is accompanied by the flip of an electron spin, separate equations have to be written down for the distribution function of spin-up and spin-down electrons. The collision term for the former is

$$\begin{aligned}\left(\frac{\partial f_\uparrow(\boldsymbol{k})}{\partial t}\right)_{\text{coll}} &= \sum_{\boldsymbol{k}'} W_{\boldsymbol{k}\uparrow,\boldsymbol{k}'\uparrow}\left\{f_\uparrow(\boldsymbol{k}')[1-f_\uparrow(\boldsymbol{k})]-f_\uparrow(\boldsymbol{k})[1-f_\uparrow(\boldsymbol{k}')]\right\} \\ &\quad+\sum_{\boldsymbol{k}'} W_{\boldsymbol{k}\uparrow,\boldsymbol{k}'\downarrow}\left\{f_\downarrow(\boldsymbol{k}')[1-f_\uparrow(\boldsymbol{k})]-f_\uparrow(\boldsymbol{k})[1-f_\downarrow(\boldsymbol{k}')]\right\}.\end{aligned} \qquad (24.4.21)$$

The transition probability W_{ab} between states a and b is given by the usual quantum mechanical formula

$$W_{ab} = \frac{2\pi}{\hbar} \left|\langle a|\mathcal{H}_{\text{s-d}}|b\rangle\right|^2 \delta(\varepsilon_a - \varepsilon_b) \,. \tag{24.4.22}$$

For spin-conserving scattering, the matrix element is

$$\langle \boldsymbol{k}' \uparrow |\mathcal{H}_{\text{s-d}}|\boldsymbol{k} \uparrow\rangle = -\frac{J}{V} S^z \,. \tag{24.4.23}$$

By determining the other matrix elements it is relatively straightforward to show that the residual resistivity is temperature independent in this case, too, just like for nonmagnetic impurities. KONDO's important observation was that, despite the weakness of the coupling, the Born approximation does not provide a satisfactory treatment of the scattering: higher-order corrections need to be taken into account as well.

A better approximation can be obtained for the transition probability by replacing the matrix element of the interaction Hamiltonian by the matrix element of the scattering matrix T in (24.4.22). When the eigenstates $|a\rangle$ and energy eigenvalues ε_a of the unperturbed system are known, the transition probability from $|a\rangle$ to $|b\rangle$ due to the interaction is

$$W_{ab} = \frac{2\pi}{\hbar} \left|\langle a|T|b\rangle\right|^2 \delta(\varepsilon_a - \varepsilon_b) \,, \tag{24.4.24}$$

where the matrix elements of T can be expressed in terms of those of the interaction Hamiltonian $\mathcal{H}_{\text{int}}$ as

$$\langle a|T|b\rangle = \langle a|\mathcal{H}_{\text{int}}|b\rangle + \sum_c \frac{\langle a|\mathcal{H}_{\text{int}}|c\rangle\langle c|\mathcal{H}_{\text{int}}|b\rangle}{\varepsilon_a - \varepsilon_c} + \dots \,. \tag{24.4.25}$$

Up to the third order in the coupling constant,

$$\begin{aligned} W_{ab} = \frac{2\pi}{\hbar}\biggl\{ &\langle a|\mathcal{H}_{\text{s-d}}|b\rangle\langle b|\mathcal{H}_{\text{s-d}}|a\rangle \qquad (24.4.26)\\ &+\sum_c \left[\frac{\langle a|\mathcal{H}_{\text{s-d}}|c\rangle\langle c|\mathcal{H}_{\text{s-d}}|b\rangle\langle b|\mathcal{H}_{\text{s-d}}|a\rangle}{\varepsilon_a - \varepsilon_c} + \text{c.c.}\right] + \dots \biggr\}\delta(\varepsilon_a - \varepsilon_b). \end{aligned}$$

First consider the scattering $\boldsymbol{k}\uparrow \to \boldsymbol{k}'\uparrow$. As shown in Fig. 24.10, two processes are possible in second order. One option is that the electron of quantum numbers $\boldsymbol{k}\uparrow$ is first scattered into an intermediate state

$$|c\rangle = |\boldsymbol{k}''\sigma\rangle = c^\dagger_{\boldsymbol{k}''\sigma}|FS\rangle \,, \tag{24.4.27}$$

and then, in a second event, the electron of quantum numbers $\boldsymbol{k}''\sigma$ is scattered to the final state $\boldsymbol{k}'\uparrow$.

The other option is that the first interaction leaves the electron of quantum numbers $\boldsymbol{k}\uparrow$ in its initial state, and the impurity spin interacts with the Fermi sea to create an electron–hole pair. The intermediate state thus contains the electron $\boldsymbol{k}\uparrow$ as well as an electron–hole pair:

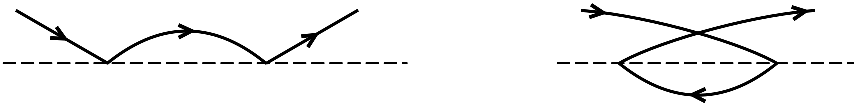


Fig. 24.10. Second-order processes of the *s*–*d* interaction. Time flows from left to right. The electron propagating backward in time corresponds to a hole

$$|c\rangle = |\boldsymbol{k}\uparrow, \boldsymbol{k}'\uparrow, \boldsymbol{k}''\sigma\rangle = c^{\dagger}_{\boldsymbol{k}\uparrow} c^{\dagger}_{\boldsymbol{k}'\uparrow} c_{\boldsymbol{k}''\sigma}|\mathrm{FS}\rangle\,. \tag{24.4.28}$$

The hole of quantum numbers $\boldsymbol{k}''\sigma$ is filled by the hitherto unperturbed incoming electron $\boldsymbol{k}\uparrow$ in the second interaction, while the electron of the created electron–hole pair survives in the final state. In the previous formulas only the states of the electrons are indicated, those of the impurity spin are not. We shall need to pay attention to this.

In order that these processes can take place, the state $\boldsymbol{k}''\sigma$ has to be initially empty in the first, and occupied in the second case. The second-order correction to the T matrix is

$$\begin{aligned} T^{(2)} = {} & \sum_{\boldsymbol{k}''\sigma} \frac{\langle \boldsymbol{k}\uparrow |\mathcal{H}_{\text{s-d}}|\boldsymbol{k}''\sigma\rangle\langle \boldsymbol{k}''\sigma|\mathcal{H}_{\text{s-d}}|\boldsymbol{k}'\uparrow\rangle}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}''}}(1 - f_0(\boldsymbol{k}'')) \\ & + \sum_{\boldsymbol{k}''\sigma} \frac{\langle \boldsymbol{k}\uparrow |\mathcal{H}_{\text{s-d}}|\boldsymbol{k}\uparrow, \boldsymbol{k}'\uparrow, \boldsymbol{k}''\sigma\rangle\langle \boldsymbol{k}\uparrow, \boldsymbol{k}'\uparrow, \boldsymbol{k}''\sigma|\mathcal{H}_{\text{s-d}}|\boldsymbol{k}'\uparrow\rangle}{\varepsilon_{\boldsymbol{k}} - (\varepsilon_{\boldsymbol{k}} + \varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}''})} f_0(\boldsymbol{k}'')\,. \end{aligned} \tag{24.4.29}$$

If the z component of the impurity spin is left unchanged by the scattering, the combined contribution of the two channels is

$$\begin{aligned} & \left(-\frac{J}{V}\right)^2 \sum_{\boldsymbol{k}''} (S^z)^2 \left(\frac{1 - f_0(\boldsymbol{k}'')}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}''}} - \frac{f_0(\boldsymbol{k}'')}{\varepsilon_{\boldsymbol{k}''} - \varepsilon_{\boldsymbol{k}'}}\right) \\ & \qquad = \left(-\frac{J}{V}\right)^2 (S^z)^2 \sum_{\boldsymbol{k}''} \frac{1}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}''}}\,. \end{aligned} \tag{24.4.30}$$

The negative sign of the second term on the left-hand side is the consequence of the Fermi–Dirac statistics. Assuming that the density of states is symmetric about the Fermi energy, the integral vanishes for the electrons on the Fermi surface, and is negligible in its vicinity. The contribution of scattering by a rigid, nonrotating spin – just like that of potential scattering – is of no interest.

On the other hand, if the spin is flipped in the intermediate state, then an electron of quantum numbers $\boldsymbol{k}''\downarrow$ appears in the first process, and the impurity spin goes over from the initial state $S^z = M$ to $S^z = M + 1$. The contribution of this process is

$$\left(-\frac{J}{V}\right)^2 \sum_{\boldsymbol{k}''} [S(S+1) - M(M+1)]\, \frac{1 - f_0(\boldsymbol{k}'')}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}''}}\,. \tag{24.4.31}$$

In the other channel, where the electron–hole pair is composed of a spin-up electron and a spin-down hole, the impurity spin has to flip down, to the state $S^z = M - 1$. The contribution of this process is therefore

$$-\left(-\frac{J}{V}\right)^2 \sum_{\boldsymbol{k}''} \left[S(S+1) - M(M-1)\right] \frac{f_0(\boldsymbol{k}'')}{\varepsilon_{\boldsymbol{k}} - (\varepsilon_{\boldsymbol{k}} + \varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}''})} . \qquad (24.4.32)$$

Since $\varepsilon_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}'}$ on account of the conservation of energy, the combined contribution is

$$\left(-\frac{J}{V}\right)^2 \sum_{\boldsymbol{k}''} \left\{ \left[S(S+1) - M^2\right] \frac{1}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}''}} - M \frac{1 - 2f_0(\boldsymbol{k}'')}{\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}''}} \right\} . \qquad (24.4.33)$$

The first term is negligible once again, however, the second term is no longer small: it even diverges logarithmically when $\varepsilon_{\boldsymbol{k}} \to \varepsilon_{\mathrm{F}}$ and $T \to 0$.

Using this form for the second-order correction to the T matrix in J, the transition probability can be determined up to third order in J. By summing over the possible spin orientations,

$$W(\boldsymbol{k} \uparrow \to \boldsymbol{k}' \uparrow) = n_{\mathrm{i}} \frac{2\pi J^2 S(S+1)}{3\hbar} \left[1 + 4Jg(\varepsilon_{\boldsymbol{k}})\right] \delta(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}'}) , \qquad (24.4.34)$$

where n_{i} is the concentration of magnetic impurities, and the singular function in the third-order correction is

$$g(\varepsilon_{\boldsymbol{k}}) = \frac{1}{V} \sum_{\boldsymbol{k}''} \frac{f_0(\boldsymbol{k}'')}{\varepsilon_{\boldsymbol{k}''} - \varepsilon_{\boldsymbol{k}}} . \qquad (24.4.35)$$

To evaluate this sum, we assume a constant density of states. This is a good approximation for electrons close to the Fermi surface, from where the singular contribution to the corresponding integral comes. This gives

$$g(\varepsilon_{\boldsymbol{k}}) \sim \begin{cases} \ln\left(k_{\mathrm{B}}T/\varepsilon_{\mathrm{F}}\right), & \text{if} \quad \varepsilon_{\boldsymbol{k}} < k_{\mathrm{B}}T , \\ \ln\left(\varepsilon_{\boldsymbol{k}}/\varepsilon_{\mathrm{F}}\right), & \text{if} \quad \varepsilon_{\boldsymbol{k}} > k_{\mathrm{B}}T . \end{cases} \qquad (24.4.36)$$

Repeating the calculation for the spin-flip scattering $\boldsymbol{k}\uparrow \to \boldsymbol{k}'\downarrow$,

$$W(\boldsymbol{k} \uparrow \to \boldsymbol{k}' \downarrow) = n_{\mathrm{i}} \frac{4\pi J^2 S(S+1)}{3\hbar} \left[1 + 4Jg(\varepsilon_{\boldsymbol{k}})\right] \delta(\varepsilon_{\boldsymbol{k}} - \varepsilon_{\boldsymbol{k}'}) . \qquad (24.4.37)$$

By substituting these formulas into the collision integral, we can define a relaxation time in the customary way. The result is

$$\frac{1}{\tau} = n_{\mathrm{i}} \frac{3\pi J^2 S(S+1) z}{2\hbar \varepsilon_{\mathrm{F}}} \left[1 + 4Jg(\varepsilon_{\boldsymbol{k}})\right] , \qquad (24.4.38)$$

where z is the number of conduction electrons per atom.

Using this energy- or temperature-dependent relaxation time in the Boltzmann equation,

$$\varrho = \varrho_0 + A \ln T \tag{24.4.39}$$

is obtained for the resistivity after some algebra. Here ϱ_0 is the resistivity of the pure sample.

Using this formula, a good fit can be obtained for the experimental results in the vicinity of the resistivity minimum. However, because of the logarithmic temperature dependence, the question immediately arises: What happens at lower temperatures? It is straightforward to show that the higher orders of the perturbation expansion give even more singular contributions, and thus perturbation theory is unable to provide a satisfactory treatment of the problem. This made the Kondo problem one of the most intensely studied problems of solid-state physics in the past decades. We shall return to it in Chapter 35 on strongly correlated electron systems.

24.4.4 Electronic Contribution to Thermal Conductivity

As was shown on page 380, the Wiedemann–Franz law, which is based on experimental findings, applies to Bloch electrons as well in the relaxation-time approximation. Therefore the electronic contribution to thermal conductivity could be determined naively but most straightforwardly using this law. Thus, the product of the thermal conductivity and the resistivity is expected to be proportional to the temperature, with a universal constant of proportionality, the Lorenz number given in (16.3.48). Using our previous results for the temperature dependence of the resistivity,

$$\varrho \sim \begin{cases} \text{const.}, & \text{if} \quad T \to 0\,, \\ T^5\,, & \text{if} \quad T < \Theta_{\mathrm{D}}\,, \\ T\,, & \text{if} \quad T > \Theta_{\mathrm{D}}\,, \end{cases} \tag{24.4.40}$$

the thermal conductivity should then behave as

$$\lambda \sim \begin{cases} T\,, & \text{if} \quad T \to 0\,, \\ T^{-4}\,, & \text{if} \quad T < \Theta_{\mathrm{D}}\,, \\ \text{const.}, & \text{if} \quad T > \Theta_{\mathrm{D}} \end{cases} \tag{24.4.41}$$

in the three temperature ranges of interest. Typical experimental results for metals are shown in Fig. 24.11.

The thermal conductivity is proportional to the temperature at low temperatures, as expected, but the constant of proportionality depends strongly on the purity of the sample. At high temperatures, too, the theoretical prediction is in agreement with the experimental findings: λ tends to a constant. However, the Wiedemann–Franz law fails in the intermediate temperature range: the thermal conductivity is proportional to $1/T^2$ rather than $1/T^4$

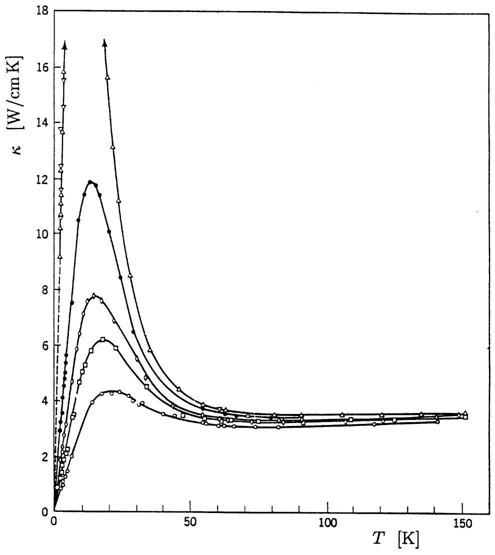


Fig. 24.11. Temperature dependence of the thermal conductivity for gold samples of different purity [J. Olsen, *Electron Transport in Metals*, Interscience Publishers, Inc., New York (1962)]

here. The experimentally observed tendency T^{-2} can be reproduced in theoretical calculations by omitting the factor $1 - \cos\theta$ in the relaxation rate. This factor played an essential role in the determination of the resistivity but seems to be unimportant in heat-conduction phenomena. This can be understood intuitively, as electron–phonon scattering cannot be considered elastic in this region, and thus the relaxation-time approximation is inapplicable. While small-angle inelastic scattering by low-momentum phonons gives a very small contribution to the decay of the electric current, the same processes contribute appreciably to the decay of the heat current, as electrons lose energy in such processes.

24.4.5 Phonon Contribution to Thermal Conductivity

Up to now we have considered only electrons as a vehicle of heat conduction. However, phonons can also contribute to thermal conductivity in solids. Note that when only normal processes are considered, in which energy and momentum are conserved simultaneously, the heat current cannot decay. If it were not for umklapp processes, solids would be perfect heat conductors.

The heat current carried by phonons can be simply written as

$$\boldsymbol{j}_Q = \sum_\lambda \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\, \hbar\omega_\lambda(\boldsymbol{q})\, \boldsymbol{v}_\lambda(\boldsymbol{q})\, g(\boldsymbol{q}) \,, \qquad (24.4.42)$$

where $\boldsymbol{v}_\lambda(\boldsymbol{q})$ is the group velocity of phonons. The Boltzmann equation governing the stationary distribution function of phonons in the presence of a temperature gradient was given in (24.2.28). Applying the relaxation-time approximation, its solution is

$$g_\lambda(\boldsymbol{q}) = g^0_\lambda(\boldsymbol{q}) - \tau(\boldsymbol{q})\frac{\partial g^0_\lambda(\boldsymbol{q})}{\partial T}\boldsymbol{v}_\lambda(\boldsymbol{q}) \cdot \boldsymbol{\nabla} T\,. \tag{24.4.43}$$

Substituting this into the current formula,

$$\boldsymbol{j}_Q = -\boldsymbol{\nabla} T \sum_\lambda \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\,\hbar\omega_\lambda(\boldsymbol{q})\,\boldsymbol{v}_\lambda(\boldsymbol{q}) \circ \boldsymbol{v}_\lambda(\boldsymbol{q})\tau(\boldsymbol{q})\frac{\partial g^0_\lambda(\boldsymbol{q})}{\partial T} \tag{24.4.44}$$

is obtained, and thus the thermal conductivity is

$$\lambda = \frac{1}{3}\sum_\lambda \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\,\hbar\omega_\lambda(\boldsymbol{q})\,v_\lambda(\boldsymbol{q})\Lambda_\lambda(\boldsymbol{q})\frac{\partial g^0_\lambda(\boldsymbol{q})}{\partial T} \tag{24.4.45}$$

in the isotropic case, where $\Lambda_\lambda(\boldsymbol{q}) = v_\lambda(\boldsymbol{q})\tau(\boldsymbol{q})$ is the mean free path. In terms of $c_\lambda(\boldsymbol{q})$, the heat capacity per mode, this can be rewritten as

$$\lambda = \frac{1}{3}\sum_\lambda \int \frac{\mathrm{d}\boldsymbol{q}}{(2\pi)^3}\,c_\lambda(\boldsymbol{q})v_\lambda(\boldsymbol{q})\Lambda_\lambda(\boldsymbol{q})\,. \tag{24.4.46}$$

Note that this is the generalization of (12.4.20),

$$\lambda = \tfrac{1}{3}c_V\bar{v}\Lambda\,, \tag{24.4.47}$$

a well-known formula of the kinetic theory of gases.

At temperatures well above the Debye temperature the mean free path of phonons is inversely proportional to the temperature, $\Lambda \sim 1/T$, and thus the thermal conductivity also decreases with increasing temperature:

$$\lambda \sim 1/T\,. \tag{24.4.48}$$

At low temperatures umklapp processes freeze out. The strength of those processes in which two phonons ($\boldsymbol{q}$ and $\boldsymbol{q}'$) merge into one ($\boldsymbol{q}''$) is proportional to the occupation of the initial states, given by

$$g_\lambda(\boldsymbol{q})g_{\lambda'}(\boldsymbol{q}') \sim \exp\left(-\frac{\hbar\omega_\lambda(\boldsymbol{q})}{k_\mathrm{B}T}\right)\,\exp\left(-\frac{\hbar\omega_{\lambda'}(\boldsymbol{q}')}{k_\mathrm{B}T}\right). \tag{24.4.49}$$

Because of the conservation of energy,

$$g_\lambda(\boldsymbol{q})g_{\lambda'}(\boldsymbol{q}') \sim \exp\left(-\frac{\hbar\omega_{\lambda''}(\boldsymbol{q}'')}{k_\mathrm{B}T}\right) \sim \exp\left(-\frac{\Theta_\mathrm{D}}{T}\right) \tag{24.4.50}$$

for phonon states close to the zone boundary at low temperatures. The temperature dependence of the thermal conductivity is then

$$\lambda \sim T^n \exp\left(\frac{\Theta_{\mathrm{D}}}{T}\right), \tag{24.4.51}$$

where the exponent n cannot be determined from our previous considerations. At low temperatures, where umklapp processes become less and less probable, the thermal conductivity would increase beyond bounds in the $T \to 0$ limit. However, when the mean free path becomes comparable to the size D of the sample, this tendency is reversed, and

$$\lambda \sim T^3 D, \tag{24.4.52}$$

where T^3 comes from the phonon specific heat. This can be readily observed in experiments. Phonons may also be scattered because of the different mass of isotopes. The corresponding, observable, contribution is

$$\lambda \sim \frac{M}{\delta M}\frac{1}{T^{1/2}}. \tag{24.4.53}$$

The left-hand side of Fig. 24.12 shows these characteristic ranges of the thermal conductivity vs. temperature graph on a typical curve, while the right-hand side is a log–log plot of the experimental results for sodium fluoride.

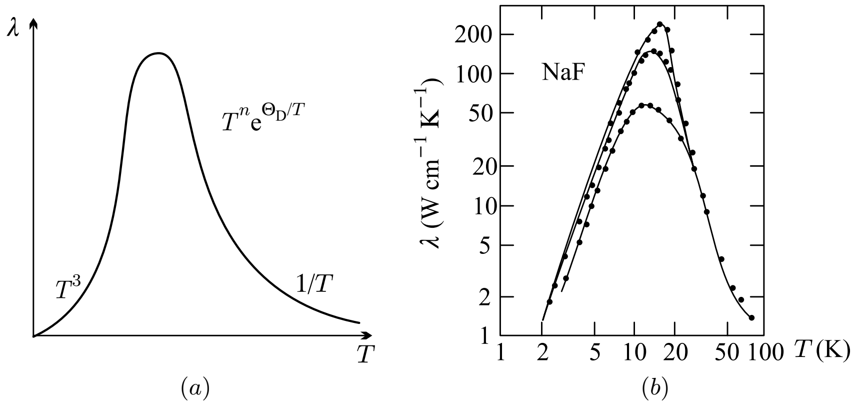


Fig. 24.12. (*a*) Typical temperature dependence of the thermal conductivity for insulators. (*b*) The measured thermal conductivity for three differently grown samples of the ionic crystal NaF [Reprinted with permission from H. E. Jackson et al., *Phys. Rev. Lett.* **25**, 26 (1970). ©1970 by the American Physical Society]

24.4.6 Transport Coefficients in Semiconductors

The formulas obtained in the relaxation-time approximation for a gas of free electrons in Chapter 16 will serve as the starting point for our study of transport phenomena in semiconductors. Expressed in terms of K_0, defined in (16.3.33), the electrical conductivity is

$$\sigma = e^2 \int \frac{\mathrm{d}\boldsymbol{k}}{4\pi^3} \left(-\frac{\partial f_0}{\partial \varepsilon_{\boldsymbol{k}}} \right) \tau(\varepsilon_{\boldsymbol{k}}) \frac{1}{3} \boldsymbol{v}_{\boldsymbol{k}}^2 \,. \tag{24.4.54}$$

There are two essential differences with metals. Firstly, the classical Maxwell-Boltzmann statistics can be applied instead of quantum statistics; secondly, the contributions of conduction-band electrons and valence-band holes have to be treated separately.

If conduction electrons at the bottom of the conduction band are treated as free particles of effective mass m_{n}^*, $\boldsymbol{v}_{\boldsymbol{k}}^2$ can be expressed in terms of the energy, and the $\boldsymbol{k}$-sum can be transformed into an energy integral using the density of states. We then have

$$\sigma = \frac{2e^2}{3m_{\mathrm{n}}^*} \int\limits_{\varepsilon_{\mathrm{c}}}^{\infty} \tau(\varepsilon) \left(-\frac{\partial f_0}{\partial \varepsilon} \right) (\varepsilon - \varepsilon_{\mathrm{c}}) \rho_{\mathrm{c}}(\varepsilon) \,\mathrm{d}\varepsilon \,. \tag{24.4.55}$$

For the density of states we shall use the formula (20.3.1),

$$\rho_{\mathrm{c}}(\varepsilon) = \frac{1}{2\pi^2} \left(\frac{2m_{\mathrm{n}}^*}{\hbar^2} \right)^{3/2} \sqrt{\varepsilon - \varepsilon_{\mathrm{c}}} \,. \tag{24.4.56}$$

The applicability of classical statistics means that f_0 can be approximated by

$$f_0(\varepsilon) = \frac{1}{\mathrm{e}^{(\varepsilon - \mu)/k_{\mathrm{B}}T} + 1} \approx \mathrm{e}^{-(\varepsilon_{\mathrm{c}} - \mu)/k_{\mathrm{B}}T} \mathrm{e}^{-(\varepsilon - \varepsilon_{\mathrm{c}})/k_{\mathrm{B}}T} \,. \tag{24.4.57}$$

According to (20.3.17) and (20.3.18), the density of excited electrons and the chemical potential are related by

$$n(T) = \frac{1}{4} \left(\frac{2m_{\mathrm{n}}^* k_{\mathrm{B}}T}{\pi\hbar^2} \right)^{3/2} \mathrm{e}^{-(\varepsilon_{\mathrm{c}} - \mu)/k_{\mathrm{B}}T} \,, \tag{24.4.58}$$

and so

$$f_0(\varepsilon) \approx 4n(T) \left(\frac{2m_{\mathrm{n}}^* k_{\mathrm{B}}T}{\pi\hbar^2} \right)^{-3/2} \mathrm{e}^{-(\varepsilon - \varepsilon_{\mathrm{c}})/k_{\mathrm{B}}T} \,. \tag{24.4.59}$$

Substituting this into the conductivity formula,

$$\sigma = e^2 \frac{4n(T)}{3m_{\mathrm{n}}^* \pi^{1/2}} (k_{\mathrm{B}}T)^{-5/2} \int\limits_{\varepsilon_{\mathrm{c}}}^{\infty} \mathrm{d}\varepsilon \mathrm{e}^{-(\varepsilon - \varepsilon_{\mathrm{c}})/k_{\mathrm{B}}T} \tau(\varepsilon) (\varepsilon - \varepsilon_{\mathrm{c}})^{3/2} \,. \tag{24.4.60}$$

By introducing the new variable $x = (\varepsilon - \varepsilon_{\mathrm{c}})/k_{\mathrm{B}}T$,

$$\sigma = e^2 \frac{4n(T)}{3m_{\mathrm{n}}^* \pi^{1/2}} \int\limits_{0}^{\infty} \tau(\varepsilon) \mathrm{e}^{-x} x^{3/2} \,\mathrm{d}x \,. \tag{24.4.61}$$

Writing the conductivity in the customary Drude form

$$\sigma = \frac{ne^2}{m_\mathrm{n}^*}\langle\tau\rangle\,, \tag{24.4.62}$$

the mean relaxation time is

$$\langle\tau\rangle = \frac{4}{3\pi^{1/2}}\int_0^\infty \tau(\varepsilon)x^{3/2}\mathrm{e}^{-x}\,\mathrm{d}x\,. \tag{24.4.63}$$

This can be rewritten as

$$\langle\tau\rangle = \frac{2}{3k_\mathrm{B}T}\frac{\displaystyle\int(\varepsilon-\varepsilon_\mathrm{c})\tau(\varepsilon)\mathrm{e}^{-(\varepsilon-\varepsilon_\mathrm{c})/k_\mathrm{B}T}\rho(\varepsilon)\,\mathrm{d}\varepsilon}{\displaystyle\int \mathrm{e}^{-(\varepsilon-\varepsilon_\mathrm{c})/k_\mathrm{B}T}\rho(\varepsilon)\,\mathrm{d}\varepsilon}\,. \tag{24.4.64}$$

The conductivity of semiconductors is customarily given in terms of the mobility μ defined by $\boldsymbol{v}_\mathrm{n} = -\mu_\mathrm{n}\boldsymbol{E}$. On account of the relationship $\boldsymbol{j}_\mathrm{n} = -en\boldsymbol{v}_\mathrm{n}$ between the current and the drift velocity,

$$\sigma = ne\mu_\mathrm{n}\,, \tag{24.4.65}$$

while the mobility and the relaxation time are related by

$$\mu_\mathrm{n} = \frac{e}{m_\mathrm{n}^*}\langle\tau\rangle\,. \tag{24.4.66}$$

Analogous formulas apply to the hole conductivity. The total conductivity is then

$$\sigma = ne\mu_\mathrm{n} + pe\mu_\mathrm{p}\,. \tag{24.4.67}$$

The temperature dependence of the conductivity (resistivity) is therefore determined by the number of charge carriers and their mobilities.

At low temperatures impurity scattering is the dominant scattering mechanism in semiconductors, too, however these impurities are usually charged. In Chapter 29 of Volume 3 we shall see that the Coulomb potential of external charges is screened by mobile electrons, and thus the $1/r$ Coulomb potential is replaced by the exponentially screened Yukawa potential:

$$U(r) \propto \frac{\mathrm{e}^{-qr}}{r}\,, \tag{24.4.68}$$

where, owing to the low electron density in semiconductors, the inverse of the screening length can be specified using the classical Debye–Hückel theory of screening: $q_\mathrm{DH}^2 = n_\mathrm{e}e^2/(\epsilon k_\mathrm{B}T)$, where ϵ is the permittivity. When the matrix element in the transition probability is determined using this potential, the transport relaxation time is given by

$$\frac{1}{\tau(\varepsilon)} \sim \varepsilon^{-3/2} \tag{24.4.69}$$

if only the energy dependence of the dominant term is retained. The temperature dependence of the mobility is then

$$\mu \sim \langle \tau \rangle \sim (k_{\mathrm{B}}T)^{3/2} \,. \tag{24.4.70}$$

The deformation potential offers a particularly well adapted approach to studying the scattering by acoustic phonons. The scattering matrix element leads to an energy- and temperature-dependent relaxation time,

$$\frac{1}{\tau(\varepsilon, T)} \sim \varepsilon^{1/2} k_{\mathrm{B}}T \,, \tag{24.4.71}$$

which, in conjunction with (24.4.63), gives

$$\mu \sim \langle \tau \rangle \sim (k_{\mathrm{B}}T)^{-3/2} \,. \tag{24.4.72}$$

Studying the interaction with optical phonons is not so straightforward as the involved processes are not elastic. At low temperatures ($k_{\mathrm{B}}T \ll \hbar\omega_0$) the crucial factor is the number of thermally excited optical phonons,

$$\frac{1}{\tau(\varepsilon)} \sim \frac{1}{\mathrm{e}^{\hbar\omega_0/k_{\mathrm{B}}T} - 1} \sim \mathrm{e}^{-\hbar\omega_0/k_{\mathrm{B}}T} \,, \tag{24.4.73}$$

and thus

$$\mu \sim \langle \tau \rangle \sim \mathrm{e}^{\hbar\omega_0/k_{\mathrm{B}}T} \,, \tag{24.4.74}$$

whereas at high temperatures the result is the same as for acoustic phonons:

$$\frac{1}{\tau(\varepsilon)} \sim k_{\mathrm{B}}T\varepsilon^{1/2} \qquad \text{and} \qquad \mu \sim \langle \tau \rangle \sim T^{-3/2} \,. \tag{24.4.75}$$

24.5 Quantum Hall Effect

The semiclassical calculation lead to the conclusion that the Hall resistance, which is related to an off-diagonal component of the conductivity tensor, has to be proportional to the magnetic field, whereas ϱ_{xx} has to be practically independent of it. However, in their experiments on silicon-based MOSFETs, von Klitzing and his coworkers[5] found that in strong magnetic fields and at low temperatures (at most 1 to 2 kelvins) ϱ_{xy} does not increase proportionally with B but plateaux appear, and the longitudinal resistivity ϱ_{xx} is not constant but oscillates violently and vanishes at the plateaux. Similar behavior is observed when, instead of tuning the magnetic field strength to control which Landau level is filled partially, the gate voltage is changed to modulate the number of charge carriers. This is shown in Fig. 24.13. At or above liquid-helium temperature practically nothing remains of this anomalous behavior.

[5] K. v. Klitzing, G. Dorda, and M. Pepper, 1980. See the footnote on page 6 of Volume 1.

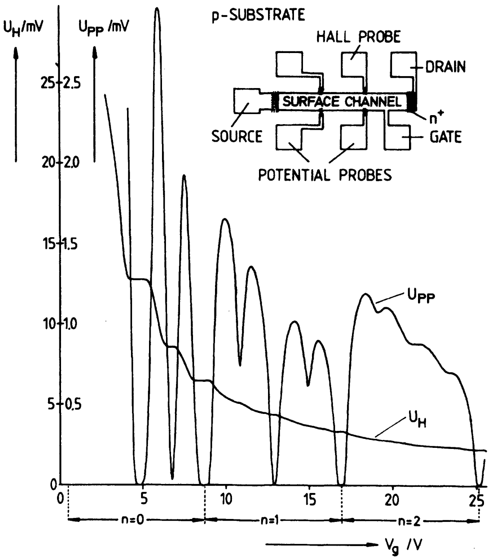


Fig. 24.13. The first experimental results of von Klitzing and coworkers for the quantum Hall effect [Reprinted with permission from *Phys. Rev. Lett.* **45**, 494 (1980). ©1980 by the American Physical Society]

As will be discussed in Chapter 27 on semiconductor devices, an inversion layer may appear at the insulator–semiconductor interface, in which the conduction band is pushed below the chemical potential. The electric field, which is perpendicular to the surface, attracts the electrons into this layer. The thickness of the layer and the number of charge carriers within it can be controlled by the gate voltage V_g. Since the typical width of this layer is 3 to 5 nm, an essentially two dimensional electron gas (2DEG)[6] appears, and the motion in the perpendicular direction freezes out. As backed up by a multitude of other experiments, the appearance of the plateaux is indeed related to the two-dimensional character of the motion of electrons, and impurities play an important role in it.

Applying the results obtained for the Hall effect to the two-dimensional case, the conductivity tensor is

$$\sigma_{\alpha\beta}(B) = \sigma_0 \begin{pmatrix} \dfrac{1}{1+(\omega_\text{c}\tau)^2} & \dfrac{-\omega_\text{c}\tau}{1+(\omega_\text{c}\tau)^2} \\ \dfrac{\omega_\text{c}\tau}{1+(\omega_\text{c}\tau)^2} & \dfrac{1}{1+(\omega_\text{c}\tau)^2} \end{pmatrix}, \qquad (24.5.1)$$

where σ_0 is the Drude conductivity. Its inverse is the resistivity tensor,

[6] The abbreviation 2DES for *two-dimensional electron system* is also commonly used.

$$\varrho_{\alpha\beta}(B) = \varrho_0 \begin{pmatrix} 1 & \omega_c\tau \\ -\omega_c\tau & 1 \end{pmatrix} = \begin{pmatrix} \varrho_0 & B/n_e e \\ -B/n_e e & \varrho_0 \end{pmatrix} . \tag{24.5.2}$$

Apparently, the two-dimensional case does not differ essentially from the previously studied case: in an isotropic system the longitudinal resistivity is independent of the magnetic field, and the Hall resistance increases in direct proportion to the applied filed. In contrast, measurements show that the Hall resistance of a doped two-dimensional electron system[7] features plateaux at

$$\varrho_{xy}^{(\nu)} = \frac{1}{\nu}\frac{h}{e^2} , \tag{24.5.3}$$

where ν is an integer. Therefore the effect is called *integer quantum Hall effect* (IQHE). When $1/\varrho_{xy}$ is plotted against $1/B$, the plateaux are regularly spaced. Moreover, the value of the off-diagonal element of the resistance tensor at the plateaux agrees to a relative accuracy of 10^{-9} with $1/\nu$ times the value $R_K = 25.812\,807\,572\,\mathrm{k\Omega}$, which is calculated from $R_K = h/e^2$. This quantity has since been termed the *von Klitzing constant.* Owing to this extraordinary precision, the quantum Hall effect was adopted in 1990 to establish a new standard for the electrical resistance. By agreement, the conventional value $R_K = 25\,812.807\,\Omega$ was chosen for the Hall resistance of the plateau of label $\nu = 1$.

The xx component of the resistivity tensor, ϱ_{xx}, exhibits strong Shubnikov–de Haas oscillations, and drops to zero wherever the Hall resistance has a plateau. Since

$$\sigma_{xx} = \frac{\varrho_{xx}}{\varrho_{xx}^2 + \varrho_{xy}^2} , \qquad \sigma_{xy} = -\frac{\varrho_{xy}}{\varrho_{xx}^2 + \varrho_{xy}^2} , \tag{24.5.4}$$

when ϱ_{xx} vanishes, so does σ_{xx}, and σ_{xy} also takes a quantized value:

$$\sigma_{xy} = -\nu\frac{e^2}{h} . \tag{24.5.5}$$

If the semiclassical expression for the off-diagonal element of the resistivity tensor were used, the equality

$$\frac{B}{n_e e} = \frac{1}{\nu}\frac{h}{e^2} \tag{24.5.6}$$

would be obtained. This would be satisfied for integer values of ν if the electron density n_e and the magnetic induction B were related by

$$n_e = \nu\frac{eB}{h} . \tag{24.5.7}$$

[7] In two-dimensional systems the off-diagonal components of the resistance and the resistivity tensors are equal: $R_{xy} = \varrho_{xy}$. This is because $E_y = -\varrho_{xy} j_x$ leads to $V_y = -R_{xy} I_x$ via the integration of j_x and E_y in the direction perpendicular to the current flow, y.

As discussed in Section 22.1.2, the degree of degeneracy per unit surface area of the Landau levels in strong magnetic fields is given by (22.1.26), and thus

$$n_{\mathrm{e}} = \nu \frac{N_{\mathrm{p}}}{L_x L_y} . \tag{24.5.8}$$

Spelled out: the quantized value of the Hall resistance measured on the plateau of index ν is obtained in the semiclassical description at a particular electron density for which the lowest ν Landau levels are completely filled and all others are empty. Since elastic scattering is impossible for completely filled Landau levels, $\tau \to \infty$, and the resistivity vanishes: $\varrho_{xx} = 0$.

However, the extended character of the plateaux cannot be understood in the above picture, as the highest filled Landau level is completely filled only at very precise values of the magnetic field. Even a tiny change in the field strength modifies the degree of degeneracy of the Landau levels. With the electron number fixed, either empty states appear on the previously completely filled Landau level, or a previously empty Landau level becomes partially filled.

The resolution of this problem lies in the observation that the width of the plateaux depends strongly on the purity of the sample. In contrast to common situations, in this case the higher the impurity concentration the better. In the presence of impurities the Landau levels are broadened, as shown in Fig. 24.14. Even more important is that a part of the states are localized. These localized states do not participate in conduction: they serve as reservoirs, so the chemical potential can go continuously over from one Landau level to the other. If the field is such that the chemical potential is inside a region where the states are localized, ϱ_{xx} vanishes because these states do not contribute to the conductivity.

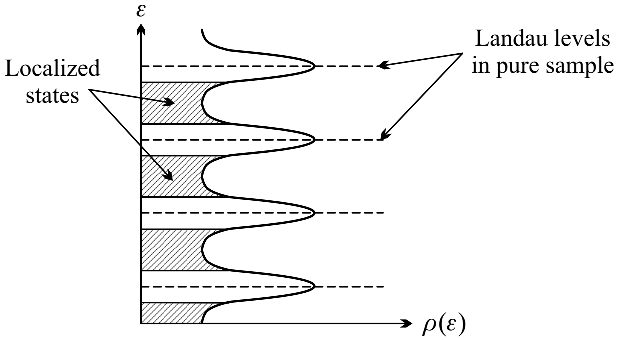


Fig. 24.14. Broadening of Landau levels and the appearance of localized states in the spectrum of a two-dimensional electron gas, due to impurities

More intriguing is the question why ϱ_{xy} takes very accurate quantized values at the same fields. To find the answer, we have to return to Chapter 22, and continue the analysis of the electron states at the sample boundaries in strong magnetic fields. As shown in Fig. 22.7, in contrast to bulk states – whose energy depends only on the quantum number n of the Landau level and is independent of the quantum number k_y that is related to the coordinate x_0 of the oscillator centers –, the energy of the edge states (whose coordinate x_0 is close to the sample boundary) is higher. The number of branches crossing the Fermi energy is the same as the number of bulk Landau levels below the Fermi energy. Moreover, these edge states remain extended even in the presence of impurities. As their velocity is strictly directed along the edge, and they all move around the sample in the same direction, impurities cannot cause backscattering. It is as if electrons propagated in one-dimensional channels along the sample edges, without collisions.

Taking the geometry that is customarily used in the measurements of the Hall effect, the current propagates in the x-direction, and the Hall voltage is measured in the y-direction, thus the finite width of the sample is important in this direction. The gauge that is best adapted to this geometry is the Landau gauge with $\boldsymbol{A} = (-By, 0, 0)$. Choosing the ansatz

$$\psi(x,y) = \mathrm{e}^{\mathrm{i}k_x x} u(y) \tag{24.5.9}$$

for the wavefunction, the equation for $u(y)$ is

$$-\frac{\hbar^2}{2m^*}\frac{\mathrm{d}^2 u(y)}{\mathrm{d}y^2} + \frac{1}{2}m^*\omega_\mathrm{c}^2(y - y_{k_x})^2 u(y) + U(y)u(y) = \varepsilon u(y)\,, \tag{24.5.10}$$

where $y_{k_x} = \hbar k_x/eB = k_x l_0^2$, and the potential $U(y)$ due to the finite width L_y of the sample is also taken into account. Treating this potential as a perturbation, in the first order we have

$$\varepsilon(n, k_x) = \left(n + \tfrac{1}{2}\right)\hbar\omega_\mathrm{c} + U(y_{k_x})\,. \tag{24.5.11}$$

The electron velocities are each other's opposite for the electron states along the two edges $y_{k_x} = 0$ and $y_{k_x} = L_y$, as

$$v_x(n, k_x) = \frac{1}{\hbar}\frac{\partial\varepsilon(n,k_x)}{\partial k_x} = \frac{1}{\hbar}\frac{\partial U(y)}{\partial y}\frac{\partial y_{k_x}}{\partial k_x} = \frac{1}{eB}\frac{\partial U(y)}{\partial y}\,. \tag{24.5.12}$$

As we shall see in Chapter 27, electrons propagating in a one-dimensional channel contribute to the conductance by e^2/h. If there are ν Landau levels below the Fermi energy, then electrons can propagate in the same number of channels, and thus their current is

$$I_x = \nu\frac{e^2}{h}V_x\,. \tag{24.5.13}$$

Because of the free propagation of the electrons there is no potential drop along the sample edge. Each contact is at the same electrochemical potential as the reservoir from which the electrons arrive. That is why ϱ_{xx} is measured to be zero in the usual setup. For the same reason, the potential difference V_y between the two sides is the same as $-V_x$, and so

$$R_{xy} = -\frac{V_y}{I_x} = \frac{1}{\nu}\frac{h}{e^2}\,, \tag{24.5.14}$$

in agreement with experimental findings.

It was found in measurements performed more recently on two-dimensional electron gases produced in AlGaAs/GaAs heterojunctions, which feature much higher mobilities, that the Hall resistance can take quantized values $R_{xy} = h/\nu e^2$ not only with integer values of ν but also for certain simple fractional numbers $\nu = p/q$, that is, for special fractional fillings of the highest partially filled Landau level (usually the lowest Landau level). However, the fractional quantum Hall effect (FQHE) cannot be interpreted in terms of the one-particle picture: the role of the electron–electron interaction is crucial. We shall discuss this in detail in Chapter 32.

Further Reading

1. F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill Book Company, New York (1968).
2. J. S. Dugdale, *The Electrical Properties of Metals and Alloys*, Edward Arnold, London (1977).
3. *The Quantum Hall Effect*, Edited by R. E. Prange and S. M. Girvin, Springer-Verlag, New York (1987).
4. H. Smith and H. H. Jensen, *Transport Phenomena*, Oxford Science Publications, Clarendon Press, Oxford (1989).
5. J. Ziman, *Electrons and Phonons, The Theory of Transport Phenomena in Solids*, Oxford Classic Texts in the Physical Sciences, Oxford University Press, Oxford (2001).

25

Optical Properties of Solids

In the previous chapter on transport properties we studied the behavior of solids in static or low-frequency electromagnetic fields. As a continuation, we shall now investigate what happens to them when subjected to electromagnetic radiation in or close to the optical region.

One of our reasons for using the semiclassical approximation in the foregoing was that at low frequencies the quanta of the electromagnetic field, photons, are not energetic enough to cause interband transitions. This is no longer true for visible light, whose wavelength ranges from 400 to 800 nm. The energy of the photons of red light ($\lambda = 620\,\text{nm}$) is about 2 eV, while that of blue light ($\lambda = 470\,\text{nm}$) is about 2.6 eV; both are sufficient to induce interband transitions. It is even easier with photons in the higher-frequency ultraviolet (UV) region, whose energy ranges up to about 100 eV. When speaking about the optical properties of solids, we must also consider infrared (IR) radiation, whose wavelength is longer than that of visible light. Since the energy of IR photons ranges from 10^{-3} eV to about 1.6 eV, they cannot usually induce interband transitions, nonetheless their interaction with solids can be treated on the same footing as that of visible light.

There are no sharp dividing lines between the infrared, visible, and ultraviolet portions of the spectrum, or the far, mid and near infrared (FIR, MIR, NIR) regions. Table 25.1 contains typical wavelength, frequency, wave number and phonon energy values for different portions of the optical region of the electromagnetic spectrum. Their conversion formulas are given in Appendix A.

Since optical wavelengths, ranging from 10 nm to 1 mm, are much larger than the atomic dimensions, solids interacting with light can often be treated as continuous media, and their atomic structure neglected. Consequently, there are two levels of studying optical properties. The classical treatment of electromagnetic radiation, based on the Maxwell equations, offers an adequate description for a wide range of phenomena. Solids are then treated as dielectric media, and characterized by optical constants. To provide a microscopic foundation to this phenomenological description, the optical constants have to be related to the dielectric constant and the optical conductivity. The latter

Table 25.1. The wavelength, frequency, wave number and energy for different portions of the optical region of the electromagnetic spectrum

	Wavelength	Frequency (THz)	Wave number (cm^{-1})	Energy
FIR	25–1000 µm	0.1–10	10–400	0.5–50 meV
MIR	2.5–25 µm	10–120	400–4000	50–500 meV
NIR	0.8–2.5 µm	120–400	4×10^3–12×10^3	0.5–1.6 eV
Visible	400–800 nm	400–800	12×10^3–24×10^3	1.6–3 eV
UV	10–400 nm	800–32 000	24×10^3–10^6	3–120 eV

quantities need to be derived from first principles, using quantum mechanics to describe the ground state and excited states of the electron system and the vibrating lattice, as well as their transitions induced by the electromagnetic field. However, if a more accurate description of electromagnetic absorption is required, we can no longer neglect that the radiation field is quantized, and that photons of well-defined energy and momentum interact with both the collective elementary excitations of solids (e.g., phonons) and the one-particle excitations (Bloch electrons). The importance of optical measurements lies in the fact that the so-called optical constants may in fact depend strongly on the wavelength of the radiation incident on the sample. Since a very wide frequency window is open to experiments, the response of the system's electrons and phonons to radiation can be studied over a vast energy range.

In this chapter we shall first explore the reflection, transmission, and absorption of light in solids using the classical approach based on the Maxwell equations, and then determine the dielectric constant for some simple cases: free electrons, bound electrons, and a system of vibrating atoms. As we shall see, this allows us to gain insight into the possible states of electrons and phonons. At the end of the chapter we shall also outline the more rigorous quantum mechanical treatment.

25.1 Interaction of Solids with the Classical Radiation Field

When a solid is illuminated by light – or more generally: when a solid is exposed to electromagnetic radiation whose frequency is in the optical region –, a part of the incident light is reflected, another part may emerge as an attenuated beam on the other side of the sample, and the rest is absorbed in the sample. In optical measurements the material properties of the sample are often inferred from absorption, transmission, and reflection data.

The simplest way to describe the optical properties of solids and to treat the interaction with the electromagnetic field is based on the Maxwell equations of classical electrodynamics. Material properties are included through the dielectric constant and the conductivity. This approach was also used in Chapter 16, where the high-frequency behavior was studied in the free-electron model of metals. In addition to the nearly free conduction electrons, we shall also take into account bound electrons as well as the contribution of phonons in this chapter.

25.1.1 Propagation and Absorption of Light in a Dielectric

Written in their conventional form, the Maxwell equations[1] of classical electrodynamics read

$$\begin{aligned} &\text{I.}\ \operatorname{curl}\boldsymbol{B} = \mu_0\Big(\boldsymbol{j} + \epsilon_0\frac{\partial \boldsymbol{E}}{\partial t}\Big)\,, && \text{II.}\ \operatorname{curl}\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}\,, \\ &\text{III.}\ \operatorname{div}\boldsymbol{E} = \rho/\epsilon_0\,, && \text{IV.}\ \operatorname{div}\boldsymbol{B} = 0\,, \end{aligned} \tag{25.1.1}$$

where ρ is the total charge density and $\boldsymbol{j}$ is the total current density.

In vacuum, where neither charges nor currents are present, the Maxwell equations in conjunction with $\mu_0 = 1/\epsilon_0 c^2$ lead simply to

$$\operatorname{curl}\operatorname{curl}\boldsymbol{E} = -\epsilon_0\mu_0\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = -\frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}\,. \tag{25.1.2}$$

Using the operator identity

$$\operatorname{curl}\operatorname{curl} = \operatorname{grad}\operatorname{div} - \boldsymbol{\nabla}^2 \tag{25.1.3}$$

and the homogeneous Maxwell equation $\operatorname{div}\boldsymbol{E} = 0$, the wave equation

$$\boldsymbol{\nabla}^2\boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} \tag{25.1.4}$$

is obtained. A similar equation applies to the magnetic induction (magnetic flux density). These equations describe transverse electromagnetic waves that propagate at the speed of light:

$$\boldsymbol{E} = \boldsymbol{E}_0\exp(-\mathrm{i}\omega[t - \hat{\boldsymbol{q}}\cdot\boldsymbol{r}/c])\,, \qquad \boldsymbol{B} = \boldsymbol{B}_0\exp(-\mathrm{i}\omega[t - \hat{\boldsymbol{q}}\cdot\boldsymbol{r}/c])\,, \tag{25.1.5}$$

where $\hat{\boldsymbol{q}}$ is the unit vector in the propagation direction of the wave, $\boldsymbol{E}_0 \perp \hat{\boldsymbol{q}}$, $\boldsymbol{B}_0 \perp \hat{\boldsymbol{q}}$, and $\boldsymbol{E}_0 \perp \boldsymbol{B}_0$, moreover $\boldsymbol{q}$, $\boldsymbol{E}_0$, and $\boldsymbol{B}_0$ span a right-handed Cartesian coordinate system, and they are related by

$$\boldsymbol{B}_0 = \frac{1}{c}\,\hat{\boldsymbol{q}}\times\boldsymbol{E}_0\,. \tag{25.1.6}$$

Longitudinal waves are not allowed, because if no external charges are present, $\operatorname{div}\boldsymbol{E} = 0$ and $\operatorname{div}\boldsymbol{B} = 0$, and thus $\boldsymbol{E}$ and $\boldsymbol{B}$ can have nonvanishing components only perpendicular to the propagation direction.

[1] J. C. Maxwell, 1873.

The situation is more complicated in solids exposed to electromagnetic radiation. Under the influence of the electric field $\boldsymbol{E}$, the oppositely directed motion of positive and negative charges gives rise to a dipole moment, which can be characterized by the polarization $\boldsymbol{P}$. Likewise, a magnetic field $\boldsymbol{B}$ leads to a magnetic polarization that can be specified by the magnetization $\boldsymbol{M}$. In addition to describing the effects of internal charges in terms of the polarization and magnetization, external charges can also be injected into the solid, and currents can also be set up externally. By using the relations

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P} \qquad \text{and} \qquad \boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M}\,, \tag{25.1.7}$$

and denoting the densities of external charges and currents by ρ_{ext} and $\boldsymbol{j}_{\text{ext}}$, the Maxwell equations in dielectric media take the form

$$\begin{aligned} &\text{I.}\ \ \operatorname{curl} \boldsymbol{H} = \boldsymbol{j}_{\text{ext}} + \frac{\partial \boldsymbol{D}}{\partial t}\,, \qquad &&\text{II.}\ \ \operatorname{curl} \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}\,, \\ &\text{III.}\ \ \operatorname{div} \boldsymbol{D} = \rho_{\text{ext}}\,, &&\text{IV.}\ \ \operatorname{div} \boldsymbol{B} = 0\,. \end{aligned} \tag{25.1.8}$$

Naturally, these equations need to be complemented by the constitutive relations between the electric displacement and electric field, the magnetic induction and magnetic field, and the current density and electric field.

The relationships between $\boldsymbol{P}$ and $\boldsymbol{E}$, and $\boldsymbol{M}$ and $\boldsymbol{H}$ are usually nonlocal. For weak fields, where the relationships are linear, simple equations can be written down for the Fourier components. For isotropic systems and cubic crystals the electric and magnetic susceptibilities are defined by

$$\frac{1}{\epsilon_0}\boldsymbol{P}(\boldsymbol{q},\omega) = \chi_{\text{el}}(\boldsymbol{q},\omega)\boldsymbol{E}(\boldsymbol{q},\omega)\,, \quad \boldsymbol{M}(\boldsymbol{q},\omega) = \chi_{\text{m}}(\boldsymbol{q},\omega)\boldsymbol{H}(\boldsymbol{q},\omega)\,. \tag{25.1.9}$$

Substituted into the equations for the Fourier components of the electric displacement and magnetic induction,

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{q},\omega) =& \epsilon_0 \boldsymbol{E}(\boldsymbol{q},\omega) + \boldsymbol{P}(\boldsymbol{q},\omega)\,, \\ \boldsymbol{B}(\boldsymbol{q},\omega) =& \mu_0 \left[\boldsymbol{H}(\boldsymbol{q},\omega) + \boldsymbol{M}(\boldsymbol{q},\omega)\right], \end{aligned} \tag{25.1.10}$$

we obtain

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{q},\omega) =& \epsilon_0[1 + \chi_{\text{el}}(\boldsymbol{q},\omega)]\boldsymbol{E}(\boldsymbol{q},\omega) = \epsilon_0\, \epsilon_{\text{r}}^{\text{tot}}(\boldsymbol{q},\omega)\boldsymbol{E}(\boldsymbol{q},\omega)\,, \\ \boldsymbol{B}(\boldsymbol{q},\omega) =& \mu_0[1 + \chi_{\text{m}}(\boldsymbol{q},\omega)]\boldsymbol{H}(\boldsymbol{q},\omega) = \mu_0\, \mu_{\text{r}}^{\text{tot}}(\boldsymbol{q},\omega)\boldsymbol{H}(\boldsymbol{q},\omega)\,. \end{aligned} \tag{25.1.11}$$

These equations relate the susceptibilities to the permittivity and permeability. The label "tot" in the superscript indicates that the quantities contain the contributions of bound as well as free electrons. We shall also assume that Ohm's law applies, i.e., the current density is proportional to the electric field:

$$\boldsymbol{j} = \sigma \boldsymbol{E}\,, \tag{25.1.12}$$

where $\boldsymbol{j}$ also contains the current induced by the external field. We shall discuss this in more detail in Chapter 29.

Note that for anisotropic crystals the constitutive relations usually have to be written as

$$\begin{aligned} D_\alpha &= \epsilon_{\alpha\beta} E_\beta \,, \\ B_\alpha &= \mu_{\alpha\beta} H_\beta \,, \qquad (\alpha, \beta = x, y, z) \\ j_\alpha &= \sigma_{\alpha\beta} E_\beta \,, \end{aligned} \tag{25.1.13}$$

where $\epsilon_{\alpha\beta}$, $\mu_{\alpha\beta}$, and $\sigma_{\alpha\beta}$ are the $\alpha\beta$ components of the permittivity, permeability, and conductivity tensors. Below we shall ignore the possibility of anisotropy, and work with scalar quantities ϵ, μ, and σ, which is justified for cubic crystals as well.

If no external charges are injected and no external currents are applied, and the electromagnetic field is assumed to be of a definite wave number and frequency,

$$\boldsymbol{E} = \boldsymbol{E}_0 \exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r} - \mathrm{i}\omega t) \,, \qquad \boldsymbol{B} = \boldsymbol{B}_0 \exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r} - \mathrm{i}\omega t) \,, \tag{25.1.14}$$

similar wave equations are obtained inside dielectric media as in empty space:

$$\boldsymbol{\nabla}^2 \boldsymbol{E}(\boldsymbol{q},\omega) = \frac{\epsilon_{\mathrm{r}}^{\mathrm{tot}}(\boldsymbol{q},\omega)\mu_{\mathrm{r}}^{\mathrm{tot}}(\boldsymbol{q},\omega)}{c^2} \frac{\partial^2 \boldsymbol{E}(\boldsymbol{q},\omega)}{\partial t^2} \,. \tag{25.1.15}$$

A similar equation applies to $\boldsymbol{H}$ (or $\boldsymbol{B}$). Consequently, only waves satisfying

$$q^2 = \frac{\omega^2}{c^2} N^2 \tag{25.1.16}$$

may propagate in the system, where

$$N = \sqrt{\epsilon_{\mathrm{r}}^{\mathrm{tot}} \mu_{\mathrm{r}}^{\mathrm{tot}}} \tag{25.1.17}$$

is the refractive index (index of refraction). Denoting the unit vector along the wave propagation direction once again by $\hat{\boldsymbol{q}}$,

$$\boldsymbol{E} = \boldsymbol{E}_0 \exp\left[\mathrm{i}\frac{\omega N}{c}\hat{\boldsymbol{q}}\cdot\boldsymbol{r} - \mathrm{i}\omega t\right] , \quad \boldsymbol{B} = \boldsymbol{B}_0 \exp\left[\mathrm{i}\frac{\omega N}{c}\hat{\boldsymbol{q}}\cdot\boldsymbol{r} - \mathrm{i}\omega t\right] , \tag{25.1.18}$$

where $\hat{\boldsymbol{q}}$, $\boldsymbol{E}_0$, and $\boldsymbol{B}_0$ are mutually perpendicular, and they span a right-handed coordinate system. The amplitudes of the electric and magnetic components of an electromagnetic wave propagating in a dielectric medium are now related by

$$\boldsymbol{B}_0 = N\sqrt{\epsilon_0\mu_0}\,\hat{\boldsymbol{q}} \times \boldsymbol{E}_0 = \frac{1}{c} N\,\hat{\boldsymbol{q}} \times \boldsymbol{E}_0 \,. \tag{25.1.19}$$

In this chapter we shall assume that the sample is weakly magnetizable ($\mu_{\mathrm{r}}^{\mathrm{tot}} \approx 1$), so we can focus on the relative permittivity alone. Then

$$N \approx \sqrt{\epsilon_{\mathrm{r}}^{\mathrm{tot}}}\,. \tag{25.1.20}$$

If $\epsilon_{\mathrm{r}}^{\mathrm{tot}}$ is real, then the propagation velocity of electromagnetic waves in the medium is equal to the speed of light divided by the refractive index: $v_{\mathrm{prop}} = c/N$.

This assumption is usually incorrect, and $\epsilon_{\mathrm{r}}^{\mathrm{tot}}$ is complex. Then the complex index of refraction N can be separated into real and imaginary parts,

$$N = n + \mathrm{i}\kappa\,, \tag{25.1.21}$$

and the electromagnetic field propagating inside a dielectric medium is described by the formulas

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}_0 \exp\left(-\frac{\omega\kappa}{c}\hat{\boldsymbol{q}}\cdot\boldsymbol{r}\right)\exp\left(\mathrm{i}\frac{\omega n}{c}\hat{\boldsymbol{q}}\cdot\boldsymbol{r} - \mathrm{i}\omega t\right), \\ \boldsymbol{B} &= \boldsymbol{B}_0 \exp\left(-\frac{\omega\kappa}{c}\hat{\boldsymbol{q}}\cdot\boldsymbol{r}\right)\exp\left(\mathrm{i}\frac{\omega n}{c}\hat{\boldsymbol{q}}\cdot\boldsymbol{r} - \mathrm{i}\omega t\right). \end{aligned} \tag{25.1.22}$$

Here, too, $\boldsymbol{B}_0$ and $\boldsymbol{E}_0$ are related by (25.1.19), but since N is complex, $\boldsymbol{E}_0$ and $\boldsymbol{B}_0$ cannot be chosen real simultaneously, as a complex phase factor appears between them. The spatial variations of the electric and magnetic fields indicate that n is related to dispersion, and κ to absorption. Thus n is the real refractive index, while κ is the extinction coefficient (absorption index). The physical reason behind absorption is precisely the phase difference between the electric and magnetic fields. According to the Lambert–Beer law[2] the intensity of the electromagnetic field decays exponentially in the sample. The linear absorption coefficient α is defined as

$$\alpha = -\frac{1}{I}\frac{\mathrm{d}I}{\mathrm{d}r}\,. \tag{25.1.23}$$

Its reciprocal is the distance over which the intensity drops to 1/e times the initial value. By using (25.1.22) for the field amplitude,

$$\alpha = \frac{2\kappa\omega}{c} = \frac{4\pi\kappa}{\lambda}\,. \tag{25.1.24}$$

The two optical constants, n and κ – which are, in fact, functions of ω – are not independent of one another. This is because the complex refractive index is related to the complex dielectric constant ($N^2 = \epsilon_{\mathrm{r}}^{\mathrm{tot}}$), whose real and imaginary parts satisfy the Kramers–Kronig relation (16.1.65). The connection between n and κ can be derived in an even more straightforward manner by realizing that the analiticity of $N(\omega)$ implies that its real and imaginary parts – just like those of the complex dielectric constant – should satisfy the Kramers–Kronig relation:

[2] J. H. Lambert, 1760 and A. Beer, 1854. It is also known as the Beer–Bouguer–Lambert law, since it was first discovered by P. Bouguer in 1729.

$$
\begin{aligned}
n(\omega) - 1 &= \frac{2}{\pi} P \int\limits_0^\infty \frac{\omega' \kappa(\omega')}{\omega'^2 - \omega^2} \, \mathrm{d}\omega' , \\
\kappa(\omega) &= -\frac{2\omega}{\pi} P \int\limits_0^\infty \frac{n(\omega') - 1}{\omega'^2 - \omega^2} \, \mathrm{d}\omega' .
\end{aligned}
\tag{25.1.25}
$$

By measuring the absorption of a beam traversing a thin layer, $n(\omega)$ could be determined – at least, in principle – if precise measurements could be performed over the entire frequency range.

25.1.2 Reflection and Refraction at an Interface

The other possibility for measuring the optical constants is based on the observation of the beam reflected from the surface of the sample. To determine the reflectivity, we choose a geometry in which a plane interface separates the vacuum in the $z > 0$ region from a medium of refractive index N in the $z < 0$ region. The propagation direction of the incident electromagnetic radiation is in the xz-plane, and makes an angle θ with the normal of the interface. We shall denote the parallel component of the electric field (which is in the plane of incidence) by E_p, and the perpendicular component by E_s.[3] It is therefore customary to speak of s- and p-polarized waves.[4] A part of the radiation is reflected, and another penetrates into the sample. We shall denote the angle of refraction by θ'', and the components of the electric field in the reflected and refracted rays that are parallel and perpendicular to the plane of incidence by E'_p and E'_s, and E''_p and E''_s. They are shown in Fig. 25.1.

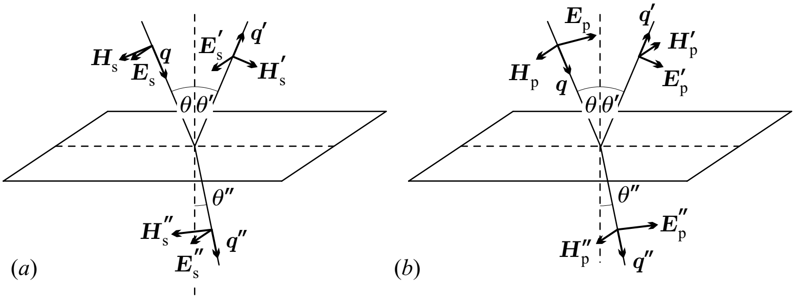


Fig. 25.1. The electric and magnetic field vectors of the incident, reflected, and refracted waves at a vacuum–solid interface for (*a*) TE and (*b*) TM polarization

[3] The label "s" comes from the German word for perpendicular, *senkrecht.*

[4] S-polarization is also called transverse electric (TE) polarization, as the electric field vector of the wave is perpendicular to the plane of incidence. Likewise, p-polarization is also known as transverse magnetic (TM) polarization.

The Maxwell equations imply that the tangential component of the electric and magnetic fields and the perpendicular component of the electric displacement and magnetic induction must be continuous across the interface of two media of different indices of refraction – provided there are no surface charges and surface currents. Denoting the surface normal by $\hat{\boldsymbol{n}}$, the continuity condition can be generally formulated as

$$\begin{aligned} (\boldsymbol{E}+\boldsymbol{E}'-\boldsymbol{E}'')\times\hat{\boldsymbol{n}}=0\,, \quad (\boldsymbol{D}+\boldsymbol{D}'-\boldsymbol{D}'')\cdot\hat{\boldsymbol{n}}=0\,,\\ (\boldsymbol{H}+\boldsymbol{H}'-\boldsymbol{H}'')\times\hat{\boldsymbol{n}}=0\,, \quad (\boldsymbol{B}+\boldsymbol{B}'-\boldsymbol{B}'')\cdot\hat{\boldsymbol{n}}=0\,. \end{aligned} \tag{25.1.26}$$

Thus, for an s-polarized wave we have

$$\begin{aligned} E_{\mathrm{s}}+E'_{\mathrm{s}} &= E''_{\mathrm{s}}\,,\\ H_{\mathrm{s}}\cos\theta - H'_{\mathrm{s}}\cos\theta' &= H''_{\mathrm{s}}\cos\theta''\,, \end{aligned} \tag{25.1.27}$$

where

$$\sin\theta'' = \frac{\sin\theta}{N} \tag{25.1.28}$$

because of Snell's law of refraction, and, as usual, $\theta' = \theta$ is assumed for reflection. Since (25.1.19) implies that

$$H_{\mathrm{s}} = \sqrt{\frac{\epsilon_0}{\mu_0}}E_{\mathrm{s}}\,, \quad H'_{\mathrm{s}} = \sqrt{\frac{\epsilon_0}{\mu_0}}E'_{\mathrm{s}}\,, \quad H''_{\mathrm{s}} = N\sqrt{\frac{\epsilon_0}{\mu_0}}E''_{\mathrm{s}}\,, \tag{25.1.29}$$

the second equation in (25.1.27) yields

$$(E_{\mathrm{s}} - E'_{\mathrm{s}})\cos\theta = NE''_{\mathrm{s}}\cos\theta''\,. \tag{25.1.30}$$

For a p-polarized wave, in which the magnetic field is parallel to the interface, we have

$$\begin{aligned} E_{\mathrm{p}}\cos\theta + E'_{\mathrm{p}}\cos\theta &= E''_{\mathrm{p}}\cos\theta''\,,\\ H_{\mathrm{p}} - H'_{\mathrm{p}} &= H''_{\mathrm{p}}\,. \end{aligned} \tag{25.1.31}$$

By expressing H_{p} in the second equation in terms of E_{p},

$$E_{\mathrm{p}} - E'_{\mathrm{p}} = NE''_{\mathrm{p}}\,. \tag{25.1.32}$$

These are the same equations as in classical optics, provided N is replaced by the real refractive index n. By taking the ratios of the amplitudes, we obtain the Fresnel formulas[5] of reflection and refraction:

$$\begin{aligned} r_{\mathrm{p}} &= \frac{E'_{\mathrm{p}}}{E_{\mathrm{p}}} = \frac{\cos\theta'' - N\cos\theta}{N\cos\theta + \cos\theta''}\,, \qquad & t_{\mathrm{p}} &= \frac{E''_{\mathrm{p}}}{E_{\mathrm{p}}} = \frac{2\cos\theta}{N\cos\theta + \cos\theta''}\,,\\ r_{\mathrm{s}} &= \frac{E'_{\mathrm{s}}}{E_{\mathrm{s}}} = \frac{\cos\theta - N\cos\theta''}{\cos\theta + N\cos\theta''}\,, \qquad & t_{\mathrm{s}} &= \frac{E''_{\mathrm{s}}}{E_{\mathrm{s}}} = \frac{2\cos\theta}{\cos\theta + N\cos\theta''}\,. \end{aligned} \tag{25.1.33}$$

[5] A. J. Fresnel, 1815–1818.

The quantities r and t are called the *reflection* and *transmission coefficients*.[6] Using Snell's law, they are usually written as

$$\begin{aligned} r_\mathrm{p} &= -\frac{\tan(\theta-\theta'')}{\tan(\theta+\theta'')}\,, \qquad & t_\mathrm{p} &= \frac{2\cos\theta\sin\theta''}{\sin(\theta+\theta'')\cos(\theta-\theta'')}\,, \\ r_\mathrm{s} &= -\frac{\sin(\theta-\theta'')}{\sin(\theta+\theta'')}\,, \qquad & t_\mathrm{s} &= \frac{2\cos\theta\sin\theta''}{\sin(\theta+\theta'')}\,. \end{aligned} \tag{25.1.34}$$

When the incident light is reflected from the interface of an optically denser medium ($n > 1$), the amplitude reflection coefficient is negative, indicating a 180° phase shift.

These equations formally apply to absorbing media as well. Compared to dielectrics, the index of refraction is complex, therefore the quantity θ'' determined from Snell's law does not have a direct physical meaning as the angle that specifies the direction of the refracted wave. Instead, it is related to the phase shift between the reflected and refracted waves.

The formulas become much simpler for normal incidence. TM- and TE-polarized waves are the same then, the amplitude of the reflected beam is

$$\frac{E'}{E} = \frac{1-N}{1+N}\,, \tag{25.1.35}$$

and the *reflectance* (or *reflectivity*) is

$$R = \left|\frac{1-N}{1+N}\right|^2 = \frac{(n-1)^2+\kappa^2}{(n+1)^2+\kappa^2}\,. \tag{25.1.36}$$

Optical measurements are often performed on thin layers. A part of the refracted wave then emerges at the other interface (transmitted wave), and another part becomes reflected, thereby augmenting the intensity of the backscattered wave and giving rise to multiple scattering as well. When the thickness d of the layer is much larger than the wavelength, and thus interference effects can be neglected, the *transmittance* is

$$\langle T\rangle = \frac{(1-R)^2\mathrm{e}^{-Kd}}{1-R^2\mathrm{e}^{-2Kd}}\,, \tag{25.1.37}$$

while the reflectance is

$$\langle R\rangle = R\bigl[1+\langle T\rangle\mathrm{e}^{-Kd}\bigr]. \tag{25.1.38}$$

By measuring the absorption and reflection, the real and imaginary parts of the complex refractive index can be determined. However, as mentioned on page 416, n and κ are not independent of one another, so reflection itself gives a complete characterization of the system, provided its value is known at all frequencies.

[6] Since they are obtained from the amplitudes of the electromagnetic field, they are also called amplitude reflection and transmission coefficients, to distinguish them from intensity reflection and transmission coefficients.

25.1.3 Role of Free and Bound Electrons

In the foregoing discussion each electron was treated on the same footing, whichever band it belonged in; that is why the label “tot” appeared in the superscript of ϵ and μ. However, it is often more convenient to distinguish bound electrons, which are part of the ion core and belong in deep-lying narrow bands, from the more or less free, highly mobile electrons of the bands that are closer to the Fermi energy. This separation is arbitrary in a sense. In the analysis of optical properties another, natural, separation is used, which is based on the energy associated with the highest studied frequency. An electron is then considered bound if its binding energy exceeds this predefined maximum energy. The effects of bound electrons are lumped into the polarization – i.e., a dielectric constant $\epsilon_\mathrm{r}(\infty)$ –, whereas the current carried by mobile conduction electrons is retained separately. Therefore an additional term appears in the wave equation. Exploiting the constitutive relation,

$$\nabla^2 \boldsymbol{E} = \frac{\epsilon(\infty)}{\epsilon_0 c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \frac{\sigma}{\epsilon_0 c^2}\frac{\partial \boldsymbol{E}}{\partial t}\,. \tag{25.1.39}$$

It is important to note that σ is now the optical conductivity, which is related to the current driven by the transverse field. Apart from the long-wavelength limit, it is generally different from the longitudinal conductivity.

Seeking the solution in the usual form (25.1.14), those waves can propagate for which

$$q^2 = \frac{\epsilon(\infty)}{\epsilon_0}\frac{\omega^2}{c^2} + \frac{\mathrm{i}\omega\sigma}{\epsilon_0 c^2} = \frac{\omega^2}{c^2}\left[\epsilon_\mathrm{r}(\infty) + \mathrm{i}\frac{\sigma}{\epsilon_0\omega}\right]. \tag{25.1.40}$$

Comparison with (25.1.16) and (25.1.17) gives

$$\epsilon_\mathrm{r}^\mathrm{tot} = \epsilon_\mathrm{r}(\infty) + \mathrm{i}\frac{\sigma}{\epsilon_0\omega}\,. \tag{25.1.41}$$

We may say that the real part of the complex dielectric constant comes from the dielectric constant $\epsilon_\mathrm{r}(\infty)$ of bound electrons and the imaginary part of the conductivity of free electrons, while its imaginary part comes from the real part of the conductivity of free electrons:

$$\epsilon_1 = \epsilon_\mathrm{r}(\infty) - \frac{\operatorname{Im}\sigma}{\epsilon_0\omega}\,, \qquad \epsilon_2 = \frac{\operatorname{Re}\sigma}{\epsilon_0\omega}\,. \tag{25.1.42}$$

If the real and imaginary parts of the complex refractive index are used instead of the real and imaginary parts of the dielectric constant, we have

$$n + \mathrm{i}\kappa = \left[\epsilon_\mathrm{r}(\infty) + \mathrm{i}\frac{\sigma}{\epsilon_0\omega}\right]^{1/2}, \tag{25.1.43}$$

that is,

$$n^2 - \kappa^2 = \epsilon_\mathrm{r}(\infty) - \frac{\operatorname{Im}\sigma}{\epsilon_0\omega}\,, \qquad 2n\kappa = \frac{\operatorname{Re}\sigma}{\epsilon_0\omega}\,. \tag{25.1.44}$$

25.1.4 Scattering of Light by Free Electrons

As a first example, we shall now examine the optical properties of metals in the simplest approximation. In the classical Drude model, conduction electrons move in a uniform background of unit dielectric constant, and the frequency-dependent conductivity is given by (16.1.41):

$$\sigma(\omega) = \frac{\sigma_0}{1 - \mathrm{i}\omega\tau}\,, \qquad \text{where} \qquad \sigma_0 = \frac{n_{\mathrm{e}} e^2 \tau}{m_{\mathrm{e}}}\,. \tag{25.1.45}$$

However, this formula for the conductivity cannot be applied without reservations. As already mentioned, the optical properties are governed by the optical conductivity, which, in turn, should be determined from the current induced by transverse electromagnetic fields. This is in striking contrast with the Drude model, where the longitudinal conductivity was calculated from the longitudinal current.

To determine the optical conductivity, we have to consider the equation of motion for electrons in the presence of an electromagnetic field. Even when there are no external currents, charges placed in an electromagnetic field move. Their displacement $\boldsymbol{r}$ is related to the electric field $\boldsymbol{E}$ through the equation of motion that contains the Lorentz force. Neglecting the term due to the magnetic field, as small to the second order, the motion of electrons is governed by the equation

$$m_{\mathrm{e}} \ddot{\boldsymbol{r}} = -e\boldsymbol{E}\,. \tag{25.1.46}$$

In a field of angular frequency ω the electrons oscillate with the same frequency, thus, in terms of the Fourier transforms,

$$-\,\omega^2 m_{\mathrm{e}} \boldsymbol{r} = -e\boldsymbol{E}\,. \tag{25.1.47}$$

The density of the current carried by the moving charges is

$$\boldsymbol{j} = -e n_{\mathrm{e}} \dot{\boldsymbol{r}}\,. \tag{25.1.48}$$

Using its Fourier transform in the Maxwell equations labeled I. and II.:

$$\mathrm{i}\boldsymbol{q} \times \boldsymbol{H} = -\mathrm{i}\omega\epsilon_0 \boldsymbol{E} + \mathrm{i}\omega n_{\mathrm{e}} e \boldsymbol{r}\,, \qquad \mathrm{i}\boldsymbol{q} \times \boldsymbol{E} = \mathrm{i}\omega\mu_0 \boldsymbol{H}\,. \tag{25.1.49}$$

Combining the two equations gives

$$\boldsymbol{q} \times [\boldsymbol{q} \times \boldsymbol{E}] = \omega\mu_0\, \boldsymbol{q} \times \boldsymbol{H} = -\omega^2 \mu_0 \epsilon_0 \boldsymbol{E} + \omega^2 \mu_0 n_{\mathrm{e}} e \boldsymbol{r}\,. \tag{25.1.50}$$

Transforming the left-hand side by making use of (3.3.6),

$$\boldsymbol{q} \times [\boldsymbol{q} \times \boldsymbol{E}] = \boldsymbol{q}(\boldsymbol{q} \cdot \boldsymbol{E}) - q^2 \boldsymbol{E}\,. \tag{25.1.51}$$

In the transverse case the first term vanishes. Then

$$q^2 \boldsymbol{E} = \omega^2 \mu_0 \epsilon_0 \boldsymbol{E} - \omega^2 \mu_0 n_{\mathrm{e}} e \boldsymbol{r} = \frac{\omega^2}{c^2} \boldsymbol{E} - \frac{\omega^2}{c^2} \frac{n_{\mathrm{e}} e}{\epsilon_0} \boldsymbol{r}\,, \tag{25.1.52}$$

where the identity $\epsilon_0\mu_0 = 1/c^2$ was used. The substitution of (25.1.47) into this equation leads to

$$q^2 = \frac{\omega^2}{c^2} - \frac{1}{c^2}\frac{n_e e^2}{\epsilon_0 m_e}\,, \tag{25.1.53}$$

from which the frequency of transverse oscillations can be calculated. By comparing this formula with (25.1.16) and (25.1.20), and making use of (16.1.69), the equation for the frequency of the longitudinal plasma oscillations,

$$\epsilon_r = 1 - \frac{n_e e^2}{\epsilon_0 \omega^2 m_e} = 1 - \frac{\omega_p^2}{\omega^2} \tag{25.1.54}$$

is found for the transverse dielectric constant of the free-electron gas. This expression is the same as the high-frequency formula for the longitudinal dielectric constant.

Considering the optical constants n and κ instead of the dielectric constant, $n = 0$ and κ is finite for $\omega < \omega_p$, while $\kappa = 0$ and n is finite for $\omega > \omega_p$. Thus an electromagnetic radiation of frequency $\omega < \omega_p$ that is incident perpendicular to the interface undergoes total reflection because oscillations of such frequency cannot propagate in the free-electron gas. The intensity of the electromagnetic field decays exponentially toward the interior of the sample. This can be interpreted classically by saying that the forced oscillations of free electrons in the electromagnetic field are 180 degrees out of phase with the driving field, so the field is canceled inside the sample. Waves with a frequency of less than ω_p can propagate only on the surface. It can be shown that the frequency of the surface plasma oscillations (surface plasmons) is smaller than $\omega_p/\sqrt{2}$.

For electromagnetic fields of frequency $\omega > \omega_p$ reflection becomes gradually weaker, and the metal becomes transparent. In metals that can be modeled by a free-electron gas only those transverse oscillations can propagate whose frequency and wave number satisfy

$$\omega_T^2 = \omega_p^2 + c^2 q^2\,. \tag{25.1.55}$$

The threshold frequency ω_p of transparency can be estimated by inserting a typical metallic electron density value into the plasma-frequency formula (16.1.69). The resulting frequency is in the ultraviolet; total reflection prevails over the entire visible frequency range. This explains the characteristic luster of metals. Certain metals, for example copper or gold, have characteristic colors because in addition to the quasi-free conduction electrons the d-electrons of the core also interact with the electromagnetic field, absorbing photons of well-defined wavelengths. By converting the threshold frequency (16.1.69) to wavelength using the relation $\lambda = 2\pi c/\omega_p$, the calculated values are found to be in fairly good agreement with the experimental results for simple metals. Both are listed in Table 25.2.

The reason why the measured threshold wavelength (frequency) is always larger (smaller) than the theoretical prediction can also be traced back to the

Table 25.2. Calculated and measured threshold wavelengths for the transparency of alkali metals

Metal	Li	Na	K	Rb	Cs
λ_{theor} (Å)	1500	2100	2900	3200	3600
λ_{exp} (Å)	2050	2100	3150	3600	4400

effects of core electrons. If their effects are lumped into the dielectric constant $\epsilon_{\text{r}}(\infty)$, which is thus larger than unity, the presence of core electrons leads to

$$\omega_{\text{p}} = \frac{\omega_{\text{p}}^0}{\sqrt{\epsilon_{\text{r}}(\infty)}}\,, \tag{25.1.56}$$

where ω_{p}^0 is the free-electron value of the plasma frequency determined from (16.1.69).

According to our previous calculation, the frequency ω_{p} marks a sharp boundary between the totally reflecting and transmitting regions. This result was the consequence of neglecting the scattering of electrons, which were taken into account through a relaxation time τ in the Drude model. This adds a damping term to the equation of motion:

$$m_{\text{e}}\ddot{\boldsymbol{r}} = -e\boldsymbol{E} - \frac{1}{\tau}m_{\text{e}}\dot{\boldsymbol{r}}\,, \tag{25.1.57}$$

or, for oscillations of angular frequency ω,

$$-\left(\omega^2 + \frac{\mathrm{i}\omega}{\tau}\right)m_{\text{e}}\boldsymbol{r} = -e\boldsymbol{E}\,. \tag{25.1.58}$$

Combined with (25.1.52), the equation for the dispersion relation of transverse oscillations is

$$q^2 = \frac{\omega^2}{c^2} - \frac{\omega^2}{c^2}\frac{n_{\text{e}}e^2}{m_{\text{e}}\epsilon_0}\frac{1}{\omega^2 + \mathrm{i}\omega/\tau}\,, \tag{25.1.59}$$

which can also be written as

$$q^2 = \left(\frac{\omega}{c}\right)^2\left[1 + \mathrm{i}\frac{\sigma_0}{\epsilon_0\omega}\frac{1}{1 - \mathrm{i}\omega\tau}\right] = \left(\frac{\omega}{c}\right)^2\left[1 - \frac{\omega_{\text{p}}^2}{\omega^2 + \mathrm{i}\omega/\tau}\right]. \tag{25.1.60}$$

The transverse dielectric constant for the free-electron gas is then

$$\epsilon_{\text{r}} = 1 + \mathrm{i}\frac{\sigma_0}{\epsilon_0\omega}\frac{1}{1 - \mathrm{i}\omega\tau} = 1 - \frac{\omega_{\text{p}}^2}{\omega^2 + \mathrm{i}\omega/\tau}\,. \tag{25.1.61}$$

This formula is the same as (16.1.70), obtained for the longitudinal dielectric constant – in agreement with the generally valid observation that the longitudinal and transverse dielectric constants are the same in the long-wavelength limit.

Using, once again, the real and imaginary parts of the complex refractive index instead of the dielectric constant,

$$(n+\mathrm{i}\kappa)^2 = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + \mathrm{i}\omega/\tau}\,, \tag{25.1.62}$$

and hence

$$\epsilon_1 = n^2 - \kappa^2 = 1 - \frac{\omega_{\mathrm{p}}^2\tau^2}{1+\omega^2\tau^2}\,, \qquad \epsilon_2 = 2n\kappa = \frac{\omega_{\mathrm{p}}^2\tau}{\omega(1+\omega^2\tau^2)}\,. \tag{25.1.63}$$

When the frequency dependence of n and κ are determined from these equations, the reflectance (25.1.36) can also be specified at different wavelengths. This generalization of the Drude model to the description of optical properties is called the *Drude–Zener model.* Figure 25.2 shows the frequency dependence of the real and imaginary parts of the dielectric constant, the optical constants (n, κ), as well as the reflectance (R) for typical values of ω_{p} and τ.

Two characteristic, frequency-like quantities appeared in the description of the properties of the electron gas: the inverse relaxation time $1/\tau$ and the plasma frequency ω_{p}.[7] Based on the magnitude of the frequency of the electromagnetic field relative to these characteristic frequencies, three typical regions of the optical behavior can be distinguished.

In the low-frequency region, where $\omega \ll 1/\tau$ and of course $\omega_{\mathrm{p}} \gg 1/\tau$,

$$\epsilon_{\mathrm{r}} \approx 1 - (\omega_{\mathrm{p}}\tau)^2 + \mathrm{i}\frac{(\omega_{\mathrm{p}}\tau)^2}{\omega\tau}\,. \tag{25.1.64}$$

The real part of the dielectric constant takes large negative values, and the imaginary part is even larger in magnitude, so this portion of the spectrum is called the *absorption region.* Here

$$n^2 - \kappa^2 \approx 1 - \omega_{\mathrm{p}}^2\tau^2 \approx -\omega_{\mathrm{p}}^2\tau^2\,, \qquad 2n\kappa \approx \frac{\omega_{\mathrm{p}}^2\tau}{\omega}\,. \tag{25.1.65}$$

To a fairly good approximation, n and κ are found to be much larger than unity, and of the same order of magnitude:

$$n \approx \kappa \approx \left(\frac{\omega_{\mathrm{p}}^2\tau}{2\omega}\right)^{1/2} = \left(\frac{\sigma_0}{2\epsilon_0\omega}\right)^{1/2} \gg 1\,. \tag{25.1.66}$$

Determined from (25.1.36), the reflectance is close to 100% in this FIR region, however, it shows a characteristic dependence on wavelength:

$$R = 1 - 2\left(\frac{2\omega}{\omega_{\mathrm{p}}^2\tau}\right)^{1/2} = 1 - 2\left(\frac{2\omega\epsilon_0}{\sigma_0}\right)^{1/2} = 1 - 2\left(\frac{4\pi c\epsilon_0}{\lambda\sigma_0}\right)^{1/2}. \tag{25.1.67}$$

[7] We saw that in metals τ is usually on the order of 10^{-14} to 10^{-15} s. According to the estimate obtained for the plasma frequency, $1/\tau$ is usually much smaller than ω_{p}.

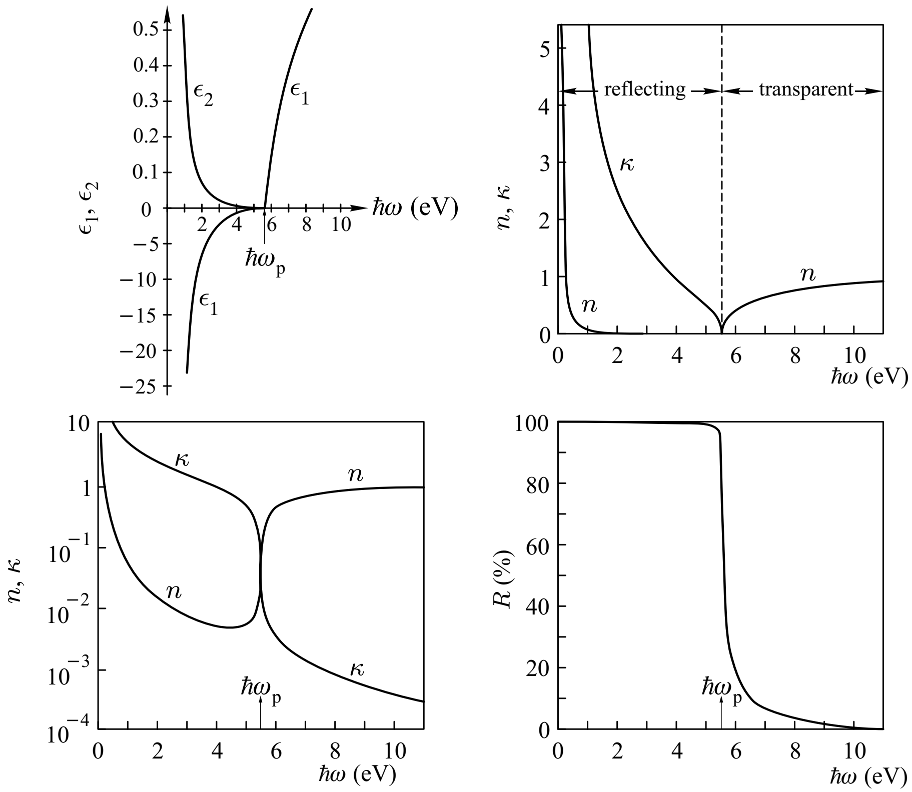


Fig. 25.2. Frequency dependence of the dielectric constant, refractive index, extinction coefficient and reflectivity of a free-electron gas for typical values of ω_p and τ (based on Wooten's book)

This is the *Hagen–Rubens relation*,[8] which is in good agreement with measurements for good conductors (such as gold, silver, and copper) at wavelengths over 30 µm.

The region $1/\tau \ll \omega$, where the period of the oscillating field is much smaller than the relaxation time, is called the *relaxation region*. Here $\omega^2\tau^2$ becomes dominant in the denominator of the dielectric constant (25.1.63):

$$\epsilon_r \approx 1 - \left(\frac{\omega_p}{\omega}\right)^2 + i\frac{\omega_p^2}{\omega^3\tau}, \tag{25.1.68}$$

or alternatively,

$$n^2 - \kappa^2 \approx 1 - \frac{\omega_p^2}{\omega^2}, \qquad 2n\kappa \approx \frac{\omega_p^2}{\omega^3\tau}. \tag{25.1.69}$$

Two regions are distinguished within the relaxation region. When $\omega < \omega_p$, the real part of the dielectric constant is still negative but the imaginary part

[8] E. Hagen and H. Rubens, 1903.

is smaller in magnitude than the real part. To a good approximation,

$$n \approx \frac{\omega_\mathrm{p}}{2\omega^2\tau}\,, \qquad \kappa \approx \frac{\omega_\mathrm{p}}{\omega}\,. \tag{25.1.70}$$

The metal remains strongly reflective throughout this region:

$$R \approx 1 - \frac{2}{\omega_\mathrm{p}\tau}\,, \tag{25.1.71}$$

while the absorption coefficient decreases as $1/\omega^2$.

When $\omega > \omega_\mathrm{p}$, which corresponds to ultraviolet frequencies, the real part of the dielectric constant becomes positive, and the extinction coefficient is small,

$$n \approx \sqrt{1 - \left(\frac{\omega_\mathrm{p}}{\omega}\right)^2} \approx 1\,, \qquad \kappa \approx \frac{\omega_\mathrm{p}^2}{2\omega^3\tau} \approx 0\,. \tag{25.1.72}$$

The reflectance is very small, on the order of a few percent, and the metal becomes transparent.

25.1.5 Reflectivity of Semiconductors

The Drude–Zener theory presented above does not only provide an approximate explanation for the optical properties of simple metals: through a slight generalization it also gives an adequate account of the wavelength dependence of the reflectance of strongly doped semiconductors. For the latter, it must be borne in mind that the polarizability of ion cores cannot be ignored in semiconductors; in fact $\epsilon_\mathrm{r}(\infty)$ can be much larger than unity. Instead of (25.1.63), the formulas

$$\epsilon_1 = n^2 - \kappa^2 = \epsilon_\mathrm{r}(\infty) - \frac{\omega_\mathrm{p}^2\tau^2}{1+\omega^2\tau^2}\,, \quad \epsilon_2 = 2n\kappa = \frac{\omega_\mathrm{p}^2\tau}{\omega(1+\omega^2\tau^2)} \tag{25.1.73}$$

have to be used for the real and imaginary parts of the dielectric constant.

The other important difference with metals is the magnitude of the relaxation time. The mean free path of electrons is of the same order in metals and semiconductors but the typical electron velocities are different: it is the Fermi velocity v_F in the former, while in the latter, where electrons can be treated classically, it is the thermal velocity – which is several orders of magnitude smaller than v_F. Therefore the relaxation time is several orders of magnitude larger in semiconductors than in metals. Consequently, the condition $\omega\tau \gg 1$ is met in the infrared region, and the absorption region ($\omega \ll 1/\tau$) is pushed out of the optical region. For the entire optical region

$$n^2 - \kappa^2 = \epsilon_\mathrm{r}(\infty) - \frac{\omega_\mathrm{p}^2}{\omega^2}\,, \qquad 2n\kappa = \frac{\omega_\mathrm{p}^2}{\omega^3\tau}\,. \tag{25.1.74}$$

At $\omega = \omega_\mathrm{p}/\sqrt{\epsilon_\mathrm{r}(\infty)}$

$$n = \kappa = \frac{\epsilon_\mathrm{r}(\infty)}{2\omega\tau} \ll 1 , \tag{25.1.75}$$

so the reflectance is still close to unity. At the slightly higher but not too distant frequency $\omega = \omega_\mathrm{p}/\sqrt{\epsilon_\mathrm{r}(\infty) - 1}$ we have $n \approx 1$ and $\kappa \ll 1$, and so the reflectance is small. This sharp drop in the reflectivity is shown in Fig. 25.3 for various values of the dopant concentration. This is the analog of the plasma edge for doped semiconductors in which the concentration of carriers in the conduction or valence band is sufficiently high for that they can be treated as an electron gas.

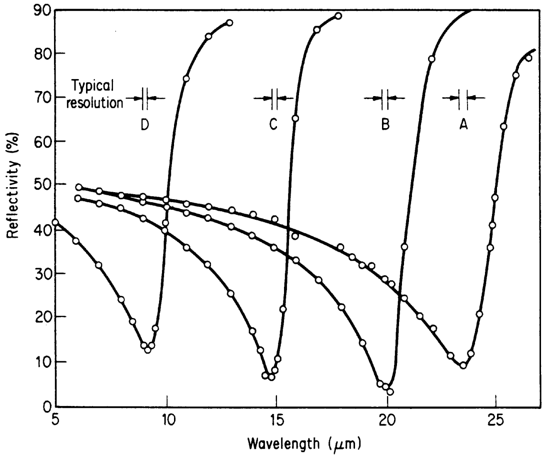


Fig. 25.3. Reflectivity of p-type PbTe samples as a function of the wavelength [Reprinted with permission from J. R. Dixon and H. R. Riedl, *Phys. Rev.* **138**, A873 (1965). ©1965 by the American Physical Society]

At even higher frequencies the dominant contribution to the dielectric constant comes from core electrons, and thus, from an optical point of view, the semiconductor behaves as a classical dielectric.

25.1.6 Interaction of Light with Bound Electrons

The results obtained in the Drude–Zener theory can be applied when only the effects of the electromagnetic field on the more or less free electrons of the conduction and valence bands have to be considered – that is, for simple metals and semiconductors. This condition is obviously not met in insulators. In the latter the interaction of light with bound electrons has to be studied. The simplest, classical discussion is based on the Lorentz model.

In this model the bound states of electrons are described in terms of classical harmonic oscillators of angular frequency ω_0. Including the corresponding term in the equation of motion (25.1.57) for electrons, we have

$$m_{\mathrm{e}}\ddot{\boldsymbol{r}} + m_{\mathrm{e}}\omega_0^2\boldsymbol{r} = -e\boldsymbol{E} - \frac{1}{\tau}m_{\mathrm{e}}\dot{\boldsymbol{r}} \,. \tag{25.1.76}$$

The equation for the Fourier component of frequency ω reads

$$m_{\mathrm{e}}\left(-\omega^2 + \omega_0^2 - \mathrm{i}\omega/\tau\right)\boldsymbol{r} = -e\boldsymbol{E} \,, \tag{25.1.77}$$

and its solution is

$$\boldsymbol{r} = -\frac{e}{m_{\mathrm{e}}}\left(\omega_0^2 - \omega^2 - \mathrm{i}\omega/\tau\right)^{-1}\boldsymbol{E} \,. \tag{25.1.78}$$

The induced dipole moment, and then the polarizability can be determined from the displacement $\boldsymbol{r}$:

$$\alpha(\omega) = \frac{-e\boldsymbol{r}}{\boldsymbol{E}} = \frac{e^2}{m_{\mathrm{e}}}\left(\omega_0^2 - \omega^2 - \mathrm{i}\omega/\tau\right)^{-1} . \tag{25.1.79}$$

When a sample of volume V contains N_{e} bound electrons with the same eigenfrequency ω_0, the polarization is

$$\boldsymbol{P} = \frac{n_{\mathrm{e}}e^2}{m_{\mathrm{e}}}\left(\omega_0^2 - \omega^2 - \mathrm{i}\omega/\tau\right)^{-1}\boldsymbol{E} \,, \tag{25.1.80}$$

and the dielectric constant is

$$\epsilon_{\mathrm{r}} = 1 + \frac{n_{\mathrm{e}}e^2}{\epsilon_0 m_{\mathrm{e}}}\frac{1}{\omega_0^2 - \omega^2 - \mathrm{i}\omega/\tau} = 1 + \frac{\omega_{\mathrm{p}}^2}{\omega_0^2 - \omega^2 - \mathrm{i}\omega/\tau} \,. \tag{25.1.81}$$

In the $\omega_0 \to 0$ limit, where bound electrons become free, the earlier results are recovered. It is obvious from the calculation that if ω is not the same for each electron but for N_j of them the eigenfrequency is ω_j and the relaxation time is τ_j, we have

$$\epsilon_{\mathrm{r}} = 1 + \frac{e^2}{\epsilon_0 m_{\mathrm{e}}}\frac{1}{V}\sum_j \frac{N_j}{\omega_j^2 - \omega^2 - \mathrm{i}\omega/\tau_j} \,. \tag{25.1.82}$$

When there is a single eigenfrequency, the real and imaginary parts of the dielectric constant are

$$\begin{aligned}\epsilon_1 &= 1 + \frac{\omega_{\mathrm{p}}^2(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + (\omega/\tau)^2} \,, \\ \epsilon_2 &= \frac{\omega_{\mathrm{p}}^2\omega/\tau}{(\omega_0^2 - \omega^2)^2 + (\omega/\tau)^2} \,.\end{aligned} \tag{25.1.83}$$

The optical constants n and κ are then given by

$$\begin{aligned}n &= \left\{\tfrac{1}{2}\left[(\epsilon_1^2 + \epsilon_2^2)^{1/2} + \epsilon_1\right]\right\}^{1/2} , \\ \kappa &= \left\{\tfrac{1}{2}\left[(\epsilon_1^2 + \epsilon_2^2)^{1/2} - \epsilon_1\right]\right\}^{1/2} .\end{aligned} \tag{25.1.84}$$

In contrast to free electrons, where the two characteristic frequencies $1/\tau$ and ω_{p} divided the optical frequency range into three parts, four parts are distinguished for bound electrons, as ω_0 is now finite.

At low frequencies $\omega \ll \omega_0$

$$\epsilon_1 \approx 1 + \frac{\omega_{\mathrm{p}}^2}{\omega_0^2} \gg 1\,, \tag{25.1.85}$$

while ϵ_2 is close to zero, so $n \approx \sqrt{\epsilon_1} > 1$, and $\kappa \approx 0$. In this region the insulator hardly absorbs any light, and its reflectance is small, so the material is transparent.

In a region of width $2/\tau$ around ω_0, where both n and κ may take large values, both absorption and reflectivity are significant. The part of the light that is not reflected by the sample is absorbed.

For $\omega \gg \omega_0$ the electrons of the insulator behave as if they were free. As long as $\epsilon_1 < 0$, the good reflectivity of metals is observed. Of course, this occurs well beyond the visible region, in the ultraviolet.

Finally, at very high frequencies ($\omega \gg \omega_{\mathrm{p}}$), ϵ_1 becomes positive. This occurs above $\omega_{\mathrm{L}} = \sqrt{\omega_0^2 + \omega_{\mathrm{p}}^2}$ in the $\tau \to \infty$ limit. The reflectance is small, and the insulator becomes transparent again.

The frequency dependence of the real and imaginary parts of the dielectric constant and of the optical constants n, κ are plotted in Fig. 25.4 for typical values of ω_0, ω_{p}, and τ, while the frequency dependence of the reflectance R in the four regions is shown in Fig. 25.5.

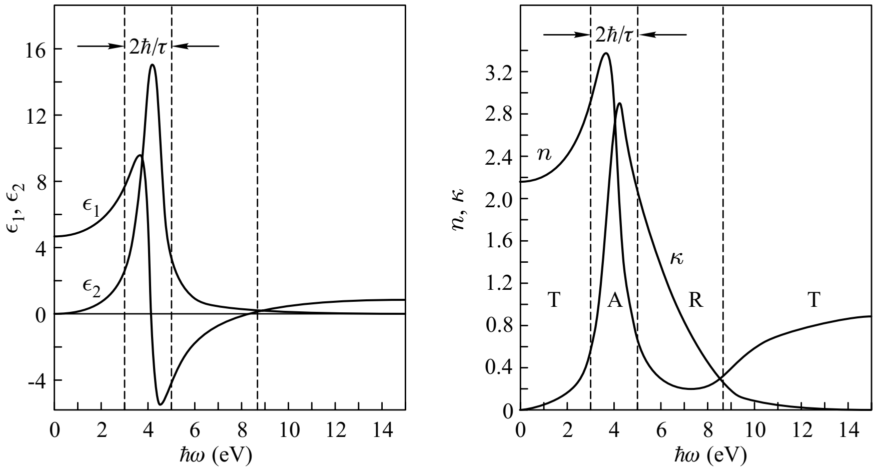


Fig. 25.4. Frequency dependence of the real and imaginary parts of the dielectric constant and of the complex refractive index in the Lorentz model for typical values of ω_0, ω_{p}, and τ (based on Wooten's book)

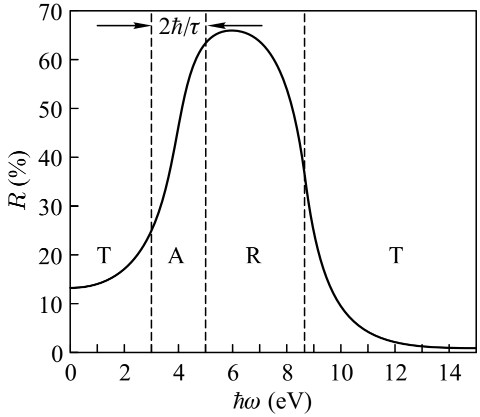


Fig. 25.5. Frequency dependence of the reflectance in the Lorentz model (based on Wooten's book)

25.1.7 Absorption and Dispersion in Ionic Crystals

Up to now we have studied how the electromagnetic field polarizes the system of electrons if the electrons can be considered free or are in bound states. In ionic crystals we have to take into account the additional polarization due to the optical vibrations of the lattice and the corresponding contribution to the dielectric constant. To understand the role of optical phonons, we shall first examine the diatomic chain discussed in Chapter 11. If the two ions are oppositely charged, the equations of motion (11.2.17) for the two kinds of atom,

$$\begin{aligned} M_1\ddot{u}_n &= -K\left[2u_n - v_n - v_{n-1}\right], \\ M_2\ddot{v}_n &= -K\left[2v_n - u_{n+1} - u_n\right], \end{aligned} \tag{25.1.86}$$

must be complemented by a term that accounts for the effects of the electromagnetic field. Taking the spatial Fourier transforms, in the long-wavelength (small q) limit, where the phase factors $\mathrm{e}^{\pm \mathrm{i}qa}$ can be approximated by unity,

$$\begin{aligned} M_1\ddot{u}_q &= -\,2K(u_q - v_q) + eE\mathrm{e}^{-\mathrm{i}\omega t}, \\ M_2\ddot{v}_q &= -\,2K(v_q - u_q) - eE\mathrm{e}^{-\mathrm{i}\omega t}. \end{aligned} \tag{25.1.87}$$

After some algebra,

$$\ddot{u}_q - \ddot{v}_q = \left[-K(u_q - v_q) + eE\mathrm{e}^{-\mathrm{i}\omega t}\right]\left(\frac{1}{M_1} + \frac{1}{M_2}\right). \tag{25.1.88}$$

Introducing the frequency ω_{TO} of long-wavelength transverse optical modes through

$$\omega_{\mathrm{TO}}^2 = K\left(\frac{1}{M_1} + \frac{1}{M_2}\right), \tag{25.1.89}$$

and assuming a harmonic time dependence, the solution of the equations of motions is

$$\left(\omega_{\mathrm{TO}}^2 - \omega^2\right)(u_q - v_q) = \frac{eE\mathrm{e}^{-\mathrm{i}\omega t}}{K}\omega_{\mathrm{TO}}^2 \,. \tag{25.1.90}$$

The polarizability of the lattice per primitive cell is then

$$e(u_q - v_q) = \frac{e^2}{K}\frac{\omega_{\mathrm{TO}}^2}{\omega_{\mathrm{TO}}^2 - \omega^2} \,. \tag{25.1.91}$$

If this were substituted directly into the dielectric constant, the resulting formula would contain the spring constant K. To avoid that, we shall write the dielectric constant as

$$\epsilon_{\mathrm{r}}(\omega) = a + b\,\frac{\omega_{\mathrm{TO}}^2}{\omega_{\mathrm{TO}}^2 - \omega^2}\,, \tag{25.1.92}$$

and express the two parameters in terms of the values of the dielectric constant at $\omega = 0$ and in the $\omega \to \infty$ limit. Since $\epsilon_{\mathrm{r}}(0) = a + b$ and $\epsilon_{\mathrm{r}}(\infty) = a$,

$$\epsilon_{\mathrm{r}}(\omega) = \epsilon_{\mathrm{r}}(\infty) + [\epsilon_{\mathrm{r}}(0) - \epsilon_{\mathrm{r}}(\infty)]\frac{\omega_{\mathrm{TO}}^2}{\omega_{\mathrm{TO}}^2 - \omega^2} \,. \tag{25.1.93}$$

Note that the dielectric constant is always real in this approximation, and for very small and very large values of ω it is positive. In between there is a frequency range,

$$\omega_{\mathrm{TO}} < \omega < \omega_{\mathrm{LO}}\,, \tag{25.1.94}$$

where

$$\omega_{\mathrm{LO}}^2 = \frac{\epsilon_{\mathrm{r}}(0)}{\epsilon_{\mathrm{r}}(\infty)}\,\omega_{\mathrm{TO}}^2\,, \tag{25.1.95}$$

in which the dielectric constant is negative and the refractive index vanishes, thus the crystal becomes perfectly reflective. Expressing $\epsilon_{\mathrm{r}}(\omega)$, as given in (25.1.93), in terms of ω_{LO},

$$\epsilon_{\mathrm{r}}(\omega) = \epsilon_{\mathrm{r}}(\infty)\,\frac{\omega_{\mathrm{LO}}^2 - \omega^2}{\omega_{\mathrm{TO}}^2 - \omega^2} \,. \tag{25.1.96}$$

The subscript “LO” of ω_{LO} is not a coincidence: according to the Lyddane–Sachs–Teller relation, this is just the frequency of longitudinal optical oscillations. The expression also shows that $\epsilon_{\mathrm{r}}(\omega)$ vanishes at this frequency, thus longitudinal modes can propagate in the system.

The fact that the reflectance is unity between ω_{TO} and ω_{LO} indicates that oscillations of such frequencies cannot propagate in ionic crystals. To determine the frequencies of the allowed modes, we have to return to (25.1.16) and (25.1.20):

$$q^2 = \frac{\omega^2}{c^2}\epsilon_{\mathrm{r}}(\omega)\,. \tag{25.1.97}$$

Substituting (25.1.96) into (25.1.97), the solutions of this equation for ω as a function of q determine the dispersion relation of the vibrations of the ionic crystal. The equation is quadratic in ω^2, and its solutions are

$$\omega^2(q) = \tfrac{1}{2}\left(\frac{c^2q^2}{\epsilon_{\mathrm{r}}(\infty)} + \omega_{\mathrm{LO}}^2\right) \pm \tfrac{1}{2}\left[\left(\frac{c^2q^2}{\epsilon_{\mathrm{r}}(\infty)} + \omega_{\mathrm{LO}}^2\right)^2 - 4\frac{c^2q^2}{\epsilon_{\mathrm{r}}(\infty)}\omega_{\mathrm{TO}}^2\right]^{1/2}. \tag{25.1.98}$$

They are shown graphically in Fig. 25.6.

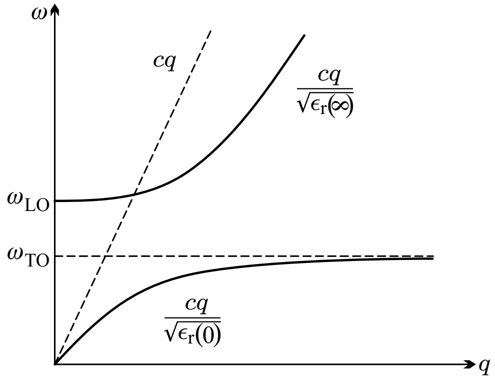


Fig. 25.6. The dispersion relation of polaritons

At small values of q one of the two branches has a light-like linear dispersion relation with a propagation velocity $c/\sqrt{\epsilon_{\mathrm{r}}(0)}$, as expected in a medium of refractive index $n = \sqrt{\epsilon_{\mathrm{r}}(0)}$, whereas the other branch corresponds to longitudinal optical modes. At larger values of q the two modes get hybridized, and become a mixture of electromagnetic and lattice vibrations. These hybrid modes are called *polaritons.*[9] At even higher frequencies we find again a mode with a linear dispersion curve and another mode whose frequency is independent of the wavelength, however, the propagation velocity for the former is $c/\sqrt{\epsilon_{\mathrm{r}}(\infty)}$, while the frequency of the latter is somewhat lower than that of transverse optical vibrations. Note that there are no modes with a frequency in the $\omega_{\mathrm{TO}} < \omega < \omega_{\mathrm{LO}}$ range. Consequently, the sample totally reflects radiations of such frequencies.

In the previous calculation we neglected damping forces, which render the lifetime of phonons finite. Just like in the Lorentz model, their inclusion in the denominator of the dielectric constant leads to the appearance of an additional term $\mathrm{i}\omega\gamma$, where γ is the inverse lifetime. Therefore

[9] J. J. Hopfield, 1958.

$$\epsilon_{\mathrm{r}}(\omega) = \epsilon_{\mathrm{r}}(\infty) + [\epsilon_{\mathrm{r}}(0) - \epsilon_{\mathrm{r}}(\infty)] \frac{\omega_{\mathrm{TO}}^2}{\omega_{\mathrm{TO}}^2 - \omega^2 - \mathrm{i}\omega\gamma} . \tag{25.1.99}$$

As shown in Fig. 25.7, the frequency dependence of the real and imaginary parts

$$\begin{aligned} \epsilon_1(\omega) &= \epsilon_{\mathrm{r}}(\infty) + [\epsilon_{\mathrm{r}}(0) - \epsilon_{\mathrm{r}}(\infty)] \frac{\omega_{\mathrm{TO}}^2(\omega_{\mathrm{TO}}^2 - \omega^2)}{(\omega_{\mathrm{TO}}^2 - \omega^2)^2 + \omega^2\gamma^2} , \\ \epsilon_2(\omega) &= [\epsilon_{\mathrm{r}}(0) - \epsilon_{\mathrm{r}}(\infty)] \frac{\omega_{\mathrm{TO}}^2 \omega\gamma}{(\omega_{\mathrm{TO}}^2 - \omega^2)^2 + \omega^2\gamma^2} \end{aligned} \tag{25.1.100}$$

is very similar to that of the Lorentz model (shown in Fig. 25.4). The imaginary part ϵ_2 has its maximum at ω_{TO}. This can be considered as a measurement instruction. The frequency of transverse oscillations is defined by the maximum of the imaginary part of the dielectric constant.

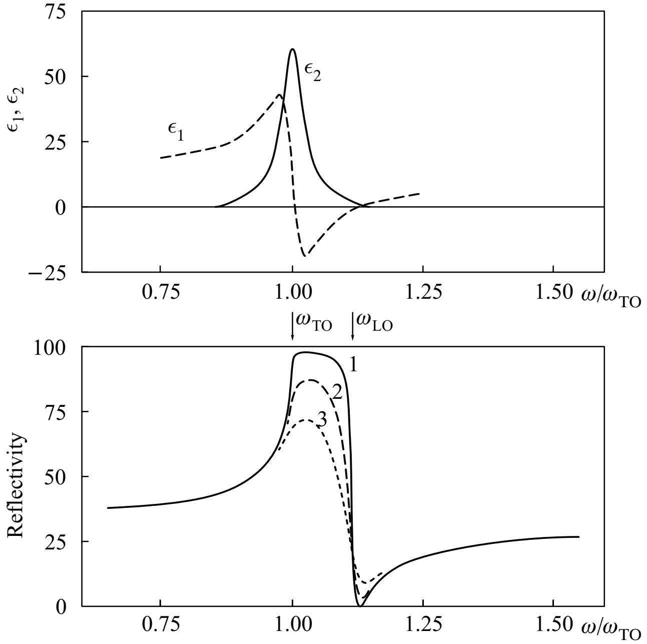


Fig. 25.7. The contribution of optical phonons to the dielectric constant and the frequency dependence of the reflectivity

The lower part of Fig. 25.7 shows the frequency dependence of the reflectance. When γ is finite, the reflection between ω_{TO} and ω_{LO} is no longer total, however, the reflectance can remain close to unity. Figure 25.8 presents the reflectivity data for aluminum antimonide. The theoretical curves fit well with the experimental results.

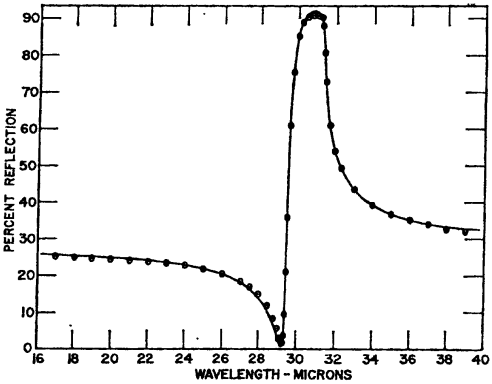


Fig. 25.8. Wavelength dependence of the reflectance of AlSb and the theoretical curve fitted to the experimental data [Reprinted with permission from W. J. Turner and W. E. Reese, *Phys. Rev.* **127**, 126 (1962). ©1962 by the American Physical Society]

Making use of the Kramers–Kronig relation for the dielectric constant,

$$\epsilon_1(\omega) = \epsilon_r(\infty) + \frac{2}{\pi} \mathrm{P} \int_0^\infty \frac{\omega' \epsilon_2(\omega')}{\omega'^2 - \omega^2} \, \mathrm{d}\omega' . \tag{25.1.101}$$

This formula clearly shows that the difference between the static and optical dielectric constants,

$$\epsilon_1(0) - \epsilon_r(\infty) = \frac{2}{\pi} \int_0^\infty \frac{\epsilon_2(\omega')}{\omega'} \, \mathrm{d}\omega' , \tag{25.1.102}$$

is due to the oscillators of finite frequency. Since the ions, bound electrons, and relatively free electrons in a solid can be associated with oscillators of highly disparate frequencies, their contributions to the dielectric constant can be simply added. At low frequencies only the contribution of the conduction electrons need to be considered. The contribution of ions gives a characteristic dispersive curve at about the optical phonon frequency. At even higher frequencies the dielectric constant seems to become saturated at a value above unity. If only frequencies up to this range were considered, this value could be taken as $\epsilon_r(\infty)$. At still higher frequencies the dispersive contribution of bound electrons can also be observed, and $\epsilon_r(\omega)$ reaches its true asymptotic value only after that. Figure 25.9 shows the frequency dependence of the real part of the dielectric constant schematically; the characteristic frequency is assumed to be 10^{12} Hz for optical phonons and 10^{15} Hz for the excitation of bound electrons. Sharp changes are observed at these frequencies. It should, nonetheless, borne in mind that the entire spectrum cannot be measured in any single experiment: we can only measure it up to a characteristic frequency ω_c, which is usually smaller than the frequency of interband transitions.

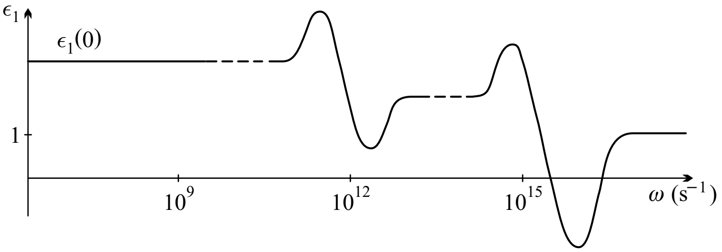


Fig. 25.9. Frequency dependence of the real part of the dielectric constant obtained by summing the contributions of electrons and lattice vibrations

25.2 Quantum Mechanical Treatment

In the previous section we determined the polarizability and dielectric constant of solids for a classically treated electron system and a classical lattice. In the quantum mechanical treatment the interaction between the electromagnetic field and a solid can often be described by the Hamiltonians

$$\mathcal{H}_{\text{int}} = -\int \boldsymbol{E}(\boldsymbol{r}) \cdot \boldsymbol{P}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{25.2.1}$$

or

$$\mathcal{H}_{\text{int}} = -\int \boldsymbol{B}(\boldsymbol{r}) \cdot \boldsymbol{M}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\,, \tag{25.2.2}$$

which correspond to the dipole approximation. Whether one or the other is used depends on the relative strength of the couplings between the electric field $\boldsymbol{E}$ and the polarization $\boldsymbol{P}$ of charged particles and between the magnetic induction $\boldsymbol{B}$ and the magnetic moment $\boldsymbol{M}$. Just like in the classical treatment, we shall focus on electrically polarizable materials, and omit the discussion of the optical properties of magnetic materials. First, we shall demonstrate that the results obtained from the quantum mechanical analysis of electronic states are very similar to the classical ones. Then we shall quantize the electromagnetic field, too, and describe the interaction of light and matter in terms of scattering between photons and the elementary excitations of the solid.

25.2.1 Dielectric Constant of the System of Electrons

Below we shall represent electrons by their wavefunctions, as usual for a quantum mechanical description, but we shall continue to treat the electromagnetic field classically for the time being. We shall assume that at $t = -\infty$ the core electrons are in the ground state of energy E_0 described by the wavefunction ϕ_0, and then an x-directed electric field of frequency ω is turned on adiabatically:

$$\boldsymbol{E}(t) = \tfrac{1}{2} E_x \hat{\boldsymbol{x}} \left(\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t} \right) \mathrm{e}^{-\delta |t|} \,. \tag{25.2.3}$$

The adiabaticity is included in the factor $\mathrm{e}^{-\delta|t|}$, where δ is infinitesimally small. Because of the applied field, the ground-state wavefunction becomes mixed with the wavefunctions of excited states. Denoting the wavefunctions of the allowed excited states by ϕ_j and the excitation energies by $\hbar\omega_{j0} = E_j - E_0$, the wavefunction of the perturbed state can be sought in the form

$$\psi(\boldsymbol{r}, t) = c_0(t)\phi_0(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}E_0 t/\hbar} + \sum_j c_j(t)\phi_j(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}E_j t/\hbar} \,, \tag{25.2.4}$$

where the coefficients $c_j(t)$ can be determined using the formulas of time-dependent perturbation theory. Up to linear order in the interaction,

$$-\frac{\hbar}{\mathrm{i}} \frac{\mathrm{d}c_j(t)}{\mathrm{d}t} = \int \phi_j^*(\boldsymbol{r}) \mathcal{H}_{\text{int}}(t) \phi_0(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}(E_j - E_0)t/\hbar} \, \mathrm{d}\boldsymbol{r} \,. \tag{25.2.5}$$

In the dipole approximation, where effects of the magnetic field are neglected, the perturbation Hamiltonian is

$$\mathcal{H}_{\text{int}}(t) = e\boldsymbol{E}(t) \cdot \boldsymbol{r} \,. \tag{25.2.6}$$

Since the time dependence appears in exponential factors, the integration of the differential equation (25.2.5) is straightforward. For times t when the field is already turned on,

$$\begin{aligned} c_j(t) &= -\tfrac{1}{2} \frac{\mathrm{i}}{\hbar} \int\limits_{-\infty}^{t} e E_x x_{j0} \left(\mathrm{e}^{\mathrm{i}\omega t'} + \mathrm{e}^{-\mathrm{i}\omega t'} \right) \mathrm{e}^{-\delta|t'|} \mathrm{e}^{\mathrm{i}(E_j - E_0)t'/\hbar} \, \mathrm{d}t' \\ &= -\tfrac{1}{2} e E_x x_{j0} \left\{ \frac{\mathrm{e}^{\mathrm{i}(\hbar\omega + E_j - E_0)t/\hbar}}{\hbar\omega + (E_j - E_0) - \mathrm{i}\delta} + \frac{\mathrm{e}^{\mathrm{i}(-\hbar\omega + E_j - E_0)t/\hbar}}{-\hbar\omega + (E_j - E_0) - \mathrm{i}\delta} \right\} \\ &= -\tfrac{1}{2} \frac{e E_x}{\hbar} x_{j0} \left\{ \frac{\mathrm{e}^{\mathrm{i}(\omega + \omega_{j0})t}}{\omega_{j0} + \omega - \mathrm{i}\delta} + \frac{\mathrm{e}^{\mathrm{i}(-\omega + \omega_{j0})t}}{\omega_{j0} - \omega - \mathrm{i}\delta} \right\} \end{aligned} \tag{25.2.7}$$

for $j \neq 0$, where x_{j0} is the dipole matrix element, given by

$$x_{j0} = \int \phi_j^*(\boldsymbol{r}) \, x \, \phi_0(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} \,. \tag{25.2.8}$$

The dipole moment of the atom in this state is

$$\begin{aligned}\langle x\rangle &= \int \psi^*(\boldsymbol{r},t)\, x\, \psi(\boldsymbol{r},t)\,\mathrm{d}\boldsymbol{r}\\ &= \sum_j \left\{x_{0j}c_j(t)\mathrm{e}^{-\mathrm{i}\omega_{j0}t} + x_{j0}c_j^*(t)\mathrm{e}^{\mathrm{i}\omega_{j0}t}\right\}\\ &= -\tfrac{1}{2}\frac{eE_x}{\hbar}\sum_j |x_{j0}|^2\left\{\frac{\mathrm{e}^{\mathrm{i}\omega t}}{\omega_{j0}+\omega-\mathrm{i}\delta} + \frac{\mathrm{e}^{-\mathrm{i}\omega t}}{\omega_{j0}-\omega-\mathrm{i}\delta}\right. \qquad (25.2.9)\\ &\qquad\qquad \left. + \frac{\mathrm{e}^{-\mathrm{i}\omega t}}{\omega_{j0}+\omega+\mathrm{i}\delta} + \frac{\mathrm{e}^{\mathrm{i}\omega t}}{\omega_{j0}-\omega+\mathrm{i}\delta}\right\}\\ &= -\tfrac{1}{2}\frac{eE_x}{\hbar}\sum_j |x_{j0}|^2\left\{\frac{2\omega_{j0}}{\omega_{j0}^2-\omega^2+\mathrm{i}\omega\delta}\mathrm{e}^{\mathrm{i}\omega t} + \frac{2\omega_{j0}}{\omega_{j0}^2-\omega^2-\mathrm{i}\omega\delta}\mathrm{e}^{-\mathrm{i}\omega t}\right\}.\end{aligned}$$

The atomic polarizability can be determined from the polarization $P = -e\langle x\rangle$. The real part of the polarizability comes from the contribution that has the same time dependence $\cos\omega t$ as the perturbing field. There is, however, an additional part, which is proportional to $\sin\omega t$ and so contains a phase shift; it is related to the absorbed energy, and can be interpreted as the imaginary part of the polarizability. This is very similar to how (3.2.99) and (3.2.101) were interpreted as the real and imaginary parts of the susceptibility in the discussion of paramagnetic resonance. Thus,

$$\alpha(\omega) = \sum_j \frac{e^2|x_{j0}|^2}{\epsilon_0\hbar}\frac{2\omega_{j0}}{\omega_{j0}^2-\omega^2-\mathrm{i}\omega\delta}\,. \qquad (25.2.10)$$

If the volume V contains N atoms, the real part of the dielectric constant is

$$\epsilon_\mathrm{r}(\omega) = 1 + \frac{N}{V}\alpha(\omega) = 1 + \frac{N}{V}\sum_j \frac{e^2|x_{j0}|^2}{\epsilon_0\hbar}\frac{2\omega_{j0}}{\omega_{j0}^2-\omega^2}\,. \qquad (25.2.11)$$

This formula is very similar to the result obtained in the $\tau \to \infty$ limit for bound electrons treated as classical oscillators. Using the same factors as in (25.1.81) and (25.1.82), the dielectric constant is customarily written as

$$\epsilon_\mathrm{r}(\omega) = 1 + \frac{e^2}{\epsilon_0 m_\mathrm{e}}\frac{N}{V}\sum_j \frac{f_j}{\omega_{j0}^2-\omega^2}\,, \qquad (25.2.12)$$

where

$$f_j = \frac{2m_\mathrm{e}}{\hbar^2}\hbar\omega_{j0}|x_{j0}|^2\,. \qquad (25.2.13)$$

On account of the analogy, this quantity is called the *oscillator strength* of the transition from the ground state to the jth excited state. It can be proved that the oscillator strengths obey a sum rule,[10]

[10] This formula is called *f-sum rule* because of the usual notation f of the oscillator strength.

$$\sum_j f_j = Z \,, \tag{25.2.14}$$

where Z is the number of excitable electrons per atom.

Since an infinite lifetime was assumed for the electron states in the foregoing, a Dirac delta peak is obtained for the imaginary part of the dielectric constant. Absorption occurs at those frequencies ω that correspond the frequency of a transition from the ground state to an excited state. However, absorption itself always leads to some broadening since the probability that the electron is in the ground state decreases exponentially, while the occupation of excited states increases. Then the inverse lifetime of the states is given by a finite $\varGamma_j$ rather than an infinitesimal δ. Consequently,

$$\epsilon_{\mathrm{r}}(\omega) = 1 + \frac{e^2}{\epsilon_0 m_{\mathrm{e}}} \frac{N}{V} \sum_j \frac{f_j}{\omega_{j0}^2 - \omega^2 - \mathrm{i}\omega\varGamma_j} \,. \tag{25.2.15}$$

The above considerations apply to bound electrons. In a system of free electrons, photons of wave vector $\boldsymbol{q}$ and frequency ω can induce transitions in which an electron of wave vector $\boldsymbol{k}$ inside the Fermi sphere is scattered to a state $\boldsymbol{k}+\boldsymbol{q}$ outside the Fermi sphere. The contribution of these processes to the dielectric constant will be determined in Chapter 29. We just quote the result here:

$$\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega) = 1 - \frac{e^2}{\epsilon_0 q^2} \frac{2}{V} \sum_{\boldsymbol{k}} \frac{f_0(\varepsilon_{\boldsymbol{k}}) - f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})}{\hbar\omega - \varepsilon_{\boldsymbol{k}+\boldsymbol{q}} + \varepsilon_{\boldsymbol{k}} + \mathrm{i}\delta} \,. \tag{25.2.16}$$

The imaginary part of the dielectric constant,

$$\epsilon_2(\boldsymbol{q},\omega) = \pi \frac{e^2}{\epsilon_0 q^2} \frac{2}{V} \sum_{\boldsymbol{k}} \left[f_0(\varepsilon_{\boldsymbol{k}}) - f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})\right] \delta(\hbar\omega - \varepsilon_{\boldsymbol{k}+\boldsymbol{q}} + \varepsilon_{\boldsymbol{k}}) \,, \tag{25.2.17}$$

which is related to absorption, is simple to interpret. The electric field of wave vector $\boldsymbol{q}$ can excite an electron of wave vector $\boldsymbol{k}$ of the Fermi sea to a state $\boldsymbol{k}+\boldsymbol{q}$ as long as the absorption of the photon can supply the required energy. The Fermi functions appear because the electron has to move from an occupied state to an initially empty one.

Using the substitution $\boldsymbol{k} \to -\boldsymbol{k}-\boldsymbol{q}$ in the term that contains $f_0(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}})$, the real part takes the form

$$\epsilon_1(\boldsymbol{q},\omega) = 1 + \frac{e^2}{\epsilon_0 q^2} \frac{2}{V} \sum_{\boldsymbol{k}} \frac{2 f_0(\varepsilon_{\boldsymbol{k}})(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{\boldsymbol{k}})}{(\varepsilon_{\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{\boldsymbol{k}})^2 - (\hbar\omega)^2} \,. \tag{25.2.18}$$

This can be easily generalized to the case where Bloch electrons moving in a periodic potential are considered and transitions between different bands (interband transition) have to be taken into account, too. The only difference is the appearance of the matrix element of the operator $\exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r})$ between the initial and final states of the electron:

$$\epsilon_{\mathrm{r}}(\boldsymbol{q},\omega) = 1 + \frac{e^2}{\epsilon_0 q^2}\frac{2}{V}\sum_{nn'\boldsymbol{k}} \frac{|\langle n\boldsymbol{k}|\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}|n'\boldsymbol{k}+\boldsymbol{q}\rangle|^2 2f_0(\varepsilon_{n\boldsymbol{k}})\,(\varepsilon_{n'\boldsymbol{k}+\boldsymbol{q}}-\varepsilon_{n\boldsymbol{k}})}{(\varepsilon_{n'\boldsymbol{k}+\boldsymbol{q}}-\varepsilon_{n\boldsymbol{k}})^2-(\hbar\omega)^2}\,, \tag{25.2.19}$$

where n and n' are band indices.

Since the wave vector associated with the transitions stimulated by light is $\boldsymbol{q} \approx 0$, the frequency-dependent dielectric function can now be written as

$$\epsilon_{\mathrm{r}}(\omega) = 1 + \frac{e^2}{\epsilon_0 m_{\mathrm{e}}}\frac{1}{V}\sum_{nn'\boldsymbol{k}} \frac{f_{nn'}(\boldsymbol{k})}{\omega^2_{nn'\boldsymbol{k}}-\omega^2}\,, \tag{25.2.20}$$

where $\hbar\omega_{nn'\boldsymbol{k}}$ is the energy of the transition between the state of energy $\varepsilon_{n\boldsymbol{k}}$ in the nth band and the state with the same wave vector in the n'th band, and $f_{nn'}(\boldsymbol{k})$ is the oscillator strength for this transition; its value is obtained from the comparison with (25.2.19).

The imaginary part

$$\epsilon_2(\omega) = \frac{\pi e^2}{\omega\epsilon_0 m_{\mathrm{e}}}\frac{1}{V}\sum_{nn'\boldsymbol{k}} f_{nn'}(\boldsymbol{k})\delta(\omega-\omega_{nn'\boldsymbol{k}}) \tag{25.2.21}$$

is usually specified in terms of the joint density of states for the two bands. Analogously to the procedure used to define the density of states in Sections 12.2.1 and 17.4.4, the $\boldsymbol{k}$-space sum can be replaced by an energy integral and an integral over the constant-energy surface. The joint density of states for bands n and n' is then defined in analogy to (17.4.26) as

$$\rho_{nn'}(\varepsilon) = \frac{2}{(2\pi)^3}\int\limits_{S(\varepsilon)} \frac{\mathrm{d}S}{|\boldsymbol{\nabla}_{\boldsymbol{k}}(\varepsilon_{n'\boldsymbol{k}}-\varepsilon_{n\boldsymbol{k}})|}, \tag{25.2.22}$$

where the factor 2 comes from the spin. The imaginary part of the dielectric constant is then

$$\epsilon_2(\omega) = \frac{\pi e^2}{\epsilon_0 m_{\mathrm{e}}\omega}\sum_{nn'} f_{nn'}(\omega)\rho_{nn'}(\hbar\omega)\,. \tag{25.2.23}$$

Since this quantity determines the strength of the absorption of electromagnetic radiation, it is immediately obvious that there is no absorption at those energies where no interband transition is possible. The joint density of states is singular not only at the minimum and maximum: Van Hove singularities appear at the energies that correspond to the saddle points in the difference of the two dispersion relations – just like for the density of states for phonons and Bloch electrons in Chapters 12 and 17. These singularities show up in the frequency dependence of the imaginary part of the dielectric constant (as illustrated for germanium in Fig. 25.10), and consequently in the absorption as well.

The previous discussion of interband transitions was based on the assumption that the electron excited to the n'th band is independent of the

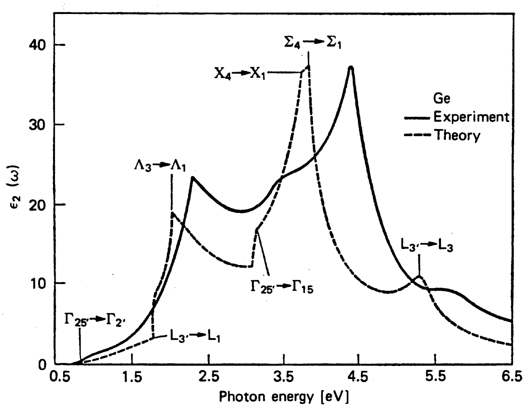


Fig. 25.10. The contribution of interband transitions to the theoretically and experimentally determined frequency-dependent dielectric constant of germanium [J. C. Phillips, *Solid State Physics*, Vol. **18**]

hole left behind in the nth band. In semiconductors the interaction between conduction-band electrons and valence-band holes cannot be neglected: they can form bound states, called *excitons*, leading to the appearance of new peaks in the absorption spectrum. We shall discuss this in more detail in the next volume.

As mentioned in Section 25.1.4, the interaction with the electromagnetic field is described in terms of the transverse dielectric constant rather than the longitudinal one, however, the two are the same in the long-wavelength limit. Just for reference, the precise formula for the imaginary part of the transverse dielectric constant, which determines the absorption, is

$$\varepsilon_2(\omega) = \frac{e^2\hbar^2}{\epsilon_0 m_\mathrm{e}^2\omega^2} \sum_{nn'} \left|\langle n', \boldsymbol{k}\,|\boldsymbol{e}\cdot\boldsymbol{\nabla}|\,n, \boldsymbol{k}\rangle\right|^2 \rho_{nn'}(\hbar\omega)\,, \tag{25.2.24}$$

where $\boldsymbol{e}$ is the unit vector in the direction of the transverse electric field, which is just the polarization vector of the photon. The above matrix element is the matrix element of the electron–photon interaction Hamiltonian that we shall discuss below.

25.2.2 Electron–Photon Interaction

In the previous calculations we used the dipole approximation (25.2.6) for the interaction between electrons and the electromagnetic field. In other cases it is more convenient to specify the electromagnetic field by a vector potential, and derive the interaction from the Hamiltonian given in (3.2.26),

$$\mathcal{H} = \sum_i \left\{ \frac{1}{2m_\mathrm{e}} \left[\boldsymbol{p}_i + e\boldsymbol{A}(\boldsymbol{r}_i)\right]^2 + U(\boldsymbol{r}_i) \right\}. \tag{25.2.25}$$

Earlier this Hamiltonian was used for the description of magnetic properties in uniform magnetic fields. By a suitable choice of the vector potential it can also serve to describe the interaction of electromagnetic radiation with solids. The usual form $\boldsymbol{p} = -\mathrm{i}\hbar\boldsymbol{\nabla}$ of the momentum operator implies that $\boldsymbol{p}\cdot\boldsymbol{A} = \boldsymbol{A}\cdot\boldsymbol{p} - \mathrm{i}\hbar\,\mathrm{div}\,\boldsymbol{A}$. In the Coulomb gauge $\mathrm{div}\,\boldsymbol{A} = 0$, so we have

$$\begin{aligned}\big[\boldsymbol{p} + e\boldsymbol{A}(\boldsymbol{r})\big]^2 &= \boldsymbol{p}^2 + e\boldsymbol{p}\cdot\boldsymbol{A} + e\boldsymbol{A}\cdot\boldsymbol{p} + e^2\boldsymbol{A}^2 \\ &= \boldsymbol{p}^2 + 2e\boldsymbol{A}\cdot\boldsymbol{p} + e^2\boldsymbol{A}^2\,. \end{aligned} \tag{25.2.26}$$

In a first approximation, the term proportional to $\boldsymbol{A}^2$ can be neglected, since for common light intensities it is much smaller than $\boldsymbol{A}\cdot\boldsymbol{p}$, so the interaction is given by

$$\mathcal{H}_{\mathrm{int}} = \sum_i \frac{e}{m_{\mathrm{e}}}\boldsymbol{A}(\boldsymbol{r}_i)\cdot\boldsymbol{p}_i = -\sum_i \frac{\mathrm{i}e\hbar}{m_{\mathrm{e}}}\boldsymbol{A}(\boldsymbol{r}_i)\cdot\boldsymbol{\nabla}_i\,. \tag{25.2.27}$$

A simple and intuitive picture of the interaction is obtained when, using the language of quantum electrodynamics, the electromagnetic field is described in terms of photons, and its interaction with the solid is pictured as scattering between photons and the elementary excitations of the solid. Therefore, we abandon the classical approach based on the Maxwell equations in favor of a description in terms of photon creation and annihilation operators. Because of the transversality of electromagnetic radiation, the electric field and the vector potential are expressed in terms of two physically interesting transverse modes:

$$\begin{aligned}\boldsymbol{E} &= -\mathrm{i}\sum_{\boldsymbol{q},\lambda}\sqrt{\frac{\hbar\omega_{\boldsymbol{q}\lambda}}{2V\epsilon_0}}\,\boldsymbol{e}_{\boldsymbol{q}\lambda}\Big(b_{\boldsymbol{q}\lambda}\mathrm{e}^{\mathrm{i}(\boldsymbol{q}\cdot\boldsymbol{r}-\omega_{\boldsymbol{q}}t)} - b^{\dagger}_{\boldsymbol{q}\lambda}\mathrm{e}^{-\mathrm{i}(\boldsymbol{q}\cdot\boldsymbol{r}-\omega_{\boldsymbol{q}}t)}\Big)\,,\\ \boldsymbol{A} &= \sum_{\boldsymbol{q},\lambda}\sqrt{\frac{\hbar}{2V\epsilon_0\omega_{\boldsymbol{q}\lambda}}}\,\boldsymbol{e}_{\boldsymbol{q}\lambda}\Big(b_{\boldsymbol{q}\lambda}\mathrm{e}^{\mathrm{i}(\boldsymbol{q}\cdot\boldsymbol{r}-\omega_{\boldsymbol{q}}t)} + b^{\dagger}_{\boldsymbol{q}\lambda}\mathrm{e}^{-\mathrm{i}(\boldsymbol{q}\cdot\boldsymbol{r}-\omega_{\boldsymbol{q}}t)}\Big)\,,\end{aligned} \tag{25.2.28}$$

where $b^{\dagger}_{\boldsymbol{q}\lambda}$ ($b_{\boldsymbol{q}\lambda}$) is the photon creation (annihilation) operator, and the polarization vectors $\boldsymbol{e}$ satisfy

$$\boldsymbol{e}_{\boldsymbol{q}\lambda}\cdot\boldsymbol{q} = 0\,, \qquad \lambda = 1, 2\,. \tag{25.2.29}$$

Substituting these into (25.2.27), and using the second-quantized representation for the electron states as well, the following types of interaction terms can appear in the Hamiltonian:

$$\mathcal{H}_{\mathrm{int}} \sim \sum_{nn'\boldsymbol{k}\boldsymbol{q}} D_{nn'\boldsymbol{k}\boldsymbol{q}}\, c^{\dagger}_{n'\boldsymbol{k}+\boldsymbol{q}\sigma} c_{n\boldsymbol{k}\sigma}\big(b_{\boldsymbol{q}} + b^{\dagger}_{-\boldsymbol{q}}\big)\,. \tag{25.2.30}$$

These terms describe processes in which an electron of wave vector $\boldsymbol{k}$ in the nth band is scattered into the n'th band with the absorption or emission of a photon. These are depicted by the first two processes in Fig. 25.11.

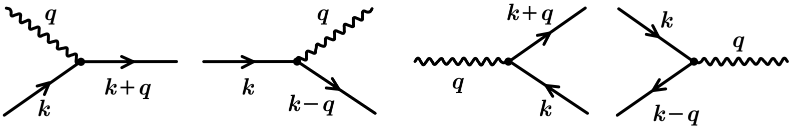


Fig. 25.11. Interaction between electrons and the radiation field: scattering with the absorption or emission of one photon, pair creation, and radiative recombination

It is more common in solid-state physics to speak about the creation of an electron–hole pair when a photon is absorbed, and about the radiative recombination of an electron–hole pair when a photon is emitted. The interaction processes reflecting this point of view are shown in the right part of the figure.

The photon energy is comparable to the typical electron energy in solids (of order eV) at wavelengths that are much larger than the atomic dimensions (or wave numbers that are much smaller than the size of the Brillouin zone). Therefore there is practically no momentum transfer ($|\boldsymbol{q}| \ll |\boldsymbol{k}|$) when a photon with an energy on the order of an eV is absorbed or emitted. Representing the interband transitions in the reciprocal space, as in Fig. 20.8, these processes correspond to vertical lines. As mentioned in connection with the band structure of semiconductors, the gap of direct-gap semiconductors can be determined from the threshold frequency of photon absorption. In indirect-gap semiconductors the electron of the electron–hole pair created by the absorption of the photon can lose energy to get to the minimum of the conduction band only by the subsequent emission of a phonon. In such indirect, two-step transitions the conservation of energy does not need to be satisfied in the intermediate state, only in the final state:

$$\boldsymbol{k}' = \boldsymbol{k} + \boldsymbol{\kappa} - \boldsymbol{q}\,, \qquad \varepsilon_{\boldsymbol{k}'} = \varepsilon_{\boldsymbol{k}} + \hbar\omega - \hbar\omega_{\boldsymbol{q}}\,, \tag{25.2.31}$$

where $\boldsymbol{\kappa}$ is the negligibly small momentum of the photon, and $\boldsymbol{q}$ is the momentum of the phonon. As discussed in Section 20.2.5, the actual indirect gap can be determined by means of such processes when the maximum of the valence band and the minimum of the conduction band are located at different wave vectors – as in germanium and silicon.

In X-ray absorption electrons are excited from deep levels because of the large photon energy. The hole left behind then acts as a scattering center. Since the associated scattering potential does not appear adiabatically but abruptly, this so-called *final-state interaction* leads to a rearrangement of the states of the electron system. It is essential to take it into account in the description of the absorption edge. However, its theoretical treatment requires the apparatus of the many-body problem.

At high intensities the term $\boldsymbol{A}^2$ can no longer be neglected. When written in the form

$$\begin{aligned}\sum_i \frac{e^2}{2m_\mathrm{e}}\boldsymbol{A}^2(\boldsymbol{r}_i) &= \int \mathrm{d}\boldsymbol{r} \sum_i \frac{e^2}{2m_\mathrm{e}}\boldsymbol{A}^2(\boldsymbol{r})\delta(\boldsymbol{r}-\boldsymbol{r}_i) \\ &= \int \mathrm{d}\boldsymbol{r} \frac{e^2}{2m_\mathrm{e}}\boldsymbol{A}^2(\boldsymbol{r})\rho(\boldsymbol{r}) ,\end{aligned} \tag{25.2.32}$$

$\boldsymbol{A}^2$ is seen to be coupled to the density $\rho(\boldsymbol{r})$ of electrons. Using the second-quantized representation for this term, too, a number of different contributions appear, for example

$$\sum_{nn'\boldsymbol{k}\boldsymbol{q}\boldsymbol{q}'} D_{nn'\boldsymbol{k}\boldsymbol{q}\boldsymbol{q}'} c^\dagger_{n'\boldsymbol{k}+\boldsymbol{q}'\sigma} c_{n\boldsymbol{k}\sigma} b^\dagger_{\boldsymbol{q}-\boldsymbol{q}'} b_{\boldsymbol{q}} , \tag{25.2.33}$$

which corresponds to the scattering of a photon, accompanied by the creation of an electron–hole pair. Terms with two-photon absorption and emission can also appear. They are shown in Fig. 25.12.

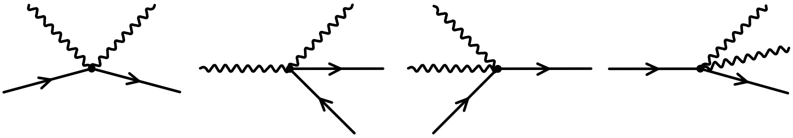


Fig. 25.12. Interaction between electrons and the radiation field: two-photon processes

An interesting manifestation of the interaction between photons and electrons in solids is photoemission. In this process the energetic electron of the created electron–hole pair leaves the solid. In the light of our previous results, the energy distribution of the emitted electrons is expected to reflect the density of states inside the solid. As mentioned in Chapter 19, information about the Fermi surface can be obtained by the ARPES method, in which the angular distribution is also measured.

25.2.3 Phonon–Photon Interaction

As mentioned in Chapter 13 on the experimental study of phonons, infrared absorption and Raman scattering provide suitable methods for measuring the energy of optical phonons at the center of the Brillouin zone, while long-wavelength acoustic phonons – in particular, their group velocity – can be studied by means of Brillouin scattering. These quantities can be derived from experimental data without knowing the precise nature of the interaction, simply by assuming the conservation of energy and quasimomentum. Below we shall give a more detailed description of the interaction, and read off the allowed absorption and scattering processes.

Consider the kinetic energy (11.1.24) of the atoms of a vibrating lattice. In the presence of an electromagnetic field described by a vector potential

$\boldsymbol{A}$ the canonical momentum $\boldsymbol{P}(m,\mu)$ is replaced by the kinetic momentum $\boldsymbol{P}(m,\mu) - q_\mu \boldsymbol{A}(\boldsymbol{r})$ in the kinetic energy formula, where q_μ is the charge of the μth ion in the primitive cell:

$$T_{\text{kin}} = \sum_{m,\mu} \frac{1}{2M_\mu} \big[\boldsymbol{P}(m,\mu) - q_\mu \boldsymbol{A}(\boldsymbol{r}(m,\mu))\big]^2 . \tag{25.2.34}$$

Up to linear order in the vector potential, the interaction between moving ions and the electromagnetic field can be written as

$$\mathcal{H}_{\text{int}} = -\sum_{m,\mu} \frac{1}{M_\mu} q_\mu \boldsymbol{P}(m,\mu) \cdot \boldsymbol{A}(\boldsymbol{r}(m,\mu)) \,. \tag{25.2.35}$$

Using the expansion (12.1.39), the momentum $\boldsymbol{P}(m,\mu) = M_\mu \dot{\boldsymbol{u}}(m,\mu)$ can also be expressed in terms of the phonon creation and annihilation operators. Likewise, the vector potential can be expressed in terms of the photon creation and annihilation operators through (25.2.28) in a very similar form. By substituting both into the interaction Hamiltonian, and evaluating the sum over the lattice points, which gives a constraint for the wave vectors,

$$\mathcal{H}_{\text{int}} = \sum_{\boldsymbol{q}} V_{\boldsymbol{q}} \left(a_{\boldsymbol{q}}^\dagger b_{\boldsymbol{q}} - a_{\boldsymbol{q}} b_{\boldsymbol{q}}^\dagger + a_{\boldsymbol{q}}^\dagger b_{-\boldsymbol{q}}^\dagger - a_{\boldsymbol{q}} b_{-\boldsymbol{q}} \right) \tag{25.2.36}$$

is obtained, where $a_{\boldsymbol{q}}^\dagger$ and $b_{\boldsymbol{q}}^\dagger$ are the phonon and photon creation operators, respectively. For simplicity, the polarization index is suppressed. The formula can be interpreted as the conversion of a photon into a phonon, or vice versa, in the interaction. Infrared absorption corresponds to the absorption of an infrared photon accompanied by the creation of a transverse optical phonon, to which the photon energy is transferred.

The full Hamiltonian of the photon–phonon system is

$$\begin{aligned} \mathcal{H} = & \sum_{\boldsymbol{q}} \hbar\omega_{\boldsymbol{q}} \left(a_{\boldsymbol{q}}^\dagger a_{\boldsymbol{q}} + \tfrac{1}{2} \right) + \sum_{\boldsymbol{q}} \hbar c |\boldsymbol{q}| \left(b_{\boldsymbol{q}}^\dagger b_{\boldsymbol{q}} + \tfrac{1}{2} \right) \\ & + \sum_{\boldsymbol{q}} V_{\boldsymbol{q}} \left(a_{\boldsymbol{q}}^\dagger b_{\boldsymbol{q}} - a_{\boldsymbol{q}} b_{\boldsymbol{q}}^\dagger + a_{\boldsymbol{q}}^\dagger b_{-\boldsymbol{q}}^\dagger - a_{\boldsymbol{q}} b_{-\boldsymbol{q}} \right) . \end{aligned} \tag{25.2.37}$$

If the system of phonons is not coupled to other degrees of freedom then this bilinear Hamiltonian can be diagonalized by means of the generalized Bogoliubov transformation

$$\alpha_{i,\boldsymbol{q}} = w_i a_{\boldsymbol{q}} + x_i b_{\boldsymbol{q}} + y_i a_{-\boldsymbol{q}}^\dagger + z_i b_{-\boldsymbol{q}}^\dagger \qquad i = 1, 2 \,, \tag{25.2.38}$$

which gives

$$\mathcal{H} = \sum_{\boldsymbol{q}} \left[\hbar\Omega_{\boldsymbol{q}}^{(1)} \left(\alpha_{1\boldsymbol{q}}^\dagger \alpha_{1\boldsymbol{q}} + \tfrac{1}{2} \right) + \hbar\Omega_{\boldsymbol{q}}^{(2)} \left(\alpha_{2\boldsymbol{q}}^\dagger \alpha_{2\boldsymbol{q}} + \tfrac{1}{2} \right) \right] . \tag{25.2.39}$$

These excitations are just the polaritons discussed in Section 25.1.7, which arise from the hybridization of optical lattice vibrations and electromagnetic radiation propagating in the solid. Their dispersion curve is shown in Fig. 25.6.

On the other hand, when phonons are also coupled to other degrees of freedom, and the interaction with photons is not the strongest, then the phonon created by the photon can, before being transformed back to a photon, transfer its energy to other degrees of freedom, while it decays, is scattered, or absorbed. Thus the absorption of the energy of the photon occurs in two steps. Because of the large disparity in the velocities, only optical phonons can be created by light. Since the energy of optical phonons is below 0.1 eV, the absorption occurs in the infrared region.

In ionic crystals the coupling to the electromagnetic field can be described alternatively in terms of the coupling between the electric field $\boldsymbol{E}$ and the polarization $\boldsymbol{P}$ due to the motion of ions. The interaction Hamiltonian is then

$$\mathcal{H}_{\text{int}} = -\boldsymbol{P} \cdot \boldsymbol{E} \,. \tag{25.2.40}$$

Expressing the electric field, through (25.2.28), in terms of the phonon creation and annihilation operators, and the polarization $\boldsymbol{P}$ in terms of the displacement of ions in the form

$$\boldsymbol{P} = \sum_{m,\mu} q_\mu^* \boldsymbol{u}_{m,\mu} = \sqrt{\frac{\hbar}{2N\omega_{\text{TO}}}} \sum_{m,\mu,\boldsymbol{q}} \frac{q_\mu^*}{\sqrt{M_\mu}} \left(\boldsymbol{e}_\mu a_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{R}_m} + \boldsymbol{e}_\mu^* a_{\boldsymbol{q}}^\dagger \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{R}_m} \right) , \tag{25.2.41}$$

where q_μ^* is the effective charge of the μth ion in the primitive cell, we obtain an expression that is similar to (25.2.36). Because of the transversality of photons, they interact only with transverse optical phonons. The absorption spectrum features a sharp peak at $\omega = \omega_{\text{TO}}$.

Several channels are open for two-phonon absorption – that is, the process in which a photon is absorbed and two phonons are created. The previous linear formula between the ionic displacement and polarization is valid only for rigid ions. Because of the displacement of neighboring ions the electron cloud becomes distorted, giving rise to a coupling between the ionic displacement of two ions and the electric field. In this way two phonons may also be created. In another channel the created phonon decays due to anharmonicity. The contributions of the two channels cannot be separated, as only the common final state is observed. Because of these processes, absorption does not occur at a single frequency but over a wide continuum.

Light can also be scattered inelastically by phonons. When a phonon is created or annihilated in the optical branch, we speak of Raman scattering. The wave number of visible light is on the order of $10^5\,\text{cm}^{-1}$, therefore light scattering is suitable to study $q \sim 0$ phonons only – provided single-phonon events alone are considered. When the photon absorbs or emits a long-wavelength acoustic phonon, we speak of Brillouin scattering. These processes are illustrated in Fig. 25.13. Only the initial and final states of the photon and phonon are shown; intermediate states are neglected.

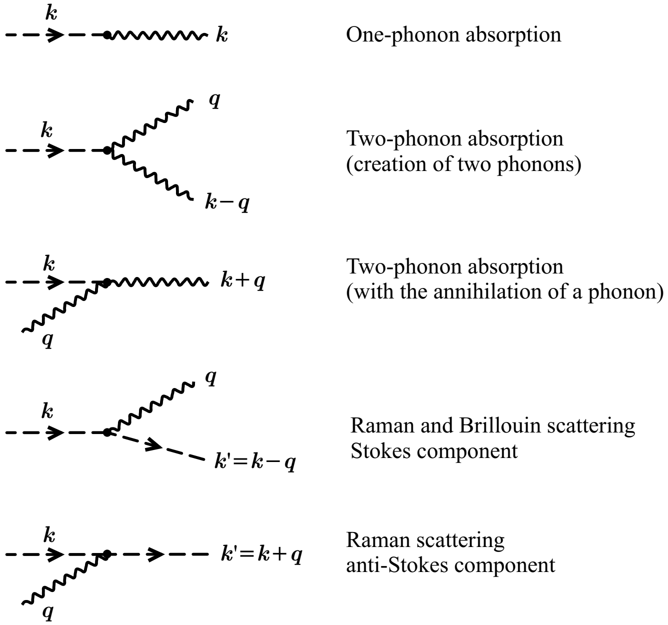


Fig. 25.13. The simplest processes of the phonon–photon interaction. Photons are represented by dashed and phonons by wavy lines

Just like for absorption, the simultaneous creation or annihilation of two phonons of wave vectors $\boldsymbol{q}$ and $-\boldsymbol{q}$ is possible in Raman scattering as well. By means of such two-phonon Raman scattering, the whole phonon spectrum can be probed. We speak of two-phonon Raman scattering even when two acoustic phonons are created or annihilated. Since the density of states is the highest at the boundary of the Brillouin zone for acoustic phonons, such measurements are particularly sensitive to phonons close to the zone boundaries.

Attention must be paid, however, to the subtlety that photons can interact directly only with optical phonons, therefore their interaction with acoustic phonons is possible only in multiple steps, through the creation of an electron–hole pair. The most probable processes are shown in Fig. 25.14.

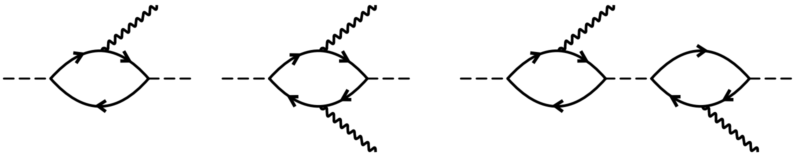


Fig. 25.14. Photon–phonon scattering processes with the emission of one and two phonons, with intermediate electron–hole pairs

In the lowest order either the hole or the electron of the created electron–hole pair emits or absorbs a phonon. In second-order processes two phonons may also be created. Two such possibilities are shown in Fig. 25.14. In the first case the Raman spectrum is continuous, while sharp lines are obtained at the sums of the two phonon frequencies in the second, since momentum conservation applies to intermediate states, too.

The strength of the individual processes can be determined only by taking into account the intermediate states, using the full quantum mechanical description. That way selection rules are also established: depending on the symmetries of the crystal, certain phonon branches are found to contribute to infrared absorption, while others to Raman scattering alone. The corresponding modes are called *infrared active* and *Raman active modes.*

Further Reading

1. M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th (expanded) Edition, reprinted with corrections, Cambridge University Press, Cambridge (2002).
2. M. Dressel and G. Grüner, *Electrodynamics of Solids, Optical Properties of Electrons in Matter*, Cambridge University Press, Cambridge (2002).
3. G R. Fowles, *Introduction to Modern Optics*, Second Edition, Holt, Rinehart, and Winston, New York (1975).
4. M. A. Fox, *Optical Properties of Solids*, Oxford Master Series in Physics, Oxford University Press, Oxford (2002).
5. C. Klingshirn, *Semiconductor Optics*, 2nd edition, Advanced Texts in Physics, Springer-Verlag, Berlin (2005).
6. H. Kuzmany, *Solid-State Spectroscopy, An Introduction*, Springer-Verlag, Berlin (1998).
7. *Optical Properties of Solids*, Edited by F. Abelès, North-Holland Publishing Company, Amsterdam (1972).
8. Y. Toyozawa, *Optical Processes in Solids*, Cambridge University Press, Cambridge (2003).
9. F. Wooten, *Optical Properties of Solids*, Academic Press, New York (1972).

26

Superconductivity

As discussed in Chapter 24, the electrical resistivity of metals decreases with decreasing temperature. In spite of zero-point vibrations, the contribution of scattering by lattice vibrations would vanish at $T = 0$, and the resistivity of an ideal crystal would be zero, whereas finite resistivities would be observed at any finite temperature. However, impurities and inhomogeneities, which are always present in the sample, reduce the transport relaxation time of electrons to a finite value even at zero temperature, and thus the conductivity remains finite at $T = 0$, too: $\sigma = n_{\mathrm{e}} e^2 \tau / m_{\mathrm{e}}$ according to the prediction of the Drude model. The liquefaction of helium (1908) opened the way to studying the resistivity of metals at much lower temperatures than before, well below the condensation point of helium at 4.22 K, also known as the liquid-helium temperature. It came as a great surprise in 1911 when H. Kamerlingh Onnes[1] observed that the resistivity of mercury (which can be purified easily in its liquid state, and is therefore regarded as the purest metal) did not decrease gradually with decreasing temperature but dropped to a very low value – zero within experimental error – around $T = 4.2$ K. This experimental finding is shown in Fig. 26.1.

Later it was confirmed that in a broad class of metals the resistivity does not decrease gradually, as a power of the temperature, but drops suddenly to zero at a finite critical temperature T_{c}. Kamerlingh Onnes coined the terms *superconductor* for such materials, and *superconductivity* for the phenomenon.[2]

In most cases the superconducting state appears at a very low temperature, however several families of materials have been discovered recently in which the transition temperature is close to or even somewhat higher than 100 K. Depending on the purity of the sample, the transition may be slightly smeared out, as in the example shown in Fig. 26.2. The resistance of the sample of

[1] See the footnote on page 2 of Volume 1.

[2] Originally, he dubbed the phenomenon “supraconductivity”, but this name was gradually replaced by the term used today.

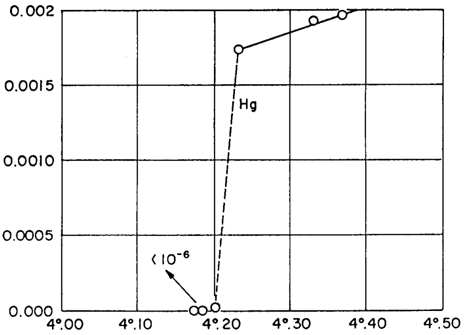


Fig. 26.1. Temperature dependence of the resistivity of mercury at low temperatures, as measured by Kamerlingh Onnes [*Comm. Phys. Lab. Univ. Leiden*, No. 120b (1911)]

composition $YBa_2Cu_3O_{7-\delta}$ (abbreviated as Y-Ba-Cu-O or YBCO) starts to decrease rapidly around 90 K, but it is finite even at 80 K.

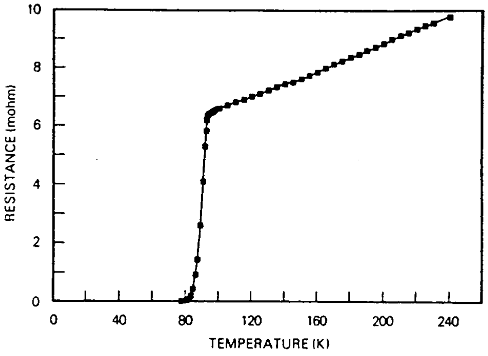


Fig. 26.2. Temperature dependence of the resistance for a high-T_c superconductor, Y-Ba-Cu-O [Reprinted with permission from M. K. Wu et al., *Phys. Rev. Lett.* **58**, 908 (1987). ©1987 by the American Physical Society]

We shall start this chapter with a brief overview of the phenomenon of superconductivity and the most characteristic properties of superconductors, and then give a phenomenological description. The microscopic theory will be presented in Chapter 34 (Volume 3), after the detailed discussion of electron–electron interactions.

26.1 Superconductivity: The Phenomenon

Even though the most striking feature of superconductors is their infinitely high conductivity, it is accompanied by a number of other interesting properties. Their discovery was a great step toward the understanding of the phenomenon. Below we shall consider them one by one.

26.1.1 Vanishing Resistance, Persistent Current

The name "superconductor" comes from the striking property of vanishing resistance below a critical temperature. A direct consequence of this vanishing resistance is the persistence of currents. In a ring-shaped superconductor a current would flow indefinitely because of the absence of resistance. According to experiments, there is no sign indicating that the current would diminish: the most precise measurements put the lifetime of the current above 10^5 years. As we shall see, this current flows on the sample surface and not in its interior.

However, superconductors cannot carry arbitrarily large currents. Above a critical current J_c the sample ceases to behave as a superconductor. In wires of 1 mm in diameter this critical current can be as high as 100 A. As we shall see, it is the magnetic field, generated by the current around the sample, that destroys superconductivity.

The statement that no current is dissipated is valid only for direct currents and low-frequency alternating currents. Current is dissipated above a threshold frequency ν, which is characteristic of the material and is usually in the microwave or infrared region. This can be seen in Fig. 26.3, which shows the frequency dependence of the real part of the optical conductivity in thin lead layers. It can be shown that the threshold frequency and the width Δ of the

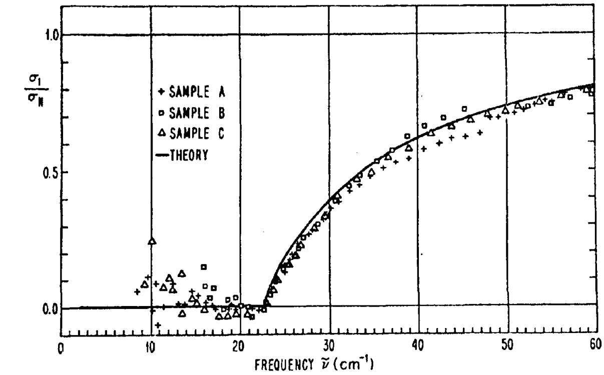


Fig. 26.3. The frequency dependence of the real part of the optical (far-infrared) conductivity in thin lead layers [Reprinted with permission from L. H. Palmer and M. Tinkham, *Phys. Rev.* **165**, 588 (1968). ©1968 by the American Physical Society]

gap in the energy spectrum (to be discussed in Section 26.1.6) are related by $h\nu \approx \Delta$.

26.1.2 Isotope Effect

In 1950 several research groups observed that the critical temperature T_c of superconductors depended on the isotopic composition: T_c was found to decrease for increasing concentrations of heavier isotopes of the same element. For mercury, which has seven stable isotopes between mass numbers 196 and 204, the transition temperature varies between 4.16 K and 4.12 K. The dependence of the critical temperature on the isotopic mass can be approximated as

$$T_\mathrm{c} \propto M^{-\alpha}\,, \tag{26.1.1}$$

where $\alpha = 1/2$ was found for mercury. In other cases the variations of the critical temperature are weaker. Table 26.1 shows the measured value of α for a number of superconductors.

Table 26.1. Measured value of exponent α of the isotope effect for a number of superconductors

	Cd	Hg	Mo	Os	Pb	Re	Ru	Sn	Tl	Zn	Zr
α	0.5	0.50	0.37	0.21	0.48	0.36	0.0	0.47	0.5	0.45	0.0

While in certain cases the exponent is zero – which corresponds to the absence of any measurable isotope effect –, it is often close to 0.5. Since different isotopes have the same electronic structure, the dependence on the isotopic composition indicates that superconductivity cannot be understood completely in terms of the electron system alone: the mass of the ions in whose field the electrons move is also important. As we shall see in Chapter 34, the microscopic theory of superconductivity – the BCS theory – provides a simple explanation for the dependence on the isotopic mass and the exponent $\alpha = 1/2$. Deviations from it can be understood in an improved theoretical framework that leads to the Eliashberg equations.

26.1.3 Meissner–Ochsenfeld Effect

Because of its infinite conductivity, the superconductor might be considered as a perfect conductor. In perfect conductors the current flow can be finite only for vanishing electric fields, $\boldsymbol{E} = 0$. However, the Maxwell equations then imply that the magnetic field cannot vary with time inside the sample. Therefore, when an external magnetic field is turned on, it cannot penetrate into the perfectly conducting region. On the other hand, if the sample is placed in a magnetic field in its normal state, and it becomes a perfect conductor only afterwards, then the magnetic field established inside the sample in the normal state would need to remain unaltered by the transition into the perfect conductor state, and the magnetic field would freeze into the sample. This means that the strength of the magnetic field in a perfect conductor would depend on the history of the sample.

Contrary to these expectations, W. Meissner and R. Ochsenfeld (1933) observed that the magnetic induction inside the superconductor was always zero in weak applied fields. When a superconductor is placed in a magnetic field, the field cannot penetrate into the sample. And if the sample is placed into the magnetic field in its normal state, and cooled below the critical temperature T_c only afterwards, the magnetic field is expelled from the sample. This is the *Meissner–Ochsenfeld effect.* The magnetic field lines are shown in Fig. 26.4 for both the normal and superconducting states.

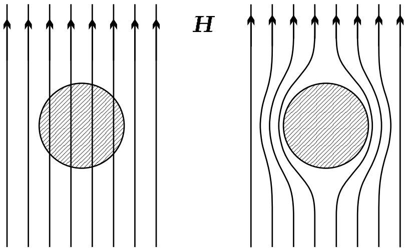


Fig. 26.4. The magnetic field lines around a sample of finite size, in the normal (left) and superconducting (right) states

To understand this behavior of superconductors, we have to assume that the magnetic field induces surface currents in a layer that is macroscopically thin but thick on the atomic scales. These cancel the applied field, and maintain the state in which the magnetic induction is zero inside the sample. In conjunction with the formula $\boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M}) = 0$, the requirement $\boldsymbol{B} = 0$ leads to

$$\boldsymbol{M} = -\boldsymbol{H} \tag{26.1.2}$$

for the magnetization of the superconductor. The susceptibility of the superconductor is therefore $\chi_\mathrm{m} = -1$, which corresponds to perfect diamagnetism.

This property of superconductors can be exploited most easily for setting up a persistent current. Consider a superconducting ring placed in a uniform axial magnetic field $\boldsymbol{H}$ above its critical temperature. With the field switched on, the sample is then cooled below its critical temperature, whereby the magnetic field $\boldsymbol{H}$ is expelled from the sample but, of course, $\boldsymbol{B}$ remains finite inside the ring. When the external field is then switched off, the field strength inside the ring must remain unchanged, since field lines cannot enter the superconducting ring. Therefore the flux through the ring is the same as the original flux of the applied field: the transient electric field generated by the switch-off induces an eddy current in the ring, and the magnetic field of the current produces the required flux. The magnetic field lines around the ring in the presence of the applied field and after its switch-off are shown in Fig. 26.5.

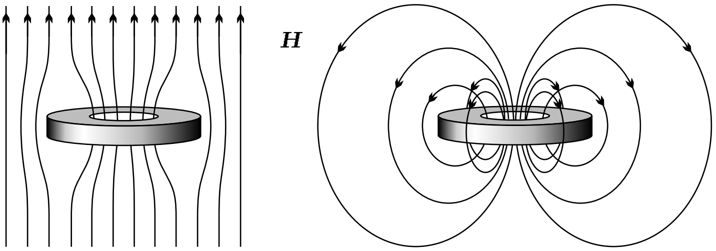


Fig. 26.5. The magnetic field lines around a ring-shaped sample in the presence of an applied field (left) and after the applied field has been switched off (right), when the superconducting current induced in the sample maintains the flux through the ring

According to the measurement results shown in Fig. 26.6, the flux through the superconducting ring cannot take any arbitrary value, only the integral multiples of the flux quantum $\Phi_0 = h/2e$. This flux quantum is exactly the half of what would be expected from the Landau quantization of electronic energy levels in strong magnetic fields.[3]

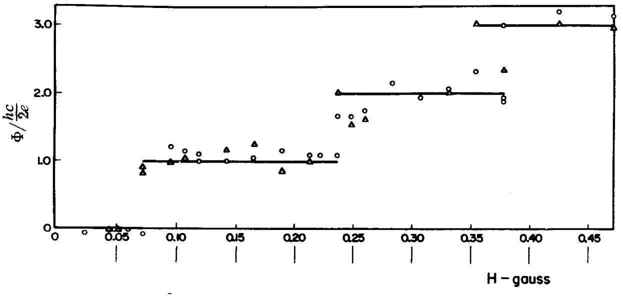


Fig. 26.6. The magnetic flux through a superconducting ring, as a function of the applied field [Reprinted with permission from B. S. Deaver, Jr. and W. M. Fairbank, *Phys. Rev. Lett.* **7**, 43 (1961). ©1961 by the American Physical Society]

[3] The flux quantum associated with the motion of electrons in a magnetic field is h/e, see page 282.

26.1.4 Critical Field

The Meissner–Ochsenfeld effect can usually be observed only in relatively weak fields. Sufficiently strong fields – which may be as low as $H \sim 10^2$ to 10^3 Oe in certain materials – can completely destroy superconductivity; the sample then shows normal metallic behavior. Depending on how the transition to the normal phase occurs, superconductors are divided into two broad classes.

In *type I superconductors* $\boldsymbol{B}$ remains zero inside the sample until the applied magnetic field reaches a temperature-dependent critical value $H_\mathrm{c}(T)$; at that point the entire sample becomes normal, that is, its conductivity jumps to a finite value. As we shall see later, the transition is first order everywhere except $T = 0$ and at $T = T_\mathrm{c}$ where $H_\mathrm{c} = 0$, therefore the magnetization curve exhibits hysteresis. When the external magnetic field is reduced, the normal state may be maintained down to a lower field strength $H_{\mathrm{c}2}$ ($H_{\mathrm{c}2} < H_\mathrm{c}$). The magnetic induction inside the sample and the magnetization of the sample are shown in Fig. 26.7.

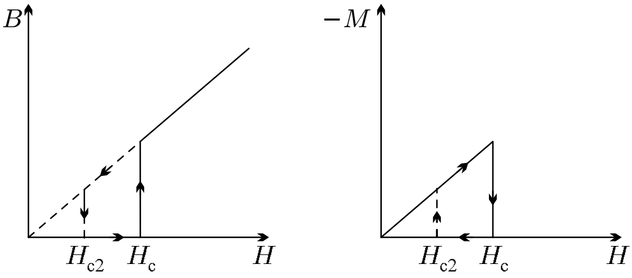


Fig. 26.7. The magnetic induction and magnetization in type I superconductors as functions of the applied magnetic field

Of course, at the critical temperature T_c the critical field is zero: $H_\mathrm{c}(T_\mathrm{c}) = 0$. At lower temperatures H_c is finite, and increases continuously as T decreases. Experimental data show that this dependence is well approximated by the function

$$H_\mathrm{c}(T) = H_\mathrm{c}(0)\left[1 - (T/T_\mathrm{c})^2\right]. \tag{26.1.3}$$

The critical field at $T = 0$ and the critical temperature are two important parameters of superconductors. The superconducting phase covers a finite region in the T–H plane; its boundary is determined by the temperature dependence of the critical field. A typical phase diagram is shown in Fig. 26.8.

However, this behavior is not common to all superconductors. In a fairly large proportion of superconductors the magnetic field starts to penetrate into the sample at the *lower critical field* $H_{\mathrm{c}1}$, however this penetration is gradual rather than abrupt. The complete penetration of the field occurs at the *upper critical field* $H_{\mathrm{c}2}$, where superconductivity is destroyed. Materials that

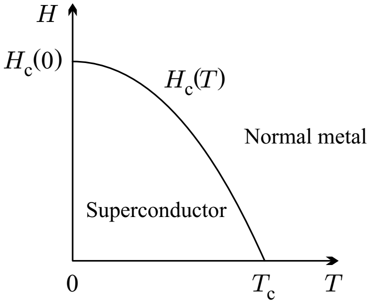


Fig. 26.8. Phase diagram of type I superconductors in the T–H plane

exhibit this behavior are called *type II superconductors*. The dependence of the magnetization and average magnetic flux density B in such superconductors on the applied magnetic field H are shown in Fig. 26.9.

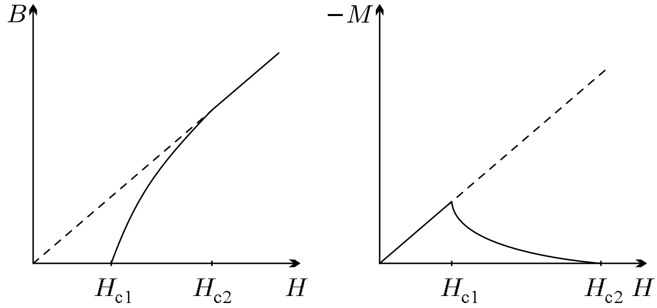


Fig. 26.9. The magnetic induction and the magnetization in type II superconductors as functions of the applied magnetic field

As opposed to type I superconductors, the penetration of the magnetic field (the disappearance of the Meissner–Ochsenfeld effect) in type II superconductors occurs at a lower field than the appearance of electrical resistivity. That is why there are two phase boundaries in the schematic phase diagram in Fig. 26.10.

When the applied field is weaker than the lower critical field H_{c1}, the sample behaves as a perfect diamagnet. This is the *Meissner phase*. When the applied field is stronger than the upper critical field H_{c2}, normal behavior is observed. For fields between H_{c1} and H_{c2}, in the *Shubnikov phase*, the sample becomes inhomogeneous: the sample interior then contains alternate superconducting and normal regions. That is why this phase is also known as the *mixed phase*. The normal phase appears inside tube-like regions ("filaments of flux") along the applied magnetic field; the magnetic field can penetrate into these parts of the sample. The magnetic flux density inside each tube is such

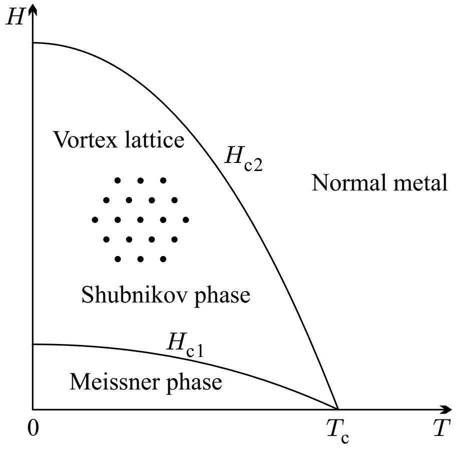


Fig. 26.10. Phase diagram of type II superconductors in the T–H plane

that the total magnetic flux through the tube is exactly one flux quantum. The current around the tube surface screens this flux, and the regions among the tubes are superconducting. Since these eddy currents show a vortex-like pattern, the tubes are called *vortices*. Defects in the crystalline order prevent the vortices from moving, and so externally applied currents may flow without any resistance in this state of the sample, too. Since H_{c2} may be as high as 10^5 Oe, type II superconductors are much more important for technological applications than type I superconductors. Among others, strong superconducting magnets can be built of them.

As we shall see, in conventional type II superconductors vortices are arranged in a regular array over the entire range between H_{c1} and H_{c2}. All high-T_c superconductors are type II materials, however their phase diagram is even more complicated, as the lattice of vortices may melt before the sample becomes a normal metal. Such a phase diagram is shown in Fig. 26.25.

26.1.5 Thermoelectric Properties

According to the Wiedemann–Franz law, good conductors are also good heat conductors. In the region where the electrical resistivity receives its dominant contribution from phonon-absorption and -emission processes, the resistivity increases with temperature, while the thermal conductivity decreases for a while and then tends to a constant value. The same behavior is observed in superconducting materials in their normal phase when a magnetic field exceeding H_c is applied. If the Wiedemann–Franz law were valid in the superconducting phase, too, then the thermal conductivity would also become infinitely large. However, λ starts to decrease at the onset of superconductivity, as shown in Fig. 26.11.

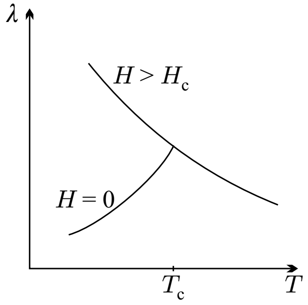


Fig. 26.11. Temperature dependence of the thermal conductivity in the vicinity of the superconducting phase transition

Another surprise is that the electric current is not accompanied by a heat current: the Peltier coefficient is zero in the superconducting phase. These phenomena indicate that the electrons responsible for superconductivity do not carry any entropy.

26.1.6 Specific Heat

As we have seen, below room temperature, the specific heat of pure metals in their normal state can be given as the sum of two terms:

$$C_\mathrm{n} = \gamma T + AT^3 \,. \tag{26.1.4}$$

The linear term, which becomes dominant at low temperatures, is the electronic contribution, while the cubic term is due to phonons. A completely different behavior is observed in the superconducting phase. The specific heat has a discontinuity at T_c, and immediately below T_c it is larger than it would be in the normal phase. On the other hand, the electronic contribution to the specific heat is exponentially small at low temperatures,

$$C_\mathrm{s} \sim \exp\left(-\frac{\Delta}{k_\mathrm{B}T}\right), \tag{26.1.5}$$

where Δ is on the order of $k_\mathrm{B}T_\mathrm{c}$. This is shown for aluminum in Fig. 26.12.

This behavior indicates that, contrary to normal metals, there are no low-energy electronic excitations in the superconducting state. A gap $\Delta \sim k_\mathrm{B}T_\mathrm{c}$ appears in the excitation spectrum. However, this gap is different in several respects from those in semiconductors or insulators. Firstly, the appearance of the gap in the latter types of material is the consequence of the periodic potential of the lattice. When electrons are added to the system, they occupy states above the gap, and thus the conductivity increases. In contrast, the energy gap is attached to the Fermi energy in superconductors. If the number of electrons were increased, the position of the gap would be shifted upward together

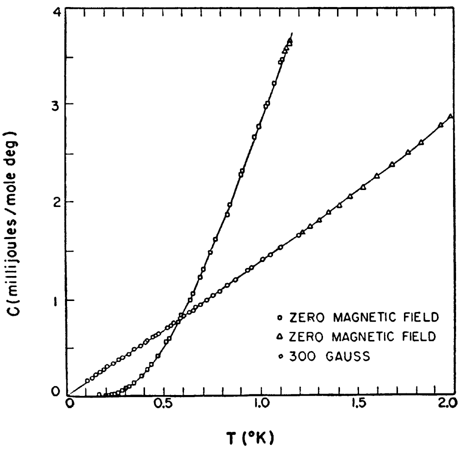


Fig. 26.12. Low-temperature specific heat of aluminum in zero magnetic field and in an external field exceeding H_c [Reprinted with permission from N. E. Phillips, *Phys. Rev.* **114**, 676 (1959). ©1959 by the American Physical Society]

with the chemical potential, and the system would remain a superconductor. Secondly, the gap depends only weakly on temperature in semiconductors, whereas it shows a strong temperature dependence in superconductors. The gap becomes narrower for increasing temperatures, and disappears at T_c.

In the absence of a magnetic field the superconducting order breaks up continuously: the phase transition is second order. However, the specific heat exhibits a discontinuity. Its magnitude can be estimated using the microscopic theory. Referred to the specific heat $C_n = \gamma T_c$ of the electron system, the relative jump of the specific heat is a universal constant:

$$\frac{C_s - C_n}{C_n} = \frac{12}{7\zeta(3)} = 1.426\,. \qquad (26.1.6)$$

As can be seen in Table 26.2, the measured value is indeed close to this number in certain materials. However, the deviation is significant in others, especially in high-T_c superconductors, where the specific-heat contribution of phonons is also important.

Table 26.2. The relative jump of the specific heat in the transition point for some superconductors

Element	Relative jump	Compound	Relative jump
Al	1.45	UPt_3	0.9
Cd	1.36	$CeRu_2Si_2$	3.5
Nb	1.93	$(TMTSF)_2ClO_4$	1.7
Pb	2.71	$YBa_2Cu_3O_7$	3.6

Using the relation between the specific heat in a uniform magnetic field and the entropy per unit volume,

$$c_H = T\left(\frac{\partial s}{\partial T}\right)_H , \tag{26.1.7}$$

the entropy can be determined. Its temperature dependence is sketched in Fig. 26.13. The entropy is lower in the superconducting phase than it would be in the normal phase. The former is therefore more ordered than the latter.

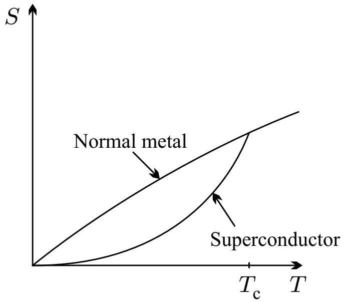


Fig. 26.13. Temperature dependence of the entropy in the superconducting and normal states

26.1.7 Tunneling in SIS and SIN Junctions

Other measurements also indicate the presence of a gap in the spectrum of electronic excitations. When a thin superconducting layer is illuminated by infrared radiation with a wavelength of a few mm ($E_{\text{photon}} \sim 10^{-3}\,\text{eV}$), absorption is observed only above a frequency threshold. The same applies to the absorption of ultrasound. As pointed out by I. GIAEVER[4] (1960), the experimental study of tunneling in junctions where two superconductors or a superconductor and a normal metal are separated by a thin insulator layer is likely to provide the most suitable method for detecting the energy gap in the superconducting phase. The first configuration is called a superconductor–insulator–superconductor (SIS) junction, and the second is a superconductor–insulator–normal metal (SIN) junction. The current–voltage characteristics of two SIS junctions are shown in Fig. 26.14. In the first case identical, and in the second case different superconducting materials are used on the two sides. In the I–V curve extrapolated to $T = 0$ the current appears only at a finite voltage, which is related to the energy gaps Δ_{L} and Δ_{R} of the superconductors on the left and right by $eV = \Delta_{\text{L}} + \Delta_{\text{R}}$.

[4] IVAR GIAEVER (1929–) was awarded the Nobel Prize in 1973 for his "experimental discoveries regarding tunneling phenomena in superconductors".

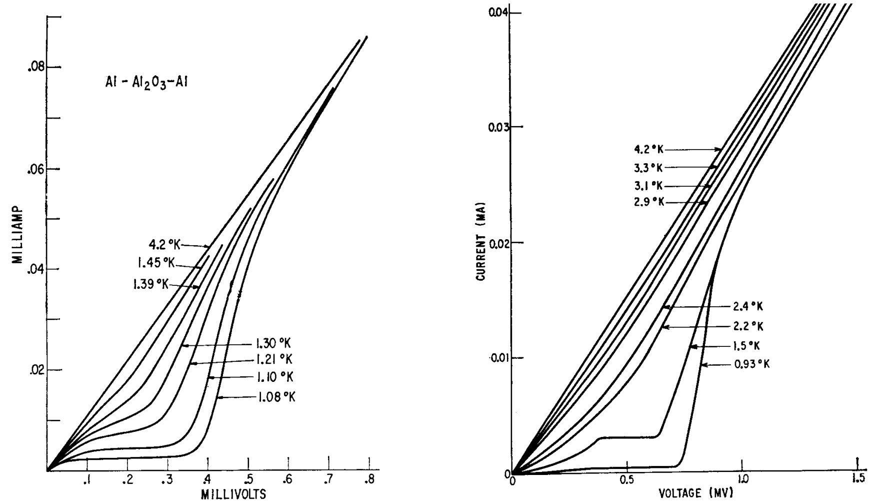


Fig. 26.14. Current–voltage characteristics of Al-Al_2O_3-Al and Al-Al_2O_3-In junctions at different temperatures [Reprinted with permission from I. Giaever and K. Megerle, *Phys. Rev.* **122**, 1101 (1961). ©1961 by the American Physical Society]

In junctions that contain high-T_c unconventional superconductors different characteristics and temperature dependence are observed. This is related to the characteristic anisotropy of the energy spectrum in such materials: the gap depends on the direction in $\boldsymbol{k}$-space.

26.2 Superconducting Materials

The first superconducting materials were discovered among elemental metals. Somewhat later, researchers turned to metallic compounds, and observed that quite a few of them became superconductors at low temperatures. By low temperatures we mean that even though the highest observed critical temperature kept growing, the record was still only 23 K in the mid-1980s. Shortly afterwards, high-T_c superconductors were discovered, with critical temperatures above 100 K. Nonetheless we are still very far from realizing the hope of finding materials that superconduct even at room temperature.

Below, we shall first consider the characteristic parameters of elemental superconductors, then turn to compound superconductors, which are the most important for applications, and finally present high-T_c superconductors.

26.2.1 Superconducting Elements

In Table 26.3 we listed those elements that become superconductors at normal pressure in the bulk. Besides the critical temperature, the critical magnetic induction B_c is listed, since this is used ever more widely in the literature instead of the critical magnetic field H_c. For type II superconductors – such as niobium, tantalum, and technetium – the magnetic induction that corresponds to the *thermodynamic critical field* H_c is given. This quantity will be defined later; for now it is enough to know that it is between H_{c1} and H_{c2}.

Table 26.3. The critical temperature T_c and the critical magnetic induction B_c of superconducting elements

Element	T_c (K)	B_c (mT)	Element	T_c (K)	B_c (mT)
Al	1.18	10.5	Pa	1.4	
Am	0.6		Pb	7.20	80.3
Be	0.03	9.9	Re	1.70	20.1
Cd	0.52	2.8	Rh	3.2×10^{-4}	5×10^{-3}
Ga	1.08	5.9	Ru	0.49	6.9
Hf	0.13	1.3	Sn	3.72	30.5
α-Hg	4.15	41.1	Ta	4.47	82.9
β-Hg	3.95	33.9	Tc	7.8	141
In	3.41	28.2	Th	1.37	16.0
Ir	0.11	1.6	Ti	0.40	5.6
α-La	4.87	80	Tl	2.38	17.6
β-La	6.06	110	U	0.68	10.0
Lu	0.1	35.0	V	5.46	140
Mo	0.92	9.7	W	0.01	0.1
Nb	9.25	206	Zn	0.86	5.4
Os	0.66	7.0	Zr	0.63	4.7

Even more confusing than the inconsistent usage of B and H, the literature also lacks unanimity in the choice of units. The CGS system is still widely used, so the critical field is often given in oersteds instead of the corresponding SI unit, A/m. Other authors speak about the magnetic field but specify it in gausses or teslas. In Table 26.3 B_c is given, in milliteslas. To obtain the value of H_c in A/m, it has to be multiplied by $10^4/4\pi = 795.8$. Oersted values are obtained through $1\,\mathrm{mT} \widehat{=} 10\,\mathrm{Oe}$.

It is worth taking a look at how these superconducting materials are distributed among the groups of the periodic table. This is shown in Table 26.4.

It is no surprise that the good semiconductors in group 14 (column IVA) and the nonmetallic elements in groups 15, 16, and 17 (columns VA, VIA and VIIA) do not become superconductors at atmospheric pressure even at very low temperatures although it should be mentioned that they do so under very high pressure – ranging from about 10 GPa to over 100 GPa for oxygen, sulfur, and bromine – at a few kelvins. For example, the critical temperature reaches 18 K at 30 GPa for phosphorus. It should also be noted that at normal pres-

Table 26.4. Superconducting elements in the periodic table. Elements shown in white on black become superconductors in the bulk at atmospheric pressure, while those in italics on gray only under high pressure or in thin films

1																	18
H	2											13	14	15	16	17	He
Li	**Be**											*B*	*C*	N	*O*	F	Ne
Na	Mg	3	4	5	6	7	8	9	10	11	12	**Al**	*Si*	*P*	*S*	Cl	Ar
K	*Ca*	*Sc*	**Ti**	**V**	*Cr*	Mn	*Fe*	Co	Ni	Cu	**Zn**	**Ga**	*Ge*	*As*	*Se*	*Br*	Kr
Rb	*Sr*	*Y*	**Zr**	**Nb**	**Mo**	**Tc**	**Ru**	**Rh**	*Pd*	Ag	**Cd**	**In**	**Sn**	*Sb*	*Te*	*I*	Xe
Cs	*Ba*	**La**	**Hf**	**Ta**	**W**	**Re**	**Os**	**Ir**	Pt	Au	**Hg**	**Tl**	**Pb**	*Bi*	Po	At	Ra
Fr	Ra	Ac	Rf	Db	Sg	Bh	Hs	Mt									

Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	**Lu**
Th	**Pa**	**U**	Np	Pu	**Am**	Cm	Bk	Cf	Es	Fm	Md	No	Lr

sure neither alkali metals, nor alkaline-earth metals (apart from beryllium), nor noble metals are superconductors – in short, none of the simplest metals to which the free-electron model can be applied successfully. Superconductors are found among the elements of groups 12 and 13 (columns IIB and IIIA), the non-semiconductor elements of group 14 (column IVA), and transition metals. However, transition metals that are magnetically ordered, or are in a sense close to becoming magnetically ordered, do not display superconductivity. But this rule is not watertight either: at high pressure iron loses its magnetization, recrystallizes in an hcp structure, and then becomes a superconductor below 2 K. Chromium exhibits superconductivity in thin films. As we shall see in Volume 3, the presence of a uniform ferromagnetic order rules out the possibility of superconductivity. In antiferromagnetic or nonuniform ferromagnetic materials the situation is not so clear: the two kinds of order usually compete and work against each other, although there are certain cases when they coexist.

Among lanthanoides, only lanthanum and lutetium are superconductors. The critical temperature is slightly different for the two crystallographic modifications of lanthanum, α-La (hcp structure) and β-La (fcc structure). To date, four actinoids have been found to exhibit superconductivity: thorium, protactinium, uranium (α and γ modifications), and americium. With the exception of niobium, technetium, and vanadium, elemental superconductors are all type I. The lower and upper critical fields are given in Table 26.5 for these exceptions.

Table 26.5. The critical temperature T_c and the lower and upper critical fields for elemental type II superconductors

Element	T_c (K)	B_{c1} (mT)	B_{c2} (mT)
Nb	9.25	173	405
Tc	7.8	120	312
V	5.4	115	296

The critical temperature may change under high pressure. The critical temperature of zirconium changes from 0.6 K at normal pressure to 11 K at 30 GPa, while the T_c of vanadium reaches 17.2 K at 120 GPa. Several other elements – such as As, B, Ba, Bi, Ca, Ce, Cs, Fe, Ge, Li, P, Sb, Sc, Se, Si, Sr, Te, and Y – that behave as normal metals or semiconductors become superconductors under pressures of order 10^3 MPa or higher. Among all elements, the highest critical temperature was observed in lithium: $T_c = 20$ K at a pressure of 50 GPa. Even Br, I, O, and S become superconductor under very high pressure.

In other cases the superconducting parameters of amorphous samples or thin films are different from those observed in bulk crystalline samples. For example, in thin films tungsten becomes a superconductor at 5.5 K instead of 0.01 K, beryllium at 9.95 K instead of 26 mK, and gallium at 8.6 K instead of 1.08 K. Another interesting finding is that carbon becomes a superconductor in its most recently discovered allotropic modification, the nanotube, and its critical temperature depends on the tube diameter. Its highest critical temperature registered to date is 15 K.

26.2.2 Superconducting Compounds

Compared to elemental superconductors, higher critical temperatures can be found in metallic alloys or compounds. A particularly interesting family of such compounds is the group of materials of composition M_3X – where M stands for niobium or vanadium –, crystallized in A15 structure (shown in Fig. 7.5). The transition temperatures for some of them are listed in Table 26.6.

Members of this family held the record of highest transition temperature for a good while. For applications, the upper critical field may be even more important; this reaches 24.7 T in Nb_3Sn and 23 T in V_3Si. The highest critical fields occur in another family, the so-called Chevrel-phase compounds. Their generic composition is $M_xMo_6X_8$, where M is a metallic element and X is either sulfur or selenium. The transition temperature may depend strongly on the number x specifying the composition, which is not necessarily one. This is ignored in Table 26.7, in which the parameters – including typical values for T_c – are listed for a number of superconductors.

Table 26.6. Transition temperature of A15 superconductors

Compound	T_c (K)	Compound	T_c (K)
V_3Au	3.0	Nb_3Au	11.5
V_3Al	9.6	Nb_3Al	19.1
V_3Ga	16.8	Nb_3Ga	14.5
V_3In	13.9	Nb_3In	9.2
V_3Si	17.1	Nb_3Si	19.0
V_3Ge	8.2	Nb_3Ge	23.2
V_3Sn	3.8	Nb_3Sn	18.1
V_3Pt	2.9	Nb_3Pt	10.9

Table 26.7. The transition temperature and upper critical magnetic induction for Chevrel-phase superconductors

Compound	T_c (K)	B_{c2} (T)	Compound	T_c (K)	B_{c2} (T)
Mo_6S_8	1.9		Mo_6Se_8	6.5	
$LaMo_6S_8$	7.1	5.4	$LaMo_6Se_8$	11.4	44.5
$Cu_2Mo_6S_8$	10.7		$Cu_2Mo_6Se_8$	5.9	
$PbMo_6S_8$	14.7	55.0	$PbMo_6Se_8$	3.8	3.8
$TlMo_6S_8$	8.7		$TlMo_6Se_8$	12.2	
$SnMo_6S_8$	11.8	34.0	$SnMo_6Se_8$	6.8	
$YbMo_6S_8$	8.6		$YbMo_6Se_8$	5.8	

It was discovered in the early 1990s that some alkali-metal-doped fullerites (which are made up of C_{60} molecules, see page 29 of Volume 1) exhibit superconductivity. Relatively high T_c was found in alkali-metal fullerides of composition M_3C_{60}. Their transition temperatures are given in Table 26.8.

Table 26.8. Critical temperature of alkali fullerides of composition M_3C_{60}

Compound	T_c (K)	Compound	T_c (K)
K_3C_{60}	19.5	Rb_3C_{60}	29.5
K_2RbC_{60}	23.0	Rb_2CsC_{60}	31
K_2CsC_{60}	24.0	$RbCs_2C_{60}$	33
KRb_2C_{60}	27.0	Cs_3C_{60}	47

It is worth noting that the increase in the critical temperature of fullerides is directly related to the increase in the lattice constant, brought about by placing larger and larger alkali atoms in the lattice.

A rather intensely studied materials of the past years has been MgB_2, in which hcp layers of manganese atoms become intertwined with honeycomb-structured (graphite-like) layers of boron atoms. The particularly keen interest is due to the highest transition temperature ever found in "conventional"[5] superconductors: $T_\mathrm{c} = 40\,\mathrm{K}$.

Owing to their physical properties, heavy-fermion superconductors occupy a special place among superconducting compounds. They will be presented in Chapter 35.

26.2.3 High-Temperature Superconductors

Well into the 1980s, newer and newer superconducting materials were discovered but the highest attained critical temperature kept increasing quite slowly. Then in 1986 an observation made by J. G. Bednorz and K. A. Müller[6] triggered an unprecedented hunt for materials with higher and higher T_c. They found that by taking the semiconducting compound $\mathrm{La}_2\mathrm{CuO}_4$, which becomes antiferromagnetically ordered at $T_\mathrm{N} = 540\,\mathrm{K}$, and replacing a part of the trivalent La^{3+} ions by divalent Ba^{2+} or Sr^{2+} ions, the resistivity of the electron-deficient material (in which hole conduction dominates) starts to drop rapidly at a higher temperature than in previously studied materials. However, the transition was not sharp, and the electrical resistance did not vanish completely, either. The Meissner–Ochsenfeld effect could not be observed, but the measurements performed in magnetic fields indicated strong diamagnetism.

By the first months of 1987 it became clear that the materials of composition $\mathrm{La}_{2-x}\mathrm{Ba}_x\mathrm{CuO}_4$become superconductors between 30 and 35 K, depending on the concentration of the Ba ions. At high pressures the critical temperature was found to be close to 40 K. Using strontium instead of barium, the critical temperature of $\mathrm{La}_{2-x}\mathrm{Sr}_x\mathrm{CuO}_4$ was observed to reach 37.5 K even at normal pressures for the composition $x \approx 0.15$.

This discovery gave a new impetus to the search of superconductors with higher and higher T_c. Still in 1987 it was found that in $\mathrm{YBa}_2\mathrm{Cu}_3\mathrm{O}_{7-\delta}$ (YBCO) the critical temperature, which depends on the oxygen content, can even reach 93 K. The transition was not sharp here, either, as shown in Fig. 26.2. The Meissner–Ochsenfeld effect could not be observed entirely: the samples did not become perfectly diamagnetic (that is, the susceptibility did not reach -1). Through the improvements in sample preparation during the past decades, the susceptibility of high-quality crystals is close to the ideal value of -1 now. When Y is replaced by a rare-earth metal, the transition temperature varies relatively little.

[5] As will be seen in Chapter 34, a superconductor is called *conventional* if the spin-singlet Cooper pairs responsible for superconductivity are formed by the electron–phonon interaction, and the order parameter exhibits s-wave symmetry, therefore the BCS theory can be used to describe it.

[6] See the footnote on page 6 of Volume 1.

The two families of materials are commonly called 214 and 123 compounds, referring to the ratios of the components. The critical temperature of some of them are listed in Table 26.9. Critical fields are not shown because they may depend on the direction of the applied field on account of the strong anisotropy of the sample. According to estimates, B_{c2} can be as high as 180 T in $YBa_2Cu_3O_{7-\delta}$.

Table 26.9. The transition temperatures for some members of the first discovered families of high-T_c superconductors

Compound	T_c (K)	Compound	T_c (K)
$La_{2-x}Ba_xCuO_4$	33	$YBa_2Cu_3O_{7-\delta}$	93
$La_{2-x}Sr_xCuO_4$	37.5	$LaBa_2Cu_3O_7$	89
$La_2CuO_{4+\delta}$	42	$NdBa_2Cu_3O_7$	96

These new superconductors share a number of common features. The characteristic layered structure of $La_{2-x}Sr_xCuO_4$ was shown in Fig. 7.23(*b*). We shall show it once again in Fig. 26.15, along with the structure of the parent compound of $YBa_2Cu_3O_{7-\delta}$.

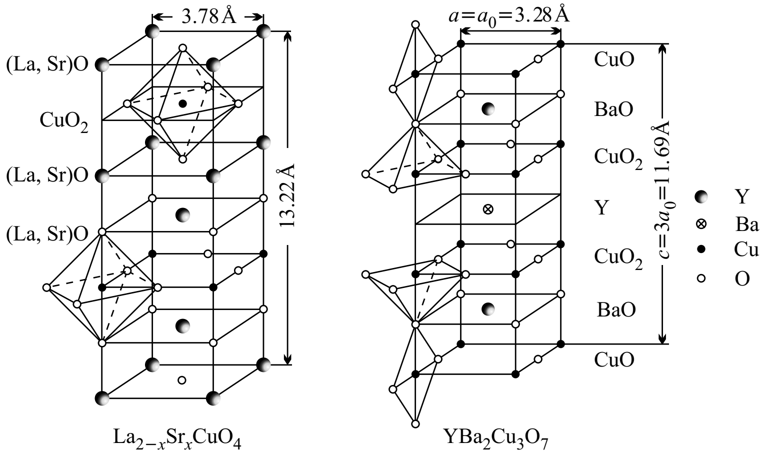


Fig. 26.15. The structure of $La_{2-x}Sr_xCuO_4$ and $YBa_2Cu_3O_7$

In both structures the copper ion and the octahedrally coordinated oxygen ions surrounding it make up Cu–O planes between the other constituents. In $La_{2-x}Sr_xCuO_4$ the La^{3+} ions sit between these planes. The Sr^{2+} (Ba^{2+}) ions only serve as a reservoir of carriers. In $YBa_2Cu_3O_{7-\delta}$ there are two Cu–O planes per primitive cell, and the oxygen deficiency provides the carriers.

More and more signs point to the conclusion that these Cu–O planes play an important role in superconductivity. That is why these families of superconductors are called cuprate superconductors.[7] In stoichiometric La_2CuO_4 the half spins of Cu^{2+} ions are coupled antiferromagnetically via superexchange through the oxygen ions, and make up an ordered antiferromagnetic structure. The situation is very similar in $YBa_2Cu_3O_{7-\delta}$ when $\delta \approx 1$. In both cases, the small variation of the concentration of one component rapidly destroys the magnetic order, and the sample becomes a superconductor. It looks as if the electrons or holes moving in the Cu–O plane, among the disordered magnetic moments, were responsible for superconductivity, and the relevant interaction between these electrons were not the same as in conventional superconductors. We shall discuss this point in detail in Chapter 34 on the microscopic theory of superconductivity.

Somewhat later appropriate technologies were developed for synthesizing material families in which the number of Cu–O planes can be controlled systematically. The materials in the series

$$\begin{gathered} HgBa_2Ca_{n-1}Cu_nO_{2n+2}, \\ TlBa_2Ca_{n-1}Cu_nO_{2n+3}, \\ Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}, \\ Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4} \end{gathered}$$

contain n Cu–O planes. The structures for $n = 1, 2, 3$ are shown for two of them in Figs. 26.16 and 26.17.

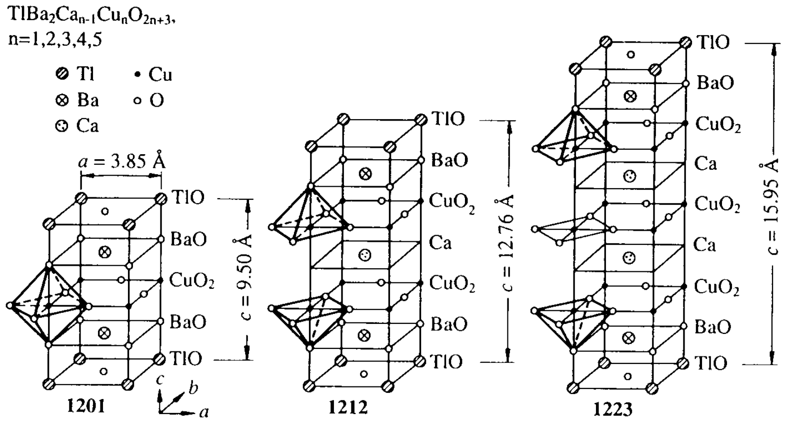


Fig. 26.16. The structure of cuprate superconductors of composition $TlBa_2Ca_{n-1}Cu_nO_{2n+3}$ for $n = 1, 2, 3$

[7] The names copper-oxide ceramic superconductors and ceramic superconductors are also used.

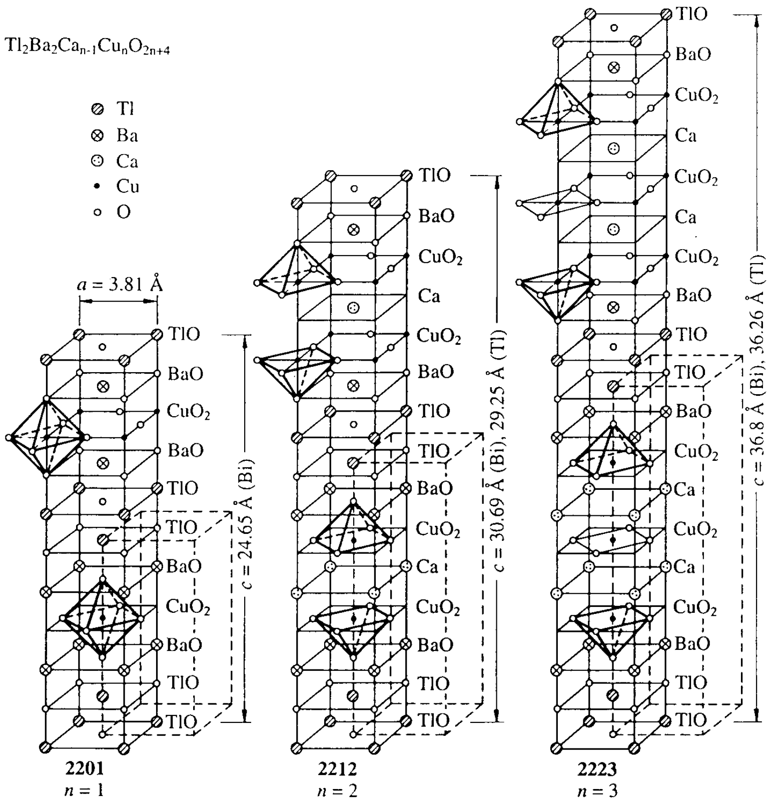


Fig. 26.17. The structure of cuprate superconductors of composition $Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$ for $n = 1, 2, 3$

As listed in Table 26.10, their transition temperatures increase for a while for increasing n, and then start to decrease again. The currently known highest T_c at atmospheric pressure, 138 K, was observed in the compound $Hg_{0.8}Tl_{0.2}Ba_2Ca_2Cu_3O_{8.33}$, which contains three Cu–O planes. At high pressure even higher transition temperatures can be reached: at 30 GPa $T_c = 157$ K was measured.

26.3 Phenomenological Description of Superconductivity

As was demonstrated in the previous sections, superconductors show different thermodynamic and electrodynamic behavior from normal metals. These differences can be fully understood only within the framework of the microscopic theory. Since that treatment draws heavily on the apparatus of the

Table 26.10. High-T_c superconducting compounds, their abbreviated notations, and their transition temperatures

Compound	Notation	T_c (K)
$Bi_2Sr_2CuO_6$	Bi-2201	9
$Bi_2Sr_2CaCu_2O_8$	Bi-2212	92
$Bi_2Sr_2Ca_2Cu_3O_{10}$	Bi-2223	110
$Tl_2Ba_2CuO_6$	Tl-2201	95
$Tl_2Ba_2CaCu_2O_8$	Tl-2212	118
$Tl_2Ba_2Ca_2Cu_3O_{10}$	Tl-2223	127
$Tl_2Ba_2Ca_3Cu_4O_{12}$	Tl-2234	109
$TlBa_2CaCu_2O_7$	Tl-1212	103
$TlBa_2Ca_2Cu_3O_9$	Tl-1223	133
$TlBa_2Ca_3Cu_4O_{11}$	Tl-1234	112
$TlBa_2Ca_4Cu_5O_{13}$	Tl-1245	<120
$HgBa_2CuO_4$	Hg-1201	95
$HgBa_2CaCu_2O_6$	Hg-1212	126
$HgBa_2Ca_2Cu_3O_8$	Hg-1223	135
$HgBa_2Ca_3Cu_4O_{10}$	Hg-1234	125
$HgBa_2Ca_4Cu_5O_{12}$	Hg-1245	101

many-body problem, it has to be deferred to Volume 3 (Chapter 34). Below we shall give a phenomenological description.

26.3.1 Thermodynamics of Superconductors

As follows from general thermodynamic considerations, the thermodynamic potential used to describe the properties of superconductors depends on the set of independent variables. If the magnetic properties are ignored, and the temperature T and volume V are given, the Helmholtz free energy

$$F = E - TS \tag{26.3.1}$$

has to be minimized. Likewise, when the temperature T and the pressure p are fixed, the state with the lowest value of the Gibbs free energy

$$G = E - TS + pV \tag{26.3.2}$$

is the equilibrium state.

The variation of the volume does not play an important role in superconductors – however, magnetic properties are essential. Then the work done by the magnetic field on the system needs to be taken into account in the internal-energy balance. The variation of the internal energy density is therefore given by

$$\mathrm{d}w = T\,\mathrm{d}s + \mu_0 \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{M}. \tag{26.3.3}$$

Henceforth we shall systematically use lowercase symbols to denote the densities of the appropriate extensive thermodynamic quantities. The contribution $\mu_0 \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{H}$ of the electromagnetic field energy has to be added to the internal energy. The variation of the internal energy density is then

$$\mathrm{d}w = T\,\mathrm{d}s + \mu_0 \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{M} + \mu_0 \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{H} = T\,\mathrm{d}s + \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{B}. \tag{26.3.4}$$

If the magnetic induction (flux density) $\boldsymbol{B}$ is the independent, natural variable, the variation of the Helmholtz free energy density f is

$$\mathrm{d}f(T, \boldsymbol{B}) = -s\,\mathrm{d}T + \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{B}. \tag{26.3.5}$$

In most experiments it is not the magnetic flux density that is controlled directly but the applied magnetic field $\boldsymbol{H}$, by means of applied currents. The relevant thermodynamic potential is then the Gibbs free energy, which is a function of T and $\boldsymbol{H}$. It is obtained from the Helmholtz free energy $F(T, \boldsymbol{B})$ by a Legendre transformation. In terms of the densities,

$$g(T, \boldsymbol{H}) = f(T, \boldsymbol{B}) - \boldsymbol{B} \cdot \boldsymbol{H}. \tag{26.3.6}$$

The behavior of the system is determined by the minimum of this quantity. The condition for thermodynamic equilibrium is

$$\left(\frac{\partial g}{\partial \boldsymbol{B}}\right)_{\boldsymbol{H}} = 0\,. \tag{26.3.7}$$

When the temperature and magnetic field are varied, the Gibbs potential changes by

$$\mathrm{d}g(T, \boldsymbol{H}) = -s\,\mathrm{d}T - \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{H}. \tag{26.3.8}$$

This implies the following formulas for the entropy and magnetic induction:

$$s = -\left(\frac{\partial g}{\partial T}\right)_{\boldsymbol{H}}, \qquad \boldsymbol{B} = -\left(\frac{\partial g}{\partial \boldsymbol{H}}\right)_{T}. \tag{26.3.9}$$

Integration of (26.3.8) gives the variation of the Gibbs potential for a sample placed in a magnetic field. In isotropic systems

$$g(T, H) = g(T, 0) - \int_0^H \boldsymbol{B}(H') \cdot \mathrm{d}\boldsymbol{H}'\,. \tag{26.3.10}$$

In the normal state, if the sample itself is not magnetically ordered, $\mu_{\mathrm{r}} \approx 1$ and $\boldsymbol{B} = \mu_0 \boldsymbol{H}$ to a good approximation, thus

$$g_{\mathrm{n}}(T, H) = g_{\mathrm{n}}(T, 0) - \tfrac{1}{2}\mu_0 H^2. \tag{26.3.11}$$

In the superconducting state of type I superconductors $\boldsymbol{B} = 0$ in the bulk of the sample, so, if the contribution of the thin surface layer can be neglected,

$$g_{\mathrm{s}}(T,H)=g_{\mathrm{s}}(T,0)\,. \tag{26.3.12}$$

The free energy in the superconducting and normal phases are plotted against the magnetic field in Fig. 26.18.

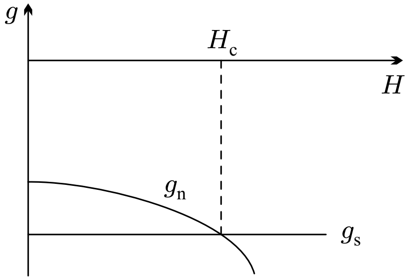


Fig. 26.18. The dependence of the Gibbs free energy on the magnetic field in the superconducting and normal phases

In the absence of an applied magnetic field the superconducting phase is stable below the critical temperature as its Gibbs free energy is lower. The difference of the free energies of the superconducting and normal phases is called the condensation energy. In the presence of a magnetic field the free energy of the normal state becomes lower, while that of the superconducting phase is left unchanged. In fields exceeding a critical strength, the Gibbs free energy is lower for the normal state than for the superconducting state. The critical magnetic field H_{c}, where the phase transition occurs, is determined by

$$g_{\mathrm{s}}(T,H_{\mathrm{c}})=g_{\mathrm{n}}(T,H_{\mathrm{c}})\,. \tag{26.3.13}$$

Using the previous formulas,

$$g_{\mathrm{s}}(T,0)=g_{\mathrm{n}}(T,0)-\tfrac{1}{2}\mu_0H_{\mathrm{c}}^2\,. \tag{26.3.14}$$

The condensation energy can thus be simply related to the critical field. This is particularly noteworthy because the condensation energy can be determined from the microscopic theory, and so the temperature dependence of the critical field can be derived.

The difference of the two Gibbs free energies in a magnetic field can also be written as

$$g_{\mathrm{s}}(T,H)=g_{\mathrm{n}}(T,H)+\tfrac{1}{2}\mu_0\left(H^2-H_{\mathrm{c}}^2\right). \tag{26.3.15}$$

The entropies of the two phases at the critical field of the transition can be determined through (26.3.9), leading to

$$s_{\mathrm{s}}-s_{\mathrm{n}}=\mu_0H_{\mathrm{c}}\,\frac{\mathrm{d}H_{\mathrm{c}}}{\mathrm{d}T}\,. \tag{26.3.16}$$

It follows from the previously derived temperature dependence of the critical field that the quantity on the right-hand side is negative, thus the entropy of the superconducting phase is indeed lower than that of the normal phase. The latent heat of the phase transition is then

$$q = T(s_\mathrm{n} - s_\mathrm{s}) = -T\mu_0 H_\mathrm{c}\,\frac{\mathrm{d}H_\mathrm{c}}{\mathrm{d}T}\,. \tag{26.3.17}$$

This quantity vanishes at $T = 0$ and $T = T_\mathrm{c}$, where $H_\mathrm{c} = 0$. The transition is second order in both points and first order everywhere else.

Calculating the specific heat from $c = T\partial s/\partial T$,

$$c_\mathrm{s} - c_\mathrm{n} = T\mu_0\left[H_\mathrm{c}\frac{\mathrm{d}^2 H_\mathrm{c}}{\mathrm{d}T^2} + \left(\frac{\mathrm{d}H_\mathrm{c}}{\mathrm{d}T}\right)^2\right]. \tag{26.3.18}$$

It seems justified to assume that the transition to the superconducting state does not modify the phonon spectrum, so the specific-heat contribution of lattice vibrations is the same in the two phases. The jump in the specific heat is related to changes in the excitation spectrum of the electron system.

Using the empirical formula (26.1.3) for the temperature dependence of the critical field,

$$c_\mathrm{s} - c_\mathrm{n} = 6\mu_0\frac{H_\mathrm{c}^2(0)}{T_\mathrm{c}}\left[-\frac{1}{3}\left(\frac{T}{T_\mathrm{c}}\right) + \left(\frac{T}{T_\mathrm{c}}\right)^3\right]. \tag{26.3.19}$$

At low temperatures the second term is much smaller than the first. Since the specific heat of the superconductor is exponentially small, we may identify the term that is proportional to the temperature with the electronic contribution $c_\mathrm{n} = \gamma T$ to the specific heat of normal metals. By making use of (26.1.3) once again, we have

$$\gamma = -\mu_0 H_\mathrm{c}\frac{\mathrm{d}^2 H_\mathrm{c}}{\mathrm{d}T^2} = 2\mu_0\left(\frac{H_\mathrm{c}(0)}{T_\mathrm{c}}\right)^2. \tag{26.3.20}$$

According to (16.2.91), γ can be related to the electronic density of states at the Fermi energy, which can thus also be determined from the measurement of H_c and T_c.

The jump in the specific heat in the critical point is given by the *Rutgers formula*[8]

$$(c_\mathrm{s} - c_\mathrm{n})_{T_\mathrm{c}} = T_\mathrm{c}\,\mu_0\left(\frac{\mathrm{d}H_\mathrm{c}}{\mathrm{d}T}\right)^2_{T_\mathrm{c}} = 4\mu_0\frac{H_\mathrm{c}^2(0)}{T_\mathrm{c}}\,. \tag{26.3.21}$$

The quantities on the two sides can be measured independently. Experimental data usually obey this formula quite well.

Thermodynamic relations can also be applied to type II superconductors, even though the *thermodynamic critical field* – which is derived from the energy difference between the normal and superconducting phases – does not

[8] A. J. Rutgers, 1936.

have a direct physical meaning. In the normal phase, where the magnetic induction inside the sample is determined by $B = \mu_0 H$,

$$\frac{\partial}{\partial H}\, g_{\mathrm{n}}(T,H) = -\mu_0 H\,. \tag{26.3.22}$$

In the superconducting phase

$$\frac{\partial}{\partial H}\, g_{\mathrm{s}}(T,H) = -B \tag{26.3.23}$$

is satisfied. By combining the two equations,

$$\frac{\partial}{\partial H}\bigl[g_{\mathrm{n}}(T,H) - g_{\mathrm{s}}(T,H)\bigr] = B - \mu_0 H = \mu_0 M\,. \tag{26.3.24}$$

Thus, when the magnetization is integrated up to the upper critical field,

$$\int\limits_0^{H_{\mathrm{c2}}} M\,\mathrm{d}H = \frac{1}{\mu_0}\bigl[g_{\mathrm{n}}(T,H_{\mathrm{c2}}) - g_{\mathrm{s}}(T,H_{\mathrm{c2}})\bigr] - \frac{1}{\mu_0}\bigl[g_{\mathrm{n}}(T,0) - g_{\mathrm{s}}(T,0)\bigr]. \tag{26.3.25}$$

At the upper critical field the Gibbs free energy is the same in the superconducting and normal phases, and for $H = 0$ this difference is the condensation energy. Writing this in the same form as for type I superconductors, we can introduce the thermodynamic critical field H_{c} through the definition

$$\tfrac{1}{2}\mu_0 H_{\mathrm{c}}^2 = g_{\mathrm{n}}(T,0) - g_{\mathrm{s}}(T,0)\,. \tag{26.3.26}$$

Then

$$\int\limits_0^{H_{\mathrm{c2}}} M\,\mathrm{d}H = -\tfrac{1}{2}H_{\mathrm{c}}^2\,. \tag{26.3.27}$$

By measuring the equilibrium magnetization curve, a measurement instruction can be given for the determination of the thermodynamic critical field.

26.3.2 Electrodynamics of Superconductors

In 1935 the brothers FRITZ and HEINZ LONDON proposed a simple system of equations for the description of the electric and magnetic properties of superconductors. They assumed that the Maxwell equations could be applied in their original form, only the constitutive relations required modification. In line with an early phenomenological theory of superconductivity, the *Gorter–Casimir two-fluid model*,[9] they considered that besides normal electrons, superconductors also contain another type of carrier, of charge $-e^*$ and mass m^*, in a number density n_{s}^*. For simplicity, we shall call them superconducting electrons in this chapter. Normal electrons participate in scattering, and

[9] C. J. GORTER and H. B. G. CASIMIR, 1934.

their relaxation time τ_n is finite, whereas superconducting electrons move in the sample without dissipation. Their current density can be written in the customary form

$$\boldsymbol{j}_\mathrm{s} = -e^* n_\mathrm{s}^* \boldsymbol{v}_\mathrm{s}\,, \tag{26.3.28}$$

where $\boldsymbol{v}_\mathrm{s}$ is the velocity of superconducting electrons. If this current is not dissipated, the superconducting electrons accelerate freely in an electric field; their equation of motion is

$$m^* \frac{\mathrm{d}\boldsymbol{v}_\mathrm{s}}{\mathrm{d}t} = -e^* \boldsymbol{E}\,. \tag{26.3.29}$$

The equation governing the variations of the current with time is then

$$\boxed{\frac{\mathrm{d}\boldsymbol{j}_\mathrm{s}}{\mathrm{d}t} = \frac{n_\mathrm{s}^* e^{*2}}{m^*} \boldsymbol{E}\,.} \tag{26.3.30}$$

This is the *first London equation*, which formulates the infinity of the conductivity.

Another relation is obtained by substituting this formula into the Maxwell equation

$$\operatorname{curl} \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}\,. \tag{26.3.31}$$

This leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{m^*}{n_\mathrm{s}^* e^{*2}} \operatorname{curl} \boldsymbol{j}_\mathrm{s}\right) = -\frac{\partial \boldsymbol{B}}{\partial t}\,, \tag{26.3.32}$$

and after some rearrangement to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\operatorname{curl} \boldsymbol{j}_\mathrm{s} + \frac{n_\mathrm{s}^* e^{*2}}{m^*} \boldsymbol{B}\right) = 0\,. \tag{26.3.33}$$

The London brothers assumed that the parenthesized expression is not only constant in time but zero, that is

$$\boxed{\operatorname{curl} \boldsymbol{j}_\mathrm{s} = -\frac{n_\mathrm{s}^* e^{*2}}{m^*} \boldsymbol{B}\,.} \tag{26.3.34}$$

This is the *second London equation.* When the magnetic induction is expressed in terms of the vector potential as $\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$,

$$\boldsymbol{j}_\mathrm{s} = -\frac{n_\mathrm{s}^* e^{*2}}{m^*} \boldsymbol{A} \tag{26.3.35}$$

is obtained, which implies a local relationship between the superconducting current and the vector potential.

By introducing the parameter

$$\lambda_{\mathrm{L}}^2 = \frac{m^*}{n_{\mathrm{s}}^* e^{*2} \mu_0}\,, \tag{26.3.36}$$

which has a dimension of length squared, the two London equations can also be written as

$$\boxed{\boldsymbol{E} = \mu_0 \lambda_{\mathrm{L}}^2 \frac{\mathrm{d}\boldsymbol{j}_{\mathrm{s}}}{\mathrm{d}t}\,, \qquad \boldsymbol{B} = -\mu_0 \lambda_{\mathrm{L}}^2 \operatorname{curl} \boldsymbol{j}_{\mathrm{s}}\,.} \tag{26.3.37}$$

In the discussion of the microscopic theory we shall see that the superconducting electrons are in fact bound electron pairs, the so-called Cooper pairs. To match this phenomenological theory with experimental results and the microscopic theory, we shall need to use $e^* = 2e$, $m^* = 2m_{\mathrm{e}}$, and $n_{\mathrm{s}}^* = n_{\mathrm{s}}/2$, where n_{s} is the actual density of superconducting electrons. Note that if the electron mass and charge, along with the actual density of electrons responsible for superconductivity were used instead of the corresponding effective parameters, a very similar formula would be obtained for λ_{L}:

$$\lambda_{\mathrm{L}}^2 = \frac{m_{\mathrm{e}}}{n_{\mathrm{s}} e^2 \mu_0}\,. \tag{26.3.38}$$

The reason behind choosing the second London equation in the form given above is that it leads naturally to the Meissner–Ochsenfeld effect. According to the Maxwell equations, in the static case

$$\frac{1}{\mu_0} \operatorname{curl} \boldsymbol{B} = \boldsymbol{j}_{\mathrm{s}}\,. \tag{26.3.39}$$

By taking the curl of both sides, and making use of the second London equation,

$$\operatorname{curl}\operatorname{curl} \boldsymbol{B} = \mu_0 \operatorname{curl} \boldsymbol{j}_{\mathrm{s}} = -\mu_0 \frac{n_{\mathrm{s}}^* e^2}{m} \boldsymbol{B} = -\frac{1}{\lambda_{\mathrm{L}}^2} \boldsymbol{B}\,. \tag{26.3.40}$$

We shall determine the solution of this equation in the special case where the $x > 0$ half-space is filled by a superconducting material and the $x < 0$ half-space by a normal metal or vacuum. By applying a uniform magnetic field in the z-direction, the magnetic induction becomes nonuniform at the superconductor side of the interface, as shown in Fig. 26.19.

As the interface is the $x = 0$ plane, all spatial variations are in the x-direction. The spatial variations of B_z are governed by

$$\frac{\mathrm{d}^2 B_z}{\mathrm{d}x^2} = \frac{1}{\lambda_{\mathrm{L}}^2} B_z\,. \tag{26.3.41}$$

We are now seeking solutions that also satisfy the auxiliary condition that the magnetic field should be B_0 in the $x < 0$ region, while deep inside the superconductor it should vanish, as required for the Meissner–Ochsenfeld effect. The result is

$$B_z(x) = \begin{cases} B_0\,, & x < 0\,, \\ B_0 \mathrm{e}^{-x/\lambda_{\mathrm{L}}}\,, & x > 0\,. \end{cases} \tag{26.3.42}$$

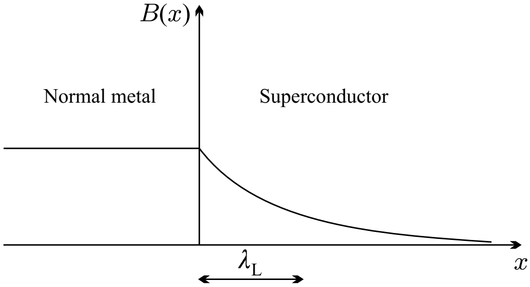


Fig. 26.19. Penetration of an applied magnetic field into the superconductor. The region is characterized by the penetration depth λ_L

Thus λ_L determines how deeply the magnetic induction can penetrate into the superconductor; only beyond that does the superconductor show the characteristic bulk behavior ($B = 0$). For this reason, λ_L is called the *London penetration depth.*

(26.3.39) implies that a surface current flows in the superconductor if the field is finite outside. The current density depends on the distance from the surface as

$$j_s = j_s^{(0)} e^{-x/\lambda_L} = \frac{1}{\mu_0 \lambda_L} B_0 e^{-x/\lambda_L} . \tag{26.3.43}$$

This surface current screens the external magnetic field inside the superconductor. The critical current – which is the highest current that can be passed through a superconducting wire of radius R without the sample becoming a normal conductor – is then straightforward to determine. Since the current can flow only close to the surface, practically within a layer of thickness λ_L, the current is approximately

$$I = 2\pi R \lambda_L j_s^{(0)} . \tag{26.3.44}$$

On the other hand, the current density on the surface cannot be larger than the value associated with the critical field outside the sample, so $j_s^{(0)}$ cannot exceed

$$j_c = \frac{1}{\mu_0 \lambda_L} B_c . \tag{26.3.45}$$

The critical current is then

$$I_c = 2\pi R B_c / \mu_0 . \tag{26.3.46}$$

As the density of superconducting electrons is temperature dependent, so is the London penetration depth. Assuming that at $T = 0$ all electrons become superconducting – that is, n_s^* can be identified with the density of conduction electrons –, λ_L is expected to be on the order of 100 nm (1000 Å) for typical metallic electron densities. The extrapolated zero-temperature value of the measured penetration depth is given in Table 26.11 for a number of superconductors. The measured values are in order-of-magnitude agreement with the estimated values given above.

Table 26.11. The experimental values for the penetration depth and coherence length in certain superconductors

Element	λ_L (nm)	ξ_0 (nm)	Compound	λ_L (nm)	ξ_0 (nm)
Al	45	1550	Nb_3Sn	65	3
Cd	110	760	Nb_3Ge	90	3
In	40	360	V_3Si	60	3
Nb	52	39	$PbMo_6S_8$	200	2
Pb	39	87	K_3C_{60}	240	2.6
Sn	42	180	UBe_{13}	400	7

According to the Gorter–Casimir two-fluid model, the density of normal electrons is proportional to the fourth power of T at finite temperatures, so the density of superconducting electrons is given by

$$n_\mathrm{s} = n_0\bigl[1 - (T/T_\mathrm{c})^4\bigr]\,. \tag{26.3.47}$$

The approximate formula for the temperature dependence of the penetration depth is therefore

$$\lambda_\mathrm{L}(T) = \lambda_\mathrm{L}(0)\bigl[1 - (T/T_\mathrm{c})^4\bigr]^{-1/2}, \tag{26.3.48}$$

in fair agreement with measurements. As the critical point is approached, the penetration depth diverges as the inverse square root of $T_\mathrm{c} - T$.

In high-T_c superconductors the penetration depth and coherence length exhibit strong anisotropy because of the layered structure. Some relevant data are listed in Table 26.12.

Table 26.12. The experimental values for the penetration depth and coherence length in two high-T_c cuprate superconductors, parallel and perpendicular to the Cu–O planes

Compound	$\lambda_\parallel$ (nm)	$\lambda_\perp$ (nm)	$\xi_\parallel$ (nm)	$\xi_\perp$ (nm)
$YBa_2Cu_3O_7$	100	500	1.2	0.3
$HgBa_2Ca_2Cu_3O_{10}$	130	3500	1.5	0.2

26.3.3 Pippard Coherence Length

The measured value of the penetration depth is often larger than the prediction of the London equation (26.3.36). The reason for this is that the second London equation assumes a local relationship between the current density of superconducting electrons and the vector potential. However, this assumption is too strong, and is not satisfied in all superconductors. In analogy to the Reuter–Sondheimer theory[10] of the anomalous skin effect, which led to a nonlocal generalization of Ohm's law, A. B. Pippard (1953) proposed to replace the local London equation with a nonlocal relationship, in which the current at point $\boldsymbol{r}$ does not depend on the vector potential in $\boldsymbol{r}$ alone but on its values $\boldsymbol{A}(\boldsymbol{r}')$ over a region of radius ξ_0 around $\boldsymbol{r}$. Based on Chambers' formula,[11]

$$\boldsymbol{j}(\boldsymbol{r}) = \frac{3\sigma}{4\pi l} \int \frac{\boldsymbol{R}[\boldsymbol{R} \cdot \boldsymbol{E}(\boldsymbol{r}')]}{R^4} \exp(-R/l)\,\mathrm{d}\boldsymbol{r}' \tag{26.3.49}$$

for the anomalous skin effect, where $\boldsymbol{R} = \boldsymbol{r} - \boldsymbol{r}'$, $R = |\boldsymbol{R}|$, σ is the macroscopic conductivity, and l is the mean free path, Pippard assumed

$$\begin{aligned} \boldsymbol{j}_{\mathrm{s}}(\boldsymbol{r}) &= -\frac{3n_{\mathrm{s}}e^2}{4\pi\xi_0 m_{\mathrm{e}}} \int \frac{\boldsymbol{R}[\boldsymbol{R} \cdot \boldsymbol{A}(\boldsymbol{r}')]}{R^4} \exp(-R/\xi_0)\,\mathrm{d}\boldsymbol{r}' \\ &= -\frac{3}{4\pi\mu_0}\frac{1}{\xi_0\lambda_{\mathrm{L}}^2} \int \frac{\boldsymbol{R}[\boldsymbol{R} \cdot \boldsymbol{A}(\boldsymbol{r}')]}{R^4} \exp(-R/\xi_0)\,\mathrm{d}\boldsymbol{r}' \end{aligned} \tag{26.3.50}$$

for superconductors. The temperature-independent characteristic length ξ_0 is the *Pippard coherence length.*

To estimate its value, we shall assume that the superconducting state is due dominantly to the electrons in the region of width $k_{\mathrm{B}}T_{\mathrm{c}}$ around the Fermi energy. The uncertainty of the momentum is then $\Delta p \approx k_{\mathrm{B}}T_{\mathrm{c}}/v_{\mathrm{F}}$. The position uncertainty

$$\Delta x \geq \hbar/\Delta p = \frac{\hbar v_{\mathrm{F}}}{k_{\mathrm{B}}T_{\mathrm{c}}} \tag{26.3.51}$$

implied by the Heisenberg uncertainty principle can then be identified with the coherence length ξ_0. According to the microscopic theory, its more accurate value can be expressed in terms of the energy gap $\mathit{\Delta}_0$ – which is of the same order of magnitude as $k_{\mathrm{B}}T_{\mathrm{c}}$:

$$\xi_0 = \frac{\hbar v_{\mathrm{F}}}{\pi \mathit{\Delta}_0} = a\frac{\hbar v_{\mathrm{F}}}{k_{\mathrm{B}}T_{\mathrm{c}}}\,, \tag{26.3.52}$$

where $a \approx 0.180$. Its experimental values for a number of materials are listed in Tables 26.11 and 26.12. Visibly, ξ_0 can be two orders of magnitude larger or smaller than the London penetration depth. This plays an important role in the classification of superconductors.

[10] G. E. H. Reuter and E. H. Sondheimer, 1948.

[11] R. G. Chambers, 1952.

If $\lambda_{\mathrm{L}} \gg \xi_0$, then $\boldsymbol{A}(\boldsymbol{r})$ varies little over the interesting part of the domain of integration in (26.3.50), and a local relationship between the current density and the vector potential expressed by the second London equation is recovered. In this case the penetration of the magnetic field is determined by λ_{L} alone. As we shall see later, such a situation is encountered in type II superconductors, which are therefore also called London superconductors.

In the opposite limit, which is typical of type I superconductors, Pippard's nonlocal relationship between the current and the vector potential has to be used. That is why type I superconductors are sometimes referred to as Pippard superconductors. The actual penetration depth – the distance over which the vector potential falls off – has to be determined self-consistently. Owing to the sharp decrease of the vector potential, the integral in (26.3.50) is reduced roughly by a factor of order λ/ξ_0:

$$\boldsymbol{j}_{\mathrm{s}}(\boldsymbol{r}) = -\frac{\lambda}{\xi_0}\frac{1}{\mu_0\lambda_{\mathrm{L}}^2}\boldsymbol{A}(\boldsymbol{r})\,. \tag{26.3.53}$$

When the penetration depth is determined from this relation using the Maxwell equations, $\lambda/(\xi_0\lambda_{\mathrm{L}}^2)$ must be identified with $1/\lambda^2$. This gives

$$\lambda = \lambda_{\mathrm{L}}^{2/3}\xi_0^{1/3}\,, \tag{26.3.54}$$

which is larger than the London value, in agreement with the measurements.

When impurities are present, and the mean free path of electrons is l because of impurity scattering, the actual coherence length is given by

$$\frac{1}{\xi} = \frac{1}{\xi_0} + \frac{1}{l}\,. \tag{26.3.55}$$

In the very imperfect (dirty) limit $l \ll \xi_0$, and thus the exponential factor in the Pippard formula is $\mathrm{e}^{-|\boldsymbol{r}|/l}$. By evaluating the integral, the current and the vector potential are now related through

$$\boldsymbol{j}_{\mathrm{s}}(\boldsymbol{r}) = -\frac{l}{\xi_0}\frac{1}{\mu_0\lambda_{\mathrm{L}}^2}\boldsymbol{A}(\boldsymbol{r})\,. \tag{26.3.56}$$

The actual penetration depth of the magnetic field is then obtained from

$$\frac{l}{\xi_0}\frac{1}{\lambda_{\mathrm{L}}^2} = \frac{1}{\lambda^2}\,, \tag{26.3.57}$$

which gives

$$\lambda = \lambda_{\mathrm{L}}\left(\frac{\xi_0}{l}\right)^{1/2}\,. \tag{26.3.58}$$

26.3.4 Flux Quantization

The condition that the current should vanish inside the superconductor immediately implies the property mentioned among the experimental findings: the magnetic flux through a ring cannot take any arbitrary value, only the integral multiples of a flux quantum $\varPhi_0$. To demonstrate this, we have to start with the Bohr quantization condition

$$\oint_C \boldsymbol{p} \cdot \mathrm{d}\boldsymbol{l} = nh \,, \tag{26.3.59}$$

where the line integral is along some closed path. As discussed in Chapter 3, the canonical momentum $\boldsymbol{p}$ and the kinetic momentum $m^* \boldsymbol{v}_{\mathrm{s}}$ that determines the kinetic energy are related by

$$\boldsymbol{p} = m^* \boldsymbol{v}_{\mathrm{s}} - e^* \boldsymbol{A} \,. \tag{26.3.60}$$

Substituting this into the quantization condition,

$$m^* \oint_C \boldsymbol{v}_{\mathrm{s}} \cdot \mathrm{d}\boldsymbol{l} - e^* \oint_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = nh \,. \tag{26.3.61}$$

Expressing the velocity in terms of the current of superconducting electrons and rearranging the terms gives

$$\frac{m^*}{n_{\mathrm{s}}^* e^{*2}} \oint_C \boldsymbol{j}_{\mathrm{s}} \cdot \mathrm{d}\boldsymbol{l} + \oint_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = \frac{|n| h}{e^*} \,. \tag{26.3.62}$$

If the ring is sufficiently thick for that no current can flow in its interior, the integration path can be chosen in such a way that the contribution of the first term on the left-hand side be zero. According to Stokes' theorem, the second term is the magnetic flux $\varPhi$ through a surface F bounded by the closed curve C:

$$\oint_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} = \int_F \operatorname{curl} \boldsymbol{A} \, \mathrm{d}F = \varPhi \,. \tag{26.3.63}$$

So the magnetic flux through the superconducting ring can take only discrete values, namely, the integral multiples of the flux quantum

$$\boxed{\varPhi_0 = \frac{h}{e^*} \,.} \tag{26.3.64}$$

To establish agreement with the measurement shown in Fig. 26.6, e^* must be chosen as twice the elementary charge.

When the region is chosen in such a way that current flows on its boundary, then the quantization condition (in units of h/e^*) does not apply to the magnetic flux but to the fluxoid

$$\Phi + \frac{m^*}{n_s^* e^{*2}} \oint_C \boldsymbol{j}_s \cdot \mathrm{d}\boldsymbol{l} \,, \tag{26.3.65}$$

which also contains the contribution of the superconducting current.

26.4 Ginzburg–Landau Theory

By complementing the general thermodynamic relations by a simple assumption about the temperature dependence of the critical field, and extending the Maxwell equations to the superconducting phase by means of the London equations, we gave a quite simple description of the thermodynamics and electrodynamics of the superconducting state in the previous section. The phase transition was studied through the comparison of the free energy of the normal and superconducting states, both of which were assumed to be homogeneous. If the sample is allowed to be inhomogeneous, i.e., normal and superconducting regions can alternate in it – a possibility that was mentioned as an empirical fact in connection with the Shubnikov phase of type II superconductors – then simple (and, as we shall see, naive) considerations straightforwardly lead to the conclusion that the homogeneous superconducting state can never be stable energetically. To demonstrate this, assume that normal and superconducting layers alternate inside the sample in such a way that the normal layers are much thinner than the superconducting ones, but even the thickness of the latter is smaller than the penetration length. In this geometry the normal regions contribute negligibly to the total free energy, however, they allow the magnetic field to penetrate into the superconducting regions, too. This way, the Gibbs free energy of homogeneous superconductors could be reduced. Moreover, superconductivity would not disappear at the critical field H_c (which is related to the condensation energy of bulk superconductors): it could persist in a configuration where thin superconducting and normal layers are stacked because the destruction of superconductivity requires a higher critical field in thin films than in bulk samples.

The existence of type I superconductors is in contradiction with this naive expectation. In type II superconductors the alternation of normal and superconducting regions is indeed observed between the upper and lower critical fields, however, homogeneous superconductivity is found below $H_{\mathrm{c}1}$ here, too. This indicates the necessity of a more precise study of normal metal–superconductor interfaces, most notably their energy. This requires a better treatment of superconductivity than the one based on the London equations.

This theory, called the *Ginzburg–Landau theory*, is the generalization of Landau's theory of second-order phase transitions to inhomogeneous superconductors in a magnetic field. The fundamental assumptions of Landau's theory were presented in Chapter 14. Using that as a starting point, we shall now derive the Ginzburg–Landau equations, and then use them to describe the macroscopic properties and the vortices that appear in the intermediate state of type II superconductors.

26.4.1 Ginzburg–Landau Equations

In 1950, before the advent of the microscopic theory of superconductivity, V. L. GINZBURG[12] and L. D. LANDAU generalized the Landau theory of second-order phase transitions to the normal–superconducting transition. They assumed that the superconducting phase can be characterized by an order parameter that is finite only in the ordered, superconducting phase, and varies continuously below the critical point, staring from zero at T_c. They also assumed that a free-energy functional can be defined in the vicinity of the phase-transition point, and it can be expanded in powers of the order parameter. The equilibrium value of the order parameter can be determined from the minimum of the functional, whereas the actual value of the free energy is given by the value of the functional at the equilibrium order parameter.

As the magnetic field is known to penetrate into type II superconductors inhomogeneously, we shall use the free-energy-density formula (14.5.17), which takes the spatial variations of the order parameter into account, too. Besides, GINZBURG and LANDAU assumed that the order parameter is somehow related to the wavefunction of superconducting electrons, and is therefore a complex quantity. Their second, even more important assumption was that the term containing the gradient of the order parameter in the series expansion of the free-energy density can be considered to be related to the kinetic energy of superconducting electrons. This physical insight then implied that, just like for an electron in a magnetic field, the effects of the magnetic field have to be taken into account by replacing the canonical momentum operator $-\mathrm{i}\hbar\boldsymbol{\nabla}$ by the kinetic momentum operator, which also contains the vector potential and the charge $-e^*$ of superconducting electrons.[13] The free-energy density of the superconducting state is then

$$f_\mathrm{s} = f_\mathrm{n} + \alpha(T)|\psi|^2 + \frac{1}{2}\beta(T)|\psi|^4 + \frac{1}{2m^*}\left|\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e^*\boldsymbol{A}\right)\psi\right|^2 + \frac{1}{2\mu_0}\boldsymbol{B}^2 \,, \qquad (26.4.1)$$

where the last term is the energy of the magnetic field $\boldsymbol{B} = \operatorname{curl}\boldsymbol{A}$. In their approach the magnetic induction (or the vector potential) is fixed, which is why the Helmholtz free energy is studied.

[12] Well after the death of LANDAU (1908–1968), who was awarded the Nobel Prize in 1962 (see footnote on page 28 of Volume 1), VITALY LAZAREVICH GINZBURG (1916–) shared the Nobel Prize with ALEXEI ALEXEEVICH ABRIKOSOV (1928–) and ANTHONY JAMES LEGGETT (1938–) in 2003 “for pioneering contributions to the theory of superconductors and superfluids”.

[13] The free energy is gauge invariant only if a universal value is taken for the charge. GINZBURG and LANDAU argued that there was no reason to choose it to be different from the electron charge. It is now known that e^* should be chosen as $2e$, a universal value for all superconductors.

Since the order parameter of the superconducting state has to vanish in the normal state – in other words, the free-energy minimum must be at $\psi = 0$ in the normal phase and at some nonzero value in the superconducting state –, the phase transition occurs at that temperature T_c for which $\alpha(T_\mathrm{c}) = 0$. The parameter α is positive above T_c and negative below it. Assuming a linear temperature dependence in the vicinity of the transition point,

$$\alpha(T) = a(T - T_\mathrm{c}) \,, \qquad a > 0 \,, \tag{26.4.2}$$

while β is chosen to be positive and temperature independent.

The spatial distribution of the order parameter and magnetic induction can be determined from the minimum of the free energy. Since the order parameter is complex, the real and imaginary parts need to be varied separately – or else, the minimum can also be sought with respect to ψ and its conjugate. The variation of the free energy is

$$\begin{aligned} \delta F_\mathrm{s} = \int \mathrm{d}\boldsymbol{r} \biggl[& \alpha\psi \, \delta\psi^* + \beta|\psi|^2\psi \, \delta\psi^* \\ & + \frac{1}{2m^*} \left(-\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \delta\psi^* \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi + \text{c.c.} \biggr] \\ & + \int \mathrm{d}\boldsymbol{r} \left[\frac{\boldsymbol{B}}{2\mu_0} \operatorname{curl} \delta\boldsymbol{A} + \frac{e^*}{2m^*} \psi^* \delta\boldsymbol{A} \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi + \text{c.c.} \right] . \end{aligned} \tag{26.4.3}$$

Integrating the terms containing the derivative of $\delta\psi^*$ and $\delta\boldsymbol{A}$ by parts,

$$\begin{aligned} \delta F_\mathrm{s} = & \int \mathrm{d}\boldsymbol{r} \biggl\{ \delta\psi^* \biggl[\alpha\psi + \beta|\psi|^2\psi + \frac{1}{2m^*} \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right)^2 \psi \biggr] + \text{c.c.} \biggr\} \\ & + \int \mathrm{d}\boldsymbol{r} \biggl\{ \delta\boldsymbol{A} \biggl[\frac{\operatorname{curl} \boldsymbol{B}}{2\mu_0} + \frac{e^*}{2m^*} \psi^* \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi \biggr] + \text{c.c.} \biggr\} . \end{aligned} \tag{26.4.4}$$

In addition to that, the integrated part gives a surface term

$$\int \mathrm{d}\boldsymbol{S} \left[\delta\psi^* \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi + \text{c.c.} \right] . \tag{26.4.5}$$

In the calculus of variations boundary conditions have to be specified, too. Customarily, ψ and its variation are both assumed to vanish at the boundaries. Instead, Ginzburg and Landau proposed the condition of vanishing current through the surface. This condition is fulfilled when the component of the integrand along the surface normal $\boldsymbol{n}$ satisfies

$$\boldsymbol{n} \cdot \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi = 0 \tag{26.4.6}$$

on the boundary. This indeed leads to the correct result for superconductor–insulator interfaces. The situation is not so simple for superconductor–normal metal interfaces. Since the wavefunction ψ can penetrate into the normal metal,[14] the condition that no current should flow through the surface could also be satisfied by the boundary condition

$$\boldsymbol{n} \cdot \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi = \mathrm{i} b \psi \,, \tag{26.4.7}$$

where b is real. An even more general choice is required for SIS junctions, in which the current can flow from one superconductor into the other through a thin insulating layer.

At the minimum of the free energy the bracketed terms in (26.4.4) must vanish. The *first Ginzburg–Landau equation* is obtained by equating the coefficient of $\delta\psi^*$ to zero:

$$\boxed{\frac{1}{2m^*} \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right)^2 \psi + \alpha \psi + \beta |\psi|^2 \psi = 0 \,.} \tag{26.4.8}$$

Formally, this equation is a Schrödinger equation for the wavefunction ψ of superconducting electrons, in which the term proportional to $|\psi|^2$ is the potential due to the other electrons. In this sense the first Ginzburg–Landau equation is a nonlinear Schrödinger equation.

The *second Ginzburg–Landau equation* is derived from the requirement that the coefficient of $\delta \boldsymbol{A}$ also vanish, using the Maxwell equation for the current. It reads

$$\boxed{\frac{1}{\mu_0} \operatorname{curl} \boldsymbol{B} = \boldsymbol{j} = -\frac{e^*}{2m^*} \psi^* \left(\frac{\hbar}{\mathrm{i}} \boldsymbol{\nabla} + e^* \boldsymbol{A} \right) \psi + \text{c.c.}} \tag{26.4.9}$$

This is just the quantum mechanical current formula, with $\boldsymbol{p}$ replaced by $\boldsymbol{p} + e^* \boldsymbol{A}$. The charge current is obtained by multiplying the quantum mechanical (particle) current by $-e^*$. In conjunction with the relation $\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$ between the vector potential and the magnetic induction, these Ginzburg–Landau equations determine the values of the order parameter ψ and magnetic induction $\boldsymbol{B}$ in the superconducting phase – and by way of it, the current as well.

The Ginzburg–Landau equations were derived phenomenologically. It was pointed out by L. P. Gorkov in 1959 that they can also be derived from the microscopic theory, and it was then established that the starred parameters have to be chosen as $e^* = 2e$ and $m^* = 2m_{\mathrm{e}}$ (as mentioned on page 476), since they are related to the Cooper pairs.

26.4.2 Gauge Symmetry Breaking

Suppose that a solution of the Ginzburg–Landau equations, the wavefunction $\psi(\boldsymbol{r})$ plus the vector potential $\boldsymbol{A}(\boldsymbol{r})$, is known. Since in both equations the operator

[14] This gives rise to the so-called *proximity effect*.

$$\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e^*\boldsymbol{A} \tag{26.4.10}$$

acts on the wavefunction $\psi(\boldsymbol{r})$,

$$\begin{aligned} \psi'(\boldsymbol{r}) &= \psi(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{r})}\,, \\ \boldsymbol{A}'(\boldsymbol{r}) &= \boldsymbol{A}(\boldsymbol{r}) - \frac{\hbar}{e^*}\boldsymbol{\nabla}\phi(\boldsymbol{r}) = \boldsymbol{A}(\boldsymbol{r}) - \frac{\Phi_0}{2\pi}\boldsymbol{\nabla}\phi(\boldsymbol{r}) \end{aligned} \tag{26.4.11}$$

is also a solution – with the same energy, magnetic induction, and current.

The transformation that connects the solutions is the well-known gauge transformation of electrodynamics, and the existence of equivalent solutions is the consequence of deriving the Ginzburg–Landau equations from a gauge-invariant free energy. Among the infinitely many solutions one is chosen by nature – just like for the rotationally symmetric Heisenberg model in which a broken-symmetry state is realized with the magnetization pointing in a well-defined direction. Similarly, gauge symmetry is broken in the superconducting state.

According to Goldstone's theorem (Section 6.3.2), the breaking of a continuous symmetry implies the existence of bosonic elementary excitations with a gapless energy spectrum. However, the theorem is valid only for short-ranged forces. Owing to the lack of screening, the Coulomb interaction remains long-ranged in superconductors, so Goldstone's theorem does not apply to superconductors.

One would think that the phase of the wavefunction, which is a typical microscopic quantum mechanical quantity, cannot be measured, and is therefore of limited importance. This is indeed so for an isolated superconductor. However, when there is a weak contact between two superconductors, which prevents the establishment of thermodynamic equilibrium but allows the transfer of electrons from one superconductor to the other, their phase difference can lead to interesting phenomena. We shall return to this point at the end of the chapter.

26.4.3 Coherence Length and Penetration Depth

Based on the London equations, we have already introduced a characteristic length of the superconducting state: the penetration depth of the magnetic field. We have also mentioned that there is another characteristic length, the coherence length. In order to interpret both of them in the framework of the Ginzburg–Landau theory, we have to examine what happens close to the surface of superconductors.

Outside the superconductor the order parameter vanishes, while deep in its interior it takes the equilibrium value. Close to the surface, its variation occurs over a region of finite width ξ. To determine this parameter, consider the first Ginzburg–Landau equation in zero magnetic field:

$$-\frac{\hbar^2}{2m^*}\boldsymbol{\nabla}^2\psi + \alpha\psi + \beta|\psi|^2\psi = 0\,. \tag{26.4.12}$$

Far from the surface, where the superconductor can be considered homogeneous, the equilibrium value of the order parameter can be obtained from

$$|\psi_0|^2 = -\frac{\alpha}{\beta} \tag{26.4.13}$$

according to (14.5.4). In terms of the dimensionless quantity $f = \psi/|\psi_0|$, (26.4.12) reads

$$-\frac{\hbar^2}{2m^*}\boldsymbol{\nabla}^2 f + \alpha f - \alpha|f|^2 f = 0\,. \tag{26.4.14}$$

It follows from the division of this equation by α that the parameter ξ of dimension length defined through

$$\xi^2 = -\frac{\hbar^2}{2m^*\alpha} \tag{26.4.15}$$

characterizes the spatial variations of f. This implies that when the order parameter varies in space, e.g., it grows from zero to its equilibrium value, all spatial variations occur on the length scale determined by ξ. In order to distinguish it from the temperature-independent Pippard coherence length ξ_0, ξ is called the Ginzburg–Landau correlation length. It can be shown in the microscopic theory that the two are not independent of one another. Although the Ginzburg–Landau theory is valid only in the vicinity of the transition point, $\xi(T)$, the characteristic length of the variations of the superconducting order parameter, and ξ_0, the parameter that appears in the electrodynamics of superconductors, can both be defined in the microscopic theory. It turns out that for pure superconductors

$$\xi(T \to 0) = \xi_0\,. \tag{26.4.16}$$

It follows from our assumption about the temperature dependence of the coefficient α in the Landau expansion that the Ginzburg–Landau coherence length diverges in the critical point:

$$\xi(T) = \sqrt{\frac{\hbar^2}{2m^* a}}\bigl(T_{\mathrm{c}} - T\bigr)^{-1/2}\,. \tag{26.4.17}$$

According to the microscopic theory,

$$\xi(T) = 0.74\,\xi_0(1 - T/T_{\mathrm{c}})^{-1/2} \tag{26.4.18}$$

for clean superconductors, while for very imperfect ones, in the “dirty” limit

$$\xi(T) = 0.855\,(\xi_0 l)^{1/2}(1 - T/T_{\mathrm{c}})^{-1/2}\,. \tag{26.4.19}$$

The spatial variations of the order parameter can be determined explicitly when the surface is an infinite plane. Choosing it as the $x = 0$ plane, all variations are along the x-direction, and the governing equation is

$$\xi^2 \frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = -f + f^3 = -f\left(1 - f^2\right) . \tag{26.4.20}$$

This equation can be integrated exactly when both sides are multiplied by $2\,\mathrm{d}f/\mathrm{d}x$. The result is

$$\xi^2 \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)^2 = \tfrac{1}{2}\left(1 - f^2\right)^2 . \tag{26.4.21}$$

Rearrangement then leads to

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{1}{\sqrt{2}\xi}\left(1 - f^2\right) , \tag{26.4.22}$$

which can be integrated once again. The final solution is

$$f(x) = \tanh \frac{x - x_0}{\sqrt{2}\xi} , \tag{26.4.23}$$

where x_0 is a constant not specified by the equations: the position coordinate of that point inside the normal metal where the order parameter would vanish if the formula valid for the superconductor were extrapolated. Toward the sample interior, the order parameter of the superconducting state indeed increases to the equilibrium value over a characteristic length ξ, as shown in Fig. 26.20.

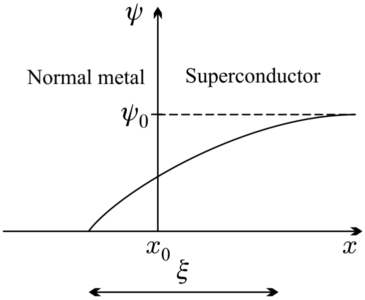


Fig. 26.20. Spatial variation of the superconducting order parameter at a normal metal–superconductor interface

The other characteristic length is the penetration depth, which was introduced and interpreted through the phenomenological London equations. We shall now show that the second London equation, which formulates the local relationship between the current and the vector potential, can be obtained from the second Ginzburg–Landau equation if the length scale ξ of the spatial variations of the wavefunction is small compared to the characteristic scale of the variations of the magnetic field. Writing $\psi(\boldsymbol{r})$ as $|\psi(\boldsymbol{r})|\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{r})}$, the superconducting current is

$$\boldsymbol{j}(\boldsymbol{r}) = -\frac{e^{*2}}{m^*}|\psi(\boldsymbol{r})|^2 \left[\boldsymbol{A}(\boldsymbol{r}) + \frac{\hbar}{e^*}\boldsymbol{\nabla}\phi(\boldsymbol{r})\right]. \tag{26.4.24}$$

This formula is very similar to (26.3.35), the formula for the local relationship between the current and the vector potential obtained from the London equation, as the second term in

$$\boldsymbol{A}(\boldsymbol{r}) + \frac{\hbar}{e^*}\boldsymbol{\nabla}\phi(\boldsymbol{r}), \tag{26.4.25}$$

which contains the gradient of the phase, can be transformed away by a gauge transformation. Instead of the density of the superconducting electrons, the previous formula contains $|\psi(\boldsymbol{r})|^2$, in complete agreement with the quantum mechanical interpretation of the wavefunction. Since ξ is assumed to be small compared to the characteristic scale of the variations of the vector potential, $|\psi(\boldsymbol{r})|$ can be approximated by the equilibrium value $|\psi_0|$. Inserting the current density into the Maxwell equations, we obtain a temperature-dependent London penetration depth defined by

$$\lambda_{\mathrm{L}}^2(T) = \frac{m^*}{\mu_0 e^{*2}|\psi_0|^2} = -\frac{m^*\beta}{\mu_0 e^{*2}\alpha} = \frac{m^*\beta}{\mu_0 e^{*2}a(T_{\mathrm{c}}-T)}. \tag{26.4.26}$$

The penetration depth diverges with the same exponent at the critical temperature as the coherence length.

26.4.4 Flux Quantization

As demonstrated in Sections 26.1.3 and 26.3.4, the magnetic flux through a superconducting ring cannot take any arbitrary value only the integral multiples of an elementary flux quantum. We shall derive this condition from the Ginzburg–Landau equations now.

Consider a non-simply connected superconductor in which the superconducting region surrounds normals regions (holes). When the sample is placed in a magnetic field, the flux lines pass through the normal region. However, the eddy currents on the boundary of the superconducting region cancel the applied field, and the condition $\boldsymbol{B} = 0$ is met inside the superconductor. The flux through the region surrounded by the superconductor is

$$\Phi = \int\limits_F \boldsymbol{B}\cdot\boldsymbol{n}\,\mathrm{d}S, \tag{26.4.27}$$

where the surface F can be chosen at will, with the sole restriction that its contour should be inside the superconducting region, and $\boldsymbol{n}$ is the unit normal of the surface element. By expressing the magnetic induction in terms of the vector potential and using Stokes' theorem, the surface integral can be transformed into the integral around the contour C:

$$\Phi = \int_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l} \,. \tag{26.4.28}$$

$\boldsymbol{A}$ can be expressed from the second Ginzburg–Landau equation as

$$\boldsymbol{A} = -\frac{m^*}{e^{*2}} \frac{1}{|\psi|^2} \boldsymbol{j} - \frac{\hbar}{2\mathrm{i}e^*} \frac{1}{|\psi|^2} \left(\psi^* \boldsymbol{\nabla} \psi - \psi \boldsymbol{\nabla} \psi^* \right) . \tag{26.4.29}$$

Sufficiently far from the surface, where no diamagnetic current flows and the superconducting order is established, the magnitude of the order parameter ψ is given by the equilibrium value but its phase may still change. Assuming that it can be written as $\psi = |\psi_0| \mathrm{e}^{\mathrm{i}\phi}$,

$$\boldsymbol{A} = -\frac{\hbar}{e^*} \boldsymbol{\nabla} \phi \,. \tag{26.4.30}$$

Using this form of $\boldsymbol{A}$ in the integral around the contour C, which is chosen in such a way that the current should vanish along it, we have

$$\Phi = -\frac{\hbar}{e^*} \int_C \boldsymbol{\nabla} \phi \, \mathrm{d}\boldsymbol{l} \,. \tag{26.4.31}$$

The absolute value of the order parameter ψ is a single-valued function of the position but the phase is not. A full turn around the normal region may change the phase by an integral multiple of 2π. Therefore the flux enclosed by the contour C is

$$\Phi = n \, 2\pi \frac{\hbar}{e^*} = n \frac{h}{e^*} \,. \tag{26.4.32}$$

Since $e^* = 2e$, the elementary flux quantum is

$$\Phi_0 = \frac{h}{2e} = 2.067 \times 10^{-15} \, \mathrm{Wb} = 2.067 \times 10^{-7} \, \mathrm{G\,cm}^2 \,. \tag{26.4.33}$$

F. London discussed flux quantization in already 1950, but he made the assumption $e^* = e$, and so his result for the flux quantum was twice as large as the correct value. The correctness of the above formula was later confirmed by the experiments shown in Fig. 26.6.

If the contour C is chosen in such a way that current can flow around it then it is the fluxoid

$$\Phi' = \Phi + \mu_0 \lambda_{\mathrm{L}}^2 \int_C \boldsymbol{j}_{\mathrm{s}} \, \mathrm{d}\boldsymbol{l} \,, \tag{26.4.34}$$

rather than the flux, that is quantized, in units of Φ_0. As mentioned in connection with the London theory, this corresponds to the Bohr–Sommerfeld quantization of the canonical momentum.

26.4.5 Energy of the Normal Metal–Superconductor Interface

To understand when normal regions can be formed inside a superconductor, the energy of the interface between a superconductor and a normal metal has to be studied. This energy depends on the relative magnitude of the coherence length and the penetration depth. Two extreme cases are shown in Fig. 26.21.

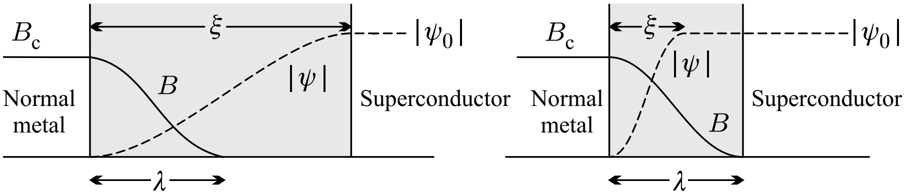


Fig. 26.21. The spatial variation of the magnetic induction and the superconducting order parameter at a normal metal–superconductor interface, in the $\lambda \ll \xi$ and $\lambda \gg \xi$ limits

The ratio $\kappa = \lambda/\xi$ of the penetration depth and coherence length, called the Ginzburg–Landau parameter, is a fundamental parameter of superconductors. When $\kappa \ll 1$, the magnetic field drops off much more rapidly than ψ grows up to its equilibrium value. There is a wide region where both $\boldsymbol{B}$ and ψ are small, i.e., both the superconducting condensate and the magnetic induction, which could reduce the energy, are absent. The surface energy of such a normal–superconductor interface is expected to be positive, so such walls are not formed spontaneously. The situation is the opposite in the $\kappa \gg 1$ case: the surface energy is expected to be negative, thus such walls are formed spontaneously. Below we shall give a better estimate for the surface energy.

When the applied magnetic field is considered as a free variable, the density of the Gibbs potential,

$$g = f - \boldsymbol{H} \cdot \boldsymbol{B}\,, \tag{26.4.35}$$

has to be minimized. Using the Ginzburg–Landau form (26.4.1) for the Helmholtz free energy, the Gibbs potential in given by

$$\begin{aligned} g_{\mathrm{n}}(T,H) &= f_{\mathrm{n}}(T,0) + \frac{1}{2\mu_0}\boldsymbol{B}^2 - \boldsymbol{H} \cdot \boldsymbol{B} \\ &= f_{\mathrm{n}}(T,0) - \tfrac{1}{2}\mu_0 H^2 \end{aligned} \tag{26.4.36}$$

in the normal state where the order parameter vanishes and $\boldsymbol{B} = \mu_0 \boldsymbol{H}$, while it is

$$\begin{aligned} g_{\mathrm{s}}(T,H) = f_n(T,0) &+ \alpha(T)|\psi|^2 + \tfrac{1}{2}\beta(T)|\psi|^4 \\ &+ \frac{1}{2m^*}\left|\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e^*\boldsymbol{A}\right)\psi\right|^2 + \frac{1}{2\mu_0}\boldsymbol{B}^2 - \boldsymbol{H} \cdot \boldsymbol{B} \end{aligned} \tag{26.4.37}$$

in the superconducting state.

When the applied magnetic field is equal to the thermodynamic critical field defined by

$$f_{\mathrm{s}}(T,0) = f_{\mathrm{n}}(T,0) - \tfrac{1}{2}\mu_0 H_{\mathrm{c}}^2\,, \tag{26.4.38}$$

then, according to (26.3.13), the Gibbs potential of the homogeneous superconductor is the same as that of the homogeneous normal phase. Their densities are given by

$$g_{\mathrm{s}}^{\mathrm{hom}}(T,H_{\mathrm{c}}) = g_{\mathrm{n}}^{\mathrm{hom}}(T,H_{\mathrm{c}}) = f_{\mathrm{n}}(T,0) - \tfrac{1}{2}\mu_0 H_{\mathrm{c}}^2\,. \tag{26.4.39}$$

When this common value – which is valid on both sides, at large distances from the interface – is subtracted from the actual density of the Gibbs potential calculated in the presence of an interface, and the difference is integrated in the direction perpendicular to the interface, the surface energy is found to be

$$\begin{aligned}\sigma_{\mathrm{ns}} &= \int\limits_{-\infty}^{\infty} \left[g_{\mathrm{s}}(T,H_{\mathrm{c}}) - g_{\mathrm{n}}^{\mathrm{hom}}(T,H_{\mathrm{c}})\right] \mathrm{d}x \\ &= \int\limits_{-\infty}^{\infty} \left\{\alpha(T)|\psi|^2 + \tfrac{1}{2}\beta(T)|\psi|^4 + \frac{1}{2m^*}\left|\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e^*\boldsymbol{A}\right)\psi\right|^2 \right. \\ &\qquad\qquad \left. + \frac{1}{2\mu_0}(B - \mu_0 H_{\mathrm{c}})^2\right\} \mathrm{d}x\,.\end{aligned} \tag{26.4.40}$$

By multiplying the first Ginzburg–Landau equation by ψ^*, and taking its integral along the x-axis, integration by parts yields

$$\int\limits_{-\infty}^{\infty} \left\{\alpha(T)|\psi|^2 + \beta(T)|\psi|^4 + \frac{1}{2m^*}\left|\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla} + e^*\boldsymbol{A}\right)\psi\right|^2\right\} \mathrm{d}x = 0\,. \tag{26.4.41}$$

Comparison with the previous equation gives the surface energy as

$$\sigma_{\mathrm{ns}} = \int\limits_{-\infty}^{\infty} \left\{-\frac{1}{2}\beta(T)|\psi|^4 + \frac{(B-\mu_0 H_{\mathrm{c}})^2}{2\mu_0}\right\} \mathrm{d}x\,. \tag{26.4.42}$$

The role of ψ and B in determining the surface energy is clear. In the $\xi \gg \lambda$ limit B and ψ are small in the transition region of width ξ, so the free energy per unit surface area is

$$\sigma_{\mathrm{ns}} \approx \tfrac{1}{2}\xi\mu_0 H_{\mathrm{c}}^2\,, \tag{26.4.43}$$

which is indeed positive. In the opposite limit, $\lambda \gg \xi$, the wavefunction almost takes the equilibrium value over a large part of the transition region of width λ, however, the magnetic induction does not drop to zero, and so

$$\sigma_{\mathrm{ns}} \approx -\tfrac{1}{2}\lambda\mu_0 H_{\mathrm{c}}^2\,, \tag{26.4.44}$$

which means that the wall energy is negative.

In general, the wall energy can be determined only numerically. The special case $\kappa = 1/\sqrt{2}$ is an exception, since the Ginzburg–Landau equations then imply

$$|\psi|^2 = \frac{\mu_0 H_{\mathrm{c}} - B}{(\mu_0 \beta)^{1/2}}, \tag{26.4.45}$$

that is, the integrand vanishes identically, and so the surface energy is zero. For $\kappa < 1/\sqrt{2}$ the surface energy is positive, thus the formation of walls is not advantageous energetically. Conversely, the surface energy is negative for $\kappa > 1/\sqrt{2}$, so the energy is lower when the system is made up of alternate normal and superconducting regions. The first case is realized in type I superconductors, and the second in type II superconductors.

26.4.6 Vortices

Type II superconductors in a thermodynamic critical field (or more generally, between the upper and lower critical fields) are thus not homogeneous, as the free energy of the system is lower if normal and superconducting regions alternate. To understand what happens in the superconducting phase in this geometry, we have to find a more accurate solution of the Ginzburg–Landau equations. As pointed out by A. A. Abrikosov[15] in 1957, the magnetic field penetrates into the superconductor in long tubes called *vortices*. This can be studied analytically for $\kappa \gg 1$, when $|\psi(\boldsymbol{r})|$ takes the asymptotic value everywhere except for a small core region. The dependence of the magnetic induction and superconducting order parameter on the distance from the center of the vortex are shown in Fig. 26.22.

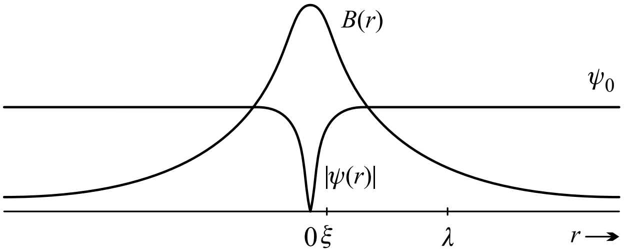


Fig. 26.22. The spatial variations of the magnetic induction and superconducting order parameter around a vortex core

(26.4.30), which was derived from the second Ginzburg–Landau equation, can be rewritten as

$$\boldsymbol{A} + \mu_0 \lambda_{\mathrm{L}}^2 \boldsymbol{j} = -\frac{1}{2\pi} \Phi_0 \boldsymbol{\nabla} \phi . \tag{26.4.46}$$

[15] See the footnotes on pages 4 of Volume 1 and 483.

By taking an arbitrary contour that does not approach the vortex core within the coherence length ξ, and expressing the current in terms of the magnetic induction through the Maxwell equations, we have

$$\int\limits_C \left(\boldsymbol{A} + \lambda_{\mathrm{L}}^2 \operatorname{curl} \boldsymbol{B}\right) \,\mathrm{d}\boldsymbol{l} = n\varPhi_0 \,. \tag{26.4.47}$$

The line integral on the left-hand side can be transformed into a surface integral by means of Stokes' theorem, yielding

$$\int\limits_F \left(\operatorname{curl} \boldsymbol{A} + \lambda_{\mathrm{L}}^2 \operatorname{curl}\operatorname{curl} \boldsymbol{B}\right) \cdot \mathrm{d}\boldsymbol{S} = n\varPhi_0 \,. \tag{26.4.48}$$

When several vortices are present in a type II superconductor, each vortex carries exactly one flux quantum in the energetically most favorable configuration. As $\xi \ll \lambda_{\mathrm{L}}$, we shall study the structure of the vortex in the $\xi \to 0$ limit. The previous equation is satisfied by an arbitrary contour if

$$\boldsymbol{B} + \lambda_{\mathrm{L}}^2 \operatorname{curl}\operatorname{curl} \boldsymbol{B} = \hat{\boldsymbol{z}}\varPhi_0 \delta_2(\boldsymbol{r}) \,, \tag{26.4.49}$$

where $\delta_2(\boldsymbol{r})$ is a δ function in the perpendicular plane, and $\hat{\boldsymbol{z}}$ is the unit vector in the direction of the magnetic field.

Compared to the London equation, an additional source term has appeared. Since $\operatorname{div} \boldsymbol{B} = 0$,

$$\boldsymbol{B} - \lambda_{\mathrm{L}}^2 \boldsymbol{\nabla}^2 \boldsymbol{B} = \varPhi_0 \hat{\boldsymbol{z}} \delta_2(\boldsymbol{r}) \,. \tag{26.4.50}$$

Changing to cylindrical coordinates, and assuming that $\boldsymbol{B}$ has only one nonvanishing component, B_z, which depends only on r, we have

$$B_z - \frac{\lambda_{\mathrm{L}}^2}{r} \frac{\mathrm{d}}{\mathrm{d}r} \left(r \frac{\mathrm{d}B_z}{\mathrm{d}r} \right) = \varPhi_0 \delta_2(\boldsymbol{r}) \,. \tag{26.4.51}$$

Outside the vortex core the solution is

$$B_z(r) = \frac{\varPhi_0}{2\pi\lambda_{\mathrm{L}}^2} K_0 \left(\frac{r}{\lambda_{\mathrm{L}}} \right) , \tag{26.4.52}$$

where K_0 is the zeroth-order modified Bessel function, while inside the vortex

$$B_z(r) = \frac{\varPhi_0}{2\pi\lambda_{\mathrm{L}}^2} K_0 \left(\frac{\xi}{\lambda_{\mathrm{L}}} \right) . \tag{26.4.53}$$

Using the form given in (C.3.59), simple analytic formulas are obtained in two limits:

$$B_z(r) = \begin{cases} \dfrac{\varPhi_0}{2\pi\lambda_{\mathrm{L}}^2} \left(\dfrac{\pi\lambda_{\mathrm{L}}}{2r} \right)^{1/2} \mathrm{e}^{-r/\lambda_{\mathrm{L}}}, & \text{if} \quad r \to \infty \,, \\ \dfrac{\varPhi_0}{2\pi\lambda_{\mathrm{L}}^2} \left(\ln \dfrac{\lambda_{\mathrm{L}}}{r} + 0.12 \right), & \text{if} \quad \xi \ll r \ll \lambda_{\mathrm{L}} \,. \end{cases} \tag{26.4.54}$$

Note that over the largest part of the Shubnikov phase the separation of vortices is smaller than the penetration depth, thus the magnetic induction is finite everywhere in the sample.

The radial variation of the current circulating around the vortex core is given by

$$j(r) = -\frac{\Phi_0}{2\pi\lambda_\mathrm{L}^3\mu_0} K_1\left(\frac{r}{\lambda_\mathrm{L}}\right) \tag{26.4.55}$$

outside the core. Far from the core of an isolated vortex, $j(r)$ decays exponentially, too, however it decreases as $1/r$ in the region $\xi \ll r \ll \lambda_\mathrm{L}$. Determined from the spatial dependence, the vortex energy per unit length is

$$E_\mathrm{vortex} = \frac{\Phi_0^2}{4\pi\mu_0\lambda_\mathrm{L}^2} \ln\left(\frac{\lambda_\mathrm{L}}{\xi}\right). \tag{26.4.56}$$

This result could have been found intuitively. Since the current of superconducting electrons is cylindrically symmetric around the vortex core, the associated velocity can be determined from the quantization condition

$$\oint \boldsymbol{p}\cdot \mathrm{d}\boldsymbol{l} = nh \tag{26.4.57}$$

for the canonical momentum $\boldsymbol{p}$ if the latter is approximated by the kinetic momentum $m^*\boldsymbol{v}_\mathrm{s}$. Taken around a circle of radius r, the integral is

$$nh = m^* \oint \boldsymbol{v}_\mathrm{s}\cdot \mathrm{d}\boldsymbol{l} = m^* v_\mathrm{s} 2\pi r\,, \tag{26.4.58}$$

and hence

$$v_\mathrm{s} = \frac{n\hbar}{m^* r}\,. \tag{26.4.59}$$

It can be assumed that the dominant part of the circulating current flows in a region into which the field can penetrate and where the order parameter has almost reached its equilibrium value – that is, where the distances r from the vortex core is larger than the coherence length ξ but smaller than the penetration depth λ_L. When the $n = 1$ solution is applied to each vortex, the kinetic energy of the superconducting electrons in a layer of unit height around the vortex core is

$$E_\mathrm{kin} = n_\mathrm{s}^* \int\limits_\xi^{\lambda_\mathrm{L}} \frac{m^* v_\mathrm{s}^2}{2} 2\pi r\,\mathrm{d}r = \frac{\pi n_\mathrm{s}^* \hbar^2}{m^*} \int\limits_\xi^{\lambda_\mathrm{L}} \frac{\mathrm{d}r}{r} = \frac{\pi n_\mathrm{s}^* \hbar^2}{m^*} \ln\kappa\,. \tag{26.4.60}$$

Using the formulas of the flux quantum and the London penetration depth, the vortex energy obtained in (26.4.56) is indeed recovered.

It can be shown that when two vortices are located at $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, separated by a distance r_{12}, their interaction energy is

$$E_{12} = \frac{\Phi_0^2}{2\pi\mu_0\lambda^2} K_0\left(\frac{r_{12}}{\lambda_{\mathrm{L}}}\right) , \tag{26.4.61}$$

that is, the interaction decays exponentially at large distances. This result lends itself to simple interpretation in terms of the magnetic moment of the vortex. The angular momentum due to the circulating current is

$$\hbar L = n_{\mathrm{s}}^* \int_{\xi}^{\lambda_{\mathrm{L}}} r m^* v_{\mathrm{s}} 2\pi r \,\mathrm{d}r = n_{\mathrm{s}}^* \pi\hbar\lambda_{\mathrm{L}}^2 \,. \tag{26.4.62}$$

The magnetic moment is obtained by multiplying it by $e^*/2m^*$:

$$\mu = \frac{\pi n_{\mathrm{s}}^* e^* \hbar}{2m^*} \lambda_{\mathrm{L}}^2 = \frac{\Phi_0}{4\mu_0} \,. \tag{26.4.63}$$

When such a magnetic moment is placed into the field $B(r_{12})$ of the other vortex, the same interaction energy is found as above.

Since the interaction between vortices is repulsive, the lowest-energy configuration for a given flux is obtained by separating the vortices as much as possible. This is realized in periodic structures. Two possibilities are shown in Fig. 26.23: a square and a triangular lattice.

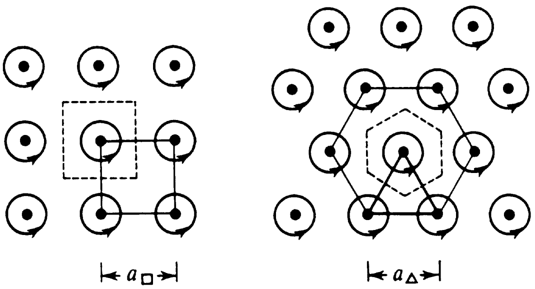


Fig. 26.23. Vortex arrangement in two-dimensional square and triangular lattices

Simple considerations show that the triangular vortex lattice in energetically more favorable. If the flux through the surface area F of the sample is $BF = \Phi = n\Phi_0$, and each vortex carries a flux quantum, then the lattice constant $a_\square$ of the square vortex lattice is given by

$$a_\square^2 B = \Phi_0 \,, \qquad a_\square = \left(\frac{\Phi_0}{B}\right)^{1/2} , \tag{26.4.64}$$

whereas for a triangular lattice

$$a_\triangle \frac{\sqrt{3}}{2} a_\triangle B = \Phi_0\,, \qquad a_\triangle = \left(\frac{4}{3}\right)^{1/4} \left(\frac{\Phi_0}{B}\right)^{1/2} = 1.075 \left(\frac{\Phi_0}{B}\right)^{1/2}, \tag{26.4.65}$$

so for a given flux $a_\triangle > a_\square$. The arrangement of vortices in such a triangular lattice is indeed observed in conventional superconductors. The image in Fig. 26.24, taken by a scanning tunneling microscope, shows the surface of a type II superconductor in the Shubnikov phase.

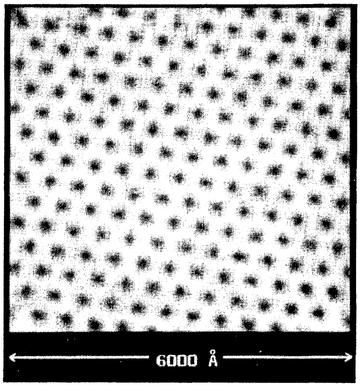


Fig. 26.24. The STM image of the vortex lattice observed in a $NbSe_2$ sample in a magnetic field of 1 T at a temperature of 1.8 K [Reprinted with permission from H. F. Hess et al., *Phys. Rev. Lett.* **62**, 214 (1989). ©1989 by the American Physical Society]

The situation is slightly different in high-T_c superconductors. When the sample is heated in an applied magnetic field H that is somewhat above the lower critical field $H_{c1}(T)$, the vortex lattice may melt before reaching the relatively high critical temperature for this field, and thus a vortex liquid can appear. It has not been confirmed by experiments whether the vortices really make up a lattice before making a transition to the liquid-like state in fields above $H_{c1}(T)$ or are disordered at lower temperatures, too. Theoretical considerations support the assumption that a statically disordered, glass-like *vortex-glass* state is realized. The corresponding phase diagram is shown in Fig. 26.25.

26.4.7 Upper and Lower Critical Fields

In the foregoing the necessity of the appearance of vortices (i.e., alternating normal and superconducting regions) was demonstrated theoretically for type II superconductors placed in the thermodynamic critical field. However, it is known from experiments that in conventional superconductors the inhomogeneous Shubnikov phase with vortices is stable for fields between the lower and upper critical fields, H_{c1} and H_{c2}. Below H_{c1}, type II superconductors are

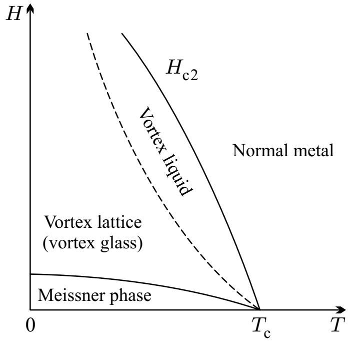


Fig. 26.25. Schematic phase diagram of high-T_c superconductors

also perfect diamagnets, fully displaying the Meissner–Ochsenfeld effect, while they make a transition into the normal state above H_{c2}. Using the previous result, we can now determine the values of the two critical fields.

For fields close to H_{c1}, where the magnetization is small, the separation of vortices becomes so large that their interaction can be neglected. Whether or not a vortex can appear is determined by the relative strength of the two competing terms in the Gibbs potential. According to (26.4.56), the vortex itself has a positive energy. In a sample of length L, the energy of the vortex is LE_{vortex}. This can be compensated by the magnetic field contribution $-\boldsymbol{B}\cdot\boldsymbol{H}$ in the Gibbs potential. Assuming that the magnetic flux through the vortex is exactly one flux quantum, the integral of the magnetic field contribution over the volume of the sample is

$$-\int\limits_V \boldsymbol{B}\cdot\boldsymbol{H}\,\mathrm{d}V = -L\,H\int\limits_F \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S} = -L\,H\varPhi_0\,. \tag{26.4.66}$$

The first vortex may appear at that value H_{c1} of the field where the two contributions are equal:

$$H_{c1}\varPhi_0 = E_{\text{vortex}}\,. \tag{26.4.67}$$

Using the vortex-energy formula, the lower critical field is

$$H_{c1} = \frac{\varPhi_0}{4\pi\mu_0\lambda_{\mathrm{L}}^2}\ln\kappa\,. \tag{26.4.68}$$

In slightly stronger fields the sudden appearance of a large number of vortices is prevented by the repulsive interaction between vortices. Therefore the number of vortices increases step by step, while their separation is still larger than λ, and the magnetic field penetrates into the sample gradually.

Now consider what happens in strong magnetic fields, when the mixed state is transformed into normal metal. Assuming that the order parameter is small in the vicinity of the transition point, the first Ginzburg–Landau equation can be linearized:

$$\frac{1}{2m^*}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+e^*\boldsymbol{A}\right)^2\psi=-\alpha\psi\,. \tag{26.4.69}$$

Formally, this is the same Schrödinger equation as (22.1.13), from which the energy levels (Landau levels) of a noninteracting electron gas placed in a strong magnetic field were determined. Adopting the results obtained there to the present case, the energy eigenvalues are

$$\varepsilon=\frac{\hbar^2k_z^2}{2m^*}+(n+\tfrac{1}{2})\hbar\omega_{\mathrm{c}}^*\,, \tag{26.4.70}$$

where, on account of $e^*=2e$,

$$\hbar\omega_{\mathrm{c}}^*=\hbar\frac{e^*B}{m^*}=\hbar\frac{2eB}{m^*}\,. \tag{26.4.71}$$

In (26.4.69) $-\alpha$ plays the role of the energy eigenvalue. At temperatures below T_{c}, where α is negative, the superconducting order parameter ψ is nonzero if $-\alpha$ is equal to one of the eigenvalues in (26.4.70). The strongest magnetic field for which the condition

$$-\alpha\ge\varepsilon_{\mathrm{min}}=\tfrac{1}{2}\hbar\omega_{\mathrm{c}}^*=\hbar\frac{eB}{m^*} \tag{26.4.72}$$

can be satisfied – and therefore a superconducting state may exist – is then given by

$$H_{\mathrm{c}2}=-\alpha\frac{m^*}{e\hbar\mu_0}\,. \tag{26.4.73}$$

The parameter α of the Ginzburg–Landau theory can be related to the coherence length ξ through (26.4.15). Making use of this connection,

$$H_{\mathrm{c}2}=\frac{\hbar}{2e\mu_0\xi^2}=\frac{\Phi_0}{2\pi\mu_0\xi^2}\,. \tag{26.4.74}$$

Close to $H_{\mathrm{c}2}$, vortices are present in a relatively high number. The previous formula can be interpreted by saying that superconductivity is destroyed at that value of the magnetic field where the separation of vortices becomes comparable to the coherence length.

When a Ginzburg–Landau-like phenomenological description is applied to anisotropic superconductors, it is quite natural to assume that an effective-mass tensor can be used in the term that contains the gradient of the order parameter. Instead of (26.4.69), the linearization of the first Ginzburg-Landau equation then leads to

$$\frac{1}{2}\sum_{\alpha\beta}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+e^*\boldsymbol{A}\right)_\alpha\left(\frac{1}{m^*}\right)_{\alpha\beta}\left(\frac{\hbar}{\mathrm{i}}\boldsymbol{\nabla}+e^*\boldsymbol{A}\right)_\beta\psi=-\alpha\psi\,. \tag{26.4.75}$$

Once again, this is a Schrödinger equation for the one-particle spectrum of an electron gas in a strong magnetic field, but the spectrum in zero magnetic field is now characterized by an effective-mass tensor. The energy of the Landau levels can again be written as (26.4.70) but now the motion in the field direction and the cyclotron frequency are determined by the components of the effective-mass tensor, in combinations that depend on the field direction. Thus the upper critical field is obtained from (26.4.72) here, too. Assuming that there are one longitudinal and two identical transverse masses, $m_\parallel^*$ and $m_\perp^*$, when the magnetic field makes an angle θ with the longitudinal direction, and the cyclotron mass in ω_c^* is taken from (21.2.42), we have

$$\omega_\mathrm{c}^* = 2eB\left(\frac{\sin^2\theta}{m_\parallel^* m_\perp^*} + \frac{\cos^2\theta}{(m_\perp^*)^2}\right)^{1/2}, \tag{26.4.76}$$

where the factor 2 comes from the double charge of the Cooper pairs. When the magnetic field is parallel or perpendicular to the symmetry axis, the upper critical field can be written as

$$H_{\mathrm{c}2\parallel} = \frac{\Phi_0}{2\pi\mu_0\xi_\parallel\xi_\perp}, \qquad H_{\mathrm{c}2\perp} = \frac{\Phi_0}{2\pi\mu_0\xi_\perp^2}, \tag{26.4.77}$$

where

$$\xi_\parallel^2 = \frac{\hbar^2}{2m_\parallel^* a(T_\mathrm{c} - T)}, \qquad \xi_\perp^2 = \frac{\hbar^2}{2m_\perp^* a(T_\mathrm{c} - T)}. \tag{26.4.78}$$

In the isotropic case the upper and lower critical fields can also be expressed in terms of the thermodynamic critical field. It is known from Landau's theory of phase transitions that the free energy decreases by $\alpha^2/2\beta$ in the ordered phase. On the other hand, the thermodynamic critical field is defined precisely by the equality of the free energy in the superconducting and normal phases. This leads to

$$\tfrac{1}{2}\mu_0 H_\mathrm{c}^2 = \frac{\alpha^2}{2\beta}. \tag{26.4.79}$$

Using the previously obtained formulas for the London penetration depth and coherence length,

$$H_\mathrm{c} = \frac{\Phi_0}{2\sqrt{2}\pi\mu_0\lambda_\mathrm{L}\xi}. \tag{26.4.80}$$

Thus

$$H_{\mathrm{c}1} = \frac{1}{\sqrt{2}}\frac{1}{\kappa}H_\mathrm{c}\ln\kappa, \qquad H_{\mathrm{c}2} = \sqrt{2}\,\kappa\,H_\mathrm{c}. \tag{26.4.81}$$

In type II superconductors, where $\kappa > 1/\sqrt{2}$, $H_{\mathrm{c}1} < H_\mathrm{c} < H_{\mathrm{c}2}$. As the applied field is increased, the Meissner phase becomes unstable below the thermodynamic critical field, and the new phase remains stable even at fields that are stronger than the thermodynamic critical field. This is possible because of the nonuniformity of the superconducting order parameter. In type I superconductors, where the upper critical field $H_{\mathrm{c}2}$ is smaller than H_c, the homogeneous superconducting order disappears and reappears at H_c as the field strength is increased and decreased. It is nevertheless possible to attribute a physical meaning to H_{c2}. Since in type I superconductors the phase transition at finite temperatures and finite magnetic fields is first order, the normal phase can be "supercooled", as shown in Fig. 26.7. H_{c2} specifies the lowest field for which the normal phase can exist when the applied magnetic field is reduced.

It is worth noting that a similar hysteresis can be observed in type II superconductors, too, at H_{c1}. The reason for this is that when field lines penetrate into the superconductor in the form of a vortex, a potential barrier has to be surmounted. The force on a vortex appearing inside the sample, close to the surface, can be described as an attraction between the vortex and its "mirror image" of opposite vorticity that is outside the sample.

26.5 Josephson Effect

It has already been mentioned that when superconducting material is placed on one or both sides of a junction, the current–voltage characteristics are not linear. This current comes from the tunneling of normal electrons, which can only be present in the system at energies above the gap. That is why the gap can be determined from the I–V characteristics. In 1962 B. D. Josephson[16] recognized that if superconductors are placed on both sides of the junction, superconducting electrons can also tunnel through the junction. We shall discuss this effect now.

26.5.1 Relation Between the Josephson Current and the Phase of the Superconductor

In order to draw a simple picture of the Josephson effect, we suppose that the width d of the insulating layer of the junction is much smaller than the coherence length: $d \ll \xi$. In line with the Ginzburg–Landau theory we assume that the system can be characterized by a complex function ψ that takes its equilibrium value far from the insulating layer, but with different phases ϕ_{L} and ϕ_{R} on the two sides. If two identical superconductors fill the regions $x < -d/2$ and $x > d/2$,

$$\psi(x) = \begin{cases} |\psi_0| \mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}} & x < -d/2\,, \\ |\psi_0| \mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}} & x > d/2\,. \end{cases} \tag{26.5.1}$$

The spatial variation in the insulating layer can be determined from the first Ginzburg–Landau equation. In the zero-field case the equation (26.4.20) for

[16] Brian David Josephson (1940–) was awarded the Nobel Prize in 1973 "for his theoretical predictions of the properties of a supercurrent through a tunnel barrier, in particular those phenomena which are generally known as the Josephson effects".

$f = \psi/|\psi_0|$ has to be considered. Based on the geometry of the system, we assume that the characteristic scale for the spatial variations of the order parameter is the thickness of the insulating layer. Then, on account of the huge factor $(\xi/d)^2$ that appears on the left-hand side, the equation can be satisfied only if $\mathrm{d}^2 f/\mathrm{d}x^2 = 0$. The requirement that the solution should match smoothly at the boundary of the insulating layer with the value in the superconductors leads to

$$f(x) \approx \left(\frac{1}{2} - \frac{x}{d}\right) \mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}} + \left(\frac{1}{2} + \frac{x}{d}\right) \mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}} \qquad -d/2 \le x \le d/2 \,. \tag{26.5.2}$$

The spatial variation of the order parameter over the Josephson junction is shown in Fig. 26.26.

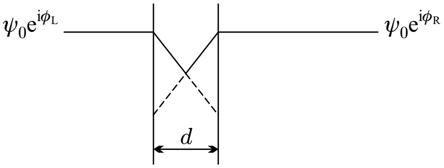

$\psi_0 \mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}}$
$\psi_0 \mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}}$
d

Fig. 26.26. Variation of the real part of the order parameter in a Josephson junction

The current determined from this wavefunction by means of the second Ginzburg–Landau equation is

$$j = \frac{2e\hbar|\psi_0|^2}{m^* d} \sin(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) \,, \tag{26.5.3}$$

that is, the supercurrent through the junction depends on the difference of the macroscopic phases of the superconductors on each side.

For the sake of later generalization, it is worth giving a brief overview of two more accurate methods. In the first approach, developed by FEYNMAN,[17] the superconductors – which are not necessarily identical – on the two sides of the junction are each described in terms of a complex order parameter, the wavefunction of the superconducting electrons,

$$\psi_{\mathrm{L}} = |\psi_{\mathrm{L}}| \mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}} \,, \qquad \psi_{\mathrm{R}} = |\psi_{\mathrm{R}}| \mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}} \,, \tag{26.5.4}$$

just like in the Ginzburg–Landau theory. In an isolated superconductor the time dependence of the wavefunction is determined by the chemical potential, as ψ satisfies the time-dependent Schrödinger equation

$$-\frac{\hbar}{\mathrm{i}} \frac{\partial \psi}{\partial t} = \mu \psi \,. \tag{26.5.5}$$

[17] R. P. FEYNMAN, 1965. See also the footnote on page 532 of Volume 1.

Because of the tunnel coupling in the Josephson junction, the wavefunction of the other side also appears in the equations. The coupling between the two sides can be taken into account by a phenomenological coupling constant T as

$$
\begin{aligned}
-\frac{\hbar}{\mathrm{i}}\frac{\partial\psi_{\mathrm{L}}}{\partial t} &= \mu_{\mathrm{L}}\psi_{\mathrm{L}} + T\psi_{\mathrm{R}}\,,\\
-\frac{\hbar}{\mathrm{i}}\frac{\partial\psi_{\mathrm{R}}}{\partial t} &= \mu_{\mathrm{R}}\psi_{\mathrm{R}} + T\psi_{\mathrm{L}}\,.
\end{aligned}
\tag{26.5.6}
$$

Substitution of the formulas in (26.5.4) leads to

$$
\begin{aligned}
\frac{\partial|\psi_{\mathrm{L}}|}{\partial t}\mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}} + \mathrm{i}|\psi_{\mathrm{L}}|\mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}}\frac{\partial\phi_{\mathrm{L}}}{\partial t} &= -\frac{\mathrm{i}}{\hbar}\mu_{\mathrm{L}}|\psi_{\mathrm{L}}|\mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}} - \frac{\mathrm{i}}{\hbar}T|\psi_{\mathrm{R}}|\mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}},\\
\frac{\partial|\psi_{\mathrm{R}}|}{\partial t}\mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}} + \mathrm{i}|\psi_{\mathrm{R}}|\mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}}\frac{\partial\phi_{\mathrm{R}}}{\partial t} &= -\frac{\mathrm{i}}{\hbar}\mu_{\mathrm{R}}|\psi_{\mathrm{R}}|\mathrm{e}^{\mathrm{i}\phi_{\mathrm{R}}} - \frac{\mathrm{i}}{\hbar}T|\psi_{\mathrm{L}}|\mathrm{e}^{\mathrm{i}\phi_{\mathrm{L}}}.
\end{aligned}
\tag{26.5.7}
$$

After some algebra we obtain

$$
\begin{aligned}
\frac{\partial|\psi_{\mathrm{L}}|}{\partial t} + \mathrm{i}|\psi_{\mathrm{L}}|\frac{\partial\phi_{\mathrm{L}}}{\partial t} &= -\frac{\mathrm{i}}{\hbar}\mu_{\mathrm{L}}|\psi_{\mathrm{L}}| - \frac{\mathrm{i}}{\hbar}T|\psi_{\mathrm{R}}|\mathrm{e}^{\mathrm{i}(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})},\\
\frac{\partial|\psi_{\mathrm{R}}|}{\partial t} + \mathrm{i}|\psi_{\mathrm{R}}|\frac{\partial\phi_{\mathrm{R}}}{\partial t} &= -\frac{\mathrm{i}}{\hbar}\mu_{\mathrm{R}}|\psi_{\mathrm{R}}| - \frac{\mathrm{i}}{\hbar}T|\psi_{\mathrm{L}}|\mathrm{e}^{-\mathrm{i}(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})}.
\end{aligned}
\tag{26.5.8}
$$

By separating the real and imaginary parts,

$$
\begin{aligned}
\frac{\partial|\psi_{\mathrm{L}}|}{\partial t} &= \frac{1}{\hbar}T|\psi_{\mathrm{R}}|\sin(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})\,,\\
\frac{\partial|\psi_{\mathrm{R}}|}{\partial t} &= -\frac{1}{\hbar}T|\psi_{\mathrm{L}}|\sin(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})\,,
\end{aligned}
\tag{26.5.9}
$$

and

$$
\begin{aligned}
\frac{\partial\phi_{\mathrm{L}}}{\partial t} &= -\frac{1}{\hbar}\mu_{\mathrm{L}} - \frac{1}{\hbar}T\frac{|\psi_{\mathrm{R}}|}{|\psi_{\mathrm{L}}|}\cos(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})\,,\\
\frac{\partial\phi_{\mathrm{R}}}{\partial t} &= -\frac{1}{\hbar}\mu_{\mathrm{R}} - \frac{1}{\hbar}T\frac{|\psi_{\mathrm{L}}|}{|\psi_{\mathrm{R}}|}\cos(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})\,.
\end{aligned}
\tag{26.5.10}
$$

The real part is related to the variations of the amplitude – that is, the variations of the number of superconducting electrons. The current density carried by the particles of charge $-e^*$ is

$$
j = -e^*\frac{\partial|\psi_{\mathrm{R}}|^2}{\partial t} = \frac{2e^*}{\hbar}T|\psi_{\mathrm{L}}||\psi_{\mathrm{R}}|\sin(\phi_{\mathrm{R}}-\phi_{\mathrm{L}})\,.
\tag{26.5.11}
$$

Thus, the supercurrent through the Josephson junction depends on the phase difference of the two superconductors, just as in the simple derivation.

Let us turn to the second method now. It was mentioned in connection with the Ginzburg–Landau equations that the usual boundary condition given by

(26.4.6) cannot be applied when current is allowed to flow through the interface. Assuming the most general Cauchy boundary condition, the values of the wavefunctions and their derivatives on the interface satisfy the simultaneous equations

$$\begin{aligned} \psi_{\mathrm{L}} &= \lambda_{11}\psi_{\mathrm{R}} + \lambda_{12}\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{\mathrm{i}e^*}{\hbar}\boldsymbol{A}\right)\psi_{\mathrm{R}} \,, \\ \left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{\mathrm{i}e^*}{\hbar}\boldsymbol{A}\right)\psi_{\mathrm{L}} &= \lambda_{21}\psi_{\mathrm{R}} + \lambda_{22}\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{\mathrm{i}e^*}{\hbar}\boldsymbol{A}\right)\psi_{\mathrm{R}} \,. \end{aligned} \tag{26.5.12}$$

As usual, the vector potential was added to the derivative to have gauge-invariant equations. Since the currents calculated from the left- and right-hand-side wavefunctions must be equal, the coefficients λ have to satisfy the auxiliary condition

$$\lambda_{11}\lambda_{22} - \lambda_{12}\lambda_{21} = 1 \,. \tag{26.5.13}$$

Determined from the second Ginzburg–Landau equation, the current is then

$$\begin{aligned} j &= \frac{\mathrm{i}e^*\hbar}{2m^*}\psi_{\mathrm{R}}^*\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{\mathrm{i}e^*}{\hbar}\boldsymbol{A}\right)\psi_{\mathrm{R}} + \text{c.c.} \\ &= \frac{\mathrm{i}e^*\hbar}{2m^*}\left\{\psi_{\mathrm{R}}^*\left[\frac{1}{\lambda_{12}}\psi_{\mathrm{L}} - \frac{\lambda_{11}}{\lambda_{12}}\psi_{\mathrm{R}}\right] - \psi_{\mathrm{R}}\left[\frac{1}{\lambda_{12}}\psi_{\mathrm{L}}^* - \frac{\lambda_{11}}{\lambda_{12}}\psi_{\mathrm{R}}^*\right]\right\} \\ &= \frac{\mathrm{i}e^*\hbar}{2m^*\lambda_{12}}\left[\psi_{\mathrm{R}}^*\psi_{\mathrm{L}} - \psi_{\mathrm{R}}\psi_{\mathrm{L}}^*\right] \\ &= \frac{e^*\hbar}{m^*\lambda_{12}}|\psi_{\mathrm{L}}|\,|\psi_{\mathrm{R}}|\sin\left(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}\right) , \end{aligned} \tag{26.5.14}$$

in agreement with the previous results.

Next, we have to examine how the phases of the superconductors on each side of the junction change with the applied voltage across or magnetic field at the tunnel junction.

26.5.2 DC Josephson Effect

An important result of Feynman's treatment is the establishment of the system of equations (26.5.10) for the variation of the phases with time. For identical superconductors, when $|\psi_{\mathrm{L}}| = |\psi_{\mathrm{R}}|$,

$$\frac{\partial\phi_{\mathrm{R}}}{\partial t} - \frac{\partial\phi_{\mathrm{L}}}{\partial t} = \frac{1}{\hbar}(\mu_{\mathrm{L}} - \mu_{\mathrm{R}}) \,. \tag{26.5.15}$$

If there is no voltage across the junction, the chemical potential is the same on both sides. Then

$$\frac{\partial}{\partial t}(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) = 0 \,, \tag{26.5.16}$$

so the phase difference is constant, and a constant supercurrent can flow through the junction. This is the *direct-current (DC) Josephson effect.*

By writing the Josephson current through the junction of cross-sectional area A as

$$I_{\mathrm{J}} = I_0 \sin(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) , \tag{26.5.17}$$

it can be shown on the basis of the microscopic theory that the maximum supercurrent that can be driven through the junction without any voltage drop is determined by the superconducting gap and the resistance R_{n} of the junction measured in its normal state. At $T = 0$

$$I_0(T = 0) = \frac{\pi \Delta}{2eR_{\mathrm{n}}} , \tag{26.5.18}$$

while at finite temperatures

$$I_0(T) = \frac{\pi \Delta}{2eR_{\mathrm{n}}} \tanh(\Delta / 2k_{\mathrm{B}}T) . \tag{26.5.19}$$

At the threshold voltage $V_{\mathrm{c}} = 2\Delta/e$, where normal current starts to flow, the value of the single-particle tunneling current is exactly the same as the maximum supercurrent. At higher voltages the current gradually reaches the value $I = V/R_{\mathrm{n}}$ that corresponds to ohmic behavior of the normal state. The current–voltage characteristic is shown in Fig. 26.27.

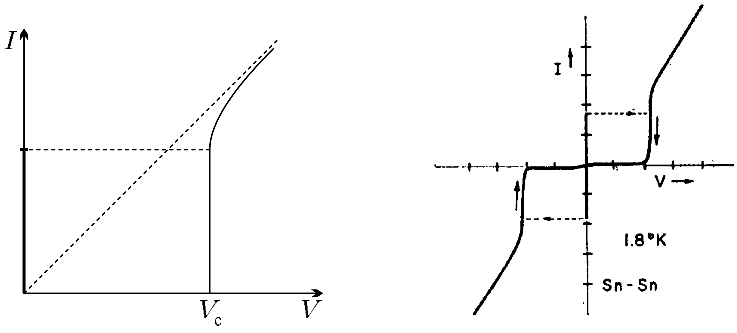


Fig. 26.27. *I–V* curve for a Josephson junction at zero temperature, and the measured supercurrent and single-particle current in an Sn-SnO$_x$-Sn junction at $T = 1.8\,\mathrm{K}$ [Reprinted with permission from R. C. Jaklevic et al., *Phys. Rev.* **140**, A 1628 (1965). ©1965 by the American Physical Society]

Experiments are in good agreement with the theoretical prediction that at low temperatures the maximum supercurrent that can be driven through the Josephson junction without any voltage drop is $\pi/4$ times the current that would flow through the junction in its normal state at the threshold voltage V_{c}. If a stronger current is passed through, a finite voltage V appears across the junction.

26.5.3 AC Josephson Effect

A voltage V applied between two superconductors gives rise to a chemical potential difference $-e^*V$ between the two sides. In the Josephson effect, where "superconducting electrons" of charge $-e^* = -2e$ tunnel, the formula

$$\mu_{\mathrm{R}} - \mu_{\mathrm{L}} = -e^*V = -2eV \tag{26.5.20}$$

has to be applied. The equation for the phase difference of the two superconductors, (26.5.15), now reads

$$\frac{\partial}{\partial t}(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) = 2\frac{eV}{\hbar}\,. \tag{26.5.21}$$

It can also be considered as the consequence of gauge invariance, as the gauge transformation (26.4.11) takes the scalar potential φ into

$$\varphi' = \varphi + \frac{\Phi_0}{2\pi}\frac{\partial\phi}{\partial t}\,, \tag{26.5.22}$$

that is, the gauge-invariant phase difference is given by

$$\Delta\phi - \frac{2\pi}{\Phi_0}\int \varphi(t')\mathrm{d}t'\,. \tag{26.5.23}$$

When the potential is denoted by V instead of φ, the previous formula is recovered.

When a finite voltage is applied, the phase difference changes with time. In particular, for a DC voltage

$$\phi_{\mathrm{R}} - \phi_{\mathrm{L}} = \delta\phi_0 + 2\frac{eV}{\hbar}t\,. \tag{26.5.24}$$

Because of the linear time dependence of the phase difference, the current through the junction oscillates sinusoidally:

$$I_{\mathrm{J}} = I_0 \sin\left(\delta\phi_0 + 2\frac{eV}{\hbar}t\right) = I_0 \sin\left(\delta\phi_0 + 2\pi\frac{V}{\Phi_0}t\right). \tag{26.5.25}$$

Thus a DC voltage gives rise to an alternating current of angular frequency $\omega = 2eV/\hbar = 2\pi V/\Phi_0$ – that is, of frequency

$$\nu = \frac{2e}{h}V\,. \tag{26.5.26}$$

This is the *alternating-current (AC) Josephson effect*, and the quantity

$$K_{\mathrm{J}} = \frac{2e}{h} = 483\,597.9\,\mathrm{GHz/V} \tag{26.5.27}$$

is called the Josephson constant. Since the above relation between the frequency and voltage is independent of the material properties, and is satisfied to a high precision, the AC Josephson effect has been used to define the voltage standard since 1990.

The phenomenon can be interpreted as follows. In order to remain a bound pair after tunneling, the superconducting pair of electrons emits a photon of energy $\hbar\omega = 2eV$, which compensates the difference between the chemical potentials on the two sides. The phenomenon was confirmed experimentally by the detection of the radiation, which is in the microwave region when the applied voltage is a few millivolts.[18]

In the *inverse AC Josephson effect* an AC voltage (a microwave field) of frequency ω and amplitude V_ω and an additional DC voltage V_0 are applied to the Josephson junction. The total voltage that determines the difference of the chemical potentials is

$$V(t) = V_0 + V_\omega \cos\omega t\,. \tag{26.5.28}$$

Substituting this into (26.5.15), the equation governing the variation of the phases with time,

$$\frac{\partial}{\partial t}(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) = \frac{2e}{\hbar}V_0 + \frac{2e}{\hbar}V_\omega \cos\omega t \tag{26.5.29}$$

is obtained. By integrating both sides,

$$\phi_{\mathrm{R}} - \phi_{\mathrm{L}} = \delta\phi_0 + \frac{2e}{\hbar}V_0 t + \frac{2e}{\hbar}\frac{V_\omega}{\omega}\sin\omega t\,, \tag{26.5.30}$$

and the current through the junction is

$$I_{\mathrm{J}} = I_0 \sin\left(\delta\phi_0 + \frac{2e}{\hbar}V_0 t + \frac{2e}{\hbar}\frac{V_\omega}{\omega}\sin\omega t\right). \tag{26.5.31}$$

This formula can be cast in a more transparent form by applying a consequence of (C.1.50):

$$\begin{aligned}\cos(a\sin\omega t) &= \sum_{n=-\infty}^{\infty} J_n(a)\cos n\omega t\,,\\ \sin(a\sin\omega t) &= \sum_{n=-\infty}^{\infty} J_n(a)\sin n\omega t\,,\end{aligned} \tag{26.5.32}$$

where J_n is the Bessel function of order n. Making use of the property

$$J_{-n}(x) = (-1)^n J_n(x) \tag{26.5.33}$$

of the Bessel functions, the current through the Josephson junction is

$$I_{\mathrm{J}} = I_0 \sum_{n=-\infty}^{\infty} (-1)^n J_n\Big(\frac{2eV_\omega}{\hbar\omega}\Big) \sin\left[\delta\phi_0 + \left(\frac{2eV_0}{\hbar} - n\omega\right)t\right]. \tag{26.5.34}$$

[18] An applied voltage of 1 mV corresponds to a frequency of 483.6 GHz.

The total current through a junction placed in a microwave cavity is the sum of the various AC components. In the absence of an applied DC voltage ($V_0 = 0$) the term $n = 0$ gives a DC component, but the factor $J_0(2eV_\omega/\hbar\omega)$ makes the amplitude smaller than the maximum current that can be driven through the junction in the absence of the microwave field. Moreover, DC components appear at all values of V_0 that satisfy

$$V_0 = n\frac{\hbar\omega}{2e}\,. \tag{26.5.35}$$

The corresponding current–voltage characteristic is shown in Fig. 26.28(a).

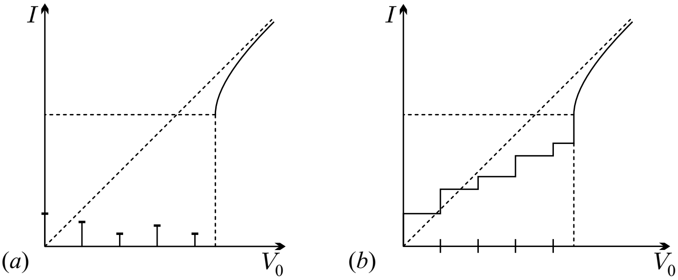


Fig. 26.28. Current–voltage characteristics for a Josephson junction in a microwave field with (a) voltage drive and (b) current drive

The situation is different when the junction is not connected to a voltage source but to a current generator, which is highly common in experiments. In this case the additional current component due to normal electrons must also be taken into account. Assuming ohmic current–voltage characteristics for this component at voltages $V > 2\Delta/e$, and adding the corresponding term to the phenomenological equations for the current, we have

$$I = I_0 \sin(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) + V/R\,, \tag{26.5.36}$$

where the phase difference and the voltage continue to be related by (26.5.21). The total current then satisfies the equation

$$I = I_0 \sin(\phi_{\mathrm{R}} - \phi_{\mathrm{L}}) + \frac{\hbar}{2eR}\frac{\partial}{\partial t}(\phi_{\mathrm{R}} - \phi_{\mathrm{L}})\,. \tag{26.5.37}$$

If this total current, rather than the voltage, is fixed externally, very different I–V curves are obtained. When the phase difference and, through it, the voltage are determined from the current, steps are found instead of discrete peaks, as shown in Fig. 26.28(b). They are called *Shapiro steps.*[19] The height of the

[19] S. Shapiro, 1963.

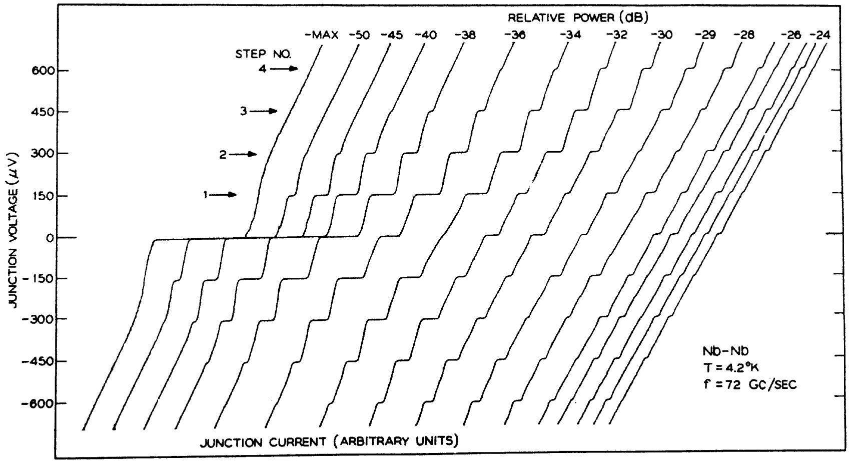


Fig. 26.29. The current–voltage characteristics for a current-driven Josephson junction placed in microwave fields of different power, according to the measurements of C. C. Grimes and S. Shapiro [Reprinted with permission from *Phys. Rev.* **169**, 397 (1968). ©1968 by the American Physical Society]

nth step is determined by the Bessel function J_n. As indicated in Fig. 26.29, measurements are in fair agreement with the theoretical description.

The situation is more complicated when coupling between the two superconductors cannot be approximated by a linear term as in (26.5.6). However, the discussion of this point would lead us too far afield.

26.5.4 Josephson Junctions in a Magnetic Field

It is known from the Ginzburg–Landau theory that the phase of the superconductor is changed by the magnetic field. It is expected to change the current through the junction, too. Consider a setup in which the sample is uniform and infinite in the y- and z-directions, and the narrow insulating oxide layer is centered at the plane $x = 0$. The magnetic field is applied in the z-direction (i.e., parallel to the surface of the oxide layer), and its strength depends on the variable x. It can be derived from an x-dependent vector potential with a nonvanishing y component, $\boldsymbol{A} = A_y(x)\hat{\boldsymbol{y}}$, which satisfies

$$B_z = \frac{\mathrm{d}A_y(x)}{\mathrm{d}x} . \tag{26.5.38}$$

$B_z(x)$ and $A_y(x)$ vary considerably only close to the oxide layer, in a region whose width is on the order of the penetration depth, as illustrated in Fig. 26.30.

In addition to the x-directed current through the junction, a y-directed surface current appears in the above-mentioned region. Its spatial variation is shown in Fig. 26.30.

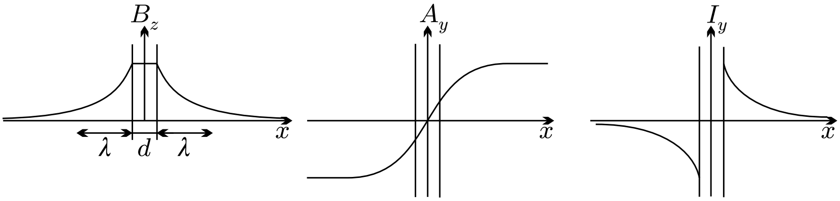


Fig. 26.30. The spatial variation of the magnetic induction, vector potential, and surface current in a Josephson junction, perpendicular to the oxide layer

Because of the presence of the vector potential, the phase of the superconductor is not homogeneous: it depends on the y coordinate. The x-directed current density at position y through the junction is determined by the phase difference $\phi(L_y) - \phi(R_y)$, where L_y and R_y are two points inside the left- and right-hand superconductors, respectively, whose distance from the interface is larger than the penetration depth. To determine this phase difference, consider the second Ginzburg–Landau equation (26.4.30). Expressed in terms of the phase of the wavefunction, the current density is

$$\boldsymbol{j} = -\frac{e^* \hbar}{m^*}|\psi|^2 \left(\boldsymbol{\nabla}\phi + \frac{e^*}{\hbar}\boldsymbol{A}\right). \tag{26.5.39}$$

Far from the insulator–superconductor interface, where no current flows,

$$\boldsymbol{\nabla}\phi = -\frac{e^*}{\hbar}\boldsymbol{A}. \tag{26.5.40}$$

The integral of this formula gives the phase difference between two points that can be connected by a path that is entirely in the superconductor.

Now consider two points, L_1 and L_2, deep inside the left-hand superconductor, as in Fig. 26.31.

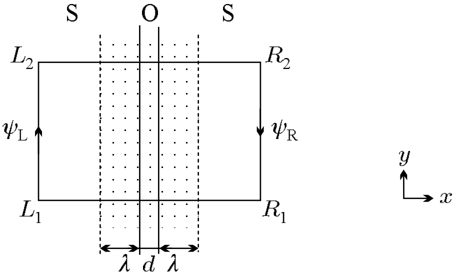


Fig. 26.31. Schematic cross section of a Josephson junction in a magnetic field that is applied in z-direction

Although the magnetic field cannot penetrate to this depth, the difference of the phases at the two points is finite because of the presence of the vector potential:

$$\phi(L_2) - \phi(L_1) = -\frac{e^*}{\hbar}\int\limits_{L_1}^{L_2} \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{l}\,. \tag{26.5.41}$$

Analogously, for two points R_1 and R_2 deep inside the right-hand side but with the same y and z coordinates as L_1 and L_2, the phase difference is

$$\phi(R_2) - \phi(R_1) = -\frac{e^*}{\hbar}\int\limits_{R_1}^{R_2} \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{l}\,. \tag{26.5.42}$$

A suitable rearrangement leads to the following expression for the variation of the phase difference between points located on opposite sides of the junction:

$$\begin{aligned}\Delta\phi(L_2 - R_2) - \Delta\phi(L_1 - R_1) &= [\phi(L_2) - \phi(R_2)] - [\phi(L_1) - \phi(R_1)]\\ &= [\phi(L_2) - \phi(L_1)] - [\phi(R_2) - \phi(R_1)]\\ &= -\frac{e^*}{\hbar}\left(\int\limits_{L_1}^{L_2} \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{l} - \int\limits_{R_1}^{R_2} \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{l}\right). \end{aligned}\tag{26.5.43}$$

Since only the y component of the vector potential is nonzero, the line integrals along the x-directed paths between L_1 and R_1, and L_2 and R_2 vanish. Owing to their vanishing contributions, these segments can be freely added to the previous integration path, and so

$$\Delta\phi(L_2 - R_2) - \Delta\phi(L_1 - R_1) = \frac{e^*}{\hbar}\oint \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{l}\,, \tag{26.5.44}$$

where the closed path is traversed counterclockwise. This can be rewritten as a surface integral for curl $\boldsymbol{A}$, which gives the flux Φ through the surface bounded by the contour. In terms of the flux quantum Φ_0 we have

$$\Delta\phi(L_2 - R_2) - \Delta\phi(L_1 - R_1) = 2\pi\frac{\Phi}{\Phi_0}\,. \tag{26.5.45}$$

Denoting the phase difference at $y = 0$ by $\Delta\phi(0)$, the phase difference at an arbitrary y is given by

$$\Delta\phi(y) = \Delta\phi(0) + 2\pi\frac{\Phi(y)}{\Phi_0}\,, \tag{26.5.46}$$

where

$$\Phi(y) = B\cdot(d + \lambda_{\mathrm{L}} + \lambda_{\mathrm{R}})y \tag{26.5.47}$$

approximately. In the previous formula d is the thickness of the oxide layer, while λ_{L} and λ_{R} are the penetration depths in the two superconductors. The Josephson current at position y is therefore

$$I_{\rm J}(y) = I_0 \sin\left(\Delta\phi(0) + 2\pi \frac{\Phi(y)}{\Phi_0}\right) . \tag{26.5.48}$$

Note that this formula allows the current to be interpreted as the sine of the gauge-invariant phase difference

$$\Delta\phi + \frac{2\pi}{\Phi_0} \int \boldsymbol{A} \cdot {\rm d}\boldsymbol{s} \tag{26.5.49}$$

in the presence of a vector potential.

The formulas also shows that the current through the junction oscillates with the height y. This is illustrated in Fig. 26.32.

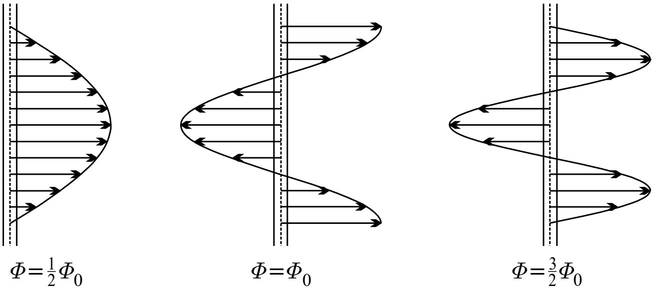


Fig. 26.32. The spatial variation of the current through the Josephson junction, perpendicular to the magnetic field, for three different values of the field strength

Because of this spatial oscillation, the total current through the junction varies with the magnetic field much like a diffraction pattern. The current is the highest in the zero-field case, since it flows in the same direction over the entire cross section of the junction then. When the field strength is such that the total magnetic flux through the junction is an integral multiple of the flux quantum, the total current is zero because the oscillatory contributions cancel out. In general, the current through the junction for a sample of width L_y is

$$\begin{aligned} I_{\rm J} &= \int_0^{L_y} I_{\rm J}(y)\,{\rm d}y = I_0 \int_0^{L_y} \sin\left(\delta\phi(0) - 2\pi\frac{\Phi(y)}{\Phi_0}\right) {\rm d}y \\ &= I_0 L_y \sin(\delta\phi(0)) \frac{\sin(\pi\Phi/\Phi_0)}{\pi\Phi/\Phi_0} , \end{aligned} \tag{26.5.50}$$

and the total flux through the junction is $\Phi = B(d + \lambda_{\rm L} + \lambda_{\rm R})L_y$. Since the direction of the current is rarely measured in experiments, the dependence of the total current on the magnetic field is similar to the Fraunhofer diffraction pattern. However, when two point-like junctions are located at two opposite points of a ring, the dependence on the magnetic field is described by the

Airy function. These two functions, along with measurement data, are shown in Fig. 26.33

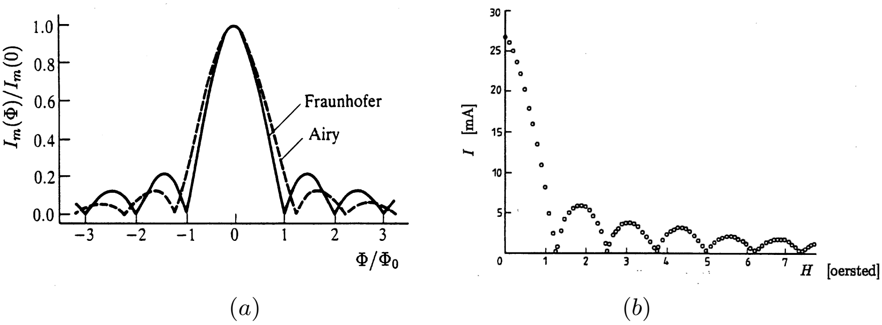


Fig. 26.33. The total current through a Josephson junction as a function of the magnetic field: (*a*) theoretical predictions; (*b*) experimental results by D. E. Langenberg et al. [*Proc. IEEE* **54**, 560 (1966)]

A more interesting interference pattern is obtained when the current is split between two arms, which are connected in parallel and enclose a finite magnetic flux.

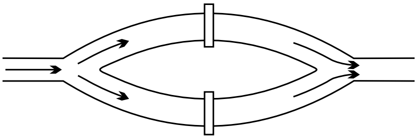


Fig. 26.34. Two Josephson junctions connected in parallel

Since the superconducting arms are broken by the insulating layers of the junction, the quantization condition does not apply to the enclosed flux: it can take any arbitrary value. Assuming that the arms are sufficiently narrow for that each junction can be characterized by a phase difference, the resultant of the currents in the two arms is

$$I_{\mathrm{J}} = I_0 \left[\sin \Delta\phi(L_1 - R_1) + \sin \Delta\phi(L_2 - R_2)\right]. \tag{26.5.51}$$

In perfect analogy with the previous results, when the two arms enclose a flux Φ, the phase differences are related by

$$\Delta\phi(L_2 - R_2) - \Delta\phi(L_1 - R_1) = 2\pi \frac{\Phi}{\Phi_0}. \tag{26.5.52}$$

Choosing the phase differences in the two arms as $\Delta\phi_0 + \pi\Phi/\Phi_0$ and $\Delta\phi_0 - \pi\Phi/\Phi_0$, the resultant current is

$$\begin{aligned} I_{\mathrm{J}} &= I_0 \left[\sin\left(\Delta\phi_0 + \pi \frac{\Phi}{\Phi_0} \right) + \sin\left(\Delta\phi_0 - \pi \frac{\Phi}{\Phi_0} \right) \right] \\ &= 2I_0 \sin\Delta\phi_0 \cdot \cos\left(\pi \frac{\Phi}{\Phi_0} \right) . \end{aligned} \qquad (26.5.53)$$

Once again, by measuring the intensity of the current but not its direction, a diffraction-pattern-like dependence is found (Fig. 26.35). The operation of SQUIDs,[20] designed to measure tiny magnetic fields,[21] is based on this principle.

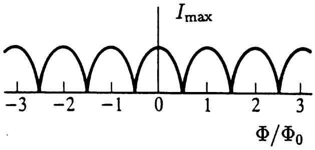


Fig. 26.35. The dependence of the maximum supercurrent on the magnetic field for two point-like Josephson junctions connected in parallel

In reality, the finite size of the Josephson junctions connected in parallel must also be taken into account. If the flux through the cross section of the junction is Φ_f then, according to (26.5.50), the previous formula for the current has to be multiplied by the appropriate factor

$$\sin\left(\pi\Phi_f/\Phi_0\right) / \left(\pi\Phi_f/\Phi_0\right) , \qquad (26.5.54)$$

at each junction. This leads to a much slower variation than the total enclosed flux Φ, and so the dependence shown in Fig. 26.36 is observed in experiments.

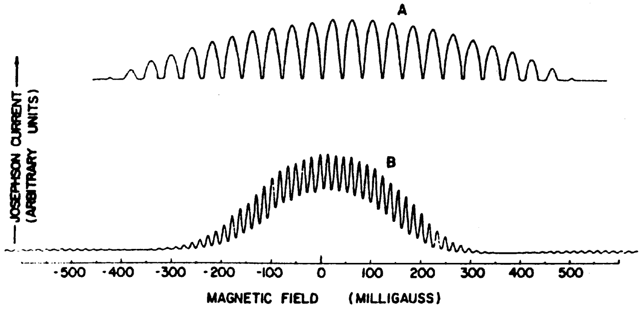


Fig. 26.36. Macroscopic interference and diffraction effects in the maximum Josephson current through two Josephson junctions coupled in parallel [Reprinted with permission from R. C. Jaklevic et al., *Phys. Rev.* **140**, A 1628 (1965). ©1965 by the American Physical Society]

[20] The acronym SQUID stands for Superconducting QUantum Interference Device.

[21] In an appropriate design, SQUIDs can also be used for the highly sensitive measurement of susceptibility and voltage.

Further Reading

1. W. Buckel and R. Kleiner, *Superconductivity: Fundamentals and Applications*, Second, Revised and Enlarged Edition, John Wiley & Sons, Ltd., New York (2004).
2. P. G. de Gennes *Superconductivity of Metals and Alloys*, W. A. Benjamin, New York (1966).
3. *Handbook of Superconductivity*, Edited by C. P. Poole, Jr., Academic Press, San Diego (2000).
4. J. B. Ketterson and S. N. Song, *Superconductivity*, Cambridge University Press, Cambridge (1999).
5. L.-P. Lévy, *Magnetism and Superconductivity*, Texts and Monographs in Physics, Springer-Verlag, Berlin (2000).
6. C. P. Poole, Jr., H. A. Farach, and R. J. Creswick, *Superconductivity*, Academic Press, San Diego (1995).
7. *Superconductivity*, Edited by R. D. Parks, Marcel Dekker, Inc., New York (1969).
8. D. R. Tilley and J. Tilley, *Superfluidity and Superconductivity*, Adam Hilger Ltd., Bristol (1986).
9. M. Tinkham, *Introduction to Superconductivity*, Second Edition, McGraw-Hill, Inc., New York (1996).

27

Transport of Carriers in Semiconductor Devices

The energy spectrum of electrons in semiconductors was discussed in Chapter 20. The modifications due to the uniform doping of pure semiconductors by donor or acceptor impurities were also analyzed in detail. However, the semiconductor devices that have become the fundamental building blocks of modern high-tech equipment consist of different regions that may differ in their doping concentration, doping type, or base material, and are separated by relatively sharp boundaries (interfaces). In other devices semiconductor regions are in contact with metals or insulators. In this chapter we shall first examine what happens to electronic states close to metal–semiconductor interfaces and in p–n junctions.

As we shall see, the charge distribution becomes nonuniform on the semiconductor side. Thus diffusion may occur, leading to the recombination and generation of charge carriers. After a brief description of these physical processes we shall investigate the consequences of applying a voltage across a metal–semiconductor junction or a p–n junction. We shall demonstrate that the current of carriers is a nonlinear function of the applied voltage. The intensity of the current may also depend strongly on its direction, and can be easily controlled through the voltage applied to one of the regions. This property is used for rectification and amplification.

Since the invention of the point-contact transistor[1] and the bipolar junction transistor[2] (whose operation is based on the physical processes taking place in p–n junctions), semiconductor technology – in particular, planar technology – has undergone a breathtaking evolution. This has led to the development and mass production of newer and newer generations of silicon-based devices. Below we shall deal only with a few simple cases, and examine those physical processes that are necessary for understanding the principles of operation.

In the description of physical processes we shall assume the applicability of the semiclassical approximation to the motion of the carriers. Only a very

[1] J. Bardeen and W. H. Brattain, 1947.

[2] W. B. Shockley, 1948. See the footnote on page 4 of Volume 1.

brief overview will be given of the quantum effects that are important in very tiny systems called nanostructures. Finally, we shall look into a recent development that aims to make use of spin (rather than charge) transport in a new type of device.

27.1 Interfaces and Junctions

The properties of inhomogeneously doped semiconductor devices can be controlled by the carriers introduced via doping. In order that the properties should indeed be determined by the introduced carriers, and be controllable by the concentration of donors and acceptors, the crystal structures on the two sides of the junction should differ as little as possible. If these structures were built by putting together two different, traditionally fabricated semiconductors, the number of surface states could be comparable to the number of carriers introduced by doping, and so the properties of the junction could be determined by the former. That is why junctions are usually produced by the nonuniform doping of single crystals. Several different techniques are used. The simplest is to change the composition of the melt from which the semiconductor crystal is grown. Another possibility is to heat one side of a ready-made crystal, and to alloy the dopants into it. A third option is to use diffusion to add dopant atoms, and create an inhomogeneous distribution in which one side of the sample is an n-type, while the other is a p-type semiconductor. More accurately controlled and more abruptly changing doping concentrations can be achieved by ion implantation and epitaxial growth.

Below we shall examine the nonuniform redistribution of charges, which depends on the amount of doping, first at the metal–semiconductor interface, and then at the interface between n- and p-type semiconductors. We shall not discuss the states around semiconductor–insulator interfaces: even though the semiconductor is in contact with the insulating oxide layer in many applications, the layer separating it from the metal is usually sufficiently narrow for that it can be treated as a simple potential barrier.

27.1.1 Metal–Semiconductor Interface

We shall first consider the interface between a metal and an n-type semiconductor. The interface with a p-type semiconductor can be treated in exactly the same way. Therefore only the results for the latter will be listed after the discussion of n-type semiconductors.

Before the contact is made, the work function – which is the distance of the chemical potential from the vacuum level (the minimum amount of energy needed to remove an electron from the metal) – may be different in the metal and the semiconductor: $\Phi_{\mathrm{m}} \gtrless \Phi_{\mathrm{s}}$. First we discuss the case $\Phi_{\mathrm{m}} > \Phi_{\mathrm{s}}$. Figure 27.1($a$) shows the occupied electronic states of the metal and semiconductor in the moment they are brought into contact, and the bottom of the conduction band (which is empty in the ground state), aligned to the vacuum level ε_{vac}. The chemical potential in semiconductors is known to depend strongly on temperature as well as the number of donors and acceptors. The figure shows a low-temperature situation in which the chemical potential of the semiconductor is between the donor level and the bottom of the conduction band.

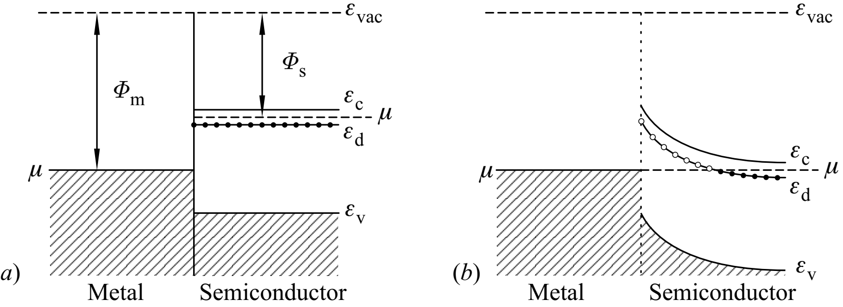


Fig. 27.1. The formation of a depletion layer at the interface of a metal and an n-type semiconductor. The energy-level diagram in the metal and semiconductor (a) at the moment of contact, (b) in thermal equilibrium, after the bending of the energy levels

As illustrated in the figure, the chemical potentials are different in the metal and the semiconductor. This situation cannot be maintained, since in thermal equilibrium the chemical potential must be the same on both sides of the interface. The equalization of the chemical potential occurs via the transfer of charges from one side to the other. When the chemical potential is lower on the metallic side (which corresponds to $\Phi_{\text{m}} > \Phi_{\text{s}}$), electrons move from the semiconductors to the metal. Since charges cannot accumulate locally in metals, the transferred electrons are distributed uniformly over the metallic side. On the other hand, the charge density may become nonuniform on the semiconductor side. This gives rise to a nonuniform electric field, which can be specified by a potential $V(\boldsymbol{r})$. Since the scale of the spatial variations of the potential is much larger than the atomic distances, the semiclassical approximation can be applied, and the position-dependent potential leads to a position-dependent shift $\Delta\varepsilon = -eV(\boldsymbol{r})$ of the energy levels. This distortion of the energy spectrum – bending of the levels – is shown in Fig. 27.1(b).

Far from the interface on the semiconductor side, where the charge distribution is already uniform, the relative location of the chemical potential with respect to the bottom of the conduction band and the donor level must be the same as in homogeneous semiconductors, which was determined in Chapter 20. On the other hand, in an absolute sense it has to be equal to the chemical potential in the metal, therefore the energy spectrum of the semiconductor is bent downward by $\Delta\varepsilon = \Phi_{\text{m}} - \Phi_{\text{s}}$. The *contact potential difference*

V_0, defined through

$$eV_0 = \Phi_{\mathrm{m}} - \Phi_{\mathrm{s}}\,, \tag{27.1.1}$$

comes from a layer of finite width around the interface, from where electrons have diffused into the metal.[3]

We now have to determine in a self-consistent manner the new, inhomogeneous charge distribution, the arising position-dependent potential $V(\boldsymbol{r})$, and the position-dependent shift of the energy levels.

At point $\boldsymbol{r}$, where the potential $V(\boldsymbol{r})$ leads to an energy shift $-eV(\boldsymbol{r})$, the density of electrons in the conduction band is not given by (20.3.17) but by

$$n(\boldsymbol{r}) = N_{\mathrm{c}}(T)\exp\left\{-\frac{\varepsilon_{\mathrm{c}} - eV(\boldsymbol{r}) - \mu}{k_{\mathrm{B}}T}\right\}, \tag{27.1.2}$$

whereas the density of ionized donor atoms is given by

$$n_{\mathrm{d}}^{+}(\boldsymbol{r}) = \frac{n_{\mathrm{d}}}{1 + 2\mathrm{e}^{(\mu - \varepsilon_{\mathrm{d}} + eV(\boldsymbol{r}))/k_{\mathrm{B}}T}} \tag{27.1.3}$$

instead of (20.5.13). The density of holes in the valence band can be expressed analogously. Because of the opposite charge of holes, the energy shift is $eV(\boldsymbol{r})$, so, instead of (20.3.20),

$$p(\boldsymbol{r}) = P_{\mathrm{v}}(T)\exp\left\{-\frac{\mu - \varepsilon_{\mathrm{v}} + eV(\boldsymbol{r})}{k_{\mathrm{B}}T}\right\} \tag{27.1.4}$$

has to be used for it.

Because of the charge redistribution on the semiconductor side, local charge neutrality is violated close to the interface, and a charge density

$$\rho(\boldsymbol{r}) = -e\left[n(\boldsymbol{r}) - n_{\mathrm{d}}^{+}(\boldsymbol{r}) - p(\boldsymbol{r})\right] \tag{27.1.5}$$

appears. In the medium of relative permittivity ϵ_{r} this charge produces an electrostatic potential $V(\boldsymbol{r})$ that satisfies the Poission equation

$$\boldsymbol{\nabla}^2 V(\boldsymbol{r}) = -\frac{1}{\epsilon}\rho(\boldsymbol{r}) = -\frac{1}{\epsilon_{\mathrm{r}}\epsilon_0}\rho(\boldsymbol{r})\,. \tag{27.1.6}$$

Since the charge distribution itself also depends on the potential, this equation has to be solved self-consistently, imposing the boundary condition that on the semiconductor side, far from the metal–semiconductor interface, charge neutrality must be satisfied locally.

To simplify calculations, consider the case in which two semi-infinite media are separated by a plane interface. Choosing the x-axis along the perpendicular direction, all variations occur in the x-direction, therefore we shall use this variable instead of $\boldsymbol{r}$.

[3] The diffusion of donor atoms can be ignored.

Suppose, furthermore, that the temperature is in the saturation range (see page 230) – that is, all donor atoms are ionized ($n_\mathrm{d}^+ \approx n_\mathrm{d}$) but hardly any holes are generated in the valence band ($p \approx 0$). As a function of the distance from the contact, the charge density is given by

$$\rho(x) = -e\left[n(x) - n_\mathrm{d}\right], \tag{27.1.7}$$

and

$$\frac{\mathrm{d}^2 V(x)}{\mathrm{d}x^2} = -\frac{1}{\epsilon}\rho(x). \tag{27.1.8}$$

In the absence of a potential, the neutrality condition implies $n \approx n_\mathrm{d}$ far from the interface in this temperature range. Referred to that, the density of electrons in the conduction band is

$$n(x) = n(\infty)\exp\left(\frac{e[V(x) - V(\infty)]}{k_\mathrm{B}T}\right) = n_\mathrm{d}\exp\left(\frac{e[V(x) - V_0]}{k_\mathrm{B}T}\right), \tag{27.1.9}$$

where $V_0 = V(\infty)$. Substituting this into (27.1.7), the Poisson equation (27.1.8) gives a closed equation for the potential:

$$\frac{\mathrm{d}^2 V(x)}{\mathrm{d}x^2} = \frac{e}{\epsilon}n_\mathrm{d}\left[\exp\left(\frac{e[V(x) - V_0]}{k_\mathrm{B}T}\right) - 1\right]. \tag{27.1.10}$$

The complete solution of this equation specifies the profile of the potential. In general, this equation cannot be solved analytically but an important property can be read off immediately. Rewriting the equation in terms of the dimensionless quantity $\widetilde{V}(x) = e[V(x) - V_0]/k_\mathrm{B}T$,

$$\frac{\mathrm{d}^2 \widetilde{V}(x)}{\mathrm{d}x^2} = \frac{e}{k_\mathrm{B}T}\frac{e}{\epsilon}n_\mathrm{d}\left[\mathrm{e}^{\widetilde{V}(x)} - 1\right]. \tag{27.1.11}$$

By introducing the quantity

$$\lambda_\mathrm{D} = \sqrt{\frac{\epsilon\, k_\mathrm{B}T}{e^2 n_\mathrm{d}}} \tag{27.1.12}$$

of dimension length, a dimensionless equation can be obtained for x/λ_D. So λ_D defines the characteristic length scale for the spatial variations of the potential – and, consequently, for the charge redistribution as well. Since λ_D is the same as the screening length in the *Debye–Hückel theory*[4] of electrolytes, it is called the *Debye length* (or *Debye screening length*). Electrons described by classical statistics in semiconductors screen the effect of the metal over such a distance.

In the temperature range where $e(V(x) - V_0)/k_\mathrm{B}T \ll 1$, the spatial variations of the potential and charge distribution can be determined explicitly. From (27.1.10) we have

[4] P. Debye and E. Hückel, 1923. This will be discussed in detail in Chapter 29.

$$\frac{\mathrm{d}^2V(x)}{\mathrm{d}x^2} \approx \frac{e^2n_\mathrm{d}}{\epsilon k_\mathrm{B}T}[V(x)-V_0] = \frac{1}{\lambda_\mathrm{D}^2}[V(x)-V_0] \,. \tag{27.1.13}$$

Its physically meaningful solution, which decreases (rather than increases) exponentially far from the interface, is

$$V(x) - V_0 = A\mathrm{e}^{-x/\lambda_\mathrm{D}} \,. \tag{27.1.14}$$

To specify the parameter A, attention must be paid to the auxiliary condition that there is no screening at $x = 0$: the relative position of energy levels is the same as in the moment of contact. Therefore

$$V(x) = V_0\left\{1 - \mathrm{e}^{-x/\lambda_\mathrm{D}}\right\}, \tag{27.1.15}$$

and the shift of energy levels is

$$\Delta\varepsilon(x) = eV_0\left\{\mathrm{e}^{-x/\lambda_\mathrm{D}} - 1\right\}, \qquad x > 0\,, \tag{27.1.16}$$

where eV_0 is defined by (27.1.1). This position-dependent shift of the energy levels is shown in Fig. 27.1(*b*). When the potential is known, the charge distribution can be determined from the Poisson equation as

$$\rho(x) = \epsilon\frac{1}{\lambda_\mathrm{D}^2}V_0\mathrm{e}^{-x/\lambda_\mathrm{D}} = \frac{e^2n_\mathrm{d}}{k_\mathrm{B}T}V_0\mathrm{e}^{-x/\lambda_\mathrm{D}} \,. \tag{27.1.17}$$

When $eV_0 = \Phi_\mathrm{m} - \Phi_\mathrm{s} > 0$, electrons diffuse from a region of depth λ_D on the semiconductor side into the metal. This region is called *depletion layer* or *depletion region*. Since the positively charged donors remain at their position, this region is in fact positively charged, which is why it is also called *space-charge region*. λ_D is also the characteristic scale for the variations of the potential, where the contact-potential drop V_0 occurs. According to our previous calculations, both the potential and charge density vary exponentially with the distance. The spatial variation of the net charge distribution and the potential are shown in Fig. 27.2.

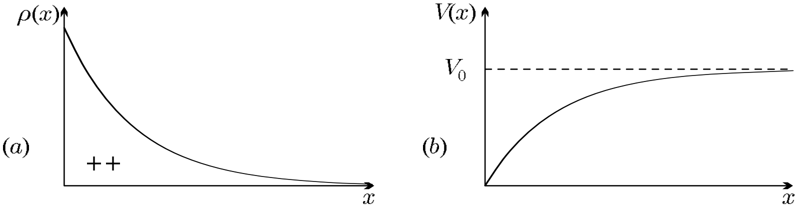


Fig. 27.2. (*a*) The net charge distribution close to the metal–semiconductor interface. (*b*) The spatial variation of the potential

By assuming a donor density of $n_\mathrm{d} = 10^{16}/\mathrm{cm}^3$ at $T = 300\,\mathrm{K}$ in germanium, $\lambda_\mathrm{D} = 4 \times 10^{-6}\,\mathrm{cm}$ is found for the characteristic length. The influence of the metal is screened over this distance. As we shall see, screening occurs much faster, over atomic distances in metals. That is why the metallic side can be considered homogeneous even after the contact has been made.

Once thermal equilibrium has been established and the depletion layer has been formed, the energy levels are bent as shown in Fig. 27.1(*b*). Electrons inside the semiconductor cannot move to the metallic side, since they would have to overcome a potential barrier known as the *Schottky barrier*,[5] and the probability for that is very low for $eV_0 \gg k_B T$. For this reason, the terms *barrier layer* and *Schottky barrier layer* are also widely used for the depletion layer.

In reality, the situation is somewhat more complicated as a charge placed close to a metallic surface is known to polarize the metal. This polarization can be described in terms of a mirror charge of the opposite sign inside the metal. Therefore the potential barrier felt by an electron approaching the interface is lowered and smoothed out in the immediate vicinity of the interface, as illustrated in Fig. 27.3. Since the lowering of the potential barrier is rather small compared to the height itself, this effect can be neglected in calculations.

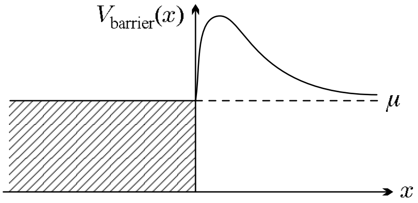


Fig. 27.3. Lowering of the Schottky barrier at a metal–semiconductor contact

Figure 27.4 shows the energy levels for the case where the difference between the work functions exceeds the gap width: $\Phi_m - \Phi_s \geq \varepsilon_g$.

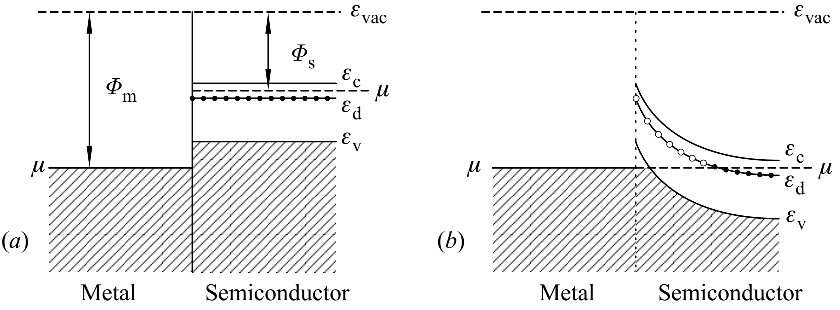


Fig. 27.4. Formation of an inversion layer at the interface of a metal with an n-type semiconductor. The energy-level diagram (*a*) before and (*b*) after the establishment of thermodynamic equilibrium

[5] W. Schottky, 1938.

Once thermal equilibrium has been established close to the interface in such a configuration, the top of the valence band is above the chemical potential in a tiny region that is much narrower than the depletion region. The states at the top of the valence band then also become empty in this small region as it is energetically more favorable for electrons to fill the electron states in the metal. Thus, close to the interface, a large number of holes appear on the semiconductor side, and the n-type semiconductor becomes a p-type semiconductor. Instead of a depletion layer, an inversion layer is formed. Because of the narrowness of the inversion layer, the holes in it can be considered to propagate freely parallel to the interface but confined to a narrow potential well in the perpendicular direction, thus their energy can take only distinct, quantized, values. Indeed, hole states appear in the inversion layer when the highest quantized level is above the Fermi energy. This requires that ε_{v} at the interface should exceed the chemical potential by more than a threshold value, and the inversion layer cannot be too narrow.

The opposite case, when the work function is larger on the semiconductor side ($\Phi_{\mathrm{m}} - \Phi_{\mathrm{s}} < 0$), is illustrated in Fig. 27.5. Following the same steps as above, it can be shown that the energy levels of the semiconductor are bent in the opposite sense as in a depletion layer. Therefore electrons move from the metal to the n-type semiconductor, and fill the states of the conduction band in a layer of thickness λ_{D}, as those states are below the chemical potential. Unlike the depletion layer, this region contains an abundant number of charges, and is accordingly called the accumulation layer. The two sides are not separated by a potential barrier.

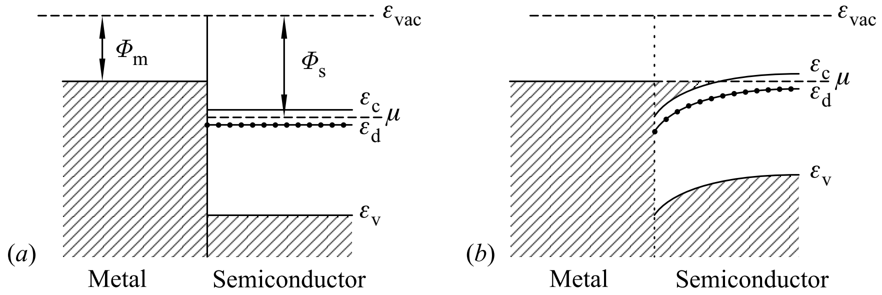

$\varepsilon_{\mathrm{vac}}$
Φ_{m}
Φ_{s}
ε_{c}
μ
ε_{d}
ε_{v}
(a)
Metal
Semiconductor
$\varepsilon_{\mathrm{vac}}$
ε_{c}
μ
ε_{d}
ε_{v}
(b)
Metal
Semiconductor

Fig. 27.5. Formation of an accumulation layer at the interface between a metal and an n-type semiconductor when $\Phi_{\mathrm{m}} - \Phi_{\mathrm{s}} < 0$. The energy-level diagram (a) before and (b) after the establishment of thermodynamic equilibrium

Accumulation, depletion, and inversion layers can also be created in p-type semiconductors; however, the condition for their appearance is reversed with respect to n-type semiconductors. As shown in Fig. 27.6, the condition for the

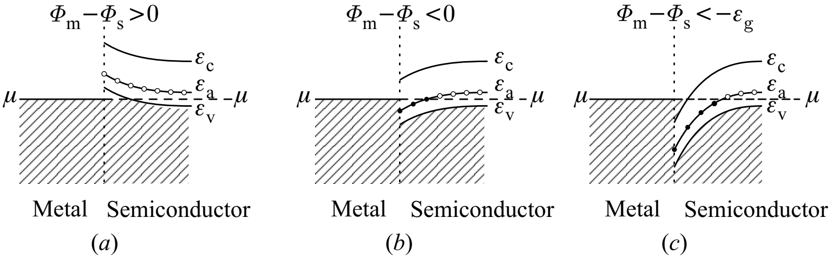


Fig. 27.6. Formation of (*a*) an accumulation, (*b*) a depletion, and (*c*) an inversion layer at the interface between a metal and a *p*-type semiconductor

formation of a depletion layer is now $\Phi_m - \Phi_s < 0$, while an inversion layer can be formed only when

$$\Phi_m - \Phi_s \leq -\varepsilon_g \tag{27.1.18}$$

is also satisfied. If $\Phi_m - \Phi_s > 0$, an accumulation layer appears.

27.1.2 MOS Structures

In devices that contain metallic as well as semiconducting regions, the metal and the semiconductor are usually not in direct contact but separated by a thin insulating layer. Such structures are called metal–insulator–semiconductor (MIS) structures. If the insulating layer is obtained by oxidizing the material on the metallic side, we speak of a metal–oxide–semiconductor (MOS) structure. In most cases this is an M-SiO$_2$-Si structure, where M stands for some metal.

The thickness of the insulating layer is usually chosen to be a few times the atomic dimension, so that the contact-potential drop should occur over a finite region. At the same time, this layer has to be sufficiently narrow for that electrons can tunnel from one side to the other. This can compensate for the possible difference in the chemical potential, since in thermal equilibrium μ has to be the same in the metal and semiconductor. This usually leads to a shift of the energy levels that depends on their distance from the interface. The MOS structure is said to be ideal if the location of the energy levels in

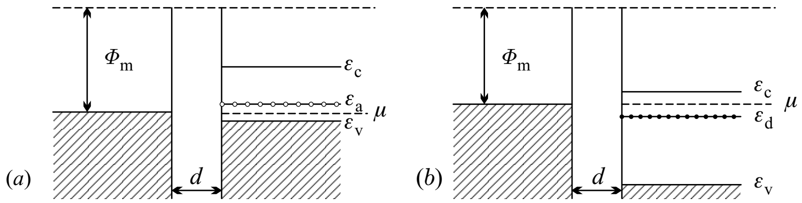


Fig. 27.7. Energy-level diagram of ideal (*a*) *p*-type and (*b*) *n*-type MOS structures

the semiconductor does not depend on the distance from the interface, as in the illustration of Fig. 27.7.

Depending on the type of the semiconductor, we speak of n-MOS and p-MOS structures.

27.1.3 *p–n* Junctions

Even more interesting than the charge redistribution at metal–semiconductor interfaces is the rearrangement at the interface of two differently doped semiconductors. In analogy to the configurations examined above, we shall now consider two semi-infinite semiconductors, a p- and an n-type, that are placed in contact in a plane, and analyze what happens to the electrons close to the interface.

As mentioned in the introduction, such junctions can be fabricated by means of the diffusion of different kinds of impurity atoms – acceptors and donors – into the two halves of an initially homogeneous semiconductor, turning them into p- and n-type semiconductors. Depending on the particular fabrication method, the boundary between the two regions can be very sharp or smeared out. For simplicity, we shall deal with *abrupt* (or *step*) *junctions.* Treating the interface as an infinite plane, variations are considered only in the perpendicular direction (x). We assume that only acceptors are present at $x < 0$ (p side) and donors at $x > 0$ (n side), in uniform distribution:

$$n_{\mathrm{d}}(x) = \begin{cases} n_{\mathrm{d}} & \text{if } x \geq 0\,; \\ 0 & \text{if } x < 0\,; \end{cases} \qquad n_{\mathrm{a}}(x) = \begin{cases} 0 & \text{if } x \geq 0\,; \\ n_{\mathrm{a}} & \text{if } x < 0\,. \end{cases} \tag{27.1.19}$$

For more realistic dopant distributions in graded junctions the calculations can be performed only numerically.

If the semiconductor were imagined to be separated into two parts by an insulating layer during the introduction of donor and acceptor atoms, the chemical potential would be at different heights with respect to the top of the valence band and bottom of the conduction band on the p- and n-sides. This is illustrated in Fig. 27.8(a).

Since the insulating layer is absent, and the n- and p-type semiconductors are in intimate contact, the chemical potential must be the same on both sides. If the chemical potentials are drawn at the same height [Fig. 27.8(b)], but the relative location of the energy levels remains unchanged on both sides, then the conduction and valence bands do not match properly across the interface. In reality, thermal equilibration and the equalization of the chemical potential occurs through the flow of electrons from the n-side to the p-side. Because of the redistribution of charges, an electron deficiency appears on the n-side, and an electron excess on the p-side. Owing to the inhomogeneous charge distribution, an electrostatic potential appears, leading to a distance-dependent bending of the energy levels close to the interface, as shown in Fig. 27.9.

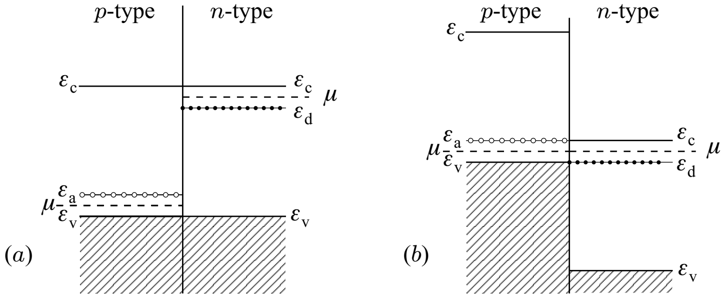


Fig. 27.8. Energy-level diagram of a p–n junction, when (a) the band edges and (b) the chemical potentials are drawn at the same height

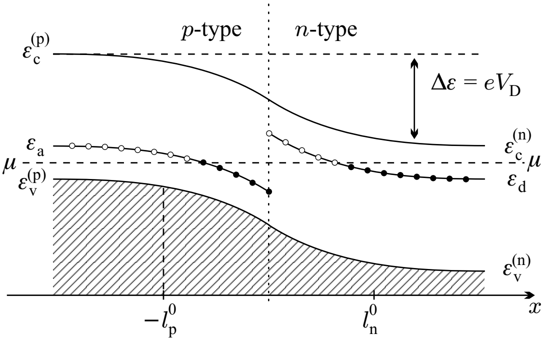


Fig. 27.9. Bending of band edges at the interface of p- and n-type semiconductors

To determine the position-dependent potential $V(\boldsymbol{r})$ and the new charge distribution, we may follow the same steps as for the metal–semiconductor interface. The potential satisfies the Poisson equation

$$\boldsymbol{\nabla}^2 V(\boldsymbol{r}) = -\frac{1}{\epsilon}\rho(\boldsymbol{r})\,. \tag{27.1.20}$$

In the most general case the spatial distribution is inhomogeneous both for donors and acceptors, so the charge density $\rho(\boldsymbol{r})$ is given by

$$\rho(\boldsymbol{r}) = -e\left[n(\boldsymbol{r}) - n_{\mathrm{d}}^{+}(\boldsymbol{r}) + n_{\mathrm{a}}^{-}(\boldsymbol{r}) - p(\boldsymbol{r})\right]. \tag{27.1.21}$$

After the contact has been made, a new thermodynamic equilibrium is established, in which the potential shifts the electron energies by $-eV(\boldsymbol{r})$, thereby changing the thermal occupation. For example, (27.1.2) has to be used for the thermal occupation of the states in the conduction band, which takes the simpler form

$$n(x) = n(\infty)\exp\left\{\frac{e[V(x) - V(\infty)]}{k_{\mathrm{B}}T}\right\} \tag{27.1.22}$$

when spatial variations are assumed to occur only in the x-direction. Rewriting the general formula (27.1.4) for the number of holes in the valence band accordingly,

$$p(x) = p(-\infty) \exp\left\{ -\frac{e[V(x) - V(-\infty)]}{k_{\mathrm{B}}T} \right\}. \tag{27.1.23}$$

It is plausible to expect that the redistribution of charges is limited to the vicinity of the interface, and far from it the previously obtained values for homogeneous n- and p-type semiconductors are recovered. We shall assume, furthermore, that the temperature is in the range where the dopant atoms are completely ionized in the homogeneous semiconductors, and they supply the dominant charge carriers, that is,

$$n(\infty) = n_{\mathrm{d}}\,, \qquad p(-\infty) = n_{\mathrm{a}}\,. \tag{27.1.24}$$

By substituting these formulas into (27.1.22) and (27.1.23), we have

$$\begin{aligned} n(x) &= n_{\mathrm{d}} \exp\left\{ \frac{e[V(x) - V(\infty)]}{k_{\mathrm{B}}T} \right\}, \\ p(x) &= n_{\mathrm{a}} \exp\left\{ \frac{-e[V(x) - V(-\infty)]}{k_{\mathrm{B}}T} \right\}. \end{aligned} \tag{27.1.25}$$

By eliminating $V(x)$ from the two equations of (27.1.25), the relative shift of the energy levels between the two sides of the junction is

$$eV(\infty) - eV(-\infty) = k_{\mathrm{B}}T \ln\left(\frac{n_{\mathrm{d}} n_{\mathrm{a}}}{n(x)p(x)} \right). \tag{27.1.26}$$

The law of mass action must be satisfied everywhere, thus, according to (20.3.27),

$$eV_{\mathrm{D}} \equiv e\Delta V = k_{\mathrm{B}}T \ln\left(\frac{n_{\mathrm{d}} n_{\mathrm{a}}}{n_{\mathrm{i}}^2} \right). \tag{27.1.27}$$

Using (20.3.24), this can be rewritten in the equivalent form

$$eV_{\mathrm{D}} = \varepsilon_{\mathrm{g}} + k_{\mathrm{B}}T \ln\left(\frac{n_{\mathrm{d}} n_{\mathrm{a}}}{N_{\mathrm{c}}(T) P_{\mathrm{v}}(T)} \right). \tag{27.1.28}$$

The last result could have been obtained very simply. As can be read off from Fig. 27.9,

$$eV_{\mathrm{D}} = \varepsilon_{\mathrm{g}} - [\varepsilon_{\mathrm{c}}(\infty) - \mu] - [\mu - \varepsilon_{\mathrm{v}}(-\infty)]\,. \tag{27.1.29}$$

Since all donor and acceptor atoms are assumed to be ionized, and far from the interface $n(\infty) = n_{\mathrm{d}}$ and $p(-\infty) = n_{\mathrm{a}}$, the combination of (20.3.17) and (20.3.20) immediately implies the previous formula for V_{D}.

This potential difference, which depends on the number of dopant atoms on each side of the sample, is called the *contact potential* or *built-in potential.*

Since this potential difference is due to the diffusion of mobile carriers, it is also called *diffusion potential*; that is why the label “D” is used.

If both $N_{\mathrm{c}}(T)$ and $P_{\mathrm{v}}(T)$ are on the order of $10^{19}/\mathrm{cm}^3$ at room temperature, and the concentrations of donors and acceptors are both between 10^{14} and $10^{18}/\mathrm{cm}^3$, then the built-in potential is found to be between 0.5 and 1 V in silicon and somewhat lower in germanium, between 0.4 and 0.6 V. These are smaller than but of the same order as the gap.

Substituting the charge-distribution formulas into (27.1.21), and the resulting expression into the Poisson equation (27.1.20), a closed formula is obtained for the potential. On the n-side ($x > 0$)

$$\begin{aligned}\frac{\mathrm{d}^2V(x)}{\mathrm{d}x^2} = \frac{e}{\epsilon}\biggl\{ & n_{\mathrm{d}} \exp\left[-e\frac{V(\infty)-V(x)}{k_{\mathrm{B}}T}\right] - n_{\mathrm{d}} \\ & - n_{\mathrm{a}} \exp\left[-e\frac{V(x)-V(-\infty)}{k_{\mathrm{B}}T}\right]\biggr\},\end{aligned} \tag{27.1.30}$$

while on the p-side ($x < 0$)

$$\begin{aligned}\frac{\mathrm{d}^2V(x)}{\mathrm{d}x^2} = \frac{e}{\epsilon}\biggl\{ & n_{\mathrm{d}} \exp\left[-e\frac{V(\infty)-V(x)}{k_{\mathrm{B}}T}\right] + n_{\mathrm{a}} \\ & - n_{\mathrm{a}} \exp\left[-e\frac{V(x)-V(-\infty)}{k_{\mathrm{B}}T}\right]\biggr\}.\end{aligned} \tag{27.1.31}$$

Even after these simplifications, the equations can usually be solved only numerically. In a simple physical picture put forward by SCHOTTKY, the distribution of extra charges is assumed to be uniform over a region of width l_{n}^0 on the n-side and l_{p}^0 on the p-side – that is, all variations occur in the region $-l_{\mathrm{p}}^0 < x < l_{\mathrm{n}}^0$. Beyond that – for $x \geq l_{\mathrm{n}}^0$ on the n-side and $x \leq -l_{\mathrm{p}}^0$ on the p-side – the situation is the same as in homogeneous semiconductors. Consequently, we shall assume that

$$n(x) = n_{\mathrm{d}}\,, \qquad V(x) = V(\infty)\,, \qquad \text{if} \qquad x \geq l_{\mathrm{n}}^0 \tag{27.1.32}$$

on the n-side and

$$p(x) = n_{\mathrm{a}}\,, \qquad V(x) = V(-\infty)\,, \qquad \text{if} \qquad x \leq -l_{\mathrm{p}}^0 \tag{27.1.33}$$

on the p-side.

As is well known, there are also holes on the n-side and electrons on the p-side, even though their concentrations are low. While the two equations in (27.1.25) are valid for any x, as long as classical statistics can be applied, the first will be used for $x \geq 0$ (n-side), and the second for $x \leq 0$ (p-side). Consequently, we shall use the notations $n_{\mathrm{n}}(x)$ and $p_{\mathrm{p}}(x)$ for the concentrations of majority carriers, indicating that they correspond to the electron concentration on the n-side and the hole concentration on the p-side. The concentration of minority carriers – holes on the n-side and electrons on the p-side – will be denoted by p_{n} and n_{p}, respectively. These are best expressed from (27.1.27) as

$$\begin{aligned} p_{\mathrm{n}}(x) &= \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}} \exp\left\{-\frac{e[V(x)-V(\infty)]}{k_{\mathrm{B}}T}\right\}, \qquad x>0\,,\\ n_{\mathrm{p}}(x) &= \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} \exp\left\{\frac{e[V(x)-V(-\infty)]}{k_{\mathrm{B}}T}\right\}, \qquad x<0\,. \end{aligned} \tag{27.1.34}$$

Beyond the transition region, where the potential can be replaced by its asymptotic value,

$$\begin{aligned} p_{\mathrm{n}}(x) &\approx \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}} \ll n_{\mathrm{d}}\,, \qquad \text{if} \qquad x \ge l_{\mathrm{n}}^0\,,\\ n_{\mathrm{p}}(x) &\approx \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} \ll n_{\mathrm{a}}\,, \qquad \text{if} \qquad x \le -l_{\mathrm{p}}^0\,. \end{aligned} \tag{27.1.35}$$

In the transition region the concentration of electrons and holes changes rapidly between the two limits. This is shown in Fig. 27.10, where the concentrations are plotted on a logarithmic scale.

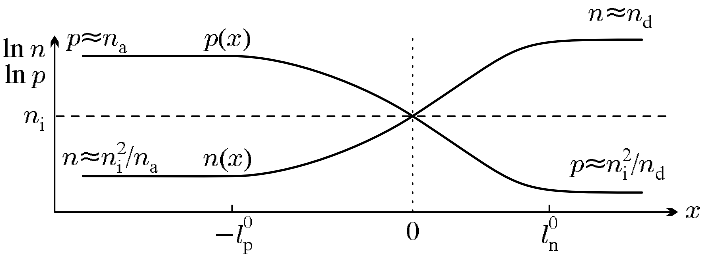


Fig. 27.10. Variations of the electron and hole concentrations in a *p*–*n* junction

The variation of $eV(x)$ in the transition region is comparable to ε_{g}. In the limit $\varepsilon_{\mathrm{g}} \gg k_{\mathrm{B}}T$, both $n(x)$ and $p(x)$ are small and can be neglected in a first approximation, as

$$\begin{aligned} n(x) &= n_{\mathrm{d}} \exp\left[-e\frac{V(\infty)-V(x)}{k_{\mathrm{B}}T}\right] \sim n_{\mathrm{d}} \exp\left[\frac{-\varepsilon_{\mathrm{g}}}{k_{\mathrm{B}}T}\right] \ll n_{\mathrm{d}}\,,\\ p(x) &= n_{\mathrm{a}} \exp\left[-e\frac{V(x)-V(-\infty)}{k_{\mathrm{B}}T}\right] \sim n_{\mathrm{a}} \exp\left[\frac{-\varepsilon_{\mathrm{g}}}{k_{\mathrm{B}}T}\right] \ll n_{\mathrm{a}}\,. \end{aligned} \tag{27.1.36}$$

Thus in the Schottky approximation all electrons supplied by the donors (holes introduced by the acceptors) are assumed to leave the region of width l_{n}^0 (l_{p}^0) on the *n*-side (*p*-side). The full width of the depletion layer is therefore $l_{\mathrm{n}}^0 + l_{\mathrm{p}}^0$. Having diffused to the other side of the contact, electrons and holes leave behind uncompensated ions in the depletion layer. The actual charge distribution is given by

$$\rho(x) = \begin{cases} en_{\mathrm{d}}\,, & 0 < x < l_{\mathrm{n}}^0\,, \\ -en_{\mathrm{a}}\,, & -l_{\mathrm{p}}^0 < x < 0\,. \end{cases} \tag{27.1.37}$$

That is why the depletion layer is also called *space-charge region* here, too. Because of the overall neutrality of the sample, the condition

$$n_{\mathrm{d}}\, l_{\mathrm{n}}^0 = n_{\mathrm{a}}\, l_{\mathrm{p}}^0 \tag{27.1.38}$$

has to be met. In customary setups one side is much more heavily doped than the other, and so the depletion layer is very narrow there. The spatial distribution of mobile carriers and the full charge density close to the interface are shown in Fig. 27.11.

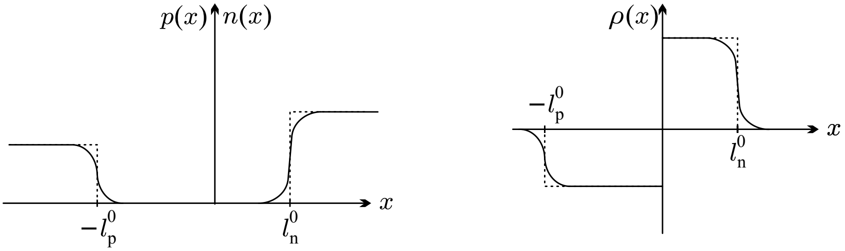


Fig. 27.11. The distribution of carriers and full charge density in a p–n junction. The dotted line indicates the Schottky approximation

Using these approximations, the equation to be solved reads

$$\frac{\mathrm{d}^2 V(x)}{\mathrm{d}x^2} = -\frac{e}{\epsilon} \begin{cases} n_{\mathrm{d}}\,, & 0 < x < l_{\mathrm{n}}^0\,, \\ -n_{\mathrm{a}}\,, & -l_{\mathrm{p}}^0 < x < 0\,. \end{cases} \tag{27.1.39}$$

By integrating it once, the field strength E is obtained. When E is required to be continuous and vanish outside the depletion layer,

$$E(x) = \begin{cases} \dfrac{en_{\mathrm{d}}}{\epsilon}(x - l_{\mathrm{n}}^0)\,, & 0 < x < l_{\mathrm{n}}^0\,, \\ -\dfrac{en_{\mathrm{a}}}{\epsilon}(x + l_{\mathrm{p}}^0)\,, & -l_{\mathrm{p}}^0 < x < 0\,. \end{cases} \tag{27.1.40}$$

From the requirement that the field strength should be continuous at $x = 0$, (27.1.38) is recovered. This is perfectly understandable, as it is known to be the consequence of charge conservation.

By integrating the field strength, too, the solution that leads to a continuous potential in l_{n}^0 and $-l_{\mathrm{p}}^0$ is

$$V(x) = \begin{cases} V(\infty) - \dfrac{en_{\mathrm{d}}}{2\epsilon}(x - l_{\mathrm{n}}^0)^2\,, & 0 < x < l_{\mathrm{n}}^0\,, \\ V(-\infty) + \dfrac{en_{\mathrm{a}}}{2\epsilon}(x + l_{\mathrm{p}}^0)^2\,, & -l_{\mathrm{p}}^0 < x < 0\,. \end{cases} \tag{27.1.41}$$

The variations of the potential and electric field are shown in Fig. 27.12.

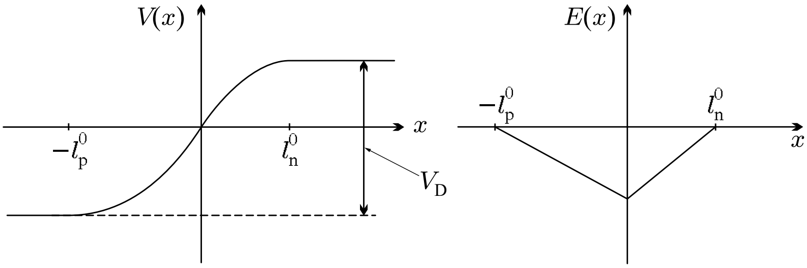


Fig. 27.12. The variations of the potential and electric field in a *p*–*n* junction

Since the potential, too, has to be continuous at $x = 0$,

$$V(\infty) - \frac{e}{2\epsilon} n_\mathrm{d} l_\mathrm{n}^{0\,2} = V(-\infty) + \frac{e}{2\epsilon} n_\mathrm{a} l_\mathrm{p}^{0\,2} \,. \tag{27.1.42}$$

This leads to a relation between the built-in potential and the depletion width:

$$V_\mathrm{D} = V(\infty) - V(-\infty) = \frac{e}{2\epsilon}\left(n_\mathrm{d} l_\mathrm{n}^{0\,2} + n_\mathrm{a} l_\mathrm{p}^{0\,2}\right) . \tag{27.1.43}$$

The width of the space-charge regions on the *n*- and *p*-sides can then be determined by making use of (27.1.38):

$$l_\mathrm{n}^0 = \left\{ \frac{2\epsilon V_\mathrm{D}}{e} \frac{n_\mathrm{a}/n_\mathrm{d}}{n_\mathrm{d} + n_\mathrm{a}} \right\}^{1/2} , \qquad l_\mathrm{p}^0 = \left\{ \frac{2\epsilon V_\mathrm{D}}{e} \frac{n_\mathrm{d}/n_\mathrm{a}}{n_\mathrm{d} + n_\mathrm{a}} \right\}^{1/2} . \tag{27.1.44}$$

The full width of the space-charge region is thus

$$l = l_\mathrm{n}^0 + l_\mathrm{p}^0 = \left\{ \frac{2\epsilon V_\mathrm{D}}{e} \frac{n_\mathrm{d} + n_\mathrm{a}}{n_\mathrm{d}\, n_\mathrm{a}} \right\}^{1/2} . \tag{27.1.45}$$

Assuming typical values between 10^{14} and $10^{18}/\mathrm{cm}^3$ for the dopant concentration and $V_\mathrm{D} = 1\,\mathrm{V}$ for the diffusion potential, the depletion width is $l_{\mathrm{n,p}}^0 \sim 10^{-6}$ to $10^{-4}\,\mathrm{cm}$. Since the drop of the diffusion potential occurs over such a distance, the electric field can be as high as 10^4 to $10^6\,\mathrm{V/cm}$ here.

27.1.4 Heterojunctions

Using epitaxial growth techniques, it is possible to grow semiconductors of different chemical composition (e.g., different doping) on top of each other – for example, $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ on GaAs – provided the mismatch of the lattice parameters is small. Depending on the type of dopants in the two materials,

n–n, n–p, p–n, and p–p heterojunctions can be fabricated.[6] Naturally, the gaps and work functions are different on the two sides of the junction. The flat-band energy-level diagram for a p–n heterostructure of p-type GaAs and n-type $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ is shown in Fig. 27.13(a). The energy difference between the tops of the valence bands is $\Delta\varepsilon_\mathrm{v}$, while that between the bottoms of the conduction bands is $\Delta\varepsilon_\mathrm{c}$. Since the chemical potential is located at different distances from the vacuum level, which cannot be maintained in thermal equilibrium, the energy levels must bend close to the interface, but in such a way that the differences $\Delta\varepsilon_\mathrm{v}$ and $\Delta\varepsilon_\mathrm{c}$ are preserved. This is illustrated in Fig. 27.13(b).

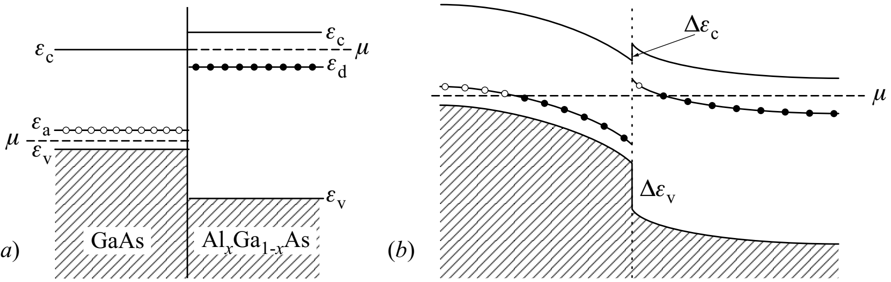


Fig. 27.13. (a) Flat-band energy-level diagram for p- and n-type semiconductors with different gaps. (b) The energy levels of a p–n heterojunction

Since the gap is larger in the n-type $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$, a well and a barrier appear in the conduction band, and a jump in the valence band. In case of heavy doping, the potential well can be shifted below the chemical potential. Then electrons move from the right-hand side to the left-hand side, and accumulate there, making up a two-dimensional electron gas (2DEG). In such configurations the particular properties of the two-dimensional electron gas can be studied. Heterojunctions are also widely used in such important technological applications as optoelectronic devices, semiconductor lasers, photodetectors, and solar cells.

Figure 27.14 shows the band structures for the three other heterojunction types: n–n, p–p, and n–p.

27.2 Generation, Motion, and Recombination of Carriers

Before turning to the study of how the charge distribution close to a metal–semiconductor or semiconductor–semiconductor interface is modified by an applied voltage, we present the physical processes that determine the current. One would expect that carriers move from one side to the other via diffusion.

[6] As usual, the first letter shows the doping type of the material with a narrower gap.

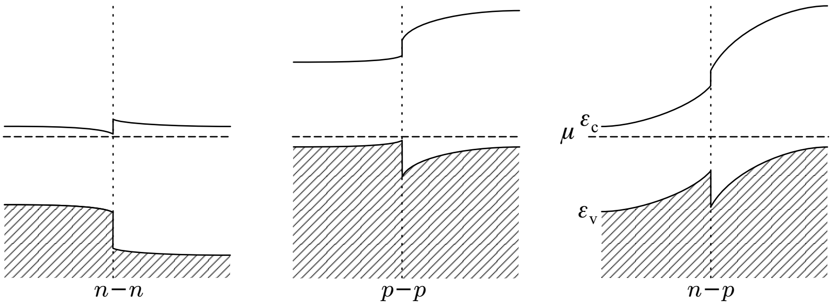


Fig. 27.14. Schematic energy-level diagrams of n–n, p–p, and n–p heterojunctions

However, this is not the only process that can change the number of carriers in space and time. Through the creation of electron–hole pairs, new carriers can be generated in the sample, and through recombination they may also disappear. Below we shall give an overview of certain features of these processes.

27.2.1 Generation and Recombination of Carriers

Thermal equilibrium is brought about by incessant collisions in semiconductors. As mentioned in Section 23.3, phonons can be absorbed while their energy and momentum are used for the creation of an electron–hole pair. The inverse process can also take place: in the annihilation of an electron and a hole, a phonon can be created. This is the recombination of carriers. The same process also exists with a photon, rather than a phonon, in the final state; it is then called radiative recombination. This is how the current is converted into light in light-emitting diodes (LEDs).[7]

If the gap is less than 0.2–0.3 eV, the direct recombination of the electron–hole pair is possible. For larger gaps the recombination occurs dominantly indirectly, through those deep levels (traps) that are located deep inside the gap and can interact both with conduction- and valence-band states. This recombination takes about $\tau \sim 10^{-3}\,\mathrm{s}$ in silicon and germanium, and only $\tau \sim 10^{-8}\,\mathrm{s}$ in GaAs. Even without going into the details of the interaction between electrons and phonons, or specifying the capture probability of traps, some general observations can be made about these generation and recombination processes.

Suppose that thermodynamic equilibrium is broken by some external disturbance, and the number densities n and p of charge carriers are different from the equilibrium values (n_0 and p_0) in the conduction and valence bands alike. Since in the recombination process an electron in the conduction band

[7] For semiconductors with a wide gap, such as GaN or GaP, the emitted light is in the visible region, whereas GaAs-based LEDs emit infrared radiation.

fills a hole in the valence band, the recombination rate (number of recombination events in unit time, R) must be proportional to the electron- and hole-number densities:

$$R = C\,n\,p\,, \tag{27.2.1}$$

where C is a proportionality factor that will be determined later.

The probability of carrier generation cannot be obtained from such simple considerations. In addition to thermally created electron–hole pairs, semiconductors also contain carriers that were introduced externally, through injection or interaction with light. We shall denote the external and thermal generation rates by G_{inj} and G_{therm}. We shall assume that the latter is independent of the actual numbers of electrons and holes – in other words, the equilibrium value can also be used in nonequilibrium states. On the other hand, in thermal equilibrium the number of carriers is constant because the increase due to thermal generation is compensated for by the decrease due to recombination. Therefore,

$$G_{\text{therm}} = R_0 = C\,n_0\,p_0\,. \tag{27.2.2}$$

In what follows, we shall ignore the possibility of carrier injection. If the number of carriers differs from the equilibrium value, the recombination and generation processes do not compensate each other. The net contribution of the two processes changes the number of carriers in such a way that thermal equilibrium should be restored. Since one electron and one hole are created (annihilated) in each generation (recombination) process, the net recombination rate of either type of carrier is

$$U = R - G_{\text{therm}} = C\,[n\,p - n_0\,p_0]\,. \tag{27.2.3}$$

The right-hand side can be rewritten as

$$U = C\,[(n - n_0)(p - p_0) + n_0(p - p_0) + p_0(n - n_0)]\,. \tag{27.2.4}$$

Assuming that the departure from equilibrium is slight, the first term, being a second-order quantity, can be neglected, and so

$$U = C\,[n_0(p - p_0) + p_0(n - n_0)]\,. \tag{27.2.5}$$

In n-type semiconductors $p_0 \ll n_0$, therefore

$$U \approx Cn_0(p - p_0)\,, \tag{27.2.6}$$

while in p-type semiconductors

$$U \approx Cp_0(n - n_0)\,, \tag{27.2.7}$$

that is, the variation in the number of carriers always depends on the excess or deficiency of minority carriers.

If no current flows in the sample, the change in the carrier numbers is exclusively due to generation and recombination processes. In n-type semiconductors

$$\frac{\partial n}{\partial t} = \frac{\partial p}{\partial t} = -Cn_0(p - p_0)\,, \tag{27.2.8}$$

and hence

$$p(t) - p_0 = \bigl[p(0) - p_0\bigr]\mathrm{e}^{-t/\tau_\mathrm{p}}\,, \tag{27.2.9}$$

where

$$\tau_\mathrm{p} = \frac{1}{Cn_0} \tag{27.2.10}$$

is the *carrier lifetime*, also called the recombination lifetime or recombination time. Its meaning is obvious from the previous exponential time dependence: τ_p is the mean recombination time for an additional hole introduced in an n-type semiconductor. Its inverse is the recombination probability in unit time.

By expressing the constant C in terms of the recombination lifetime, the variations in the densities of electrons and holes in n-type semiconductors due to recombination and generation are governed by the equations

$$\frac{\partial n}{\partial t} = -\frac{p - p_0}{\tau_\mathrm{p}}\,, \qquad \frac{\partial p}{\partial t} = -\frac{p - p_0}{\tau_\mathrm{p}}\,. \tag{27.2.11}$$

The situation in p-type semiconductors can be treated analogously. If the number of carriers differs from the value in thermal equilibrium, the variation in the number of extra electrons can be described in terms of a recombination lifetime τ_n that is related to C through

$$\tau_\mathrm{n} = \frac{1}{Cp_0}\,. \tag{27.2.12}$$

In terms of τ_n, the net recombination rate is

$$U = \frac{n - n_0}{\tau_\mathrm{n}}\,, \tag{27.2.13}$$

and

$$\frac{\partial n}{\partial t} = -\frac{n - n_0}{\tau_\mathrm{n}}\,, \qquad \frac{\partial p}{\partial t} = -\frac{n - n_0}{\tau_\mathrm{n}}\,. \tag{27.2.14}$$

Up to this point we have ignored the current that may flow in the semiconductor. If the number of carriers were conserved, the continuity equation would apply. Because of the generation and recombination processes, the continuity equation needs to be complemented by the previous terms for the variations in the particle numbers. The net variations in the numbers of electrons and holes due to generation and recombination processes are known to be determined by the number of minority carriers. Therefore in n-type semiconductors, where the minority carriers are holes, we have:

$$\frac{\partial n}{\partial t} - \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_\mathrm{n} = -\frac{p - p_0}{\tau_\mathrm{p}}\,, \qquad \frac{\partial p}{\partial t} + \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_\mathrm{p} = -\frac{p - p_0}{\tau_\mathrm{p}}\,, \tag{27.2.15}$$

where $\boldsymbol{j}_{\mathrm{n}}$ and $\boldsymbol{j}_{\mathrm{p}}$ are the currents carried by electrons and holes, respectively. Analogously, in p-type semiconductors, where the minority carriers are electrons,

$$\frac{\partial n}{\partial t} - \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\mathrm{n}} = -\frac{n-n_0}{\tau_{\mathrm{n}}}\,, \qquad \frac{\partial p}{\partial t} + \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\mathrm{p}} = -\frac{n-n_0}{\tau_{\mathrm{n}}}\,. \tag{27.2.16}$$

In the presence of a current, the number of charge carriers can remain constant in time, as the divergence of the current can compensate the loss due to recombination.

27.2.2 Diffusion and Drift of Carriers

The recombination lifetime introduced above is much longer than the mean time between collisions during the motion of an electron or hole, which characterizes the amount of time the particle spends in a particular state. The latter is about $\tau \sim 10^{-12}$ to $10^{-13}\,$s. The motion of electrons should therefore be considered as diffusion in the transition region, where the charge distribution is nonuniform. This indicates the inadequacy of the previous static picture, and the necessity of including the diffusion current of carriers in the description of the processes close to the contact, even though there is no net current. The *diffusion potential* $V(\boldsymbol{r})$, which is present because of the nonuniformity of the charge distribution, gives rise to an ohmic drift current. The two components of the current must cancel out perfectly. This leads to a self-consistent relation between the charge distribution and the potential.

Denoting the electron diffusion coefficient by D_{n}, the particle-current density driven by diffusion is

$$\boldsymbol{j}^{\mathrm{diff}}(\boldsymbol{r}) = -D_{\mathrm{n}}\operatorname{grad} n(\boldsymbol{r}) \tag{27.2.17}$$

according to *Fick's first law.*[8] Since each electron carries a charge $-e$, the diffusion current is

$$\boldsymbol{j}_{\mathrm{n}}^{\mathrm{diff}}(\boldsymbol{r}) = -e\boldsymbol{j}^{\mathrm{diff}}(\boldsymbol{r}) = eD_{\mathrm{n}}\operatorname{grad} n(\boldsymbol{r})\,. \tag{27.2.18}$$

Besides the charge distribution, the potential also varies in space in inhomogeneous specimens. Therefore the electric field $\boldsymbol{E}(\boldsymbol{r}) = -\operatorname{grad} V(\boldsymbol{r})$ does not vanish, nor does the drift current $\boldsymbol{j}_{\mathrm{n}}^{\mathrm{drift}}(\boldsymbol{r}) = -en(\boldsymbol{r})\boldsymbol{v}_{\mathrm{n}}(\boldsymbol{r})$, which is proportional to it. The velocity $\boldsymbol{v}_{\mathrm{n}}$ of electrons can be related to the electric field $\boldsymbol{E}$ through the mobility μ_{n} as $\boldsymbol{v}_{\mathrm{n}} = -\mu_{\mathrm{n}}\boldsymbol{E}$. Since electrons move against the field, the drift current is

$$\boldsymbol{j}_{\mathrm{n}}^{\mathrm{drift}}(\boldsymbol{r}) = en(\boldsymbol{r})\mu_{\mathrm{n}}\boldsymbol{E}(\boldsymbol{r}) = -en(\boldsymbol{r})\mu_{\mathrm{n}}\operatorname{grad} V(\boldsymbol{r})\,. \tag{27.2.19}$$

There is no net charge flow in thermal equilibrium as the two currents cancel out:

[8] A. Fick, 1855.

$$- en(\boldsymbol{r})\mu_{\mathrm{n}} \operatorname{grad} V(\boldsymbol{r}) + eD_{\mathrm{n}} \operatorname{grad} n(\boldsymbol{r}) = 0\,. \tag{27.2.20}$$

Since (27.1.2) implies

$$\operatorname{grad} n(\boldsymbol{r}) = \frac{e}{k_{\mathrm{B}}T} n(\boldsymbol{r}) \operatorname{grad} V(\boldsymbol{r})\,, \tag{27.2.21}$$

comparison with (27.2.20) leads to the *Einstein relation*[9]

$$\mu_{\mathrm{n}} = \frac{e}{k_{\mathrm{B}}T} D_{\mathrm{n}} \tag{27.2.22}$$

between the diffusion coefficient and the mobility. These considerations provide additional support to the picture that the nonuniform charge distribution on the two sides of the contact is due to the diffusion of electrons, and justifies the term *diffusion potential.*

The diffusion and drift currents of holes can be treated likewise. Because of the positive charge of holes, the diffusion current, which is proportional to the concentration gradient, is

$$\boldsymbol{j}_{\mathrm{p}}^{\mathrm{diff}}(\boldsymbol{r}) = -eD_{\mathrm{p}} \operatorname{grad} p(\boldsymbol{r})\,, \tag{27.2.23}$$

and the drift current, which is driven by the potential gradient, is

$$\boldsymbol{j}_{\mathrm{p}}^{\mathrm{drift}}(\boldsymbol{r}) = ep(\boldsymbol{r})\boldsymbol{v}_{\mathrm{p}}(\boldsymbol{r}) = e\mu_{\mathrm{p}} p(\boldsymbol{r})\boldsymbol{E} = -e\mu_{\mathrm{p}} p(\boldsymbol{r}) \operatorname{grad} V(\boldsymbol{r})\,, \tag{27.2.24}$$

where μ_{p} is the hole mobility, and we exploited that holes are drifted along the field.

Since the net current of holes is zero in thermal equilibrium,

$$- eD_{\mathrm{p}} \operatorname{grad} p(\boldsymbol{r}) - e\mu_{\mathrm{p}} p(\boldsymbol{r}) \operatorname{grad} V(\boldsymbol{r}) = 0\,. \tag{27.2.25}$$

As implied by (27.1.4), the density of holes satisfies

$$\operatorname{grad} p(\boldsymbol{r}) = -\frac{e}{k_{\mathrm{B}}T} p(\boldsymbol{r}) \operatorname{grad} V(\boldsymbol{r})\,. \tag{27.2.26}$$

The Einstein relation is valid for holes as well:

$$D_{\mathrm{p}} = \frac{k_{\mathrm{B}}T}{e} \mu_{\mathrm{p}}\,. \tag{27.2.27}$$

The mobility of electrons and holes, as well as the diffusion coefficients at room temperature are listed in Table 27.1 for a number of semiconductors.

Through their diffusive motion, particles move a distance $L = \sqrt{Dt}$ in time t. Thus the minority carriers injected into a p- or n-type semiconductor through an interface are expected to cover a distance

[9] A. Einstein, 1905, also known as the Einstein–Smoluchowski relation because the theory of Brownian motion was also worked out independently by M. Smoluchowski one year later.

Table 27.1. The room-temperature mobility and diffusion coefficient of electrons and holes in the transition region

Semiconductor	μ_{n} (cm^2/V s)	μ_{p} (cm^2/V s)	D_{n} (cm^2/s)	D_{p} (cm^2/s)
Si	1350	480	34.9	12.9
Ge	3900	1900	100.8	49.1
GaAs	8800	320	227.5	10.3
GaSb	3750	680		
InAs	33 000	450		
InSb	77 000	850		

$$L_{\mathrm{n}} = \sqrt{D_{\mathrm{n}}\tau_{\mathrm{n}}}\,, \quad \text{and} \quad L_{\mathrm{p}} = \sqrt{D_{\mathrm{p}}\tau_{\mathrm{p}}} \tag{27.2.28}$$

during the carrier lifetime $\tau_{\mathrm{n,p}}$. To prove this, we shall consider the equations governing the spatial and temporal variations of the density of electrons and holes, (27.2.15) and (27.2.16), in the stationary case.

When the concentrations vary in a single direction, and thus diffusion occurs only along that direction, the equation for the minority carriers (holes) in n-type semiconductors is simplified to

$$\frac{1}{e}\frac{\mathrm{d}j_{\mathrm{p}}}{\mathrm{d}x} = -\frac{p - p_0}{\tau_{\mathrm{p}}}\,. \tag{27.2.29}$$

Substitution of the diffusion current from (27.2.23) yields

$$D_{\mathrm{p}}\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} = \frac{p - p_0}{\tau_{\mathrm{p}}}\,. \tag{27.2.30}$$

In terms of L_{n} and L_{p}, defined in (27.2.28), the equation for the variations of the hole density is

$$\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} = \frac{p - p_0}{L_{\mathrm{p}}^2}\,. \tag{27.2.31}$$

In the physically meaningful solution the density of the injected holes decreases exponentially with the distance, as $\exp(-x/L_{\mathrm{p}})$. If the carriers are injected at $x = 0$, and the equilibrium concentration p_0 is to be recovered for large values of x, the solution is

$$p(x) = p_0 + [p(0) - p_0]\mathrm{e}^{-x/L_{\mathrm{p}}}\,. \tag{27.2.32}$$

Therefore L_{p} is the hole diffusion length. On the other hand, when the injected charges are removed (drained) at the end of the semiconducting sample of width d, the spatial distribution of holes is given by

$$p(x) = p_0 + [p(0) - p_0]\frac{\sinh[(d - x)/L_{\mathrm{p}}]}{\sinh(d/L_{\mathrm{p}})}\,. \tag{27.2.33}$$

In perfect analogy, the spatial variation of the density of injected electrons on the p-side is given by $\exp(x/L_{\rm n})$, where $L_{\rm n}$ is the diffusion length of electrons, which now play the role of minority carriers.

The estimated magnitude of these diffusion lengths turns out to be much larger than the width of the depletion layer. This point will be important for the operation of semiconductor devices.

27.2.3 Fundamental Equations of the Physics of Semiconductor Devices

We are now in the position to write down the equations that can be used to model the behavior of semiconductor devices mathematically. The Maxwell equations are, of course, valid, but they have to be complemented by the constitutive relations and the continuity equation.

The total current density in an arbitrary point of the semiconductor is the sum of the electron and hole currents:

$$\boldsymbol{j}(\boldsymbol{r}) = \boldsymbol{j}_{\rm n}(\boldsymbol{r}) + \boldsymbol{j}_{\rm p}(\boldsymbol{r})\,. \tag{27.2.34}$$

The current of either carrier type can be decomposed further, into a diffusion current and a drift current that is the response to the applied electric field:

$$\begin{aligned} \boldsymbol{j}_{\rm n}(\boldsymbol{r}) &= en(\boldsymbol{r})\mu_{\rm n}\boldsymbol{E} + eD_{\rm n}\boldsymbol{\nabla} n(\boldsymbol{r})\,,\\ \boldsymbol{j}_{\rm p}(\boldsymbol{r}) &= ep(\boldsymbol{r})\mu_{\rm p}\boldsymbol{E} - eD_{\rm p}\boldsymbol{\nabla} p(\boldsymbol{r})\,. \end{aligned} \tag{27.2.35}$$

The electric field $\boldsymbol{E}$ is determined by the total charge density. In addition to mobile carriers, the charged donors and acceptors must also be taken into account:

$$\epsilon\,\operatorname{div}\boldsymbol{E} = -e\left[n(\boldsymbol{r}) + n_{\rm a}^-(\boldsymbol{r}) - n_{\rm d}^+(\boldsymbol{r}) - p(\boldsymbol{r})\right]. \tag{27.2.36}$$

We shall also need the extensions of the continuity equation that contain the generation and recombination terms. Allowing for the injection of carriers,

$$\begin{aligned} \frac{\partial n}{\partial t} - \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\rm n} &= G_{\rm inj,\,n} + G_{\rm therm} - R\,,\\ \frac{\partial p}{\partial t} + \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\rm p} &= G_{\rm inj,\,p} + G_{\rm therm} - R\,. \end{aligned} \tag{27.2.37}$$

Since the thermal generation and recombination of carriers are determined by the number of minority carriers, it is useful to write down separate equations for n-type and p-type semiconductors. Following the convention that the type of the semiconductor is indicated by a subscript, the variations of the number density of electrons (n) and holes (p) are given by

$$\begin{aligned} \frac{\partial n_{\rm n}}{\partial t} &= G_{\rm inj,\,n} + D_{\rm n}\boldsymbol{\nabla}^2 n_{\rm n} + \mu_{\rm n}\operatorname{div}(n_{\rm n}\boldsymbol{E}) - \frac{p_{\rm n} - p_{\rm n0}}{\tau_{\rm p}}\,,\\ \frac{\partial p_{\rm n}}{\partial t} &= G_{\rm inj,\,p} + D_{\rm p}\boldsymbol{\nabla}^2 p_{\rm n} - \mu_{\rm p}\operatorname{div}(p_{\rm n}\boldsymbol{E}) - \frac{p_{\rm n} - p_{\rm n0}}{\tau_{\rm p}} \end{aligned} \tag{27.2.38}$$

in an n-type semiconductor, and by

$$\begin{aligned} \frac{\partial n_{\mathrm{p}}}{\partial t} &= G_{\mathrm{inj,\,n}} + D_{\mathrm{n}} \boldsymbol{\nabla}^2 n_{\mathrm{p}} + \mu_{\mathrm{n}} \operatorname{div}(n_{\mathrm{p}} \boldsymbol{E}) - \frac{n_{\mathrm{p}} - n_{\mathrm{p}0}}{\tau_{\mathrm{n}}} \,, \\ \frac{\partial p_{\mathrm{p}}}{\partial t} &= G_{\mathrm{inj,\,p}} + D_{\mathrm{p}} \boldsymbol{\nabla}^2 p_{\mathrm{p}} - \mu_{\mathrm{p}} \operatorname{div}(p_{\mathrm{p}} \boldsymbol{E}) - \frac{n_{\mathrm{p}} - n_{\mathrm{p}0}}{\tau_{\mathrm{n}}} \end{aligned} \qquad (27.2.39)$$

in a p-type semiconductor.

In principle, the solution of these equations allow us to determine the current in any system composed of arbitrarily doped semiconductor components for any applied voltage – that is, the characteristics of the semiconductor device. Below we shall consider a few simple cases.

27.3 Biased Semiconductor Junctions

In the previous section we found that the charge distribution became nonuniform near to the interface between a metal and a semiconductor or two semiconductors. According to Fick's first law, a diffusion current has to flow through the interface then. However, no net charge current is produced because an internal potential, the diffusion potential, is built up. Steady current can flow only when an external voltage is applied. Owing to the different character of the two sides, the current depends strongly on the polarity, therefore these junctions may exhibit rectifying properties. The rectifiers used in the first radios were based on metal–semiconductor contacts.

27.3.1 Biased Schottky Diodes

We shall first consider a metal–semiconductor junction in which a depletion layer is formed between the metal and the n-type semiconductor. In the previous section we solved the Poisson equation to determine the spatial variations of the static potential $V(x)$ and the electron density $n(x)$ in the depletion layer. In reality, this equilibrium state is established by two currents that compensate each other perfectly: a diffusion current toward the metallic side, and an oppositely directed drift current,

$$\boldsymbol{j}_{\mathrm{n}}(\boldsymbol{r}) = en(\boldsymbol{r})\mu_{\mathrm{n}}\boldsymbol{E} + eD_{\mathrm{n}}\boldsymbol{\nabla} n(\boldsymbol{r}) = 0\,. \qquad (27.3.1)$$

Writing the electric field as the negative gradient of the potential $V(\boldsymbol{r})$, and exploiting the Einstein relation, the solution given in (27.1.9) is recovered:

$$n(x) = n(\infty)\exp\left(\frac{e[V(x) - V(\infty)]}{k_{\mathrm{B}}T}\right). \qquad (27.3.2)$$

Let us now apply a voltage $V > 0$ to this setup (known as the *Schottky diode* or *Schottky barrier diode*) in such a way that the metal is the positive electrode. By convention, the junction is then said to be *forward biased.* When the metal is connected to the negative electrode, it is *reverse biased.* The thermal equilibrium is broken by the applied voltage, and so the chemical potential becomes different on the two sides: the energy levels on the metallic side are shifted downward by eV. Alternatively, we could just as well say that the energy levels of the semiconductor experience an upward shift of the same magnitude. However, since the relative location of the energy levels on the two sides remains unaltered (the levels are "pinned") in the point of contact, the extra potential grows to its asymptotic value gradually with the distance. This leads to a further bending of the energy levels on the semiconductor side. The energy-level diagrams in the presence of an applied voltage are shown in Fig. 27.15.

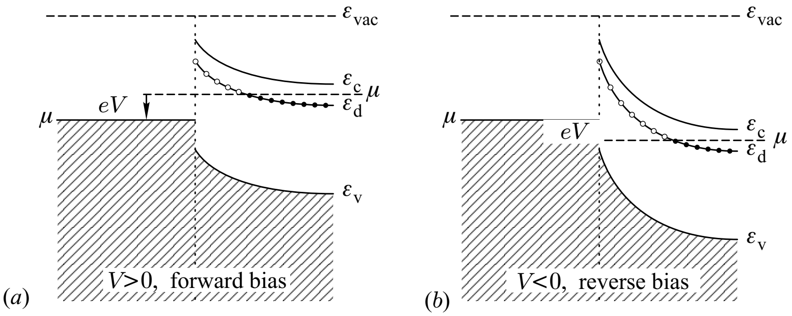


Fig. 27.15. Energy-level diagram at the interface of a metal and an n-type semiconductor in the presence of an applied voltage. (*a*) $V > 0$, (*b*) $V < 0$

When a nonzero voltage is applied, a net current flows from the semiconductor to the metal or vice versa. In n-type semiconductors the dominant contribution comes from electrons because at the relevant temperatures the valence band of the semiconductor is completely filled, just like those electron states in the metal that are at the same height.

Instead of $eV_0 = \varPhi_{\text{m}} - \varPhi_{\text{s}}$, the height of the potential barrier seen by the electrons coming from the semiconductor side is now $eV_0 - eV$. They can pass through it either classically, if their thermal energy is large enough, or by quantum mechanical tunneling. In metal–semiconductor contacts the first option is more important. Later we shall also discuss semiconductor devices in which tunneling plays the dominant role. For $V > 0$ the potential barrier is smaller than in the equilibrium state, therefore electrons start to flow from the semiconductor to the metal. The current is due to the majority carriers injected from the semiconductor to the metal.

Two situations need to be distinguished. If the electron mean free path exceeds the thickness of the depletion layer, the current can be determined from the Richardson–Dushman equation[10] of thermionic emission. Only those electrons that are able to overcome the potential barrier $V_0 - V$ can get from the semiconductor to the metal side. According to the formulas of classical statistics, if the number density of electrons inside the semiconductor is n_d, the number density of electrons whose velocity components are between v_x and $v_x + \mathrm{d}v_x$, v_y and $v_y + \mathrm{d}v_y$, v_z and $v_z + \mathrm{d}v_z$ is given by

$$\mathrm{d}n = n_\mathrm{d} \left(\frac{m_\mathrm{n}^*}{2\pi k_\mathrm{B} T} \right)^{3/2} \mathrm{e}^{-m_\mathrm{n}^*(v_x^2+v_y^2+v_z^2)/2k_\mathrm{B}T} \, \mathrm{d}v_x \mathrm{d}v_y \mathrm{d}v_z \,. \tag{27.3.3}$$

The density of the current carried by them in the $-x$-direction is

$$\mathrm{d}j_x = e v_x n_\mathrm{d} \left(\frac{m_\mathrm{n}^*}{2\pi k_\mathrm{B} T} \right)^{3/2} \mathrm{e}^{-m_\mathrm{n}^*(v_x^2+v_y^2+v_z^2)/2k_\mathrm{B}T} \, \mathrm{d}v_x \mathrm{d}v_y \mathrm{d}v_z \,. \tag{27.3.4}$$

However, only those electrons get over the potential barrier for which

$$\tfrac{1}{2} m_\mathrm{n}^* v_x^2 > e(V_0 - V) \,, \tag{27.3.5}$$

hence the total current density is

$$j_x = e n_\mathrm{d} \left(\frac{m_\mathrm{n}^*}{2\pi k_\mathrm{B} T} \right)^{3/2} \int\limits_{v_0}^{\infty} v_x \mathrm{e}^{-m_\mathrm{n}^* v_x^2/2k_\mathrm{B}T} \mathrm{d}v_x \iint\limits_{-\infty}^{\infty} \mathrm{e}^{-m_\mathrm{n}^*(v_y^2+v_z^2)/2k_\mathrm{B}T} \, \mathrm{d}v_y \mathrm{d}v_z \,, \tag{27.3.6}$$

where

$$\tfrac{1}{2} m_\mathrm{n}^* v_0^2 = e(V_0 - V) \,. \tag{27.3.7}$$

The integral gives

$$j_x = \tfrac{1}{4} n_\mathrm{d} \left(\frac{8 k_\mathrm{B} T}{\pi m_\mathrm{n}^*} \right)^{1/2} \mathrm{e}^{-eV_0/k_\mathrm{B}T} \mathrm{e}^{eV/k_\mathrm{B}T} \,. \tag{27.3.8}$$

Electrons can pass over the potential barrier in the other direction, too. However, the potential barrier felt by these electrons is not affected by the applied voltage, so the reverse current is independent of V. Since the two currents compensate each other for $V = 0$, the total current is

$$j = j_0 \left(\mathrm{e}^{eV/k_\mathrm{B}T} - 1 \right) . \tag{27.3.9}$$

A slightly different approach has to be taken when the electron mean free path is smaller than the thickness of the depletion layer. Making use of the Einstein relation, the current-density formula

[10] O. W. Richardson, 1901, S. Dushman, 1923. Owen Willans Richardson (1879–1959) was awarded the Nobel Prize in 1928 “for his work on the thermionic phenomenon and especially for the discovery of the law named after him”.

$$\boldsymbol{j}(\boldsymbol{r}) = en(\boldsymbol{r})\mu_{\mathrm{n}}\boldsymbol{E} + eD_{\mathrm{n}}\boldsymbol{\nabla}n(\boldsymbol{r}) = -en(\boldsymbol{r})\mu_{\mathrm{n}}\boldsymbol{\nabla}V(\boldsymbol{r}) + eD_{\mathrm{n}}\boldsymbol{\nabla}n(\boldsymbol{r}) \,, \quad (27.3.10)$$

which contains the drift and diffusion currents, can then be rewritten as

$$\boldsymbol{j}(\boldsymbol{r}) = eD_{\mathrm{n}}\mathrm{e}^{eV(\boldsymbol{r})/k_{\mathrm{B}}T}\boldsymbol{\nabla}\left(n(\boldsymbol{r})\mathrm{e}^{-eV(\boldsymbol{r})/k_{\mathrm{B}}T}\right). \quad (27.3.11)$$

By moving the first exponential factor to the left-hand side, integrating on the semiconductor side from the contact to the other end of the sample (i.e., infinity), and exploiting that the current is the same in all cross sections,

$$j_x \int_0^\infty \mathrm{e}^{-eV(x)/k_{\mathrm{B}}T}\,\mathrm{d}x = eD_{\mathrm{n}}n(x)\mathrm{e}^{-eV(x)/k_{\mathrm{B}}T}\Big|_{x=0}^{x=\infty} \quad (27.3.12)$$

is obtained. Far from the contact $n(x) \approx n_{\mathrm{d}}$, and the potential grows to $V_0 - V$ rather than V_0. The potential is zero at the contact because the net current is small and the density of electrons is practically the same as for $V = 0$ – that is,

$$n(0) = n_{\mathrm{d}}\mathrm{e}^{-eV_0/k_{\mathrm{B}}T}. \quad (27.3.13)$$

Consequently, the current depends on the applied voltage as

$$j = j_0\left(\mathrm{e}^{eV/k_{\mathrm{B}}T} - 1\right). \quad (27.3.14)$$

The current then increases exponentially with the voltage.

When the metallic side is the negative electrode – i.e., the Schottky diode is reverse biased – the electrons can carry the same weak current from the metal to the semiconductor, however, the current in the opposite direction becomes weaker and weaker as the electrons have to surmount an increasingly high potential barrier. Because of this property, Schottky diodes are also called rectifying contacts or blocking contacts. The nonlinear current–voltage characteristic is shown in Fig. 27.16.

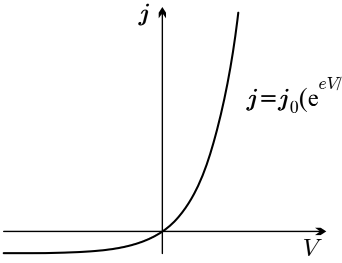


Fig. 27.16. Current–voltage characteristic of a Schottky diode

The voltage dependence of the current through the depletion layer at the contact of a p-type semiconductor and a metal can be studied in the same way. The characteristics are identical.

The situation is completely different when a bias voltage is applied to a metal–semiconductor junction with an accumulation layer. As illustrated in Fig. 27.17, electrons do not have to overcome any potential barrier, so the current–voltage characteristic is linear. Such junctions are therefore called *ohmic contacts.*

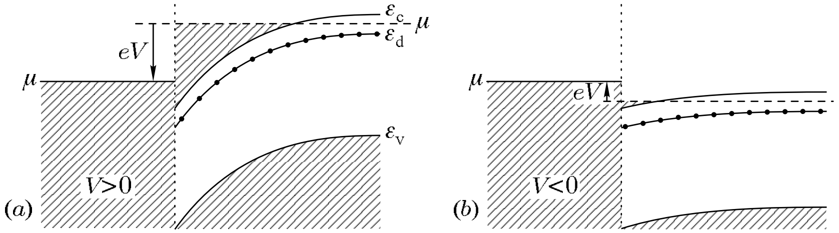


Fig. 27.17. The shift of energy levels at the interface between a metal and an n-type semiconductor with an accumulation layer when the applied voltage (*a*) $V > 0$, (*b*) $V < 0$

27.3.2 Biased MOS Structures

It was mentioned in Section 27.1.2 that, by definition, the location of the energy levels in ideal MOS structures does not depend on the distance from the interface. This is true as long as no voltage is applied across the junction. When a nonzero voltage is applied, a part of it drops across the oxide layer, and another part in a region of the semiconductor close to the interface, which makes the energy levels position dependent. The situation is shown for MOS structures with p- and n-type semiconductors in Fig. 27.18.

When a negative voltage is applied to the metallic side in a p-MOS structure, the energy bands of the semiconductor become bent upward close to the interface. Therefore the majority carriers (holes) accumulate at the interface on the semiconductor side but no current flows. When a positive voltage is applied to the metallic side, the energy bands become bent in the opposite direction, and the holes move away from the vicinity of the interface. The capacity of the system thus depends on the applied voltage. For sufficiently high positive voltages the distortion of the bands can be so large that an inversion layer is formed at the semiconductor–oxide interface. This possibility is used in field-effect transistors (FETs), which will be discussed in the Section 27.4.2.

Similar situations are encountered in MOS structures with an n-type semiconductor.

27.3.3 Current–Voltage Characteristics of p–n Junctions

Before examining the effects of an applied voltage on a p–n junction, we shall briefly return to the discussion of how the equilibrium state is established in a

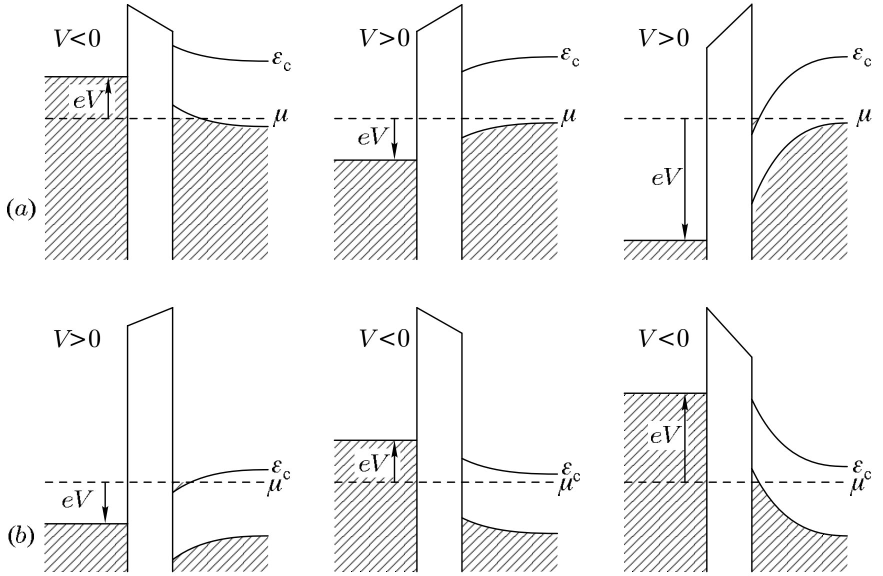


Fig. 27.18. Biased MOS structures with (*a*) a *p*-type, (*b*) an *n*-type semiconductor

p–*n* junction in the absence of any bias. When the Poisson equation was solved for the depletion layer in Section 27.1.3, we treated the system as if it were static. In reality, two oppositely directed currents flow in the transition region: the diffusion current, driven by the gradient of the carrier concentration, and the drift current, driven by the gradient of the potential. The two currents cancel out perfectly. Moreover, this cancellation applies separately for the current of electrons and holes:

$$\begin{aligned} \boldsymbol{j}_{\mathrm{n}}(\boldsymbol{r}) &= en(\boldsymbol{r})\mu_{\mathrm{n}}\boldsymbol{E} + eD_{\mathrm{n}}\boldsymbol{\nabla} n(\boldsymbol{r}) = 0\,, \\ \boldsymbol{j}_{\mathrm{p}}(\boldsymbol{r}) &= ep(\boldsymbol{r})\mu_{\mathrm{p}}\boldsymbol{E} - eD_{\mathrm{p}}\boldsymbol{\nabla} p(\boldsymbol{r}) = 0\,. \end{aligned} \tag{27.3.15}$$

Using the Einstein relation, it is straightforward to show that the solution of these equations leads to the relation (27.1.25) between the carrier density and the potential.

Let us now investigate the effects of an applied voltage V. By convention, V is considered positive and the junction forward biased if the applied voltage increases the electrostatic potential of the *p*-side. When the voltage source is connected the opposite way, the junction is reverse biased. Because of this bias, the energy levels on the *p*-side are shifted by $-eV$ with respect to the *n*-side. Besides, the thermal equilibrium is broken by the applied voltage, and the chemical potential becomes different on the two sides. Denoting its value by μ_{n} and μ_{p} on the corresponding sides, we have

$$\mu_{\mathrm{n}} - \mu_{\mathrm{p}} = eV\,. \tag{27.3.16}$$

In the first approximation we may assume that this potential drop occurs across the depletion layer. Instead of eV_{D}, the relative shift of the energy levels on the two sides is again $eV_{\mathrm{D}} - eV$. The potential difference between the two sides is therefore reduced when $V > 0$, just like the bending of the energy levels close to the interface. This is illustrated in Fig. 27.19(*a*) .

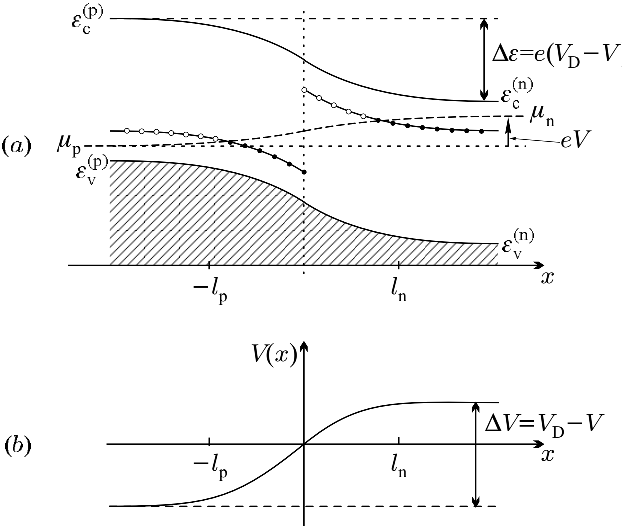

$\varepsilon_{\mathrm{c}}^{(\mathrm{p})}$
$\Delta\varepsilon = e(V_{\mathrm{D}} - V)$
$\varepsilon_{\mathrm{c}}^{(\mathrm{n})}$
μ_{n}
eV
(*a*)
μ_{p}
$\varepsilon_{\mathrm{v}}^{(\mathrm{p})}$
$\varepsilon_{\mathrm{v}}^{(\mathrm{n})}$
x
$-l_{\mathrm{p}}$
l_{n}
$V(x)$
$\Delta V = V_{\mathrm{D}} - V$
(*b*)
x
$-l_{\mathrm{p}}$
l_{n}

Fig. 27.19. (*a*) Energy-level diagram of a forward-biased *p*–*n* junction. (*b*) The local potential that leads to a shift of the energy levels

Figure 27.19(*b*) shows the spatial variation of the diffusion potential, reduced by the applied voltage. The potential drop now occurs over a shorter distance than for $V = 0$ because $V_{\mathrm{D}} - V$ (rather than V_{D}) is used in the formulas (27.1.44) for the width of the depletion layer.

Because of the reduced thickness of the depletion layer and the nonzero net current, SCHOTTKY's assumption – that all spatial variations are limited to the depletion layer of width $l_{\mathrm{p}}^0 + l_{\mathrm{n}}^0$ – is no longer valid. We shall use this observation to determine the current through a biased diode from the current flowing outside the depletion layer. The weak net current in the depletion layer cannot be determined sufficiently precisely as it is the difference of two oppositely directed large currents: the drift and diffusion currents. They are both large because the variations of the potential and concentration are practically limited to this region. Because of the former, the electric field is large, leading to a large drift current; because of the latter, the concentration gradient is large, and thus so is the diffusion current.

If there is a net current through the sample, it injects electrons into the *p*-side and holes into the *n*-side. The diffusion lengths $L_{\mathrm{n,p}}$ are much larger than the thickness of the space-charge region (see page 540). Thus, even though the variation of the charge density occurs dominantly over the depletion layer, the equilibrium distribution of carriers is not established immediately beyond it, as the injected charges can travel much farther than the depletion layer. This means that a new region appears on either side of the depletion layer, a *diffusion region* of width L_{n} on the p-side and another of width L_{p} on the n-side. The diffusion and recombination of the carriers occur primarily in these regions. The subscript indicates that the relevant process is the diffusion of minority carriers.

To determine the spatial variations of the charge density, we shall make use of the smallness of the net current in the depletion layer relative to the drift and diffusion currents. The two large terms on the right-hand side of (27.2.35) thus cancel out to a good approximation:

$$\begin{aligned} 0 &\approx en(\boldsymbol{r})\mu_{\mathrm{n}}\boldsymbol{E} + eD_{\mathrm{n}}\frac{\mathrm{d}n(x)}{\mathrm{d}x}\,, \\ 0 &\approx ep(\boldsymbol{r})\mu_{\mathrm{p}}\boldsymbol{E} - eD_{\mathrm{p}}\frac{\mathrm{d}p(x)}{\mathrm{d}x}\,. \end{aligned} \tag{27.3.17}$$

Using these equations to express the electric field in terms of the applied voltage,

$$\begin{aligned} en(x)\mu_{\mathrm{n}}\frac{\mathrm{d}V(x)}{\mathrm{d}x} &= eD_{\mathrm{n}}\frac{\mathrm{d}n(x)}{\mathrm{d}x}\,, \\ ep(x)\mu_{\mathrm{p}}\frac{\mathrm{d}V(x)}{\mathrm{d}x} &= -eD_{\mathrm{p}}\frac{\mathrm{d}p(x)}{\mathrm{d}x}\,. \end{aligned} \tag{27.3.18}$$

By exploiting the Einstein relation, the formulas given in (27.1.25) are recovered:

$$n(x) \sim \exp\frac{eV(x)}{k_{\mathrm{B}}T}\,, \qquad p(x) \sim \exp\frac{-eV(x)}{k_{\mathrm{B}}T}\,. \tag{27.3.19}$$

Far from the interface, the electron number density at normal temperatures is given by $n_{\mathrm{n}}(\infty) = n_{\mathrm{d}}$ on the n-side and by $n_{\mathrm{p}}(\infty) = n_{\mathrm{p}}$ on the p-side. This drop in the electron density gives rise to a diffusion potential V_{D} between the two sides of the p–n junction. The potential itself varies over a region of width $l = l_{\mathrm{n}}^0 + l_{\mathrm{p}}^0$, as specified by (27.1.45). When an applied voltage V is used, the potential drop across the depletion layer is reduced to $V_{\mathrm{D}} - V$, so the thickness of the layer is changed. The new values l_{n} and l_{p} are now determined by the equations

$$n_{\mathrm{d}}\, l_{\mathrm{n}} = n_{\mathrm{a}}\, l_{\mathrm{p}} \tag{27.3.20}$$

and

$$V_{\mathrm{D}} - V = \frac{e}{2\epsilon}\left(n_{\mathrm{d}}l_{\mathrm{n}}^2 + n_{\mathrm{a}}l_{\mathrm{p}}^2\right), \tag{27.3.21}$$

rather than (27.1.38) and (27.1.43). At the boundaries of the shortened depletion layer the carrier concentration does not drop to the equilibrium values but to

$$n_{\mathrm{p}}(-l_{\mathrm{p}}) = n_{\mathrm{d}} \exp\left(-\frac{e(V_{\mathrm{D}} - V)}{k_{\mathrm{B}}T}\right). \tag{27.3.22}$$

Since according to (27.1.27)

$$\frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} = n_{\mathrm{d}} \mathrm{e}^{-eV_{\mathrm{D}}/k_{\mathrm{B}}T}, \tag{27.3.23}$$

the electron density on the left-hand side of the depletion layer is

$$n_{\mathrm{p}}(-l_{\mathrm{p}}) = \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} \mathrm{e}^{eV/k_{\mathrm{B}}T} = n_{\mathrm{p}}(-\infty) \mathrm{e}^{eV/k_{\mathrm{B}}T}. \tag{27.3.24}$$

The decrease of the electron density continues over the diffusion region until the equilibrium value $n_{\mathrm{i}}^2/n_{\mathrm{a}}$ is reached.

We may say that electrons are injected from the n-side to the p-side over the potential barrier that has been reduced by the positive applied voltage. Moving to the diffusion region, the electrons appear as minority carriers. This increase in the number of minority carriers in that region of the p-side plays an important role in the operation of semiconductor devices.

It is less important but nonetheless noteworthy that, on account of the requirement of local charge neutrality, this increase in the number of the minority carriers gives rise to an increase in the number of majority carriers (holes) in the diffusion region. However, this increase is negligibly small compared to the number of thermally excited holes that appear because of the presence of acceptors.

By the same token, the number of minority carriers (holes) increases in the diffusion region of the n-side, too. At the edge of the depletion layer

$$p_{\mathrm{n}}(l_{\mathrm{n}}) = \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}} \mathrm{e}^{eV/k_{\mathrm{B}}T} = p_{\mathrm{n}}(\infty) \mathrm{e}^{eV/k_{\mathrm{B}}T}\,. \tag{27.3.25}$$

The spatial distribution of electrons and holes are shown in Fig. 27.20.

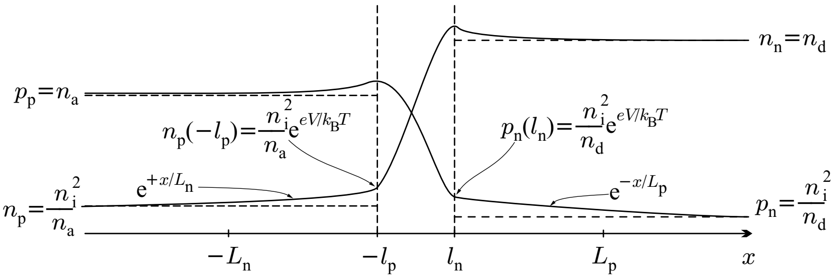

$n_n = n_d$
$p_p = n_a$
$n_p(-l_p) = \frac{n_i^2}{n_a} e^{eV/k_BT}$
$p_n(l_n) = \frac{n_i^2}{n_d} e^{eV/k_BT}$
e^{+x/L_n}
e^{-x/L_p}
$n_p = \frac{n_i^2}{n_a}$
$p_n = \frac{n_i^2}{n_d}$
$-L_n$
$-l_p$
l_n
L_p
x

Fig. 27.20. Electron and hole densities at both sides of a forward-biased p–n junction

The total current due to the applied voltage is the same across any cross section of the sample, however the relative contribution of electrons and holes varies. This is illustrated in Fig. 27.21.

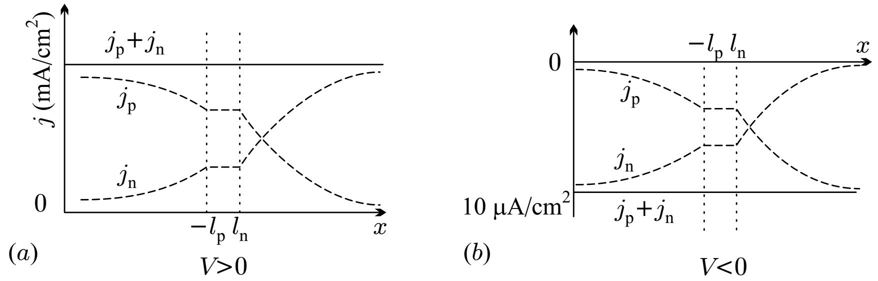


Fig. 27.21. The total current density (j) and the current densities carried by electrons (j_n) and holes (j_p) along the p–n junction, for (a) $V > 0$ and (b) $V < 0$

Outside the diffusion region, where the carrier concentration is practically constant, only a weak drift current flows. On either side of the interface, the current of electrons or holes shows a strong position dependence in the diffusion region. The electric field $\boldsymbol{E}$ is small, and therefore so is the drift current – however, the diffusion current of minority carriers is important. The condition for the neglect of the drift current can be established by the following simple consideration. During the recombination time $\tau_{n,p}$ carriers in an electric field $\boldsymbol{E}$ travel a distance $\mu_{n,p} E \tau_{n,p}$. If this characteristic length is smaller than the diffusion length $L_{n,p}$, the drift current is negligible compared to the diffusion current. The condition for this is

$$\mu_{n,p} E \tau_{n,p} \ll \sqrt{D_{n,p}\tau_{n,p}} \,. \tag{27.3.26}$$

By making use of the Einstein relation,

$$E \ll \frac{k_B T}{e} \frac{1}{\sqrt{D_{n,p}\tau_{n,p}}} = \frac{k_B T}{e L_{n,p}} \,. \tag{27.3.27}$$

Keeping in mind that the largest part of the applied voltage drops over the space-charge region, this condition is met in the diffusion region, provided the bias is not too large.

This is not the case inside the depletion layer. As we have seen, because of the large potential and concentration drops, the diffusion and drift currents are equally large but oppositely directed. The net currents of electrons and holes are therefore small, of order mA/cm^2 for customary voltages. As the depletion layer is thin compared to the diffusion length (i.e., the mean distance traveled before recombination), it is justified to assume that the generation and recombination of carriers are negligible in this region. The electron and hole currents pass through the depletion layer without attenuation, and so they are the same at l_n and $-l_p$:

$$j_n(l_n) = j_n(-l_p) \,, \qquad j_p(l_n) = j_p(-l_p) \,. \tag{27.3.28}$$

The total current density

$$j = j_{\mathrm{n}} + j_{\mathrm{p}} \tag{27.3.29}$$

can be determined from the values of the two components at the boundary of the depletion layer – at the p-side for electrons and at the n-side for holes –, that is,

$$j_{\mathrm{n}} = j_{\mathrm{n}}(-l_{\mathrm{p}})\,, \qquad j_{\mathrm{p}} = j_{\mathrm{p}}(l_{\mathrm{n}})\,. \tag{27.3.30}$$

According to our previous considerations, the electron current is almost completely diffusive at the boundary of the transition region on the p-side. By neglecting the drift term, we have

$$j_{\mathrm{n}}(-l_{\mathrm{p}}) = eD_{\mathrm{n}} \left.\frac{\mathrm{d}n_{\mathrm{p}}(x)}{\mathrm{d}x}\right|_{x=-l_{\mathrm{p}}} . \tag{27.3.31}$$

Similarly, at the boundary of the transition region on the n-side, the hole drift current can be neglected:

$$j_{\mathrm{p}}(l_{\mathrm{n}}) = -eD_{\mathrm{p}} \left.\frac{\mathrm{d}p_{\mathrm{n}}(x)}{\mathrm{d}x}\right|_{x=l_{\mathrm{n}}} . \tag{27.3.32}$$

The equation for the number density of holes on the n-side, (27.2.31), and the analogous equation for the number density of electrons on the p-side have to be solved subject to the boundary condition that the number densities of minority carriers in the homogeneous region satisfy

$$n_{\mathrm{p}}(-\infty) = \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}}\,, \qquad \text{and} \qquad p_{\mathrm{n}}(\infty) = \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}}\,. \tag{27.3.33}$$

The solutions are

$$\begin{aligned} n_{\mathrm{p}}(x) &= \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} + \left[n_{\mathrm{p}}(-l_{\mathrm{p}}) - \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} \right] \mathrm{e}^{(x+l_{\mathrm{p}})/L_{\mathrm{n}}}\,, \qquad x \le -l_{\mathrm{p}}\,, \\ p_{\mathrm{n}}(x) &= \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}} + \left[p_{\mathrm{n}}(l_{\mathrm{n}}) - \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}} \right] \mathrm{e}^{-(x-l_{\mathrm{n}})/L_{\mathrm{p}}}\,, \qquad x \ge l_{\mathrm{n}}\,. \end{aligned} \tag{27.3.34}$$

The derivatives in (27.3.31) and (27.3.32) are then

$$\begin{aligned} j_{\mathrm{n}}(-l_{\mathrm{p}}) &= e\frac{D_{\mathrm{n}}}{L_{\mathrm{n}}} \left[n_{\mathrm{p}}(-l_{\mathrm{p}}) - \frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}} \right], \\ j_{\mathrm{p}}(l_{\mathrm{n}}) &= -e\frac{D_{\mathrm{p}}}{L_{\mathrm{p}}} \left[p_{\mathrm{n}}(l_{\mathrm{n}}) - \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}} \right]. \end{aligned} \tag{27.3.35}$$

Taking the carrier densities from (27.3.24) and (27.3.25),

$$\begin{aligned} j_{\mathrm{n}}(-l_{\mathrm{p}}) &= e\frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}}\frac{D_{\mathrm{n}}}{L_{\mathrm{n}}} \left[\mathrm{e}^{eV/k_{\mathrm{B}}T} - 1 \right], \\ j_{\mathrm{p}}(l_{\mathrm{n}}) &= -e\frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}}\frac{D_{\mathrm{p}}}{L_{\mathrm{p}}} \left[\mathrm{e}^{eV/k_{\mathrm{B}}T} - 1 \right]. \end{aligned} \tag{27.3.36}$$

Using (27.3.29) and (27.3.30), the total current density through the p–n junction is

$$j = en_{\mathrm{i}}^2 \left(\frac{D_{\mathrm{n}}}{L_{\mathrm{n}} n_{\mathrm{a}}} + \frac{D_{\mathrm{p}}}{L_{\mathrm{p}} n_{\mathrm{d}}} \right) \left(\mathrm{e}^{eV/k_{\mathrm{B}}T} - 1 \right) . \tag{27.3.37}$$

The effects of a reverse bias ($V < 0$) can be analyzed in much the same way. The spatial variations of the energy levels and the potential between the two sides are shown in Fig. 27.22. The potential difference is increased, and so is the width of the depletion layer.

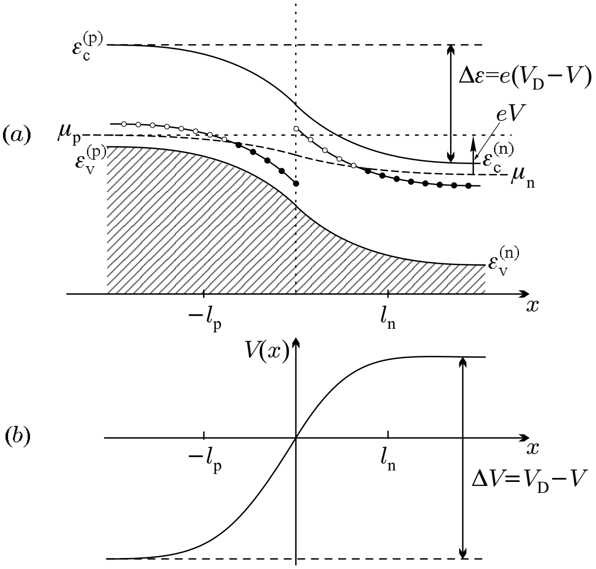

(a)
$\varepsilon_{\mathrm{c}}^{(\mathrm{p})}$
$\Delta\varepsilon = e(V_{\mathrm{D}} - V)$
eV
μ_{p}
$\varepsilon_{\mathrm{v}}^{(\mathrm{p})}$
$\varepsilon_{\mathrm{c}}^{(\mathrm{n})}$
μ_{n}
$\varepsilon_{\mathrm{v}}^{(\mathrm{n})}$
$-l_{\mathrm{p}}$
l_{n}
x
(b)
$V(x)$
$-l_{\mathrm{p}}$
l_{n}
x
$\Delta V = V_{\mathrm{D}} - V$

Fig. 27.22. (*a*) Energy-level diagram of a reverse-biased p–n junction ($V < 0$). (*b*) The local potential that is responsible for the shift of the energy levels

Since the depletion layer has become wider, the concentration of minority carriers at its boundary is smaller than the equilibrium ($V = 0$) value, therefore a diffusion current appears in the diffusion region here, too. The variations of the electron and hole currents are shown in Fig. 27.21(*b*). To obtain a better picture of the physical processes inside the junction, the spatial variations of the electron and hole densities are also plotted for this situation in Fig. 27.23.

A similar calculation shows that the formula (27.3.37) for the total current density is valid for $V < 0$, too, that is,

$$j(V) = j_0 \left[\mathrm{e}^{eV/k_{\mathrm{B}}T} - 1 \right] . \tag{27.3.38}$$

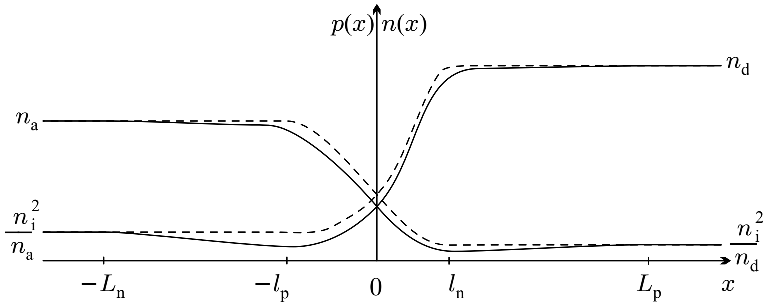


Fig. 27.23. Spatial variations of the electron and hole densities at both sides of a reverse-biased *p*–*n* junction

This is *Shockley's law.*[11] The strongly nonlinear current–voltage characteristic has the same features as the characteristic of a Schottky diode, shown in Fig. 27.16. The *p*–*n* junction behaves as a rectifier, too. To illustrate the actual current and voltage values, it should be remembered that a forward bias of a few tenths of a volt typically gives a current of order 10 mA through the *p*–*n* junction. In contrast, for a reverse bias the current is a couple of µA, and almost independent of the bias voltage. The current through germanium diodes is indeed correctly given by (27.3.38). On the other hand, the contribution of carriers generated in the space-charge region cannot be neglected in silicon, where $n_{\rm i}$ is much smaller. The previous calculation has to be refined then.

An intuitive picture can be formed about the total current by decomposing it into four parts:

$$\begin{aligned} j_{\rm n}({\rm n} \to {\rm p}) &= e\frac{n_{\rm i}^2}{n_{\rm a}}\frac{D_{\rm n}}{L_{\rm n}}{\rm e}^{eV/k_{\rm B}T}\,, \\ j_{\rm n}({\rm p} \to {\rm n}) &= -e\frac{n_{\rm i}^2}{n_{\rm a}}\frac{D_{\rm n}}{L_{\rm n}}\,, \\ j_{\rm p}({\rm p} \to {\rm n}) &= e\frac{n_{\rm i}^2}{n_{\rm d}}\frac{D_{\rm p}}{L_{\rm p}}{\rm e}^{eV/k_{\rm B}T}\,, \\ j_{\rm p}({\rm n} \to {\rm p}) &= -e\frac{n_{\rm i}^2}{n_{\rm d}}\frac{D_{\rm p}}{L_{\rm p}}\,. \end{aligned} \qquad (27.3.39)$$

The first term is the current carried from the *n*-side to the *p*-side by the majority carriers, electrons. The exponential voltage dependence is logical: the applied voltage reduces the height of the potential barrier that the conduction-band electrons on the *n*-side have to overcome in order to get through to the *p*-side. This part of the current is also called the *electron recombination current*: having passed through the potential barrier, the electrons recombine with the holes (majority carriers) of the *p*-side.

[11] W. B. Shockley, 1949.

The second term is the countercurrent of thermally excited electrons from the p-side. This is independent of the bias voltage, since the electrons (minority carriers) slide down the potential barrier to get to the n-side. This component is also called the *generation current.* To estimate its magnitude, it should be noted that those minority carriers are likely to get to the n-side that are created close to the interface, within the diffusion length L_{n}. Since the electron generation rate on the p-side is n_0/τ_{n}, the current density of the electrons generated within a distance L_{n} of the interface is

$$j_{\mathrm{n}}(\mathrm{p} \to \mathrm{n}) = -eL_{\mathrm{n}}n_0/\tau_{\mathrm{n}}\,. \tag{27.3.40}$$

Exploiting the relation $n_0 = n_{\mathrm{i}}^2/p_0$, and that $p_0 = n_{\mathrm{a}}$ on the p-side,

$$j_{\mathrm{n}}(\mathrm{p} \to \mathrm{n}) = -e\frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}}\frac{L_{\mathrm{n}}}{\tau_{\mathrm{n}}}\,. \tag{27.3.41}$$

Eliminating the carrier lifetime in favor of the diffusion coefficient, and making use of (27.2.28),

$$j_{\mathrm{n}}(\mathrm{p} \to \mathrm{n}) = -e\frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}}\frac{D_{\mathrm{n}}}{L_{\mathrm{n}}}\,. \tag{27.3.42}$$

In perfect analogy, the third and fourth equations in (27.3.39) describe the hole recombination and generation currents. For $V > 0$, the holes on the p-side have to overcome a reduced potential barrier, just as was discussed for electrons above, and there is also a countercurrent of minority carriers (holes) generated on the n-side:

$$j_{\mathrm{p}}(\mathrm{n} \to \mathrm{p}) = -e\frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}}\frac{L_{\mathrm{p}}}{\tau_{\mathrm{p}}} = -e\frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}}\frac{D_{\mathrm{p}}}{L_{\mathrm{p}}}\,. \tag{27.3.43}$$

The negative sign appears because the current flows from the $x > 0$ side to the $x < 0$ side.

27.3.4 Zener and Avalanche Breakdown in *p*–*n* Junctions

At moderately large negative voltages carriers can no longer overcome the potential barrier, and the saturation value of the current is given by

$$j_0 = -e\left[\frac{n_{\mathrm{i}}^2}{n_{\mathrm{a}}}\frac{D_{\mathrm{n}}}{L_{\mathrm{n}}} + \frac{n_{\mathrm{i}}^2}{n_{\mathrm{d}}}\frac{D_{\mathrm{p}}}{L_{\mathrm{p}}}\right]. \tag{27.3.44}$$

When larger negative voltages are applied to the p–n junction, the current is observed to increase sharply at a threshold value that is characteristic of the diode, and can range from a few volts to a thousand volts. The current can reach very high values without an increase in the voltage. This phenomenon is called the breakdown of the diode, and the threshold value of the voltage is the breakdown voltage. The corresponding current–voltage characteristic is shown in Fig. 27.24.

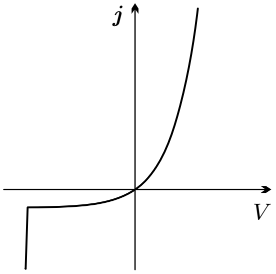


Fig. 27.24. The current–voltage characteristic of a Zener diode

The breakdown can occur for several reasons. For example, at sufficiently large voltages the heat generated by the reverse carriers can give rise to thermal instability in the diode. The device becomes hot because of the produced Joule heat, and therefore the number of minority carriers is increased. This leads to an increased current, which causes a further rise in the temperature, and so on. The acceleration of this mechanism can result in enormous currents without an increase in the voltage. However, other mechanisms proved to be more important in practical applications.

The thickness of the depletion layer depends on the doping of both sides. In highly doped diodes carriers can also tunnel through the thin depletion layer. This tunnel current starts to flow when the reverse bias is large enough (a few V) to bring the bottom of the conduction band on the n-side to the same level as the valence band on the p-side (Fig. 27.25), since the electrons in the valence band can then tunnel through the potential barrier, to an empty state of the conduction band. This sharp increase in the current is the *Zener breakdown* or *Zener effect.*[12]

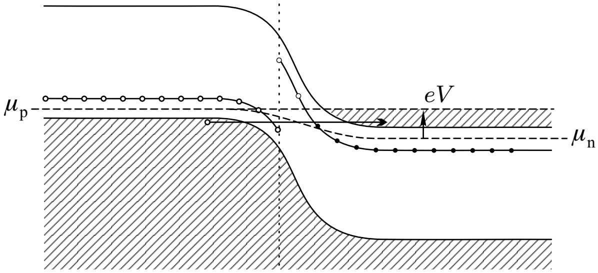


Fig. 27.25. The location of the energy levels in a Zener diode for large negative applied voltages

Those p–n junctions that are fabricated expressly in such a way that their current–voltage characteristics should feature a breakdown at some large neg-

[12] C. Zener, 1934.

ative voltage in order to achieve voltage control are called Zener diodes. However, breakdown in Zener diodes is often not caused by tunneling but by avalanches brought about by collisional ionization. Since the potential drop occurs dominantly across the depletion layer, the electric field can reach very high values, even $10^6\,\mathrm{V/cm}$, in the very narrow depletion layers of heavily doped *p*–*n* junctions. Such field strengths may require kilovolt bias voltages in less heavily dopes devices. Electrons accelerated by such fields can excite further, covalently bound electrons into the conduction band via collisional ionization. As holes are left behind in the valence band, this mechanism generates electron–hole pairs. Accelerated by the electric field, these new carriers generate further electron–hole pairs, and so on. This *avalanche effect* leads to *avalanche breakdown.*

27.3.5 Tunnel Diodes

In Zener diodes the electrons in the valence band of the *p*-side tunnel to the conduction band on the *n*-side when a sufficiently large negative voltage is applied. A new situation is encountered in *p*–*n* junctions made up of two heavily doped semiconductors. Such devices are called *tunnel* or *Esaki diodes.*[13]

In such configurations the number of holes on the *p*-side and the number of electrons on the *n*-side become so large that the chemical potential moves from the gap into the valence and conduction bands, respectively. Such semiconductors were termed degenerate semiconductors in Chapter 20. The location of bands in the equilibrium state ($V = 0$) is shown in Fig. 27.26.

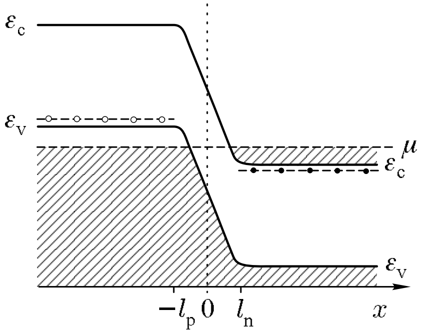


Fig. 27.26. The location of the energy levels in the equilibrium state in heavily doped *p*–*n* junctions

Because of the heavy doping, the depletion layer is narrower than in usual *p*–*n* junctions, therefore the current comes dominantly from quantum mechanical tunneling rather than the drift driven by the built-in potential or

[13] L. Esaki, 1958. Leo Esaki (1925–) was awarded the Nobel Prize in 1973 for his "experimental discoveries regarding tunneling phenomena in semiconductors".

diffusion. However, without bias, the oppositely directed currents cancel out. When a bias is applied, the bands become shifted, as illustrated in Fig. 27.27.

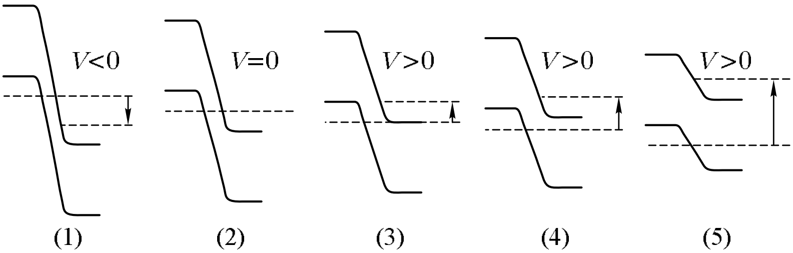


Fig. 27.27. The location of the energy levels in heavily doped p–n junctions when the bias is zero (2), negative (1), positive with increasing values (3–5)

When a negative voltage is applied, the chemical potential is higher on the p-side than on the n-side. Since the electrons at the top of the valence band on the p-side are facing empty states on the n-side, more electrons can tunnel from the p-side to the n-side than in the opposite direction. The net current is then proportional to the voltage. The situation is similar for low positive voltages as long as the shift of the bands is sufficiently small for that the electrons on the n-side whose energy is equal to the chemical potential can tunnel to empty states in the valence band on the p-side. The reverse process can also take place, but globally more electrons tunnel from the n-side to the p-side than the other way around. At a higher positive voltage the chemical potential on the n-side reaches the height of the top of the valence band on the p-side. This is shown in the middle figure. For even higher voltages the current decreases as fewer and fewer electrons find a hole into which they could tunnel. This corresponds to the region of negative differential resistance in the current–voltage characteristic in Fig. 27.28.

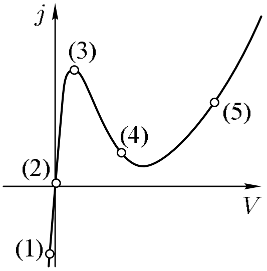


Fig. 27.28. Current–voltage characteristic of an Esaki diode. Numbers correspond to the voltage regimes marked in the previous figure

At even higher voltages the tunnel current can completely vanish. At the same time, the diffusion and drift currents become larger, since the relative position of the bands is the same now as in a usual p–n junction, leading to an exponential voltage dependence of the current, in line with Shockley's law.

27.4 Simple Semiconductor Devices

As mentioned in the introduction, the increasingly high number of semiconductor devices can be put down to the enormous development of planar technology and photolithographic techniques, allowing manufacturers to build dedicated configurations of metal–semiconductor, metal–insulator–semiconductor, or semiconductor–semiconductor junctions with well-defined properties. The operation of such devices is nonetheless governed by the simple physical principles outlined above. We shall present a few simple examples below.

27.4.1 Bipolar Transistors

In bipolar transistors[14] two semiconductor junctions of opposite polarity are juxtaposed. Depending on the order of the p- and n-type layers the transistor is either p–n–p or n–p–n-type. Below we shall briefly discuss the basic notions of the operation of p–n–p transistors. The same considerations apply to n–p–n translators as well, only the current and voltage directions are reversed.

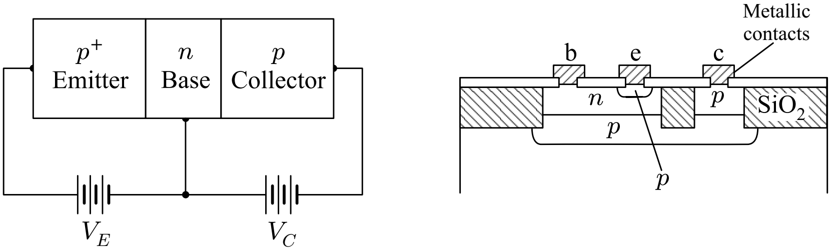


Fig. 27.29. Schematic structure of bipolar p–n–p transistors, and the sectional view of a bipolar transistor fabricated using the planar technology

As illustrated in the schematic diagram, the heavily doped p-region on the left-hand side (emitter) and the moderately doped p-region on the right-hand side (collector) are separated by a thin, lightly doped n-type layer (base), whose width is smaller than the diffusion length. In transistors fabricated using the planar technology the spatial arrangement of the n- and p-type

[14] The name "transistor" was coined by J. R. Pierce from the word "transresistance" to fit in with the names of other devices, such as varistor and thermistor.

regions is different but the principles of operation are the same. To obtain a simple picture about the operation of transistors, consider the location of the energy levels when a nonzero voltage is applied. This is shown in Fig. 27.30.

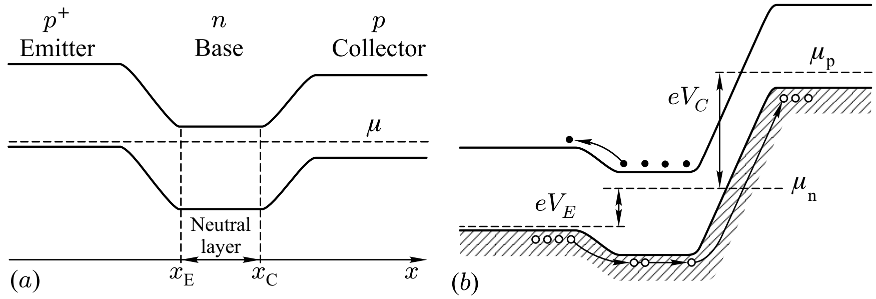


Fig. 27.30. Energy-level diagram of a bipolar transistor in the (a) absence (b) presence of a bias voltage

Consider the p–n junction on the left. Since $n_\mathrm{a} \gg n_\mathrm{d}$, we have $p_\mathrm{p} \gg n_\mathrm{n}$ for the majority carriers. The law of mass action then implies $n_\mathrm{p} \ll p_\mathrm{n}$ for the minority carriers. Applying a voltage $V_\mathrm{E} > 0$ across the emitter–base junction, the current is dominantly carried by holes injected from the emitter into the base. If the formula derived for a p–n junction is used to determine the current in a first approximation,

$$j_\mathrm{E} = \frac{e p_\mathrm{p} D_\mathrm{p}}{L_\mathrm{p}} \left[\mathrm{e}^{eV_\mathrm{E}/k_\mathrm{B}T} - 1 \right] . \tag{27.4.1}$$

To obtain a more accurate result, the finite width of the base has to be taken into account. The hole concentrations at the boundaries of the depletion layers next to the emitter and collector sides are then given by

$$p(x_\mathrm{E}) = p_\mathrm{E} \mathrm{e}^{eV_\mathrm{E}/k_\mathrm{B}T} , \qquad p(x_\mathrm{C}) = p_\mathrm{C} \mathrm{e}^{eV_\mathrm{C}/k_\mathrm{B}T} , \tag{27.4.2}$$

where p_E and p_C are the equilibrium hole concentrations in the emitter and collector. If these boundary conditions are imposed, and the equilibrium concentration is taken as p_B, the solution of the equations inside the base yields

$$\begin{aligned} p(x) = p_\mathrm{B} &+ (p_\mathrm{E} - p_\mathrm{B}) \frac{\sinh[(x_\mathrm{C} - x)/L_\mathrm{B}]}{\sinh[(x_\mathrm{C} - x_\mathrm{B})/L_\mathrm{B}]} \\ &+ (p_\mathrm{C} - p_\mathrm{B}) \frac{\sinh[(x - x_\mathrm{E})/L_\mathrm{B}]}{\sinh[(x_\mathrm{C} - x_\mathrm{B})/L_\mathrm{B}]} , \end{aligned} \tag{27.4.3}$$

where L_B is the hole diffusion length in the base. Neglecting the width of the depletion layer compared to the width $W = x_\mathrm{C} - x_\mathrm{E}$ of the neutral region of the base, the concentration of the electrons (minority carriers) on the two p-sides is given by

$$\begin{aligned} n(x) &= n_\mathrm{E} + n_\mathrm{E}\left[\mathrm{e}^{eV_\mathrm{E}/k_\mathrm{B}T} - 1\right]\mathrm{e}^{(x-x_\mathrm{E})/L_\mathrm{E}}\,, && x < x_\mathrm{E}\,,\\ n(x) &= n_\mathrm{C} + n_\mathrm{C}\left[\mathrm{e}^{eV_\mathrm{C}/k_\mathrm{B}T} - 1\right]\mathrm{e}^{-(x-x_\mathrm{C})/L_\mathrm{C}}\,, && x > x_\mathrm{C}\,. \end{aligned} \tag{27.4.4}$$

The total current density between the emitter and base is then

$$\begin{aligned} j_\mathrm{E} = {}& \frac{ep_\mathrm{B}D_\mathrm{B}}{L_\mathrm{B}}\frac{1}{\sinh(W/L_\mathrm{B})}\\ & \times\left\{\cosh(W/L_\mathrm{B})\left[\mathrm{e}^{eV_\mathrm{E}/k_\mathrm{B}T} - 1\right] - \left[\mathrm{e}^{eV_\mathrm{C}/k_\mathrm{B}T} - 1\right]\right\}\\ & + \frac{en_\mathrm{E}D_\mathrm{E}}{L_\mathrm{E}}\left[\mathrm{e}^{eV_\mathrm{E}/k_\mathrm{B}T} - 1\right]. \end{aligned} \tag{27.4.5}$$

If the base is sufficiently thin, most holes pass through it without recombination and reach the collector. When a negative voltage is applied between the base and collector, the motion of holes gives a collector current j_C. Its magnitude can be determined in the same way:

$$\begin{aligned} j_\mathrm{C} = {}& \frac{ep_\mathrm{B}D_\mathrm{B}}{L_\mathrm{B}}\frac{1}{\sinh(W/L_\mathrm{B})}\\ & \times\left\{\left[\mathrm{e}^{eV_\mathrm{E}/k_\mathrm{B}T} - 1\right] - \coth(W/L_\mathrm{B})\left[\mathrm{e}^{eV_\mathrm{C}/k_\mathrm{B}T} - 1\right]\right\}\\ & - \frac{en_\mathrm{C}D_\mathrm{C}}{L_\mathrm{C}}\left[\mathrm{e}^{eV_\mathrm{C}/k_\mathrm{B}T} - 1\right]. \end{aligned} \tag{27.4.6}$$

Because of the loss in the base, the base current

$$j_\mathrm{B} = j_\mathrm{E} - j_\mathrm{C} \tag{27.4.7}$$

is much smaller than the collector current. When the base current is increased, the neutrality of the base can be preserved only if the collector current also increases. The ratio $j_\mathrm{C}/j_\mathrm{B}$ is the *current gain* of the transistor, which can be much larger than unity if the width of the base is indeed smaller than the hole diffusion length.

27.4.2 Field-Effect Transistors

The principle of the field effect – namely, the possibility of modulating the conductivity of solids by a transverse electric field, and thereby controlling the electric current – was proposed as early as 1925 by J. E. Lilienfeld. The theory of the field effect was put forward by W. Shockley in 1952, but the first experimental confirmation came only in 1960 with the fabrication of the first field-effect transistor (FET). Its simplest version is the junction field-effect transistor (JFET) shown in Fig. 27.31. Between two heavily doped p-type gates there is an n-type layer with metallic contacts at each end that serve as current source and drain. By applying negative voltage to the gates, the thickness of the depletion layers increases in the p–n junctions, reducing

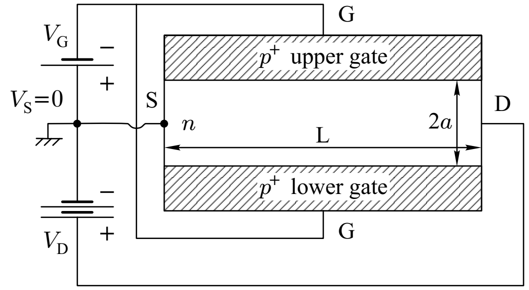


Fig. 27.31. The schematic structure of field-effect transistors

the cross section of the n-type channel through which the current can flow from the source to the drain. The resistance of field-effect transistors can thus be controlled by the gate voltage.

In MOSFETs (metal–oxide–semiconductor field-effect transistors) the resistance is controlled electrostatically via the modification of the charge distribution at the oxide–semiconductor interfaces. Their schematic structure is shown in Fig. 27.32.

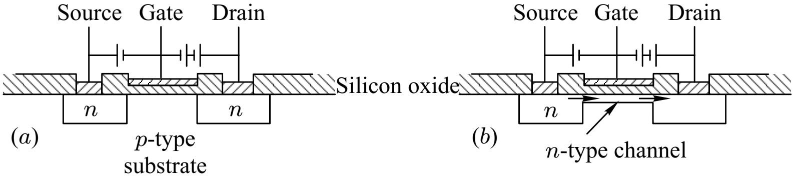


Fig. 27.32. (a) The schematic structure of MOSFETs. (b) The formation of a conducting channel at a large gate voltage

The active region of the MOSFET is made up of a semiconductor substrate (body) – in our example, p-type – with an insulating layer of silicon oxide grown on it, and a metal contact evaporated on top. Depending on the voltage applied between the metal contact (gate) and the body electrode, an accumulation or depletion of the majority carriers can occur on the semiconductor side, or even minority carriers can form an inversion layer in this metal–insulator–semiconductor junction. On either side of the gate, a small hole is etched into the silicon oxide layer using photolithographic techniques. By introducing large quantities of dopant through the holes, island-like regions of the opposite (n) type semiconductor are created. The metal contacts of these heavily doped islands serve as source and drain.

As long as the voltage applied to the control electrode (gate) is small, no current flows between the n-type source and drain through the p-type substrate. However, when the gate voltage reaches a sufficiently large positive threshold value, an inversion layer is formed at the interface between the oxide film and the p-type substrate, and an n-type channel opens through which current can flow from the source to the drain.

The thickness of the inversion layer increases with increasing gate voltage. The current does not grow linearly with the drain–source voltage but more slowly because, on account of the potential drop in the channel, the voltage between the channel and the gate depends on the position along the channel. Consequently, the width of the channel and the carrier density in the channel decrease toward the drain.

When the difference between the gate-to-source voltage (input control voltage) and the drain-to-source voltage (output voltage) reaches the threshold value, the inversion layer at the drain disappears. The current cannot be increased further by applying even higher drain voltages. The resulting current–voltage characteristics are shown in Fig. 27.33.

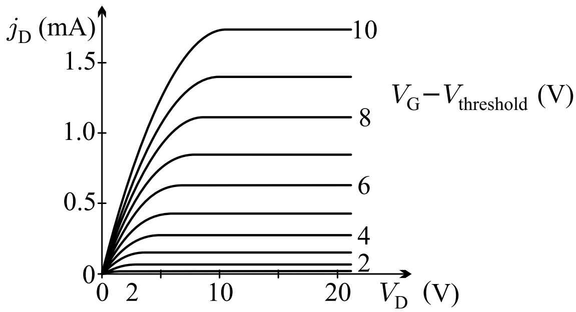


Fig. 27.33. The current–voltage characteristics of MOSFETs: the drain current j_D as a function of the drain-to-source voltage V_D for different gate-to-source voltages V_G

27.4.3 Semiconductor Lasers and Solar Cells

The operation of lasers[15] is known to be based on the principle of population inversion: under external excitation, the occupation of the higher-lying energy level exceeds that of the lower level, and through an avalanche-like process of stimulated emission a coherent light beam emerges. If the intensity of this beam is larger than the intensity loss due to absorption, a high-intensity coherent beam is obtained. Population inversion can also occur in highly doped p–n junctions, so such semiconductor devices can be used as laser sources.

As demonstrated in connection with the tunnel diode, in p–n junctions made up of two heavily doped (degenerate) semiconductors the chemical potential is located in the valence band on the p-side and the conduction band on the n-side. Applying a sufficiently large positive voltage across this junction, such that the condition

[15] Laser is the acronym for Light Amplification by Stimulated Emission of Radiation.

$$eV \geq \varepsilon_{\mathrm{g}} \tag{27.4.8}$$

is met, a narrow layer appears next to the interface, in which there are occupied electron states in the conduction band and empty states in the valence band. This situation is illustrated in Fig. 27.34.

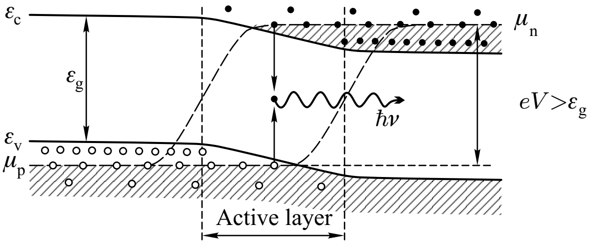


Fig. 27.34. Energy-band diagram of a semiconductor laser

The population inversion in the transition layer is the consequence of the injection of electrons from the n-side to the p-side and holes in the opposite direction by the current through the diode. As an electron–hole pair is annihilated, the diode emits light. The emitted light is in the infrared ($\lambda \approx 1200\,\mathrm{nm}$) for a gap of 1 eV, whereas it is visible ($\lambda \approx 500\,\mathrm{nm}$) for a gap of 2.5 eV.

Below a current threshold, absorption is more probable than stimulated emission because of the losses due to diffusion and recombination, and so the diode emits incoherent light. Above the threshold, the p–n junction operates as a laser, and emits a coherent light beam in a narrow frequency range. However, the efficiency of this coherent light emission by the p–n junction is quite poor. The losses can be reduced, and laser operation can be obtained at a lower threshold current in the double heterojunction configuration shown in Fig. 27.35, where an epitaxially grown GaAs layer is inserted between the

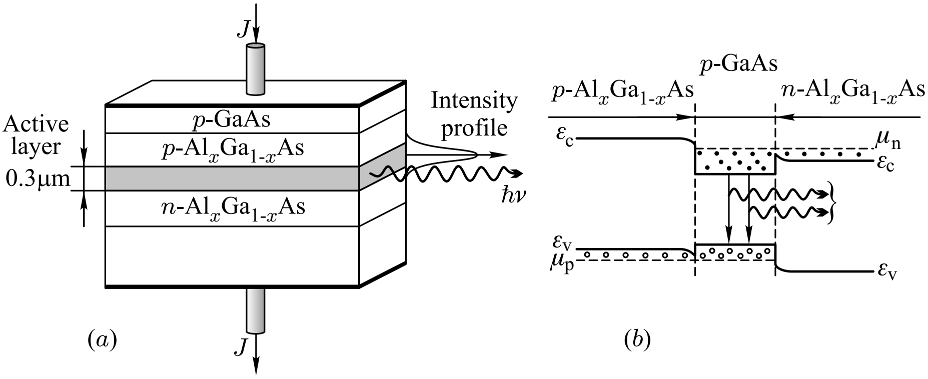


Fig. 27.35. (a) Schematic structure and (b) energy-band diagram of a double heterojunction (DH) laser

p- and n-type $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ layers. That is why such devices are called DH (double heterojunction) lasers. Since the gap is smaller in the middle layer, this is where population inversion occurs, and so this is the active region of the laser.

As electric current passes through a LED, light is emitted. The inverse process is also possible in semiconductor junctions: incident light can generate carriers, and thus induce a current or voltage. This is the operation principle of solar cells.

27.5 Semiconductor Quantum Devices

In the previous section we were primarily concerned with systems in which the metal or semiconductor on either side of the interface was much thicker than the transition region, therefore the former could be considered infinite in size. Moreover, since the conditions of the semiclassical approximation were satisfied, it was possible to calculate the current induced in the system by an applied voltage using classical equations. The development of materials technology, and miniaturization in particular, has led to the fabrication of ever smaller semiconductor devices. The linear size of the components on chips is typically one micron. This is the realm of microelectronics. In the past decades even tinier structures were produced, in which the linear size of the sample in one, two, or even three spatial dimensions is on the order of the de Broglie wavelength of the current-carrying electrons.[16] For electrons confined to such small regions quantum effects and the quantized nature of the energy spectrum become important. Therefore those structures in which the free propagation of electrons is confined in one, two, and all three directions are called *quantum wells*, *quantum wires*, and *quantum dots*. Since the linear sizes of the confinement region are smaller than a micron, typically a few hundred nanometers (and the relevant wavelength in semiconductors can be almost comparable, over 10 nm), the body of physical phenomena observed in such systems is usually referred to as *nanophysics*. Since these dimensions are halfway between the typical length scales of bulk (macroscopic) and atomic (microscopic) samples, the terms *mesoscopic systems* and *mesoscopic physics* are also commonly used.[17]

Two more length scales are important in such systems. The classical theory of electron transport, based on the Boltzmann equation, assumes that electrons collide frequently, and the mean free path l is much smaller than the size of the sample, thus the motion of electrons is diffusive. It is then possible to define a local electrochemical potential that varies continuously in the sample.

[16] Since the energy of the participating electrons is the Fermi energy, the relevant size is the Fermi wavelength.

[17] The Greek words $\mu\varepsilon\sigma o\varsigma$ (mesos) and $\nu\alpha\nu o\varsigma$ (nanos) mean middle and dwarf, respectively.

Another essential assumption is that the phase coherence of the wave packet is lost in collisions. This loss occurs over a length l_ϕ, called the phase-coherence length, which is characteristic of the material and depends on temperature. The mean free path l and the phase-coherence length l_ϕ are two different quantities. The former is determined by inelastic electron–electron collisions, electron–phonon interactions that involve phonon absorption or emission, and spin-flip magnetic impurity scattering processes, while the latter dominantly comes from elastic scattering by the static potential of impurities. At low temperatures $l_\phi \gg l$. Classical transport theory can be used as long as the characteristic linear dimension L of the sample is much larger than l_ϕ and l. If $l \ll L < l_\phi$, the motion of the electron is still diffusive, however interference effects can no longer be neglected. We shall investigate this in Chapter 36 of Volume 3. If, however, the linear dimension L is comparable to or smaller than the mean free path l, the electron propagates ballistically. Then local thermodynamic equilibrium exists only at the ohmic contacts, and the concept of the electrochemical potential is meaningful only there. Below we shall give a very sketchy overview of the physics of such systems, focusing on the energy spectrum and some simple cases of carrier transport. The detailed discussion of this very rapidly evolving field is far beyond the scope of this book. This striking evolution is motivated only in part by the desire of understanding the new physical phenomena. Even more important is the driving force of technology: if the present miniaturization trend in the semiconductor industry (as expressed by Moore's law) continues, such nanostructure devices are expected to be widely used in one or two decades.

27.5.1 Electron Spectrum of Quantum Wells

Using epitaxial growth techniques, it is possible to produce structures in which a thin layer of semiconductor with a narrow gap is inserted between two identical semiconductors with a large gap. For example, a thin layer of GaAs can be inserted between two layers of $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$. Such structures are increasingly important for applications. The location of the edge of the conduction and valence bands in such a sandwich structure is shown in Fig. 27.36(a).

Electrons at the bottom of the conduction band of GaAs feel a narrow potential well of width d. Assuming that the well is sufficiently deep for that the potential barrier can be taken infinitely high in the calculation of the lowest states, the boundary conditions imposed on the wavefunction implies, in line with the discussion in Chapter 16, that the electrons can be in the states of energy

$$\varepsilon_\perp = \frac{\hbar^2\pi^2}{2m_{\mathrm{n}}^* d^2} n^2 \tag{27.5.1}$$

measured from the bottom of the band, where m_{n}^* is the effective mass of the electrons. If the depth of the potential well is finite, the wavefunction does

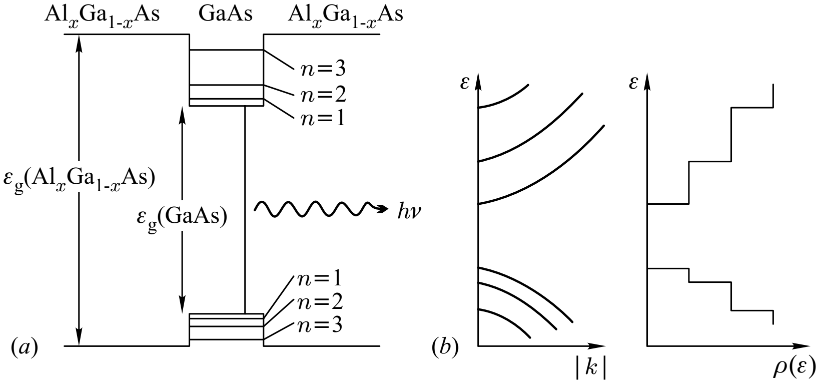


Fig. 27.36. (*a*) The location of the edge of the conduction and valence bands in a GaAs quantum well. (*b*) The dispersion relation for the states in a quantum well and their density of states

not vanish identically outside the potential well. By requiring the continuity of the wavefunction and its derivative across the boundary of the well, only a finite number of bound states may appear; their energies are smaller than the height of the potential barrier, and their wavefunctions decay exponentially outside the well. The separation of the levels is not exactly proportional to n^2 but we shall neglect the difference below.

When the motion parallel to the potential barrier is also taken into account, the allowed energies are

$$\varepsilon = \varepsilon_{\mathrm{c}} + \frac{\hbar^2 k_{\parallel}^2}{2m_{\mathrm{n}}^*} + \frac{\hbar^2\pi^2}{2m_{\mathrm{n}}^* d^2} n^2 \,. \tag{27.5.2}$$

It can be shown by the same token that holes can occupy states of energy

$$\varepsilon = \varepsilon_{\mathrm{v}} - \frac{\hbar^2 k_{\parallel}^2}{2m_{\mathrm{p}}^*} - \frac{\hbar^2\pi^2}{2m_{\mathrm{p}}^* d^2} n^2 \tag{27.5.3}$$

in the valence band. As long as the width d of the quantum well is sufficiently small compared to the transverse dimensions, the energies of the states of different quantum numbers n are fairly well separated, and form subbands. This spectrum and the corresponding density of states are shown in Fig. 27.36(*b*). To calculate the latter, we made use of the property that within each subband the electron system can be considered two-dimensional, and the density of states of a 2DES is constant. If the well is narrow enough (nanoscale), the subbands of adjacent quantum numbers n are sufficiently separated for that the behavior should be governed by a single subband. Such a quantum well is ideally adapted to the investigation of the properties of two-dimensional electron systems.

An important application of quantum wells is the quantum-well laser (QW laser). This is a double-heterojunction laser with a nanoscale active layer; the light emitted in the transition between its quantized energy levels is amplified coherently. The wavelength of the emitted light can be controlled by the thickness of the active layer. Even better lasers can be produced from structures with multiple quantum wells, in which GaAs and $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ layers alternate.

In a special configuration wide-gap and narrow-gap semiconductors alternate in one direction periodically. If the thickness of the wide-gap semiconductor layers is small enough for that the wavefunctions extending beyond the wells formed in the narrow-gap layers overlap, the carriers are not localized to one particular well but can propagate in the superlattice of quantum wells by tunneling.

27.5.2 Quantum Wires and Quantum Dots

If the free motion of electrons in a 2DEG at the semiconductor interface is further limited by gates to a few hundred nm or less in one direction, a practically one dimensional electron system is obtained. Assuming that the potential well at the cross section of such a quantum wire is deep and sharp, the state of the electron can be described in terms of a propagating plane wave along and two standing waves perpendicular to the axis of the wire:

$$\psi(x,y,z) = \frac{1}{\sqrt{L}} \mathrm{e}^{\mathrm{i}k_x x} \left(\frac{4}{L_y L_z} \right)^{1/2} \sin \frac{m\pi y}{L_y} \sin \frac{n\pi z}{L_z} \,. \tag{27.5.4}$$

The energy of the state is then

$$\varepsilon_{k_x,m,n} = \frac{\hbar^2 k_x^2}{2m^*} + \frac{m^2\hbar^2\pi^2}{2m^* L_y^2} + \frac{n^2\hbar^2\pi^2}{2m^* L_z^2} \,, \tag{27.5.5}$$

and the density of states is the sum of terms with the typical inverse-square-root singularity of one-dimensional systems:

$$\rho(\varepsilon) = \left(\frac{m^*}{\pi\hbar^2} \right) \left(\frac{\hbar^2}{2m^*} \right)^{1/2} \sum_{m,n} (\varepsilon - \varepsilon_{m,n})^{-1/2} \theta(\varepsilon - \varepsilon_{m,n}) \,, \tag{27.5.6}$$

where $\theta(x)$ is the Heaviside step function and

$$\varepsilon_{m,n} = \frac{m^2\hbar^2\pi^2}{2m^* L_y^2} + \frac{n^2\hbar^2\pi^2}{2m^* L_z^2} \,. \tag{27.5.7}$$

If the wire is sufficiently thin, the branches are fairly well separated in energy, and only a few of them are partially filled at low temperatures. Assuming that the states in the branches of quantum numbers m and n are not mixed by the infrequent collisions, each branch can be considered as an independent channel for the propagation of electrons.

If the length L of the wire is smaller than the electron mean free path l, the propagation of electrons is not diffusive (with frequent collisions) but ballistic. Such relatively short and narrow constrictions between two wider electron reservoirs are also called *quantum point contacts* (QPCs). Investigating the electron transport in and conductance of such systems, LANDAUER[18] found that when a sample of transmission coefficient T is placed between two ideally conducting reservoirs, and the same contacts are used for the voltage and current measurements, the conductance is

$$G = 2\frac{e^2}{h}T\,. \tag{27.5.8}$$

This is the *Landauer formula.* The factor 2 comes from spin. This value of the conductance can be understood intuitively from the formula $\rho(\varepsilon_{\mathrm{F}}) = 1/(\pi\hbar v_{\mathrm{F}})$ for the density of states per spin orientation at the Fermi energy in a one-dimensional electron gas. Considering only particles that propagate in the same direction, the density of states is $1/(2\pi\hbar v_{\mathrm{F}})$. When a voltage V is applied to the sample, electrons in a region of width eV carry the current. Thus, if each electron reaches the other side without collision, the current per spin orientation is

$$I = ev_{\mathrm{F}}\frac{1}{2\pi\hbar v_{\mathrm{F}}}eV = \frac{e^2}{h}V\,. \tag{27.5.9}$$

The velocity in the current formula is canceled by the factor $1/v_{\mathrm{F}}$ in the density of states, which leads to a conductance of e^2/h per spin orientation. When electrons undergo collisions, the transmission coefficient is smaller than unity, and the conductance is also proportionally smaller.

If electrons can propagate in several independent channels, the total transmission coefficient in the conductance formula is the sum of the transmission coefficients of individual channels. Naturally, only partially filled channels below the Fermi energy need to be considered. The transmission coefficient T can generally be determined using the methods of quantum mechanics. If the propagation of electrons is indeed ballistic, the transmission coefficient is unity, and thus, because of the spin degeneracy, the conductance is quantized in units of $G_0 = 2e^2/h = 7.748 \times 10^{-5}\,\Omega^{-1}$ called the conductance quantum.[19] This is shown in Fig. 27.37. The voltage between the two ends of the wire is kept constant in the measurement setup, and the number of electrons in the wire and the number of open channels in which electrons can propagate are controlled by the negative gate voltage. Then steps appear as a function of the gate voltage. The gate voltage plays a double role in that. Firstly, it changes the electrostatic potential of the wire with respect to the source and drain contacts, secondly, as the magnitude of the negative voltage is increased, the

[18] R. LANDAUER, 1957.

[19] In terms of the resistance: R can only be $1/n$th of $h/2e^2 = 12.906\,\mathrm{k}\Omega$.

region in which electrons can propagate becomes narrower, and thus the energy difference between the channels becomes larger. Both effects lead to a decrease in the number of open channels.

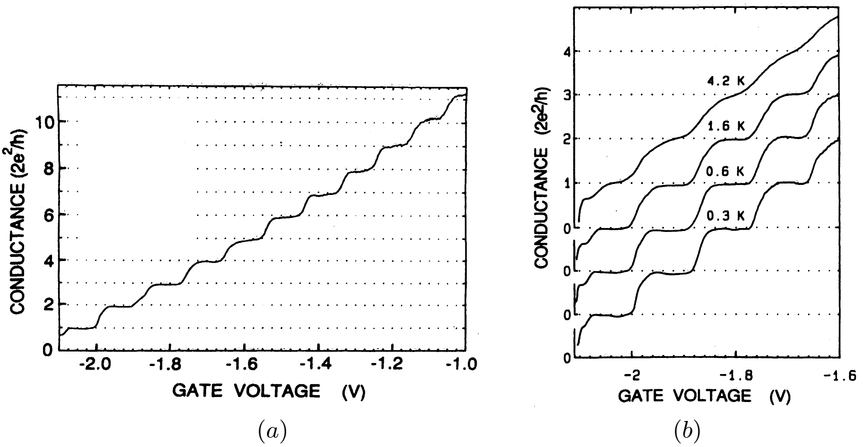


Fig. 27.37. (a) Quantized conductance in a quantum wire, at 0.6 K. (b) The quantization is smeared out at higher temperatures [Reprinted with permission from B. J. van Wees et al., *Phys. Rev. Lett.* **60**, 848 (1988) and *Phys. Rev. B* **43**, 12431 (1991). ©1991 by the American Physical Society]

In quantum dots the motion of charge carriers is confined in all three spatial dimensions. The linear dimension of the allowed region is less than a μm in semiconductors, whereas in metallic samples it is on the order of the Fermi wavelength. This can be achieved in practice by confining the two-dimensional electron gas of a semiconductor device to a small region by means of voltages applied to the gates on the surface. This way a practically zero dimensional electron gas is obtained.

Within a region of such dimensions, the number of mobile electrons that participate in conduction is at most a few thousand. Owing to the small capacity of the dot, the charging energy becomes important on this scale. Adding charge Q to a system of capacity C requires an energy Q^2/C. If the energy needed for the addition of a single electron exceeds the thermal energy, then the Coulomb repulsion of the electrons on the dot prevents it. This phenomenon is called the *Coulomb blockade.* The quantized nature of charge then becomes important, and electrons can be moved one by one in such systems. The new phenomena arising from this property can be studied thoroughly in a nanostructure in which the quantum dot is connected via tunnel junctions to two metallic electron reservoirs, a source and a drain, and its voltage is controlled by a third electrode (gate). For reasons that will be understood later, such structures are also called single-electron transistors (SETs). Their schematic structure is shown in Fig. 27.38.

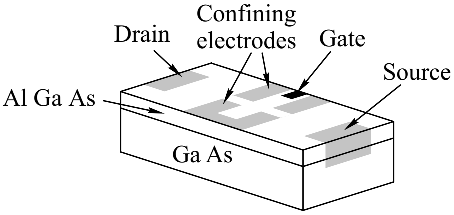


Fig. 27.38. The schematic structure of single-electron transistors

If there are N mobile electrons ($Q = -eN$) on the dot, the total electrostatic energy is the sum of the term QV_g due to the gate voltage V_g and the charging energy $Q^2/2C$:

$$E(Q) = QV_\mathrm{g} + \frac{Q^2}{2C}\,. \tag{27.5.10}$$

The dependence of this energy on the number of electrons is shown in Fig. 27.39.

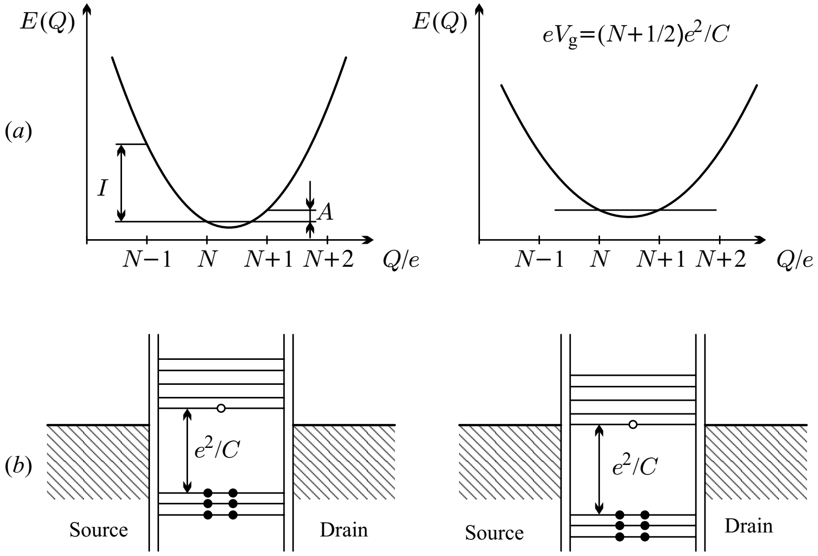


Fig. 27.39. (a) The electrostatic energy of a quantum dot as a function of the number of electrons in the Coulomb blockade region and in the degeneracy point. (b) The spectrum associated with the addition or removal of an electron in the two cases. Full circles indicate the N electrons on the dot; the empty circle is the $N+1$th electron

If N changed continuously, the energy minimum would be at $N = CV_\mathrm{g}/e$. Since the particle number is a discrete variable, the actual electron number is the integer N that is closest to that minimum. Referred to this discrete minimum, the lowest energies of the systems with $N+1$ and $N-1$ electrons are A and I higher, respectively, where A is the electron affinity and I is the ionization energy:

$$\begin{aligned} A &= E(N+1) - E(N) = -eV_{\mathrm{g}} + \frac{e^2}{C}\bigl(N + \tfrac{1}{2}\bigr) , \\ I &= E(N-1) - E(N) = eV_{\mathrm{g}} - \frac{e^2}{C}\bigl(N - \tfrac{1}{2}\bigr) . \end{aligned} \tag{27.5.11}$$

Obviously, $A + I = e^2/C$. By choosing the gate voltage $V_{\mathrm{g}} = eN/C$ in such a way that the minimum be at an integer charge state, both the addition and removal of an electron require an energy $e^2/2C$. The energy needed to add or remove an electron is finite at other gate voltages, too. This nonzero energy difference prevents electrons from jumping from one contact to the dot and then on to the other contact. The only exception is

$$V_{\mathrm{g}} = e\bigl(N \pm \tfrac{1}{2}\bigr)/C , \tag{27.5.12}$$

when either A or I vanishes, and thus the electrostatic energy is the same for N and $N+1$ or N and $N-1$ charges. The charge of the dot can fluctuate between the two values. If a small external voltage is applied between the source and drain, a current can flow through the quantum dot, since electrons can jump from one contact to the dot and then on to the other contact. By varying the gate voltage, this situation occurs for all values V_{g} for which the previous condition is satisfied by an integer N – that is, finite peaks separated by regular intervals $\delta V_{\mathrm{g}} = e/C$ appear in the conductance. The conductance oscillates with V_{g}. This is the Coulomb blockade oscillation.

The number of electrons on the dot, the magnitude of the electron affinity and ionization energy (that are finite because of the charging energy), and, through them, the current between the source and drain can thus be controlled by the gate voltage. This explains why the device is called a single-electron transistor.

In addition to the quantized nature of charge, the discreteness of the energy levels must also be taken into account in semiconductor quantum dots. By taking a complete set of one-particle states of energy ε_λ, the dot can be described in terms of the Hamiltonian

$$\mathcal{H}_{\mathrm{dot}} = \sum_\lambda \bigl(\varepsilon_\lambda - eV_{\mathrm{g}}\bigr) a_\lambda^\dagger a_\lambda + \bigl(\hat{\mathcal{N}} e\bigr)^2/2C , \tag{27.5.13}$$

where $\hat{\mathcal{N}} = \sum_\lambda a_\lambda^\dagger a_\lambda$ is the particle-number operator for the electrons on the dot. When there are N electrons on the dot, the energy required to add the $(N+1)$th is not ε_{N+1} but the electrostatic energy difference between the configurations with $N+1$ and N electrons:

$$\begin{aligned} \Delta E &= \varepsilon_{N+1} + \bigl[- (N+1)eV_{\mathrm{g}} + (N+1)^2 e^2/2C\bigr] - \bigl[- NeV_{\mathrm{g}} + N^2 e^2/2C\bigr] \\ &= \varepsilon_{N+1} - eV_{\mathrm{g}} + (N + \tfrac{1}{2}) e^2/C . \end{aligned} \tag{27.5.14}$$

Tunneling occurs when this energy is the same as the chemical potential of the contact:

$$\mu = \varepsilon_{N+1} - eV_g + (N + \tfrac{1}{2})e^2/C\,, \tag{27.5.15}$$

that is, at the gate voltage

$$V_g = (N + \tfrac{1}{2})e/C + \bigl(\varepsilon_{N+1} - \mu\bigr)/e\,. \tag{27.5.16}$$

Therefore the separation of the conductance peaks is not exactly regular but

$$\delta V_g = e/C + \bigl(\varepsilon_{N+1} - \varepsilon_N\bigr)/e\,, \tag{27.5.17}$$

and their amplitudes are not equal, either. Figure 27.40 shows the Coulomb blockade oscillations in a quantum dot in a GaAs/AlGaAs heterojunction. The capacity decreases and the peak separation increases for decreasing dot size. The thermal broadening can be interpreted in terms of the Fermi–Dirac statistics for the occupation of the levels.

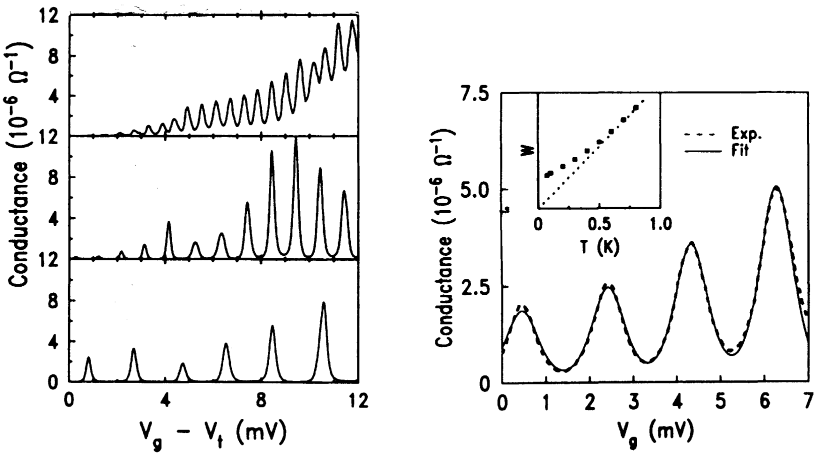


Fig. 27.40. Coulomb blockade oscillations in quantum dots of different sizes in GaAs/AlGaAs heterojunctions for progressively shorter distances between the two constrictions, with a corresponding increase in the period at low temperature ($T =$ 50 mK), and the thermal broadening of the peaks at a higher temperature ($T =$ 800 mK) [Reprinted with permission from U. Meirav et al., *Phys. Rev. Lett.* **65**, 771 (1990). ©1990 by the American Physical Society]

As shown in Fig. 27.41, similar oscillations had been observed earlier in quantum wires, too. It has been established that the underlying mechanism is the same: the Coulomb blockade. The reason for this is that quantum wires are often disordered, made up of weakly coupled smaller parts (islands), between which electrons can travel by tunneling. The charging energy of individual islands can be high, and then the conductance of the quantum wire is similar to that of a quantum dot.

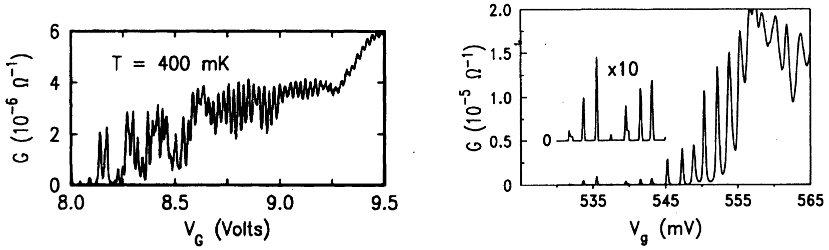


Fig. 27.41. Coulomb blockade oscillations in the conductance of disordered quantum wires [Reprinted with permission from J. H. F. Scott-Thomas et al., *Phys. Rev. Lett.* **62**, 583 (1989), and U. Meirav et al., *Phys. Rev. B* **40**, 5871 (1989). ©1989 by the American Physical Society]

It should be noted that if V_g is fixed and the potential difference between the source and drain is varied, the measured current–voltage characteristics are nonlinear. This is once again due to the Coulomb blockade. The current increases step by step each time the potential difference reaches the threshold for a new electron to jump to the quantum dot. This pattern is called the *Coulomb staircase.*

27.6 Basics of Spintronics

Just like for classical semiconductor devices, only the motion of charges and the current carried by them were examined in the semiconductor quantum devices presented in the previous section. Since the standard devices are nonmagnetic, it was taken for granted that the current carried by electrons is the same for both spin orientations. The role of spin was adding a factor 2 to the formulas of extensive quantities compared to the spinless case. The recent discovery that the electron spin can play a much more important role gave birth to a new branch of electronics called *spin-based electronics* or *spintronics.* It is concerned with the construction of quantum devices by making use of the spin-selective motion of electrons and the interaction of the spin of charge carriers with magnetic materials. If the quantum state defined by the spin remains coherent for a sufficiently long time, such devices can be used for storing and transmitting information.

One of the most important examples of spin-dependent transport is the giant magnetoresistance[20] (GMR) observed in systems made up of two or more magnetic layers separated by thin nonmagnetic layers. If the mobile electrons

[20] The adjective “giant” indicates that the variation of the resistance may be considerably larger than ten percent in such layered structures, as opposed to the customary few percent in metals placed in magnetic fields. The colossal magnetoresistance (CMR) of certain perovskite manganites, in which the variation of the resistance is several orders of magnitude larger, is certainly caused by some other, as-yet not well understood, mechanism.

in the nonmagnetic layer between two adjacent ferromagnetic layers mediate antiferromagnetic coupling via the RKKY interaction (pages 466 of Volume 1 and 605) then the magnetization alternates in subsequent layers. As the research groups of A. Fert and P. Grünberg[21] demonstrated independently in 1988, the resistance is larger in this state than in the configuration where the moments of subsequent ferromagnetic layers are aligned in the same direction in a strong magnetic field. To understand this phenomenon, it should be borne in mind that in scattering processes without spin flip – which are usually much stronger than spin-flip processes – the transition probability is proportional to the density of final states, and this density of states is very different for the two spin orientations in ferromagnetic materials. Consequently, ferromagnetic layers act as polarizators: the transmission probability is much higher for electrons whose magnetic moment is parallel to the ferromagnetic moment of the layer than for the opposite spin orientation. If the two ferromagnetic layers are oppositely magnetized (antiferromagnetically coupled), electrons of both spin orientations are scattered in the same way as they pass through both layers. (For either spin orientation, the passage is almost without collisions through one but not the other layer.) However, when the magnetization of the ferromagnetic layers is parallel, half of the electrons – those with a favorable spin orientation – pass through both layers easily, while the other half (with the opposite spin) are strongly scattered in both layers. Since the currents carried by the electrons of the two spin orientations flow in parallel, the resistance is smaller in the last case than for antiferromagnetically coupled layers. Such configurations in which the resistance is controlled by the magnetic field via reversing the magnetization of a layer are called spin valves. Nowadays such layered structures are widely used in the read heads of hard disks. Writing and reading processes are based on giant magnetoresistance because a higher information density can be achieved this way.

Nevertheless, research in spintronics is mostly concerned with semiconductors rather than metallic systems – firstly, because such components can be integrated more easily with standard semiconductor devices, and secondly, because signal amplification can be achieved only by means of semiconductor devices. The description of the behavior of such systems requires the generalization of the fundamental equations of semiconductor devices (Section 27.2.3) to the spin-polarized case. The total current can now be decomposed into four parts:

$$\boldsymbol{j}(\boldsymbol{r}) = \boldsymbol{j}_{\mathrm{n}\uparrow}(\boldsymbol{r}) + \boldsymbol{j}_{\mathrm{n}\downarrow}(\boldsymbol{r}) + \boldsymbol{j}_{\mathrm{p}\uparrow}(\boldsymbol{r}) + \boldsymbol{j}_{\mathrm{p}\downarrow}(\boldsymbol{r}) . \tag{27.6.1}$$

Just like in (27.2.35), the electron current for each spin orientation is given by the equations

$$\begin{aligned} \boldsymbol{j}_{\mathrm{n}\uparrow}(\boldsymbol{r}) &= e n_\uparrow(\boldsymbol{r}) \mu_{\mathrm{n}\uparrow} \boldsymbol{E} + e D_{\mathrm{n}\uparrow} \boldsymbol{\nabla} n_\uparrow(\boldsymbol{r}) , \\ \boldsymbol{j}_{\mathrm{n}\downarrow}(\boldsymbol{r}) &= e n_\downarrow(\boldsymbol{r}) \mu_{\mathrm{n}\downarrow} \boldsymbol{E} + e D_{\mathrm{n}\downarrow} \boldsymbol{\nabla} n_\downarrow(\boldsymbol{r}) , \end{aligned} \tag{27.6.2}$$

[21] Albert Fert (1938–) and Peter Grünberg (1939–) were awarded the Nobel Prize in 2007 “for the discovery of giant magnetoresistance”.

where the mobility and the diffusion coefficient can be spin dependent. Similar equations apply to the hole current:

$$\begin{aligned} \boldsymbol{j}_{\mathrm{p}\uparrow}(\boldsymbol{r}) &= e p_\uparrow(\boldsymbol{r})\mu_{\mathrm{p}\uparrow}\boldsymbol{E} - eD_{\mathrm{p}\uparrow}\boldsymbol{\nabla} p_\uparrow(\boldsymbol{r})\,, \\ \boldsymbol{j}_{\mathrm{p}\downarrow}(\boldsymbol{r}) &= e p_\downarrow(\boldsymbol{r})\mu_{\mathrm{p}\downarrow}\boldsymbol{E} - eD_{\mathrm{p}\downarrow}\boldsymbol{\nabla} p_\downarrow(\boldsymbol{r})\,. \end{aligned} \tag{27.6.3}$$

The Poisson equation relating the gradient of the electric field to the charge density takes the form

$$\epsilon \operatorname{div} \boldsymbol{E} = -e\left[n_\uparrow(\boldsymbol{r}) + n_\downarrow(\boldsymbol{r}) + n_{\mathrm{a}}^-(\boldsymbol{r}) - n_{\mathrm{d}}^+(\boldsymbol{r}) - p_\uparrow(\boldsymbol{r}) - p_\downarrow(\boldsymbol{r})\right] \tag{27.6.4}$$

if the sample contains donors and acceptors.

Finally, we have to generalize of the continuity equation to establish the relationship between the spatial variation of the charge density and the currents, keeping track of the spin dependence of the generation and recombination of carriers. The formula (27.2.1) can be used for recombination, with the obvious auxiliary condition that the electron and hole that take part in the process must be of opposite spins. Additionally, a new term appears, the spin relaxation. Spins can be reversed with a characteristic time $\tau_{\uparrow\downarrow}$; its reciprocal is the spin-flip rate $\Gamma_{\uparrow\downarrow}$. Such spin-flip processes couple the numbers of spin-up and spin-down electrons and holes. Therefore, instead of the system of equation (27.2.37), its generalization

$$\begin{aligned} \frac{\partial n_\uparrow}{\partial t} - \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\mathrm{n}\uparrow} &= G_{\mathrm{inj,\,n}\uparrow} + G_{\mathrm{therm}\uparrow} - C_\uparrow n_\uparrow p_\downarrow + \Gamma_{\mathrm{n}\uparrow\downarrow} n_\downarrow - \Gamma_{\mathrm{n}\downarrow\uparrow} n_\uparrow\,, \\ \frac{\partial n_\downarrow}{\partial t} - \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\mathrm{n}\downarrow} &= G_{\mathrm{inj,\,n}\downarrow} + G_{\mathrm{therm}\downarrow} - C_\downarrow n_\downarrow p_\uparrow - \Gamma_{\mathrm{n}\uparrow\downarrow} n_\downarrow + \Gamma_{\mathrm{n}\downarrow\uparrow} n_\uparrow\,, \\ \frac{\partial p_\uparrow}{\partial t} + \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\mathrm{p}\uparrow} &= G_{\mathrm{inj,\,p}\uparrow} + G_{\mathrm{therm}\uparrow} - C_\downarrow n_\downarrow p_\uparrow + \Gamma_{\mathrm{p}\uparrow\downarrow} p_\downarrow - \Gamma_{\mathrm{p}\downarrow\uparrow} p_\uparrow\,, \\ \frac{\partial p_\downarrow}{\partial t} + \frac{1}{e}\boldsymbol{\nabla}\boldsymbol{j}_{\mathrm{p}\downarrow} &= G_{\mathrm{inj,\,p}\downarrow} + G_{\mathrm{therm}\downarrow} - C_\uparrow n_\uparrow p_\downarrow - \Gamma_{\mathrm{p}\uparrow\downarrow} p_\downarrow + \Gamma_{\mathrm{p}\downarrow\uparrow} p_\uparrow \end{aligned} \tag{27.6.5}$$

has to be used. The real difficulty in understanding spin-polarized transport lies is the quantum mechanical determination of the parameters in the previous phenomenological equations. Its detailed discussion cannot be given here; we just mention that, among others, we would need to clarify how good a quantum number the spin is for characterizing the state of holes in the presence of the spin–orbit interaction. Note that in establishing the previous set of equations we assumed that carriers undergo frequent collisions, and transport can be conceived as a diffusive motion. In small samples, nanostructures, this assumption is not valid: the motion can be ballistic. Just like in quantum devices based on charge transport, this requires another description.

The generation and injection of spin-polarized electrons are among the basic challenges of spintronics. Obviously, spin polarization can be achieved most simply in ferromagnetic materials. The existence of ferromagnetic semiconductors was an interesting discovery of the past decades. For applications, the most important are the structures based on III–V semiconductors (e.g., InAs or GaAs) and a magnetic element (for example manganese), which substitutes partially (at most at a few atomic percent level) the element from column III. Curie temperatures above 100 K have been achieved this way. Using such materials facilitates the fabrication of semiconductor junctions in which one side is a ferromagnetic and the other a nonmagnetic semiconductor. The quest for the most effective way of injecting spin-polarized electrons from a ferromagnet into a nonmagnetic semiconductor – so that spin current could flow in the latter – has been a topic of particularly keen interest.

The polarized electrons injected into the nonmagnetic component are not in equilibrium. Relaxation processes would lead to the establishment of thermal equilibrium but such processes are relatively slow (the spin-flip time is on the order of 10^{-9} s), so electrons can pass through a nanostructure with their phase preserved. However, the spin can be manipulated by means of the spin–orbit interaction, which opens the way to new phenomena.

As mentioned in Chapter 3, the spin–orbit coupling can be described by the Rashba term (3.1.36) in two-dimensional systems, where the coupling strength α is proportional to the gradient of the potential, and can therefore be controlled by the voltage applied to the system. During their motion electrons feel a spin-dependent magnetic field, so spin-up and spin-down electrons propagate differently. An interesting consequence of this difference is the experimentally confirmed *spin Hall effect.* The Rashba term gives rise to a pure spin current that is perpendicular to the electric current through the sample, and the spin polarization will be different on the two sides.

The operation of the spin-field-effect transistor (spin-FET) proposed by DATTA and DAS[22] (Fig. 27.42) is based on the Rashba-type spin–orbit coupling using polarized electrons injected from the ferromagnet into the nonmagnetic semiconductor.

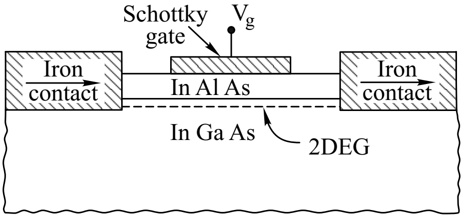


Fig. 27.42. Schematic structure of the Datta–Das spin-field-effect transistor [S. Datta and B. Das, *Appl. Phys. Lett.* **56**, 665 (1990)]

The two-dimensional electron gas at the interface of a heterojunction is connected to a ferromagnetic source and a drain. The spin-polarized charges injected from the source (emitter) can travel along the interface and become

[22] S. DATTA and B. DAS, 1990.

absorbed at the other side, provided the polarization of the drain (collector) is the same. However, when the electron spins are flipped in the channel by some mechanism, the electrons can no longer be absorbed, and there is no current. When a voltage is applied to the gate, the spins of the electrons traveling in the channel start to rotate. The current through the channel can therefore be controlled by the gate voltage.

Among the many open issues of spintronics the most intriguing is the possibility of using the two states of a spin as a quantum bit (qubit) for storing and transmitting information, and thus for building a spin-based quantum computer. Hopefully, we have a not so long way to go before finding the answer to this problem as well as the other questions mentioned above.

Further Reading

1. M. Balkanski and R. E. Wallis, *Semiconductor Physics and Applications*, Oxford University Press, Oxford (2000).
2. K. W. Böer, *Survey of Semiconductor Physics*, Vol. 2. *Barriers, Junctions, Surfaces, and Devices*, Van Nostrand Reinhold, New York (1992).
3. D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge Studies in Semiconductor Physics and Microelectronic Engineering, Cambridge University Press, Cambridge (1997).
4. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley and Sons, New York (1967).
5. *Handbook on Semiconductors*, Completely Revised and Enlarged Edition, Series Editor T. S. Moss, Volume 4. *Device Physics*, Volume Editor C. Hilsum, North-Holland, Amsterdam (1993).
6. *Modern Semiconductor Device Physics*, Edited by S. M. Sze, John Wiley & Sons, Inc., New York (1998).
7. K. K. Ng, *Complete Guide to Semiconductor Devices*, McGraw-Hill, New York (1995).
8. D. J. Roulston, *An Introduction to the Physics of Semiconductor Devices*, Oxford University Press, Oxford (1999).
9. *Semiconductor Spintronics and Quantum Computation*, Eds. D. D. Awschalom, D. Loss and N. Samarth, NanoScience and Technology, Springer-Verlag Berlin (2002).
10. S. M. Sze, *Physics of Semiconductor Devices*, Second Edition, Wiley-Interscience, New York (1981).
11. S. M. Sze, *Semiconductor Devices, Physics and Technology*, Wiley-Interscience, New York (1985).
12. S. Wang *Fundamentals of Semiconductor Theory and Device Physics*, Prentice Hall, Inc., Englewood Cliffs (1989).

G

Quantum Mechanical Perturbation Theory

Quantum mechanical perturbation theory is a widely used method in solid-state physics. Without the details of derivation, we shall list a number of basic formulas of time-independent (stationary) and time-dependent perturbation theory below. For simplicity, we shall use the Dirac notation for wavefunctions and matrix elements.

G.1 Time-Independent Perturbation Theory

Assume that the complete solution (eigenfunctions and eigenvalues) of the Schrödinger equation

$$\mathcal{H}_0 |\psi_i^{(0)}\rangle = E_i^{(0)} |\psi_i^{(0)}\rangle \tag{G.1.1}$$

is known for a system described by a simple Hamiltonian $\mathcal{H}_0$. If the system is subject to a time-independent (stationary) perturbation described by the Hamiltonian $\mathcal{H}_1$ – which can be an external perturbation or the interaction between the components of the system –, the eigenvalues and eigenfunctions change. The method for determining the new ones depends on whether the unperturbed energy level in question is degenerate or not.

G.1.1 Nondegenerate Perturbation Theory

We now introduce a fictitious coupling constant λ, whose value will be treated as a parameter in the calculations and set equal to unity in the final result, and write the full Hamiltonian $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1$ as

$$\mathcal{H} = \mathcal{H}_0 + \lambda \mathcal{H}_1 \,. \tag{G.1.2}$$

The parameter λ is purely a bookkeeping device to keep track of the relative order of magnitude of the various terms, since the energy eigenvalues and eigenfunctions will be sought in the form of an expansion in powers of λ:

$$\begin{aligned} \left|\psi_i\right\rangle &= \left|\psi_i^{(0)}\right\rangle + \sum_{n=1}^{\infty} \lambda^n \left|\psi_i^{(n)}\right\rangle, \\ E_i &= E_i^{(0)} + \sum_{n=1}^{\infty} \lambda^n E_i^{(n)}. \end{aligned} \tag{G.1.3}$$

The series is convergent if the perturbation is weak, that is, in addition to the formally introduced parameter λ, the interaction Hamiltonian itself contains a small parameter, the physical coupling constant.

By substituting this expansion into the Schrödinger equation and collecting the same powers of λ from both sides, we obtain

$$\begin{aligned} \mathcal{H}_0\left|\psi_i^{(0)}\right\rangle &= E_i^{(0)}\left|\psi_i^{(0)}\right\rangle, \\ \mathcal{H}_0\left|\psi_i^{(1)}\right\rangle + \mathcal{H}_1\left|\psi_i^{(0)}\right\rangle &= E_i^{(0)}\left|\psi_i^{(1)}\right\rangle + E_i^{(1)}\left|\psi_i^{(0)}\right\rangle, \\ \mathcal{H}_0\left|\psi_i^{(2)}\right\rangle + \mathcal{H}_1\left|\psi_i^{(1)}\right\rangle &= E_i^{(0)}\left|\psi_i^{(2)}\right\rangle + E_i^{(1)}\left|\psi_i^{(1)}\right\rangle + E_i^{(2)}\left|\psi_i^{(0)}\right\rangle \end{aligned} \tag{G.1.4}$$

and similar equations for higher-order corrections. The corrections to the energy and wavefunction of any order are related to the lower-order ones by the recursion formula

$$\begin{aligned} &(\mathcal{H}_0 - E_i^{(0)})\left|\psi_i^{(n)}\right\rangle + (\mathcal{H}_1 - E_i^{(1)})\left|\psi_i^{(n-1)}\right\rangle \\ &\qquad - E_i^{(2)}\left|\psi_i^{(n-2)}\right\rangle - \ldots - E_i^{(n)}\left|\psi_i^{(0)}\right\rangle = 0\,. \end{aligned} \tag{G.1.5}$$

Multiplying the second equation in (G.1.4) (which comes from the terms that are linear in λ) by $\left\langle\psi_i^{(0)}\right|$ from the left, the first-order correction to the energy is

$$E_i^{(1)} = \left\langle\psi_i^{(0)}\right|\mathcal{H}_1\left|\psi_i^{(0)}\right\rangle. \tag{G.1.6}$$

To determine the correction to the wavefunction, the same equation is multiplied by $\left\langle\psi_j^{(0)}\right|$ $(j \neq i)$:

$$E_j^{(0)}\left\langle\psi_j^{(0)}\middle|\psi_i^{(1)}\right\rangle + \left\langle\psi_j^{(0)}\right|\mathcal{H}_1\left|\psi_i^{(0)}\right\rangle = E_i^{(0)}\left\langle\psi_j^{(0)}\middle|\psi_i^{(1)}\right\rangle. \tag{G.1.7}$$

Since the eigenfunctions of $\mathcal{H}_0$ make up a complete set, the functions $\left|\psi_i^{(n)}\right\rangle$ can be expanded in terms of them:

$$\left|\psi_i^{(n)}\right\rangle = \sum_j C_{ij}^{(n)}\left|\psi_j^{(0)}\right\rangle. \tag{G.1.8}$$

The coefficients $C_{ii}^{(n)}$ are not determined by the previous equations: their values depend on the normalization of the perturbed wavefunction. Substituting the previous formula into (G.1.7), we have

$$E_j^{(0)} C_{ij}^{(n)} + \left\langle\psi_j^{(0)}\right|\mathcal{H}_1\left|\psi_i^{(0)}\right\rangle = E_i^{(0)} C_{ij}^{(n)}\,, \tag{G.1.9}$$

and hence

$$\big|\psi_i^{(1)}\big\rangle = \sum_{j\neq i} \frac{\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{E_i^{(0)} - E_j^{(0)}} \big|\psi_j^{(0)}\big\rangle \,. \tag{G.1.10}$$

The second-order correction to the energy is then

$$E_i^{(2)} = \sum_{j\neq i} \frac{\Big|\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle\Big|^2}{E_i^{(0)} - E_j^{(0)}} \,, \tag{G.1.11}$$

and the second-order correction to the wavefunction is

$$\begin{aligned}\big|\psi_i^{(2)}\big\rangle = &\sum_{j\neq i}\sum_{k\neq i} \frac{\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_k^{(0)}\big\rangle\big\langle\psi_k^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)\big(E_i^{(0)} - E_k^{(0)}\big)} \psi_j^{(0)} \\ &- \sum_{j\neq i} \frac{\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)^2} \psi_j^{(0)} \,.\end{aligned} \tag{G.1.12}$$

However, this wavefunction is not normalized to unity. Proper normalization is ensured by the choice

$$\begin{aligned}\big|\psi_i^{(2)}\big\rangle = &\sum_{j\neq i}\sum_{k\neq i} \frac{\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_k^{(0)}\big\rangle\big\langle\psi_k^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)\big(E_i^{(0)} - E_k^{(0)}\big)} \big|\psi_j^{(0)}\big\rangle \\ &- \sum_{j\neq i} \frac{\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)^2} \big|\psi_j^{(0)}\big\rangle \\ &- \tfrac{1}{2}\sum_{j\neq i} \frac{\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_j^{(0)}\big\rangle\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)^2} \big|\psi_i^{(0)}\big\rangle \,.\end{aligned} \tag{G.1.13}$$

Finally, the third-order correction to the energy is

$$\begin{aligned}E_i^{(3)} = &\sum_{j\neq i}\sum_{k\neq i} \frac{\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_j^{(0)}\big\rangle\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_k^{(0)}\big\rangle\big\langle\psi_k^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)\big(E_i^{(0)} - E_k^{(0)}\big)} \\ &- \sum_{j\neq i} \frac{\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_j^{(0)}\big\rangle\big\langle\psi_j^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(0)}\big\rangle}{\big(E_i^{(0)} - E_j^{(0)}\big)^2} \,.\end{aligned} \tag{G.1.14}$$

In this Rayleigh–Schrödinger perturbation theory the explicit form of higher-order corrections becomes increasingly complicated. A relatively simple recursion formula can be obtained by introducing the projection operator

$$P_i = 1 - \big|\psi_i^{(0)}\big\rangle\big\langle\psi_i^{(0)}\big| = \sum_{j\neq i} \big|\psi_j^{(0)}\big\rangle\big\langle\psi_j^{(0)}\big| \tag{G.1.15}$$

which projects onto the subspace that is orthogonal to the state $\big|\psi_i^{(0)}\big\rangle$. The nth-order energy correction can then be written as

$$E_i^{(n)} = \big\langle\psi_i^{(0)}\big|\mathcal{H}_1\big|\psi_i^{(n-1)}\big\rangle\,, \tag{G.1.16}$$

where the matrix element is to be taken with the wavefunction

$$\begin{aligned}\big|\psi_i^{(n)}\big\rangle = \frac{1}{E_i^{(0)}-\mathcal{H}_0}P_i\Big[\big(\mathcal{H}_1 - E_i^{(1)}\big)\big|\psi_i^{(n-1)}\big\rangle \\ -E_i^{(2)}\big|\psi_i^{(n-2)}\big\rangle - \ldots - E_i^{(n-1)}\big|\psi_i^{(1)}\big\rangle\Big]\,,\end{aligned} \tag{G.1.17}$$

which is in the subspace mentioned above.

Formally simpler expressions can be obtained when the Brillouin–Wigner perturbation theory is used. The perturbed wavefunction in the Schrödinger equation

$$\big(\mathcal{H}_0 + \mathcal{H}_1\big)\big|\psi_i\big\rangle = E_i\big|\psi_i\big\rangle \tag{G.1.18}$$

is then chosen in the form

$$\big|\psi_i\big\rangle = C_0\big|\psi_i^{(0)}\big\rangle + \big|\Delta\psi_i\big\rangle\,, \tag{G.1.19}$$

where $\big|\Delta\psi_i\big\rangle$ is orthogonal to $\big|\psi_i^{(0)}\big\rangle$, and C_0 takes care of the appropriate normalization. After some algebra, the eigenvalue equation reads

$$\big(\mathcal{H}_0 - E_i\big)\big|\Delta\psi_i\big\rangle + \mathcal{H}_1\big|\psi_i\big\rangle = C_0\big(E_i - E_i^{(0)}\big)\big|\psi_i^{(0)}\big\rangle\,. \tag{G.1.20}$$

By applying the projection operator P_i, and exploiting the relations

$$P_i\big|\psi_i^{(0)}\big\rangle = 0\,, \qquad P_i\big|\psi_i\big\rangle = \big|\Delta\psi_i\big\rangle \tag{G.1.21}$$

as well as the commutation of P_i and $\mathcal{H}_0$,

$$\big(E_i - \mathcal{H}_0\big)\big|\Delta\psi_i\big\rangle = P_i\mathcal{H}_1\big|\psi_i\big\rangle \tag{G.1.22}$$

is obtained. Its formal solution is

$$\big|\psi_i\big\rangle = C_0\big|\psi_i^{(0)}\big\rangle + \frac{P_i}{E_i - \mathcal{H}_0}\mathcal{H}_1\big|\psi_i\big\rangle\,. \tag{G.1.23}$$

Iteration then yields

$$\big|\psi_i\big\rangle = C_0\sum_{n=0}^{\infty}\left(\frac{P_i}{E_i - \mathcal{H}_0}\mathcal{H}_1\right)^n\big|\psi_i^{(0)}\big\rangle\,, \tag{G.1.24}$$

and

$$\Delta E_i = \sum_{n=0}^{\infty}\big\langle\psi_i^{(0)}\big|\mathcal{H}_1\left(\frac{P_i}{E_i - \mathcal{H}_0}\mathcal{H}_1\right)^n\big|\psi_i^{(0)}\big\rangle \tag{G.1.25}$$

for the energy correction. In this method the energy denominator contains the perturbed energy E_i rather than the unperturbed one $E_i^{(0)}$. To first order in the interaction,

$$E_i = E_i^{(0)} + \langle \psi_i^{(0)} | \mathcal{H}_1 | \psi_i^{(0)} \rangle \,, \tag{G.1.26}$$

while to second order,

$$\begin{aligned} E_i = {} & E_i^{(0)} + \langle \psi_i^{(0)} | \mathcal{H}_1 | \psi_i^{(0)} \rangle \\ & + \sum_{j \neq i} \frac{\langle \psi_i^{(0)} | \mathcal{H}_1 | \psi_j^{(0)} \rangle \langle \psi_j^{(0)} | \mathcal{H}_1 | \psi_i^{(0)} \rangle}{E_i - E_j^{(0)}} + \dots \,. \end{aligned} \tag{G.1.27}$$

It is easy to show that by rearranging the energy denominator and expanding it as

$$\frac{1}{E_i - \mathcal{H}_0} = \frac{1}{E_i^{(0)} - \mathcal{H}_0 + \Delta E_i} = \frac{1}{E_i^{(0)} - \mathcal{H}_0} \sum_{n=0}^{\infty} \left(\frac{-\Delta E_i}{E_i^{(0)} - \mathcal{H}_0} \right)^n , \tag{G.1.28}$$

the results of the Rayleigh–Schrödinger perturbation theory are recovered.

The formulas of time-dependent perturbation theory can also be used to determine the ground-state energy and wavefunction of the perturbed system, provided the interaction is assumed to be turned on adiabatically. The appropriate formulas are given in Section G.2.

G.1.2 Degenerate Perturbation Theory

In the previous subsection we studied the shift of nondegenerate energy levels due to the perturbation. For degenerate levels a slightly different method has to be used because the formal application of the previous formulas would yield vanishing energy denominators.

Assuming that the ith energy level of the unperturbed system is p-fold degenerate – that is, the same energy $E_i^{(0)}$ belongs to each of the states $|\psi_{i_1}^{(0)}\rangle$, $|\psi_{i_2}^{(0)}\rangle$, $\dots$, $|\psi_{i_p}^{(0)}\rangle$ –, any linear combination of these degenerate eigenstates is also an eigenstate of $\mathcal{H}_0$ with the same energy. We shall use such linear combinations to determine the perturbed states. We write the wavefunctions of the states of the perturbed system that arise from the degenerate states as

$$|\psi\rangle = \sum_k c_{i_k} |\psi_{i_k}^{(0)}\rangle + \sum_{n \neq i} c_n |\psi_n^{(0)}\rangle \,, \tag{G.1.29}$$

where the c_{i_k} are of order unity, whereas the other coefficients c_n that specify the mixing with the unperturbed eigenstates whose energy is different from $E_i^{(0)}$ are small, proportional to the perturbation. By substituting this form into the Schrödinger equation, and multiplying both sides by $\langle \psi_{i_j}^{(0)}|$ from the left,

$$\Delta E c_{i_j} = \sum_k \langle \psi^{(0)}_{i_j} | \mathcal{H}_1 | \psi^{(0)}_{i_k} \rangle c_{i_k} + \sum_{n \neq i} \langle \psi^{(0)}_{i_j} | \mathcal{H}_1 | \psi^{(0)}_{n} \rangle c_n \tag{G.1.30}$$

is obtained. Since the coefficients c_n are small, the second term on the right-hand side can be neglected in calculating the leading-order energy correction, which is given by

$$\sum_k \left[\langle \psi^{(0)}_{i_j} | \mathcal{H}_1 | \psi^{(0)}_{i_k} \rangle - \delta_{jk} \Delta E \right] c_{i_k} = 0 \,. \tag{G.1.31}$$

This homogeneous system of equations has nontrivial solutions if the determinant of the coefficient matrix vanishes:

$$\det \left(\langle \psi^{(0)}_{i_j} | \mathcal{H}_1 | \psi^{(0)}_{i_k} \rangle - \delta_{jk} \Delta E \right) = 0 \,. \tag{G.1.32}$$

The solutions of this pth-order equation – that is, the eigenvalues of the matrix made up of the matrix elements $\langle \psi^{(0)}_{i_j} | \mathcal{H}_1 | \psi^{(0)}_{i_k} \rangle$ – specify the eventual splitting of the initially p-fold degenerate level, i.e., the shift of the perturbed levels with respect to the unperturbed one. Thus the interaction Hamiltonian needs to be diagonalized on the subspace of the degenerate states of $\mathcal{H}_0$. In general, the degeneracy is lifted at least partially by the perturbation. As discussed in Appendix D on group theory, the symmetry properties of the full Hamiltonian determine which irreducible representations appear, and what the degree of degeneracy is for each new level.

G.2 Time-Dependent Perturbation Theory

If the perturbation depends explicitly on time, no stationary states can arise. We may then be interested in the evolution of the system: What states can be reached at time t from an initial state $|\psi^{(0)}_i\rangle$ if the perturbation is turned on suddenly at time t_0? The answer lies in the solution of the time-dependent Schrödinger equation

$$\left[\mathcal{H}_0 + \lambda \mathcal{H}_1(t) \right] |\psi_i(t)\rangle = -\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial t} |\psi_i(t)\rangle \,. \tag{G.2.1}$$

The wavefunction $|\psi_i(t)\rangle$ is sought in the form

$$|\psi_i(t)\rangle = \sum_j c_{ij}(t) |\psi^{(0)}_j\rangle \, \mathrm{e}^{-\mathrm{i} E^{(0)}_j t/\hbar} \,, \tag{G.2.2}$$

subject to the initial condition

$$c_{ij}(t_0) = \delta_{ij} \,. \tag{G.2.3}$$

Since the time dependence of the unperturbed state has been written out explicitly, the functions $c_{ij}(t)$ are expected to vary slowly in time. Expanding the coefficients once again into powers of λ,

$$c_{ij}(t) = c_{ij}^{(0)}(t) + \sum_{r=1}^{\infty} \lambda^r c_{ij}^{(r)}(t)\,, \tag{G.2.4}$$

where, naturally, the zeroth-order term is a constant:

$$c_{ij}^{(0)}(t) = \delta_{ij}\,. \tag{G.2.5}$$

Substituting this series expansion into the Schrödinger equation, we find

$$-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t} c_{ij}^{(r)}(t) = \sum_k \mathrm{e}^{\mathrm{i}(E_j^{(0)} - E_k^{(0)})t/\hbar} \big\langle \psi_j^{(0)} \big| \mathcal{H}_1(t) \big| \psi_k^{(0)} \big\rangle\, c_{ik}^{(r-1)}(t)\,. \tag{G.2.6}$$

The explicit formulas for the first two terms obtained through iteration are

$$c_{ij}^{(1)}(t) = -\frac{\mathrm{i}}{\hbar} \int_{t_0}^{t} \big\langle \psi_j^{(0)} \big| \mathcal{H}_1(t_1) \big| \psi_i^{(0)} \big\rangle\, \mathrm{e}^{\mathrm{i}(E_j^{(0)} - E_i^{(0)})t_1/\hbar} \mathrm{d}t_1\,, \tag{G.2.7}$$

and

$$\begin{aligned} c_{ij}^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_0}^{t_1} \mathrm{d}t_2 \sum_k \big\langle \psi_j^{(0)} \big| \mathcal{H}_1(t_1) \big| \psi_k^{(0)} \big\rangle\, \mathrm{e}^{\mathrm{i}(E_j^{(0)} - E_k^{(0)})t_1/\hbar} \\ \times \big\langle \psi_k^{(0)} \big| \mathcal{H}_1(t_2) \big| \psi_i^{(0)} \big\rangle\, \mathrm{e}^{\mathrm{i}(E_k^{(0)} - E_i^{(0)})t_2/\hbar} \end{aligned} \tag{G.2.8}$$

In the interaction picture the time dependence of an arbitrary operator O is given by

$$\hat{O}(t) = \mathrm{e}^{\mathrm{i}\mathcal{H}_0 t/\hbar} O \mathrm{e}^{-\mathrm{i}\mathcal{H}_0 t/\hbar}\,. \tag{G.2.9}$$

Using this form for the Hamiltonian, which may have an intrinsic time dependence as well, the first two coefficients $c_{ij}^{(n)}$ can be written in terms of the operators

$$\hat{\mathcal{H}}_1(t) = \mathrm{e}^{\mathrm{i}\mathcal{H}_0 t/\hbar} \mathcal{H}_1(t) \mathrm{e}^{-\mathrm{i}\mathcal{H}_0 t/\hbar} \tag{G.2.10}$$

as

$$c_{ij}^{(1)}(t) = -\frac{\mathrm{i}}{\hbar} \int_{t_0}^{t} \big\langle \psi_j^{(0)} \big| \hat{\mathcal{H}}_1(t_1) \big| \psi_i^{(0)} \big\rangle \mathrm{d}t_1 \tag{G.2.11}$$

and

$$\begin{aligned} c_{ij}^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_0}^{t_1} \mathrm{d}t_2 \sum_k \big\langle \psi_j^{(0)} \big| \hat{\mathcal{H}}_1(t_1) \big| \psi_k^{(0)} \big\rangle \\ \times \big\langle \psi_k^{(0)} \big| \hat{\mathcal{H}}_1(t_2) \big| \psi_i^{(0)} \big\rangle\,. \end{aligned} \tag{G.2.12}$$

Since the intermediate states $\big|\psi_k^{(0)}\big\rangle$ constitute a complete set, the previous formula simplifies to

$$c_{ij}^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_0}^{t_1} \mathrm{d}t_2 \big\langle \psi_j^{(0)} \big| \hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2) \big| \psi_i^{(0)} \big\rangle \,. \tag{G.2.13}$$

The same result is obtained when the double integral on the t_1, t_2 plane is evaluated in reverse order:

$$c_{ij}^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_0}^{t} \mathrm{d}t_2 \int_{t_2}^{t} \mathrm{d}t_1 \big\langle \psi_j^{(0)} \big| \hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2) \big| \psi_i^{(0)} \big\rangle, \tag{G.2.14}$$

or by swapping the notation of the two time variables:

$$c_{ij}^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_1}^{t} \mathrm{d}t_2 \big\langle \psi_j^{(0)} \big| \hat{\mathcal{H}}_1(t_2)\hat{\mathcal{H}}_1(t_1) \big| \psi_i^{(0)} \big\rangle. \tag{G.2.15}$$

Using these two formulas, the coefficient can also be written as

$$c_{ij}^{(2)}(t) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \frac{1}{2} \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_0}^{t} \mathrm{d}t_2 \big\langle \psi_j^{(0)} \big| T\big\{\hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2)\big\} \big| \psi_i^{(0)} \big\rangle \,, \tag{G.2.16}$$

where T is the time-ordering operator, which orders the operators in a product in descending order of their time argument. Its action can be written in terms of the Heaviside step function as

$$T\big\{\hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2)\big\} = \theta(t_1 - t_2)\hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2) + \theta(t_2 - t_1)\hat{\mathcal{H}}_1(t_2)\hat{\mathcal{H}}_1(t_1) \,. \tag{G.2.17}$$

Generalizing this to arbitrary orders, and setting $\lambda = 1$,

$$\begin{aligned} c_{ij}(t) = \delta_{ji} + \sum_{n=1}^{\infty} \left(-\frac{\mathrm{i}}{\hbar}\right)^n \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_0}^{t_1} \mathrm{d}t_2 \ldots \int_{t_0}^{t_{n-1}} \mathrm{d}t_n \\ \times \big\langle \psi_j^{(0)} \big| \hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2)\ldots\hat{\mathcal{H}}_1(t_n) \big| \psi_i^{(0)} \big\rangle, \end{aligned} \tag{G.2.18}$$

or, in time-ordered form,

$$\begin{aligned} c_{ij}(t) = \delta_{ji} + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\mathrm{i}}{\hbar}\right)^n \int_{t_0}^{t} \mathrm{d}t_1 \int_{t_0}^{t} \mathrm{d}t_2 \ldots \int_{t_0}^{t} \mathrm{d}t_n \\ \times \big\langle \psi_j^{(0)} \big| T\big\{\hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2)\ldots\hat{\mathcal{H}}_1(t_n)\big\} \big| \psi_i^{(0)} \big\rangle. \end{aligned} \tag{G.2.19}$$

The time evolution of the wavefunction between times t_0 and t is therefore governed by the operator $S(t, t_0)$:

$$\big|\psi(t)\big\rangle = S(t, t_0)\big|\psi(t_0)\big\rangle \,, \tag{G.2.20}$$

where

$$S(t,t_0)=\sum_{n=0}^{\infty}\left(-\frac{\mathrm{i}}{\hbar}\right)^n\int\limits_{t_0}^{t}\mathrm{d}t_1\int\limits_{t_0}^{t_1}\mathrm{d}t_2\ldots\int\limits_{t_0}^{t_{n-1}}\mathrm{d}t_n\,\hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2)\ldots\hat{\mathcal{H}}_1(t_n)\,, \tag{G.2.21}$$

or

$$S(t,t_0)=\sum_{n=0}^{\infty}\frac{1}{n!}\left(-\frac{\mathrm{i}}{\hbar}\right)^n\int\limits_{t_0}^{t}\mathrm{d}t_1\int\limits_{t_0}^{t}\mathrm{d}t_2\ldots\int\limits_{t_0}^{t}\mathrm{d}t_n\,T\bigl\{\hat{\mathcal{H}}_1(t_1)\hat{\mathcal{H}}_1(t_2)\ldots\hat{\mathcal{H}}_1(t_n)\bigr\}\,. \tag{G.2.22}$$

Now consider a system whose initial wavefunction at $t_0=0$ is $\bigl|\psi_i^{(0)}\bigr\rangle$, and a constant perturbation that acts for a finite period of time. According to (G.2.7), the amplitude of the state $\bigl|\psi_j^{(0)}\bigr\rangle$ that becomes admixed to the initial state at time t is given by

$$c_{ij}^{(1)}=-\bigl\langle\psi_j^{(0)}\bigl|\mathcal{H}_1\bigr|\psi_i^{(0)}\bigr\rangle\frac{\mathrm{e}^{\mathrm{i}(E_j^{(0)}-E_i^{(0)})t/\hbar}-1}{E_j^{(0)}-E_i^{(0)}} \tag{G.2.23}$$

in the lowest order of perturbation theory. The transition probability from state $\bigl|\psi_i^{(0)}\bigr\rangle$ to $\bigl|\psi_j^{(0)}\bigr\rangle$ is then

$$W_{i\to j}=\bigl|c_{ij}^{(1)}\bigr|^2=2\Bigl|\bigl\langle\psi_j^{(0)}\bigl|\mathcal{H}_1\bigr|\psi_i^{(0)}\bigr\rangle\Bigr|^2\frac{1-\cos\bigl((E_j^{(0)}-E_i^{(0)})t/\hbar\bigr)}{\bigl(E_j^{(0)}-E_i^{(0)}\bigr)^2} \tag{G.2.24}$$

in the same order. For large values of the time, the formula on the right-hand side gives significant probabilities only for states whose energy difference is at most of order $2\pi\hbar/t$. Since

$$\lim_{t\to\infty}\frac{1-\cos(x-x_0)t}{(x-x_0)^2}=\pi t\delta(x-x_0)\,, \tag{G.2.25}$$

the transition rate in the $t\to\infty$ limit is

$$w_{i\to j}=\frac{\mathrm{d}W_{i\to j}}{\mathrm{d}t}=\frac{2\pi}{\hbar}\Bigl|\bigl\langle\psi_j^{(0)}\bigl|\mathcal{H}_1\bigr|\psi_i^{(0)}\bigr\rangle\Bigr|^2\delta(E_j^{(0)}-E_i^{(0)})\,. \tag{G.2.26}$$

As mentioned in the previous section, the formulas of time-dependent perturbation theory can also be used to specify the energy shifts due to a stationary perturbation, provided the interaction is assumed to be turned on adiabatically at $t_0=-\infty$. Inserting a factor $\exp(-\alpha|t|)$ in the interaction Hamiltonian, which specifies the adiabatic switch-on by means of an infinitesimally small α, we have

$$\hat{\mathcal{H}}_1(t)=\mathrm{e}^{\mathrm{i}\mathcal{H}_0t/\hbar}\mathcal{H}_1\mathrm{e}^{-\mathrm{i}\mathcal{H}_0t/\hbar}\mathrm{e}^{-\alpha|t|}\,. \tag{G.2.27}$$

in the interaction picture. Using the ground state $|\Psi_0\rangle$ of energy E_0 of the unperturbed system, the energy correction due to the perturbation is

$$\Delta E = \frac{\langle\Psi_0|\mathcal{H}_1 S(0,-\infty)|\Psi_0\rangle}{\langle\Psi_0|S(0,-\infty)|\Psi_0\rangle}, \tag{G.2.28}$$

and the wavefunction is

$$|\Psi\rangle = \frac{S(0,-\infty)|\Psi_0\rangle}{\langle\Psi_0|S(0,-\infty)|\Psi_0\rangle}. \tag{G.2.29}$$

As J. Goldstone (1957) pointed out, the same result may be formulated in a slightly different way. Considering a many-particle system with a nondegenerate ground state, the contribution of each term in the perturbation expansion can be represented by time-ordered diagrams that show the intermediate states through which the system gets back to the ground state. This representation contains terms in which some of the particles participating in the intermediate processes are in no way connected to the incoming and outgoing particles. It can be demonstrated that the contributions of the disconnected parts are exactly canceled by the denominator in (G.2.28) and (G.2.29), so

$$\begin{aligned}\Delta E &= \sum_{n=0}^{\infty} \left\langle\Psi_0\middle|\mathcal{H}_1 \left(\frac{1}{E_0-\mathcal{H}_0}\mathcal{H}_1\right)^n \middle|\Psi_0\right\rangle_{\text{con}}, \\ |\Psi\rangle &= \sum_{n=0}^{\infty} \left(\frac{1}{E_0-\mathcal{H}_0}\mathcal{H}_1\right)^n |\Psi_0\rangle_{\text{con}},\end{aligned} \tag{G.2.30}$$

where the label "con" indicates that only the contribution of connected diagrams need to be taken into account. It should be noted that instead of Goldstone's time-ordered diagrams, the perturbation series for the ground-state energy can also be represented in terms of Feynman diagrams, which are more commonly used in the many-body problem. Only connected diagrams need to be considered in that representation, too.

Reference

1. C. Cohen-Tannoudji, B. Diu and F. Laloë, *Quantum Mechanics*, John Wiley & Sons, New York (1977).

H

Second Quantization

The quantum mechanical wavefunction is most often considered as the function of the space and time variables when the solutions of the Schrödinger equation are sought. In principle, this approach is applicable even when the system is made up of a large number of interacting particles. However, it is then much more convenient to use the occupation-number representation for the wavefunction. We shall introduce the creation and annihilation operators, and express the Hamiltonian in terms of them, too.

H.1 Occupation-Number Representation

It was mentioned in Chapter 12 on the quantum mechanical treatment of lattice vibrations that the eigenstates of the harmonic oscillator can be characterized by the quantum number n that can take nonnegative integer values. Using the linear combinations of the position variable x and its conjugate momentum, it is possible to construct operators $a^\dagger$ and a that increase and decrease this quantum number. We may say that when these ladder operators are applied to an eigenstate, they create an additional quantum or annihilate an existing one. Consequently, these operators are called the creation and annihilation operators of the elementary quantum or excitation. States can be characterized by the number of quanta they contain – that is, by the occupation number. Using DIRAC's notation, the state ψ_n of quantum number n – which can be constructed from the ground state of the oscillator by the n-fold application of the creation operator $a^\dagger$, and thus contains n quanta – will henceforth be denoted by $|n\rangle$. The requirement that such states should also be normalized to unity leads to

$$a|n\rangle = \sqrt{n}\,|n-1\rangle\,, \qquad a^\dagger|n\rangle = \sqrt{n+1}\,|n+1\rangle\,. \tag{H.1.1}$$

The operators a and $a^\dagger$ of the quantum mechanical oscillator satisfy the bosonic commutation relation.

This occupation-number representation can be equally applied to many-particle systems made up of fermions (e.g., electrons) or bosons (e.g., phonons, magnons). Any state of an interacting system consisting of N particles can be expanded in terms of the complete set of states of the noninteracting system. The eigenstates of the noninteracting many-particle system can, in turn, be expressed in terms of the one-particle eigenstates. When the one-particle problem is solved for the noninteracting system, the complete set $\phi_1(\xi), \phi_2(\xi), \dots, \phi_i(\xi), \dots$ of one-particle states is obtained, where the collective notation ξ is used for the spatial variable $\boldsymbol{r}$ and spin s of the particles: $\xi = (\boldsymbol{r}, s)$.

Such complete sets are the set of eigenfunctions for the harmonic oscillator, and the system of plane waves, Bloch functions, or Wannier functions for electrons. The construction of the complete set of many-particle functions from one-particle functions is different for bosons and fermions: for bosons, several particles may be in the same state, whereas this possibility is excluded by the Pauli principle for fermions. The two cases must therefore be treated separately.

Bosons

For noninteracting bosons the states of the many-particle system are described by means of those combinations of the one-particle states that are completely symmetric with respect to the interchange of the space and spin variables. If an N-particle system contains n_1, n_2, $\dots, n_k, \dots$ particles in the states ϕ_1, $\phi_2, \dots, \phi_k, \dots$, where

$$\sum_k n_k = N, \tag{H.1.2}$$

the wavefunction with the required symmetry properties is

$$\varPhi_{n_1,n_2,\dots,n_k,\dots} = \left(\frac{n_1! n_2! \dots n_k! \dots}{N!} \right)^{1/2} \sum_P \phi_{p_1}(\xi_1)\phi_{p_2}(\xi_2) \dots \phi_{p_\mathrm{N}}(\xi_\mathrm{N}) \,, \tag{H.1.3}$$

where $1, 2, \dots, k, \dots$ occurs among the indices p_i exactly $n_1, n_2, \dots, n_k, \dots$ times, and summation is over all possible permutations of the indices.

It turns out practical to introduce a more concise notation. If the wavefunctions of the one-particles states are known, the wavefunction $\varPhi$ is unambiguously characterized by the numbers $n_1, n_2, \dots, n_k, \dots$ that specify the occupation of each one-particle state, therefore the above-defined state can be concisely denoted by

$$\varPhi_{n_1,n_2,\dots,n_k,\dots} \equiv |n_1, n_2, \dots, n_k, \dots\rangle \,. \tag{H.1.4}$$

This is the occupation-number representation, while the vector space spanned by the set of all such basis states with nonnegative integers n_k for bosons is called the Fock space.

We shall now define the creation and annihilation operators that act in the Fock space and increase and decrease the occupation number by one:

$$\begin{aligned} a_k^\dagger |n_1, n_2, \dots, n_k, \dots\rangle &= f_\mathrm{c} |n_1, n_2, \dots, n_k + 1, \dots\rangle\,, \\ a_k |n_1, n_2, \dots, n_k, \dots\rangle &= f_\mathrm{a} |n_1, n_2, \dots, n_k - 1, \dots\rangle\,. \end{aligned} \tag{H.1.5}$$

If the normalization factors for bosons are chosen the same way as for harmonic oscillators, that is,

$$\begin{aligned} a_k^\dagger |n_1, n_2, \dots, n_k, \dots\rangle &= \sqrt{n_k + 1} |n_1, n_2, \dots, n_k + 1, \dots\rangle\,, \\ a_k |n_1, n_2, \dots, n_k, \dots\rangle &= \sqrt{n_k} |n_1, n_2, \dots, n_k - 1, \dots\rangle\,, \end{aligned} \tag{H.1.6}$$

then $\hat{n}_k = a_k^\dagger a_k$ is the number operator that gives the occupation number of the state of index k, since

$$a_k^\dagger a_k |n_1, n_2, \dots, n_k, \dots\rangle = n_k |n_1, n_2, \dots, n_k, \dots\rangle\,, \tag{H.1.7}$$

and the commutation relations are the usual ones for bosons:

$$a_k a_{k'}^\dagger - a_{k'}^\dagger a_k = \delta_{kk'}\,. \tag{H.1.8}$$

Any state $|n_1, n_2, \dots, n_k, \dots\rangle$ can be constructed from the vacuum by means of creation operators:

$$\varPhi_{n_1, n_2, \dots, n_k, \dots} = \frac{1}{\sqrt{n_1! n_2! \dots n_k! \dots}} \left(a_1^\dagger\right)^{n_1} \left(a_2^\dagger\right)^{n_2} \dots \left(a_k^\dagger\right)^{n_k} \dots |0\rangle. \tag{H.1.9}$$

As has been mentioned, these states make up a complete set, and the wavefunctions of interacting many-particle system can be expressed as linear combinations of them.

Fermions

A similar approach can be adopted for fermions, however, the many-particle wavefunction has to be chosen as

$$\varPhi_{n_1, n_2, \dots, n_k, \dots} = \frac{1}{\sqrt{N!}} \sum_P (-1)^P \phi_{p_1}(\xi_1) \phi_{p_2}(\xi_2) \dots \phi_{p_\mathrm{N}}(\xi_\mathrm{N}) \tag{H.1.10}$$

to meet the requirement of complete antisymmetrization. This is equivalent to building a Slater determinant from the one-particle wavefunctions:

$$\varPhi_{n_1, n_2, \dots, n_k, \dots} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_{p_1}(\xi_1) & \phi_{p_2}(\xi_1) & \dots & \phi_{p_\mathrm{N}}(\xi_1) \\ \phi_{p_1}(\xi_2) & \phi_{p_2}(\xi_2) & \dots & \phi_{p_\mathrm{N}}(\xi_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{p_1}(\xi_\mathrm{N}) & \phi_{p_2}(\xi_\mathrm{N}) & \dots & \phi_{p_\mathrm{N}}(\xi_\mathrm{N}) \end{vmatrix}. \tag{H.1.11}$$

Since each one-particle state can occur at most once, when the states are ordered in some arbitrary way, the product in (H.1.10) with indices

$$p_1 < p_2 < \cdots < p_{\mathrm{N}} \tag{H.1.12}$$

is chosen with positive sign, and the signs for other configurations follow from the parity of the permutation.

The occupation-number representation can be used for fermions as well. The wavefunction is then written in Fock space as

$$\Phi_{n_1,n_2,\ldots,n_k,\ldots} \equiv |n_1,n_2,\ldots,n_k,\ldots\rangle\,, \tag{H.1.13}$$

where n_i can be either 0 or 1. The creation and annihilation operators must be introduced in such a way that the equations

$$\begin{aligned} a_k^\dagger|n_1,n_2,\ldots,n_k,\ldots\rangle &= 0\,, \qquad \text{if} \qquad n_k = 1\,,\\ a_k|n_1,n_2,\ldots,n_k,\ldots\rangle &= 0\,, \qquad \text{if} \qquad n_k = 0 \end{aligned} \tag{H.1.14}$$

be satisfied. After ordering the one-particle states, the normalization of the states obtained by the application of the creation and annihilation operators are chosen as

$$\begin{aligned} a_k^\dagger|n_1,n_2,\ldots,n_k,\ldots\rangle &= \sqrt{1-n_k}(-1)^{S_k}|n_1,n_2,\ldots,n_k+1,\ldots\rangle\,,\\ a_k|n_1,n_2,\ldots,n_k,\ldots\rangle &= \sqrt{n_k}(-1)^{S_k}|n_1,n_2,\ldots,n_k-1,\ldots\rangle\,, \end{aligned} \tag{H.1.15}$$

where

$$S_k = \sum_{i<k} n_i\,. \tag{H.1.16}$$

With this choice $\hat{n}_k = a_k^\dagger a_k$ is the number operator for fermions as well, since when it acts on the state $|n_1,n_2,\ldots,n_k,\ldots\rangle$

$$a_k^\dagger a_k|n_1,n_2,\ldots,n_k,\ldots\rangle = n_k|n_1,n_2,\ldots,n_k,\ldots\rangle\,. \tag{H.1.17}$$

In reverse order, however,

$$a_k a_k^\dagger|n_1,n_2,\ldots,n_k,\ldots\rangle = \sqrt{(1+n_k)(1-n_k)}|n_1,n_2,\ldots,n_k,\ldots\rangle\,. \tag{H.1.18}$$

Since n_k can only take the values 0 and 1, the eigenvalue of $a_k a_k^\dagger$ is $1-n_k$, and thus the operator identity

$$a_k a_k^\dagger + a_k^\dagger a_k = 1 \tag{H.1.19}$$

holds. By taking states of different quantum numbers, if the state of quantum number k precedes the state with quantum number k' in the order, we have

$$\begin{aligned} &a_k a_{k'}^\dagger|n_1,n_2,\ldots,n_k,\ldots,n_{k'},\ldots\rangle \\ &\quad = (-1)^{S_k}(-1)^{S_{k'}}\sqrt{n_k}\sqrt{1-n_{k'}}|n_1,n_2,\ldots,n_k-1,\ldots,n_{k'}+1,\ldots\rangle, \end{aligned} \tag{H.1.20}$$

while for the reverse order of the operators the -1 factors due to antisymmetrization are different:

$$\begin{aligned} & a^\dagger_{k'} a_k |n_1, n_2, \dots, n_k, \dots, n_{k'}, \dots\rangle \\ & \quad = (-1)^{S_k} (-1)^{S_{k'}-1} \sqrt{n_k} \sqrt{1-n_{k'}} |n_1, n_2, \dots, n_k - 1, \dots, n_{k'}+1, \dots\rangle , \end{aligned} \tag{H.1.21}$$

and thus

$$[a_k, a^\dagger_{k'}]_+ \equiv a_k a^\dagger_{k'} + a^\dagger_{k'} a_k = \delta_{kk'} \, , \tag{H.1.22}$$

where $[A, B]_+$ is the anticommutator of the two operators. Likewise, it can be shown that

$$[a_k, a_{k'}]_+ = 0 \, , \qquad [a^\dagger_k, a^\dagger_{k'}]_+ = 0 \, . \tag{H.1.23}$$

The state Φ in which the one-particle states of index $p_1 < p_2 < \dots < p_\mathrm{N}$ are filled can be written as

$$\Phi = a^\dagger_{p_1} a^\dagger_{p_2} \dots a^\dagger_{p_\mathrm{N}} |0\rangle \tag{H.1.24}$$

in terms of the creation operators, where $|0\rangle$ is the vacuum state.

H.2 Second-Quantized Form of Operators

In the discussion of many-particle systems we mostly encounter operators that are the sums of terms acting on individual particles or contain the variables of two particles. The kinetic energy of a system and the interaction with an applied field are examples for the first, while pair interaction between the particles is an example of the second. Below we shall show that the one- and two-particle operators can be expressed in simple forms in terms of the creation and annihilation operators. Equivalence is based on the requirement that their action on the wavefunctions given in occupation-number representation lead to the same matrix elements as the usual representation.

H.2.1 Second-Quantized Form of One-Particle Operators

We shall first discuss one-particle operators. In complete generality, they can be written as

$$F^{(1)} = \sum_{i=1}^{N} f(\xi_i) \, . \tag{H.2.1}$$

The operator f either leaves the particle in the same state or takes it into another. We shall first consider diagonal matrix elements. For bosons, each particle gives the same contribution because of symmetrization. By selecting a particle and assuming that it is in the state of label l,

$$\begin{aligned}\iint \dots \int \Phi^*_{n_1,n_2,\dots,n_k,\dots} \sum_{i=1}^{N} f(\xi_i) \Phi_{n_1,n_2,\dots,n_k,\dots} \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 \dots \mathrm{d}\xi_{\mathrm{N}} \\ = N \frac{n_1! n_2! \dots n_k! \dots}{N!} \sum_l \int \phi_l^*(\xi) f(\xi) \phi_l(\xi) \, \mathrm{d}\xi \qquad \text{(H.2.2)} \\ \times \sum_{P'} \int \dots \int \phi^*_{p_1}(\xi_2) \dots \phi^*_{p_N}(\xi_N) \phi_{p_1}(\xi_2) \dots \phi_{p_N}(\xi_N) \, \mathrm{d}\xi_2 \, \dots \, \mathrm{d}\xi_N \, .\end{aligned}$$

To calculate the factor that remains after the separation of the matrix element of the state l, only those states need to be considered in the permutation P' that contain the state of label l only $n_l - 1$ times. Owing to the orthonormality of the one-particle states, the value of the previous formula is

$$\sum_l n_l \int \phi_l^*(\xi) f(\xi) \phi_l(\xi) \, \mathrm{d}\xi \, . \qquad \text{(H.2.3)}$$

In the off-diagonal terms nonzero matrix elements are obtained between those states $\Phi_{n_1,n_2,\dots,n_k,\dots,n_l,\dots}$ and $\Phi_{n_1,n_2,\dots,n_k+1,\dots,n_l-1,\dots}$ for which the occupation numbers of two one-particle states differ by one unit each. Then

$$\iint \dots \int \Phi^*_{n_1,n_2,\dots,n_k+1,\dots,n_l-1,\dots} \sum_i f(\xi_i) \Phi_{n_1,n_2,\dots,n_k,\dots,n_l,\dots} \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 \dots \mathrm{d}\xi_N \, . \qquad \text{(H.2.4)}$$

Because of the normalization factors of the two wavefunctions the matrix element is proportional to

$$I = \left(\frac{n_1! n_2! \dots (n_k+1)! \dots (n_l-1)!}{N!} \right)^{1/2} \left(\frac{n_1! n_2! \dots n_k! \dots n_l!}{N!} \right)^{1/2} . \qquad \text{(H.2.5)}$$

Since each particle contributes by the same amount, the matrix element is

$$\begin{aligned}\iint \dots \int \Phi^*_{n_1,n_2,\dots,n_k+1,\dots,n_l-1,\dots} \sum_i f(\xi_i) \Phi_{n_1,n_2,\dots,n_k,\dots,n_l,\dots} \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 \dots \mathrm{d}\xi_{\mathrm{N}} \\ = N I \sum_{kl} \int \phi_k^*(\xi) f(\xi) \phi_l(\xi) \, \mathrm{d}\xi \qquad \text{(H.2.6)} \\ \times \sum_{P'} \int \dots \int \phi^*_{p_1}(\xi_2) \dots \phi^*_{p_N}(\xi_N) \phi_{p_1}(\xi_2) \dots \phi_{p_N}(\xi_N) \, \mathrm{d}\xi_2 \, \dots \, \mathrm{d}\xi_N \, .\end{aligned}$$

After the separation of the integral for the selected particle, the remaining terms correspond to a state that contains $N-1$ particles, with occupation numbers $n_1, n_2, \dots, n_k, \dots, n_l - 1, \dots$. Since there are

$$\frac{(N-1)!}{n_1! n_2! \dots n_k! \dots (n_l-1)! \dots} \qquad \text{(H.2.7)}$$

such states, the separation of the ξ-integral leaves behind a factor $\sqrt{n_k+1}\sqrt{n_l}$, so the matrix element is

$$\sqrt{n_k+1}\sqrt{n_l}\int \phi_k^*(\xi) f(\xi)\phi_l(\xi)\,\mathrm{d}\xi\,. \tag{H.2.8}$$

The same expressions are obtained for the diagonal and off-diagonal matrix elements if the states are specified in occupation-number representation, the operator $F^{(1)}$ is chosen as

$$F^{(1)} = \sum_{kl} a_k^\dagger f_{kl} a_l\;, \tag{H.2.9}$$

where

$$f_{kl} = \int \phi_k^*(\xi) f(\xi)\phi_l(\xi)\,\mathrm{d}\xi\,, \tag{H.2.10}$$

and the previously obtained relations for the action of the creation and annihilation operators are used in the calculation of the matrix element. Therefore the operator given in (H.2.9), which acts in the Fock space, is the second-quantized form of one-particle operators for bosons. Note that while the sum is over N particles in the first-quantized formula (H.2.1) of the one-particle operator, it is over the quantum numbers of the one-particle states in the second-quantized formula.

The intermediate steps are slightly different for fermions, since a Slater determinant wavefunction is specified in terms of the occupation numbers, and the normalization factors are also different – nevertheless the final result is the same: the one-particle operators for fermions can again be represented as (H.2.9) in terms of creation and annihilation operators.

H.2.2 Second-Quantized Form of Two-Particle Operators

This approach can be extended to the two-body interaction term in the Hamiltonian and similar operators that are the sums of terms containing the coordinates of two particles:

$$F^{(2)} = \sum_{ij} f(\xi_i,\xi_j)\,. \tag{H.2.11}$$

Since the variables of two particles appear in each term, such an operator has a nonvanishing matrix element only between states for which the occupation numbers of at most four one-particle states change: two decrease and two others increase by one. The matrix element to be evaluated is thus

$$\begin{aligned}\iint\cdots\int \Phi^*_{n_1,\dots,n_k+1,\dots,n_l+1,\dots,n_m-1,\dots,n_n-1,\dots} \sum_{ij} f(\xi_i,\xi_j)&\\ \Phi_{n_1,\dots,n_k,\dots,n_l,\dots,n_m,\dots,n_n,\dots}\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2\dots\mathrm{d}\xi_N\,.&\end{aligned} \tag{H.2.12}$$

Intuitively, we may say that one particle is taken from state $\phi_m(\xi_i)$ to $\phi_k(\xi_i)$, and the other from state $\phi_n(\xi_j)$ to $\phi_l(\xi_j)$. Because of the indistinguishability of the particles, other combinations occur with the same weight. Besides, there are nonvanishing matrix elements for processes in which the occupation changes for three or just two states. A lengthy but straightforward calculation shows that the same matrix elements are obtained if the wavefunction is chosen in the occupation-number representation and the two-particle operator is written in the form

$$F^{(2)} = \sum_{klmn} f_{klmn} a_k^\dagger a_l^\dagger a_m a_n \,, \tag{H.2.13}$$

where

$$f_{klmn} = \int \phi_k^*(\xi_1)\phi_l^*(\xi_2) f(\xi_1,\xi_2)\phi_m(\xi_2)\phi_n(\xi_1)\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 \,. \tag{H.2.14}$$

Therefore the operator (H.2.13), which acts in the Fock space, is the second-quantized form of two-particle operators for bosons.

Once again, the intermediate steps are slightly different for fermions because there are no processes with doubly occupied states. Nevertheless the second-quantized form of two-particle operators remains the same as above.

H.2.3 Field Operators

Using the one-particle wavefunction $\phi_k(\xi)$ of the state of quantum number k, it is customary to introduce the field operators

$$\hat{\psi}(\xi) = \sum_k \phi_k(\xi) a_k \,, \qquad \hat{\psi}^\dagger(\xi) = \sum_k \phi_k^*(\xi) a_k^\dagger \tag{H.2.15}$$

that are defined in real space. The commutation relations of creation and annihilation operators and the completeness relations of the one-particle functions imply

$$\begin{aligned} \left[\hat{\psi}(\xi), \hat{\psi}^\dagger(\xi')\right] &= \sum_{k,k'} \phi_k(\xi)\phi_{k'}^*(\xi')\left[a_k, a_{k'}^\dagger\right] \\ &= \sum_k \phi_k(\xi)\phi_k^*(\xi') = \delta(\xi-\xi') \end{aligned} \tag{H.2.16}$$

for bosons and

$$\begin{aligned} \left[\hat{\psi}(\xi), \hat{\psi}^\dagger(\xi')\right]_+ &= \sum_{k,k'} \phi_k(\xi)\phi_{k'}^*(\xi')\left[a_k, a_{k'}^\dagger\right]_+ \\ &= \sum_k \phi_k(\xi)\phi_k^*(\xi') = \delta(\xi-\xi') \end{aligned} \tag{H.2.17}$$

for fermions. It can also be shown that by taking a commutator for bosons and an anticommutator for fermions,

$$\left[\hat{\psi}(\xi), \hat{\psi}(\xi')\right]_{\mp} = \left[\hat{\psi}^{\dagger}(\xi), \hat{\psi}^{\dagger}(\xi')\right]_{\mp} = 0\,. \tag{H.2.18}$$

When the spin variable is separated,

$$\left[\hat{\psi}_{\alpha}(\boldsymbol{r}), \hat{\psi}^{\dagger}_{\beta}(\boldsymbol{r}')\right]_{\mp} = \delta_{\alpha\beta}\delta(\boldsymbol{r}-\boldsymbol{r}')\,, \tag{H.2.19}$$

and

$$\left[\hat{\psi}_{\alpha}(\boldsymbol{r}), \hat{\psi}_{\beta}(\boldsymbol{r}')\right]_{\mp} = 0\,, \qquad \left[\hat{\psi}^{\dagger}_{\alpha}(\boldsymbol{r}), \hat{\psi}^{\dagger}_{\beta}(\boldsymbol{r}')\right]_{\mp} = 0\,. \tag{H.2.20}$$

The creation and annihilation operators $a_k^{\dagger}$ and a_k change the occupation of a one-particle state of a given quantum number. Below we shall see that field operators can be interpreted as operators that create and annihilate a particle at ξ. In other words, the state $\hat{\psi}^{\dagger}_{\sigma}(\boldsymbol{r})|0\rangle$ contains a particle of spin quantum number σ at point $\boldsymbol{r}$ of the real space. Similarly, $\hat{\psi}^{\dagger}_{\sigma}(\boldsymbol{r})\hat{\psi}_{\sigma}(\boldsymbol{r})$ is the density operator of spin-σ particles at $\boldsymbol{r}$.

The one- and two-particle operators can then be rewritten as

$$F^{(1)} = \int \mathrm{d}\xi\, \hat{\psi}^{\dagger}(\xi) f^{(1)}(\xi)\hat{\psi}(\xi)\,, \tag{H.2.21}$$

and

$$F^{(2)} = \int \mathrm{d}\xi_1 \int \mathrm{d}\xi_2 \hat{\psi}^{\dagger}(\xi_1)\hat{\psi}^{\dagger}(\xi_2) f^{(2)}(\xi_1,\xi_2)\hat{\psi}(\xi_2)\hat{\psi}(\xi_1)\,. \tag{H.2.22}$$

In this representation the one-particle operator has exactly the same form as the expectation value of the one-particle operator in first quantization, except that the wavefunction and it complex conjugate are replaced by the field operator and its Hermitian adjoint – hence the name second quantization.

H.2.4 Second-Quantized Form of the Electronic Hamiltonian

Among the terms of the Hamiltonian, the kinetic energy and the potential are one-particle operators, and the pair interaction is a two-particle operator. For electrons, when the potential and the pair interaction are spin independent,

$$\mathcal{H} = \sum_i \mathcal{H}_i^{(1)} + \tfrac{1}{2}\sum_{ij} U^{(2)}(\boldsymbol{r}_i, \boldsymbol{r}_j)\,, \tag{H.2.23}$$

where the one-particle part contains

$$\mathcal{H}_i^{(1)} = -\frac{\hbar^2}{2m_{\mathrm{e}}}\boldsymbol{\nabla}_i^2 + U(\boldsymbol{r}_i)\,. \tag{H.2.24}$$

By taking a complete set of one-particle states, and denoting, as customary for electrons, the creation and annihilation operators of a particle in state $\phi_k(\xi)$ by $c_k^{\dagger}$ and c_k instead of $a_k^{\dagger}$ and a_k, we have

$$\mathcal{H} = \sum_{kl} H_{kl} c_k^\dagger c_l + \tfrac{1}{2} \sum_{klmn} U_{klmn}^{(2)} c_k^\dagger c_l^\dagger c_m c_n \,, \tag{H.2.25}$$

where

$$H_{kl} = \int \phi_k^*(\xi) \left(-\frac{\hbar^2}{2m_\mathrm{e}} \boldsymbol{\nabla}^2 + U(\boldsymbol{r}) \right) \phi_l(\xi) \, \mathrm{d}\xi \,, \tag{H.2.26}$$

and

$$U_{klmn}^{(2)} = \iint \mathrm{d}\xi \, \mathrm{d}\xi' \, \phi_k^*(\xi) \phi_l^*(\xi') U^{(2)}(\boldsymbol{r}, \boldsymbol{r}') \phi_m(\xi') \phi_n(\xi) \,. \tag{H.2.27}$$

In general, the states are chosen in such a way that the one-particle part be diagonal. This is the case when the Bloch functions determined in the presence of a periodic potential are used as a complete basis set. However, this is not the only option. In the Hubbard model the Wannier states are used, and so the one-particle term in the Hamiltonian that describes the hopping of electrons between lattice points is not diagonal.

Using the field operators instead of the creation and annihilation operators,

$$\begin{aligned} \mathcal{H} = & \int \hat{\psi}^\dagger(\xi) \left(-\frac{\hbar^2}{2m_\mathrm{e}} \boldsymbol{\nabla}^2 + U(\boldsymbol{r}) \right) \hat{\psi}(\xi) \, \mathrm{d}\xi \\ & + \iint \mathrm{d}\xi \, \mathrm{d}\xi' \, \hat{\psi}^\dagger(\xi) \hat{\psi}^\dagger(\xi') U^{(2)}(\boldsymbol{r}, \boldsymbol{r}') \hat{\psi}(\xi') \hat{\psi}(\xi) \,. \end{aligned} \tag{H.2.28}$$

Writing out the spin variable explicitly, the spin independence of the potential and of the interaction implies

$$\begin{aligned} \mathcal{H} = & \sum_\sigma \int \hat{\psi}_\sigma^\dagger(\boldsymbol{r}) \left(-\frac{\hbar^2}{2m_\mathrm{e}} \boldsymbol{\nabla}^2 + U(\boldsymbol{r}) \right) \hat{\psi}_\sigma(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} \\ & + \sum_{\sigma\sigma'} \iint \mathrm{d}\boldsymbol{r} \, \mathrm{d}\boldsymbol{r}' \, \hat{\psi}_\sigma^\dagger(\boldsymbol{r}) \hat{\psi}_{\sigma'}^\dagger(\boldsymbol{r}') U^{(2)}(\boldsymbol{r}, \boldsymbol{r}') \hat{\psi}_{\sigma'}(\boldsymbol{r}') \hat{\psi}_\sigma(\boldsymbol{r}) \,. \end{aligned} \tag{H.2.29}$$

The description is highly simplified by choosing the plane waves as the complete set. The one-particle states are then characterized by the wave vector $\boldsymbol{k}$ and the spin quantum number σ. The usual formula

$$\mathcal{H}_\mathrm{kin} = -\sum_i \frac{\hbar^2}{2m_\mathrm{e}} \frac{\partial^2}{\partial \boldsymbol{r}_i^2} \tag{H.2.30}$$

for the kinetic energy can be rewritten in second-quantized form as

$$\mathcal{H}_\mathrm{kin} = \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma\sigma'} c_{\boldsymbol{k}\sigma}^\dagger H_{\sigma\sigma'}(\boldsymbol{k}, \boldsymbol{k}') c_{\boldsymbol{k}'\sigma'} \,, \tag{H.2.31}$$

where

$$\begin{aligned} H_{\sigma\sigma'}(\boldsymbol{k},\boldsymbol{k}') &= \frac{1}{V}\int \mathrm{d}\boldsymbol{r}\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\left(-\frac{\hbar^2}{2m_\mathrm{e}}\right)\frac{\partial^2}{\partial\boldsymbol{r}^2}\mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}}\delta_{\sigma\sigma'} \\ &= \frac{\hbar^2\boldsymbol{k}^2}{2m_\mathrm{e}}\delta_{\boldsymbol{k}\boldsymbol{k}'}\delta_{\sigma\sigma'}\,, \end{aligned} \tag{H.2.32}$$

and so

$$\mathcal{H}_\mathrm{kin} = \sum_{\boldsymbol{k}\sigma}\frac{\hbar^2\boldsymbol{k}^2}{2m_\mathrm{e}}c^\dagger_{\boldsymbol{k}\sigma}c_{\boldsymbol{k}\sigma}\,. \tag{H.2.33}$$

The second-quantized form of the one-particle potential $U(\boldsymbol{r})$ contains the Fourier transform of the potential:

$$\mathcal{H}_U = \frac{1}{V}\sum_{\boldsymbol{k}\boldsymbol{k}'\sigma}U(\boldsymbol{k}-\boldsymbol{k}')c^\dagger_{\boldsymbol{k}\sigma}c_{\boldsymbol{k}'\sigma} = \frac{1}{V}\sum_{\boldsymbol{k}\boldsymbol{q}}U(\boldsymbol{q})c^\dagger_{\boldsymbol{k}+\boldsymbol{q}\sigma}c_{\boldsymbol{k}\sigma}\,. \tag{H.2.34}$$

For a spin-independent two-particle interaction $U^{(2)}(\boldsymbol{r}_i-\boldsymbol{r}_j)$ the second-quantized form is

$$\mathcal{H}_\mathrm{int} = \tfrac{1}{2}\sum_{\substack{\boldsymbol{k}_1\boldsymbol{k}_2\boldsymbol{k}_3\boldsymbol{k}_4\\ \sigma\sigma'}} U^{(2)}(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4)c^\dagger_{\boldsymbol{k}_1\sigma}c^\dagger_{\boldsymbol{k}_2\sigma'}c_{\boldsymbol{k}_3\sigma'}c_{\boldsymbol{k}_4\sigma}\,, \tag{H.2.35}$$

where

$$\begin{aligned} U^{(2)}(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4) &= \frac{1}{V^2}\int\mathrm{d}\boldsymbol{r}_1\int\mathrm{d}\boldsymbol{r}_2\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_1\cdot\boldsymbol{r}_1}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_2\cdot\boldsymbol{r}_2} \\ &\qquad\times U^{(2)}(\boldsymbol{r}_1-\boldsymbol{r}_2)\mathrm{e}^{\mathrm{i}\boldsymbol{k}_3\cdot\boldsymbol{r}_2}\mathrm{e}^{\mathrm{i}\boldsymbol{k}_4\cdot\boldsymbol{r}_1} \\ &= \frac{1}{V^2}\int\mathrm{d}\boldsymbol{r}_1\int\mathrm{d}\boldsymbol{r}_2\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_1\cdot\boldsymbol{r}_1}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_2\cdot\boldsymbol{r}_2} \\ &\qquad\times\frac{1}{V}\sum_{\boldsymbol{q}}U^{(2)}(\boldsymbol{q})\mathrm{e}^{\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{r}_1-\boldsymbol{r}_2)}\mathrm{e}^{\mathrm{i}\boldsymbol{k}_3\cdot\boldsymbol{r}_2}\mathrm{e}^{\mathrm{i}\boldsymbol{k}_4\cdot\boldsymbol{r}_1} \\ &= \frac{1}{V}\sum_{\boldsymbol{q}}U^{(2)}(\boldsymbol{q})\delta_{\boldsymbol{k}_1,\boldsymbol{k}_4+\boldsymbol{q}}\delta_{\boldsymbol{k}_2,\boldsymbol{k}_3-\boldsymbol{q}}\,. \end{aligned} \tag{H.2.36}$$

By renaming the indices, the interaction term can be written as

$$\mathcal{H}_\mathrm{int} = \frac{1}{2V}\sum_{\substack{\boldsymbol{k}\boldsymbol{k}'\boldsymbol{q}\\ \sigma\sigma'}}U^{(2)}(\boldsymbol{q})c^\dagger_{\boldsymbol{k}+\boldsymbol{q}\sigma}c^\dagger_{\boldsymbol{k}'-\boldsymbol{q}\sigma'}c_{\boldsymbol{k}'\sigma'}c_{\boldsymbol{k}\sigma}\,. \tag{H.2.37}$$

If the one-particle periodic potential is taken into account by using Bloch states instead of plane waves, and the corresponding creation and annihilation operators $c^\dagger_{n\boldsymbol{k}\sigma}$ and $c_{n\boldsymbol{k}\sigma}$, then the entire one-particle part of the Hamiltonian – the kinetic energy plus the one-particle potential – can be diagonalized. This leads to

$$\mathcal{H}_\mathrm{kin}+\mathcal{H}_U = \sum_{n\boldsymbol{k}\sigma}\varepsilon_{n\boldsymbol{k}}c^\dagger_{n\boldsymbol{k}\sigma}c_{n\boldsymbol{k}\sigma}\,, \tag{H.2.38}$$

where $\varepsilon_{n\boldsymbol{k}}$ is the energy of Bloch electrons in the presence of the periodic potential. The interaction is not restricted to electrons in the same band. Electrons from different bands can be scattered to other bands provided the quasimomentum is conserved to within an additive reciprocal-lattice vector.

H.2.5 Number-Density and Spin-Density Operators

In the first-quantized formulation the number density of spinless particles is given by

$$n(\boldsymbol{r}) = \sum_l \delta(\boldsymbol{r} - \boldsymbol{r}_l)\,. \tag{H.2.39}$$

Its Fourier transform is

$$n(\boldsymbol{q}) = \int \mathrm{d}\boldsymbol{r}\, n(\boldsymbol{r}) \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} = \sum_l \int \mathrm{d}\boldsymbol{r}\, \delta(\boldsymbol{r} - \boldsymbol{r}_l) \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} = \sum_l \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}_l}\,. \tag{H.2.40}$$

In terms of plane-wave-creation and -annihilation operators, the general rule for one-particle operators implies

$$n(\boldsymbol{q}) = \sum_{\boldsymbol{k},\boldsymbol{k}'} c^\dagger_{\boldsymbol{k}} n(\boldsymbol{k}, \boldsymbol{k}') c_{\boldsymbol{k}'}\,, \tag{H.2.41}$$

where

$$n(\boldsymbol{k}, \boldsymbol{k}') = \frac{1}{V} \int \mathrm{d}\boldsymbol{r}\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\cdot\boldsymbol{r}} = \delta_{\boldsymbol{k}',\boldsymbol{k}+\boldsymbol{q}}\,. \tag{H.2.42}$$

A part of the sum can then be evaluated; it yields

$$n(\boldsymbol{q}) = \sum_{\boldsymbol{k}} c^\dagger_{\boldsymbol{k}} c_{\boldsymbol{k}+\boldsymbol{q}}\,. \tag{H.2.43}$$

Using an inverse Fourier transform it can be shown that the density operator in real space can be expressed particularly simply in terms of the field operator:

$$n(\boldsymbol{r}) = \hat{\psi}^\dagger(\boldsymbol{r}) \hat{\psi}(\boldsymbol{r})\,. \tag{H.2.44}$$

For particles with spin, an additional sum over the spin quantum number appears:

$$n(\boldsymbol{q}) = \sum_{\boldsymbol{k}\sigma} c^\dagger_{\boldsymbol{k}\sigma} c_{\boldsymbol{k}+\boldsymbol{q}\sigma}\,, \tag{H.2.45}$$

while the number-density operator is given in real space by

$$n(\boldsymbol{r}) = \sum_\sigma \hat{\psi}^\dagger_\sigma(\boldsymbol{r}) \hat{\psi}_\sigma(\boldsymbol{r})\,. \tag{H.2.46}$$

We can now show that the field operator $\hat{\psi}^\dagger_\sigma(\boldsymbol{r})$ indeed adds a spin-σ particle to the system at $\boldsymbol{r}$. To this end, we shall rewrite the operator $n(\boldsymbol{r})$ as

$$n(\boldsymbol{r}) = \sum_{\sigma} \int \mathrm{d}\boldsymbol{r}'' \hat{\psi}_{\sigma}^{\dagger}(\boldsymbol{r}'') \delta(\boldsymbol{r} - \boldsymbol{r}'') \hat{\psi}_{\sigma}(\boldsymbol{r}) , \tag{H.2.47}$$

and examine its action on the state $\hat{\psi}_{\sigma}^{\dagger}(\boldsymbol{r}')|0\rangle$. Making use of the commutation relation of field operators,

$$n(\boldsymbol{r}) \hat{\psi}_{\sigma}^{\dagger}(\boldsymbol{r}')|0\rangle = \delta(\boldsymbol{r} - \boldsymbol{r}') \hat{\psi}_{\sigma}^{\dagger}(\boldsymbol{r}')|0\rangle , \tag{H.2.48}$$

which means that there is indeed a particle at $\boldsymbol{r} = \boldsymbol{r}'$.

Constructing the row and column vector

$$\hat{\boldsymbol{\psi}}^{\dagger} = (\hat{\psi}_{\uparrow}^{\dagger} \; \hat{\psi}_{\downarrow}^{\dagger}) \qquad \text{and} \qquad \hat{\boldsymbol{\psi}} = \begin{pmatrix} \hat{\psi}_{\uparrow} \\ \hat{\psi}_{\downarrow} \end{pmatrix} \tag{H.2.49}$$

from the field operators, we have

$$n(\boldsymbol{r}) = \hat{\boldsymbol{\psi}}^{\dagger}(\boldsymbol{r}) \hat{\boldsymbol{\psi}}(\boldsymbol{r}) . \tag{H.2.50}$$

The particle-number operator is the integral of $n(\boldsymbol{r})$ over the entire space:

$$N = \int n(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} , \tag{H.2.51}$$

which is the same as the $\boldsymbol{q} = 0$ component of the quantity $n(\boldsymbol{q})$:

$$N = \sum_{\boldsymbol{k}\sigma} c_{\boldsymbol{k}\sigma}^{\dagger} c_{\boldsymbol{k}\sigma} . \tag{H.2.52}$$

If the field operators are not expanded in a plane-wave basis but in terms of the Bloch states, and a single band is considered,

$$\begin{aligned} \hat{\psi}_{\sigma}(\boldsymbol{r}) &= \frac{1}{\sqrt{V}} \sum_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{\boldsymbol{k}}(\boldsymbol{r}) c_{\boldsymbol{k}\sigma} , \\ \hat{\psi}_{\sigma}^{\dagger}(\boldsymbol{r}) &= \frac{1}{\sqrt{V}} \sum_{\boldsymbol{k}} \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_{\boldsymbol{k}}^{*}(\boldsymbol{r}) c_{\boldsymbol{k}\sigma}^{\dagger} , \end{aligned} \tag{H.2.53}$$

and hence

$$n(\boldsymbol{r}) = \frac{1}{V} \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma} \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{r}} u_{\boldsymbol{k}}^{*}(\boldsymbol{r}) u_{\boldsymbol{k}'}(\boldsymbol{r}) c_{\boldsymbol{k}\sigma}^{\dagger} c_{\boldsymbol{k}'\sigma} . \tag{H.2.54}$$

By taking its Fourier transform,

$$\begin{aligned} n(\boldsymbol{q}) &= \int n(\boldsymbol{r}) \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}} \mathrm{d}\boldsymbol{r} \\ &= \frac{1}{V} \sum_{\boldsymbol{k}\boldsymbol{k}'\sigma} \int \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}+\boldsymbol{q}-\boldsymbol{k}')\cdot\boldsymbol{r}} u_{\boldsymbol{k}}^{*}(\boldsymbol{r}) u_{\boldsymbol{k}'}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} \, c_{\boldsymbol{k}\sigma}^{\dagger} c_{\boldsymbol{k}'\sigma} . \end{aligned} \tag{H.2.55}$$

Because of the lattice periodicity of the functions $u_{\boldsymbol{k}}(\boldsymbol{r})$ the integral vanishes unless $\boldsymbol{k}' = \boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}$, where the reciprocal-lattice vector $\boldsymbol{G}$ ensures that both $\boldsymbol{k}$ and $\boldsymbol{k}'$ are in the first Brillouin zone. By separating the integral into two parts, an integral over the primitive cell and a sum over cells,

$$n(\boldsymbol{q}) = \frac{N}{V} \sum_{\boldsymbol{k}\boldsymbol{G}\sigma} \int\limits_{v} u^*_{\boldsymbol{k}}(\boldsymbol{r}) u_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}}(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r}\, c^\dagger_{\boldsymbol{k}\sigma} c_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G},\sigma} \,. \tag{H.2.56}$$

In terms of the Pauli matrices

$$\sigma^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad \sigma^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{H.2.57}$$

the density of magnetic moment can be written as

$$\boldsymbol{m}(\boldsymbol{r}) = \tfrac{1}{2} g_{\mathrm{e}} \mu_{\mathrm{B}} \sum_i \boldsymbol{\sigma}_i \delta(\boldsymbol{r} - \boldsymbol{r}_i) \,. \tag{H.2.58}$$

Using, once again, the row and column vectors of field operators, we obtain the second-quantized operator

$$\boldsymbol{m}(\boldsymbol{r}) = \tfrac{1}{2} g_{\mathrm{e}} \mu_{\mathrm{B}} \hat{\boldsymbol{\psi}}^\dagger(\boldsymbol{r}) \boldsymbol{\sigma} \hat{\boldsymbol{\psi}}(\boldsymbol{r}) \,. \tag{H.2.59}$$

Just like for the particle density, the Fourier transform can again be expressed in a particularly simple form in terms of the creation and annihilation operators of plane-wave states. When written in components, the operators

$$\sigma^+ = \tfrac{1}{2}\left(\sigma^x + \mathrm{i}\sigma^y\right) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma^- = \tfrac{1}{2}\left(\sigma^x - \mathrm{i}\sigma^y\right) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \tag{H.2.60}$$

and the corresponding

$$m^\pm = m_x \pm \mathrm{i} m_y \tag{H.2.61}$$

are customarily used instead of the x and y components of the magnetic moment. Explicitly,

$$\begin{aligned} m^z(\boldsymbol{q}) &= \tfrac{1}{2} g_{\mathrm{e}} \mu_{\mathrm{B}} \sum_{\boldsymbol{k}} \left[c^\dagger_{\boldsymbol{k}\uparrow} c_{\boldsymbol{k}+\boldsymbol{q}\uparrow} - c^\dagger_{\boldsymbol{k}\downarrow} c_{\boldsymbol{k}+\boldsymbol{q}\downarrow} \right], \\ m^+(\boldsymbol{q}) &= g_{\mathrm{e}} \mu_{\mathrm{B}} \sum_{\boldsymbol{k}} c^\dagger_{\boldsymbol{k}\uparrow} c_{\boldsymbol{k}+\boldsymbol{q}\downarrow} \,, \\ m^-(\boldsymbol{q}) &= g_{\mathrm{e}} \mu_{\mathrm{B}} \sum_{\boldsymbol{k}} c^\dagger_{\boldsymbol{k}\downarrow} c_{\boldsymbol{k}+\boldsymbol{q}\uparrow} \,. \end{aligned} \tag{H.2.62}$$

References

1. A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems*, McGraw-Hill Book Company, New York (1971).
2. G. D. Mahan, *Many-Particle Physics*, Third Edition, Kluwer Academic/ Plenum Publishers New York (2000).

I

Canonical Transformation

Instead of tackling the quantum mechanical eigenvalue problem directly, it is often more practical to perform a unitary canonical transformation on the Hamiltonian that leaves the energy spectrum unaltered. This can be achieved either by transforming away some degrees of freedom, and generating an effective interaction among the remaining ones, or by transforming the Hamiltonian directly to a diagonal form. Below we shall present both approaches.

I.1 Derivation of an Effective Hamiltonian

It is a recurrent situation in solid-state physics that a system is made up of two distinct parts whose components interact but we are interested only in the properties of one subsystem. The effects of the other subsystem – i.e., its degrees of freedom (or at least some of them) – can then be transformed away by means of a canonical transformation. This is the case for an interacting system of electrons and phonons when the effective interaction between the electrons mediated by the phonons is studied, as discussed in Chapter 23. Below we shall first treat the method of canonical transformation generally, and then present some other applications as well.

I.1.1 General Formulation of the Problem

By separating an unperturbed part $\mathcal{H}_0$ – whose energy eigenstates can be calculated exactly – from the interaction part $\mathcal{H}_{\text{int}}$ of the Hamiltonian, and by formally introducing a coupling constant λ, the Hamiltonian can be written in the generic form

$$\mathcal{H} = \mathcal{H}_0 + \lambda \mathcal{H}_1 \,. \tag{I.1.1}$$

We shall now demonstrate that by means of a canonical transformation

$$\widetilde{\mathcal{H}} = \mathrm{e}^{S} \mathcal{H} \mathrm{e}^{-S} \,, \tag{I.1.2}$$

where

$$S^\dagger = -S \tag{I.1.3}$$

because of the unitarity of the transformation, the effects of the perturbation on the space of eigenstates of $\mathcal{H}_0$ can be taken into account by an equivalent interaction term instead of $\mathcal{H}_1$.

When an arbitrary operator $\mathcal{O}$ and its transform $\widetilde{\mathcal{O}}$ are considered, the series expansion of the unitary operator $\exp(\pm S)$ and the subsequent rearrangement of the terms of the same powers of S gives

$$\widetilde{\mathcal{O}} = \mathrm{e}^{S}\mathcal{O}\mathrm{e}^{-S} = \mathcal{O} + [S,\mathcal{O}] + \tfrac{1}{2}\,[S,[S,\mathcal{O}]] + \tfrac{1}{3!}\,[S,[S,[S,\mathcal{O}]]] + \dots\,. \tag{I.1.4}$$

Applying this formula to the Hamiltonian (I.1.1),

$$\widetilde{\mathcal{H}} = \mathcal{H}_0+\lambda\mathcal{H}_1+[S,\mathcal{H}_0]+\lambda\,[S,\mathcal{H}_1]+\tfrac{1}{2}\,[S,[S,\mathcal{H}_0]]+\tfrac{1}{2}\lambda\,[S,[S,\mathcal{H}_1]]+\dots\,. \tag{I.1.5}$$

The direct interaction term $\mathcal{H}_1$ can be eliminated by a suitable choice of S by requiring that

$$\lambda\mathcal{H}_1 + [S,\mathcal{H}_0] = 0\,. \tag{I.1.6}$$

The operator S is thus proportional to λ. Eliminating $\mathcal{H}_1$ from the transformed Hamiltonian (I.1.5) by means of this equation,

$$\widetilde{\mathcal{H}} = \mathcal{H}_0 - \tfrac{1}{2}\,[S,[S,\mathcal{H}_0]] - \tfrac{1}{3}\,[S,[S,[S,\mathcal{H}_0]]] + \dots \tag{I.1.7}$$

to third order in the coupling constant. The Hamiltonian of the effective interaction is thus

$$\mathcal{H}_{\text{eff}} = -\tfrac{1}{2}\,[S,[S,\mathcal{H}_0]] - \tfrac{1}{3}\,[S,[S,[S,\mathcal{H}_0]]] + \dots\,. \tag{I.1.8}$$

Alternatively, it can be written as

$$\mathcal{H}_{\text{eff}} = \tfrac{1}{2}\,[S,\lambda\mathcal{H}_1] + \tfrac{1}{3}\,[S,[S,\lambda\mathcal{H}_1]] + \dots\,. \tag{I.1.9}$$

In most cases only the first (leading) term is taken into account.

As we shall see, the operator S generating the canonical transformation can sometimes be given explicitly. In other cases we shall content ourselves with specifying the matrix elements of the transformed Hamiltonian between any initial and final states ($|i\rangle$ and $|f\rangle$, of energy E_i and E_f) of the unperturbed system. By keeping only the first term on the right-hand side of (I.1.9) and inserting a complete set of intermediate states by making use of the property $\sum_j |j\rangle\langle j| = 1$,

$$\langle f|\mathcal{H}_{\text{eff}}|i\rangle = \tfrac{1}{2}\sum_j \left[\langle f|S|j\rangle\langle j|\lambda\mathcal{H}_1|i\rangle - \langle f|\lambda\mathcal{H}_1|j\rangle\langle j|S|i\rangle\right]. \tag{I.1.10}$$

Taking the matrix elements of (I.1.6) between the intermediate states,

$$\langle j|\lambda\mathcal{H}_1|j'\rangle + \langle j|S\mathcal{H}_0 - \mathcal{H}_0 S|j'\rangle = 0\,. \tag{I.1.11}$$

If these states are eigenstates of the unperturbed system with an energy E_j then the previous equation implies

$$\langle j|S|j'\rangle = \frac{\langle j|\lambda\mathcal{H}_1|j'\rangle}{E_j - E_{j'}} \,. \tag{I.1.12}$$

By substituting this form of the matrix element into (I.1.10), and taking into account that the initial and final states are also eigenstates of the unperturbed Hamiltonian, we find

$$\langle f|\mathcal{H}_{\text{eff}}|i\rangle = \tfrac{1}{2}\sum_j \langle f|\lambda\mathcal{H}_1|j\rangle\langle j|\lambda\mathcal{H}_1|i\rangle \left[\frac{1}{E_f - E_j} - \frac{1}{E_j - E_i}\right]. \tag{I.1.13}$$

When elastic transitions are considered, and the common energy $E_i = E_f$ is denoted by E_0,

$$\langle f|\mathcal{H}_{\text{eff}}|i\rangle = -\sum_j \frac{\langle f|\lambda\mathcal{H}_1|j\rangle\langle j|\lambda\mathcal{H}_1|i\rangle}{E_j - E_0} \,. \tag{I.1.14}$$

It is often not necessary to know these matrix elements over the entire Hilbert space of the system's states; using physical considerations it may be sufficient to know them over a subspace. It is then often possible to find explicitly an effective Hamiltonian that gives the same matrix elements in that subspace.

In Chapter 23 we showed how the effective electron–electron interaction can be derived from the electron–phonon interaction. Below we shall first derive the effective interaction between magnetic moments in an electron system, and then demonstrate that even the interaction between the magnetic moment and the electron system can be considered as an effective interaction, and obtained from the Anderson model that describes the interaction between conduction electrons and d-electrons that are "bound" to the atom.

I.1.2 RKKY Interaction

It was mentioned in Chapter 14 that a localized spin $\boldsymbol{S}_1$ placed in a system of free electrons interacts with them through its magnetic moment, and – provided $\boldsymbol{S}_1$ is fixed – it can polarize the electron system around itself. If a second spin $\boldsymbol{S}_2$ is placed at a distance r from the first, its orientation will not be arbitrary but determined by the local value of the spin density generated by the first spin. Since in reality the first spin is not fixed, interactions mediated by the mobile electrons may eventually lead to processes in which the two localized spins of magnitude S mutually flip each other. By choosing the kinetic energy of the mobile electrons as the unperturbed Hamiltonian, and the interaction between the conduction electrons and the localized spins as a perturbation, the canonical transformation is chosen in such a way that this direct interaction is replaced by an effective interaction between the two spins.

The interaction between the localized moment of d-electrons and the itinerant s-electrons is determined by the spin density of the s-electrons at the localized moment. Just like in the Heisenberg model, the scalar product of the two spins is taken with an exchange coupling constant J. This is the so-called s–d interaction. Writing the spin density of electrons in terms of the field operators as in (H.2.59), the interaction between the spin $\boldsymbol{S}_l$ at $\boldsymbol{R}_l$ and the conduction electrons is

$$\begin{aligned}\mathcal{H}_{\text{s–d}} &= -2J\boldsymbol{S}_l \cdot \boldsymbol{s}(\boldsymbol{R}_l) = -J\boldsymbol{S}_l \cdot \left(\hat{\boldsymbol{\psi}}^\dagger(\boldsymbol{R}_l)\boldsymbol{\sigma}\hat{\boldsymbol{\psi}}(\boldsymbol{R}_l)\right) \\ &= -J\sum_{\alpha\beta} \boldsymbol{S}_l \cdot \boldsymbol{\sigma}_{\alpha\beta}\hat{\psi}^\dagger_\alpha(\boldsymbol{R}_l)\hat{\psi}_\beta(\boldsymbol{R}_l)\,.\end{aligned} \tag{I.1.15}$$

Using the creation and annihilation operators in momentum representation instead of the field operators, we have

$$\begin{aligned}\mathcal{H}_{\text{s–d}} = -\frac{J}{V}\sum_{\boldsymbol{k}\boldsymbol{k}'} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{R}_l} \Big\{ & S_l^+ c^\dagger_{\boldsymbol{k}'\downarrow}c_{\boldsymbol{k}\uparrow} + S_l^- c^\dagger_{\boldsymbol{k}'\uparrow}c_{\boldsymbol{k}\downarrow} \\ & + S_l^z\left(c^\dagger_{\boldsymbol{k}'\uparrow}c_{\boldsymbol{k}\uparrow} - c^\dagger_{\boldsymbol{k}'\downarrow}c_{\boldsymbol{k}\downarrow}\right)\Big\}\,.\end{aligned} \tag{I.1.16}$$

The total spin is conserved in the interaction but there can be an exchange between conduction electrons and localized spins. The different terms in the previous formula describe the processes that increase, reduce, and preserve the z component of the localized spin.

A somewhat more general formula can be obtained by assuming that the interaction is not strictly local. The coupling strength then depends on what state $\boldsymbol{k}'$ the electron of wave vector $\boldsymbol{k}$ is scattered into.

$$\begin{aligned}\mathcal{H}_{\text{s–d}} = -\frac{1}{V}\sum_{\boldsymbol{k}\boldsymbol{k}'} J_{\boldsymbol{k}'\boldsymbol{k}}\mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{R}_l} \Big\{ & S_l^+ c^\dagger_{\boldsymbol{k}'\downarrow}c_{\boldsymbol{k}\uparrow} + S_l^- c^\dagger_{\boldsymbol{k}'\uparrow}c_{\boldsymbol{k}\downarrow} \\ & + S_l^z\left(c^\dagger_{\boldsymbol{k}'\uparrow}c_{\boldsymbol{k}\uparrow} - c^\dagger_{\boldsymbol{k}'\downarrow}c_{\boldsymbol{k}\downarrow}\right)\Big\}\,.\end{aligned} \tag{I.1.17}$$

In what follows, we shall assume that this $\boldsymbol{k}$-dependence can be ignored.

If the system contains N_{i} localized spins $\boldsymbol{S}_1, \boldsymbol{S}_2, \dots$ at $\boldsymbol{R}_1, \boldsymbol{R}_2, \dots$, the interaction with the conduction electrons is determined by the interaction Hamiltonian

$$\begin{aligned}\mathcal{H}_{\text{s–d}} = -\frac{J}{V}\sum_{l=1}^{N_{\mathrm{i}}}\sum_{\boldsymbol{k}\boldsymbol{k}'} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{R}_l} \Big\{ & S_l^+ c^\dagger_{\boldsymbol{k}'\downarrow}c_{\boldsymbol{k}\uparrow} + S_l^- c^\dagger_{\boldsymbol{k}'\uparrow}c_{\boldsymbol{k}\downarrow} \\ & + S_l^z\left(c^\dagger_{\boldsymbol{k}'\uparrow}c_{\boldsymbol{k}\uparrow} - c^\dagger_{\boldsymbol{k}'\downarrow}c_{\boldsymbol{k}\downarrow}\right)\Big\}\,.\end{aligned} \tag{I.1.18}$$

This can be interpreted by the following picture: the interaction of a localized spin with the electron system gives rise to the creation or annihilation of an electron–hole pair. Since the strength of the interaction depends on the initial state of the localized spin, and spin-flip processes are allowed, the electron–hole pair carries information about the state of localized spins. The net result of two subsequent scattering processes – in which an electron–hole pair created in the vicinity of one spin is annihilated at another– is that the state of the first spin affects the state of the second spin, that is, an effective coupling arises between them.

The effective Hamiltonian has nonvanishing matrix elements between states for which the state of the electron system remains the same and only the orientations of the localized spins change. This is described by the effective interaction of the two spins. (I.1.14) also has nonzero matrix elements between states that differ by two electron–hole pairs – but we shall neglect them below.

The electron system is assumed to be initially in its ground state denoted by $|\mathrm{FS}\rangle$. The state of the spins can be characterized by the value M_l of their z component. Consequently, the initial state of the full system is

$$|i\rangle = |\mathrm{FS}\rangle|\{M_l\}\rangle \,. \tag{I.1.19}$$

As intermediate states we must take states in which an electron–hole pair has been created from the Fermi sea in addition to the possible spin flip, that is,

$$|j\rangle = c^{\dagger}_{\boldsymbol{k}'\sigma'} c_{\boldsymbol{k}\sigma}|\mathrm{FS}\rangle|\{M_l''\}\rangle \,. \tag{I.1.20}$$

Since the spin flip does not require any energy, the energy of the intermediate state is

$$E_j = \varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}} + E_0 \,. \tag{I.1.21}$$

After the second scattering event, the electron system may get back into its initial state but the localized spins may be flipped,

$$|f\rangle = |\mathrm{FS}\rangle|\{M_l'\}\rangle \,, \tag{I.1.22}$$

where the conservation of the spin z component implies

$$M_1 + M_2 + \cdots = M_1' + M_2' + \dots \,. \tag{I.1.23}$$

Using these states in (I.1.14),

$$\begin{aligned}\langle f|\mathcal{H}_{\mathrm{eff}}|i\rangle = -\sum_{\boldsymbol{k}\boldsymbol{k}'\sigma\sigma'} \sum_{\{M_l''\}} \frac{1}{\varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}}} &\langle\{M_l'\}|\langle\mathrm{FS}|\mathcal{H}_{\mathrm{s\text{–}d}} c^{\dagger}_{\boldsymbol{k}'\sigma'} c_{\boldsymbol{k}\sigma}|\mathrm{FS}\rangle|\{M_l''\}\rangle \\ &\times\langle\{M_l''\}|\langle\mathrm{FS}|c^{\dagger}_{\boldsymbol{k}\sigma} c_{\boldsymbol{k}'\sigma'}\mathcal{H}_{\mathrm{s\text{–}d}}|\mathrm{FS}\rangle|\{M_l\}\rangle \,.\end{aligned} \tag{I.1.24}$$

The previous formula gives a nonvanishing contribution when the electron of the electron–hole pair created by the s–d interaction is outside the Fermi sphere, while the hole is inside. This restriction can be incorporated by means of the factor $f_0(\varepsilon_{\boldsymbol{k}})[1 - f_0(\varepsilon_{\boldsymbol{k}'})]$.

Allowing all possible orientations for the spin of the electron–hole pair (and thus for the localized spin), the total contribution is

$$\begin{aligned}\langle f|\mathcal{H}_{\text{eff}}|i\rangle = &-\sum_{\boldsymbol{k}\boldsymbol{k}'}\frac{f_0(\varepsilon_{\boldsymbol{k}})[1-f_0(\varepsilon_{\boldsymbol{k}'})]}{\varepsilon_{\boldsymbol{k}'}-\varepsilon_{\boldsymbol{k}}}\left(\frac{J}{V}\right)^2\sum_{ll'}\mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot(\boldsymbol{R}_{l'}-\boldsymbol{R}_l)}\\ &\times\sum_{\{M_l''\}}\left[2\langle\{M_l'\}|S_l^z|\{M_l''\}\rangle\langle\{M_l''\}|S_{l'}^z|\{M_l\}\rangle\right.\\ &+\langle\{M_l'\}|S_l^+|\{M_l''\}\rangle\langle\{M_l''\}|S_{l'}^-|\{M_l\}\rangle\\ &\left.+\langle\{M_l'\}|S_l^-|\{M_l''\}\rangle\langle\{M_l''\}|S_{l'}^+|\{M_l\}\rangle\right].\end{aligned}\tag{I.1.25}$$

Evaluating the sum for the complete set of intermediate spin states,

$$\begin{aligned}\langle f|\mathcal{H}_{\text{eff}}|i\rangle = &-\left(\frac{J}{V}\right)^2\sum_{ll'}\sum_{\boldsymbol{k}\boldsymbol{k}'}\frac{f_0(\varepsilon_{\boldsymbol{k}})[1-f_0(\varepsilon_{\boldsymbol{k}'})]}{\varepsilon_{\boldsymbol{k}'}-\varepsilon_{\boldsymbol{k}}}\mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot(\boldsymbol{R}_{l'}-\boldsymbol{R}_l)}\\ &\times 2\langle\{M_l'\}|\left[S_l^zS_{l'}^z+\tfrac{1}{2}\left(S_l^+S_{l'}^-+S_l^-S_{l'}^+\right)\right]|\{M_l\}\rangle.\end{aligned}\tag{I.1.26}$$

This can be considered as the matrix element of the operator

$$\mathcal{H} = -\sum_{ll'}J(\boldsymbol{R}_l-\boldsymbol{R}_{l'})\boldsymbol{S}_l\cdot\boldsymbol{S}_{l'}\,,\tag{I.1.27}$$

thus indirect exchange can be described in terms of an effective Hamiltonian that has the same form as the Hamiltonian of direct exchange. To determine its strength, the notation $\boldsymbol{r}=\boldsymbol{R}_1-\boldsymbol{R}_2$ is introduced, and the sum

$$I = \left(\frac{1}{V}\right)^2\sum_{\boldsymbol{k}\boldsymbol{k}'}\frac{f_0(\varepsilon_{\boldsymbol{k}})[1-f_0(\varepsilon_{\boldsymbol{k}'})]}{\varepsilon_{\boldsymbol{k}'}-\varepsilon_{\boldsymbol{k}}}\mathrm{e}^{-\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{r}}\tag{I.1.28}$$

has to be evaluated. Replacing the sum by an integral, the angular integrals are readily calculated:

$$\begin{aligned}I = &\frac{1}{(2\pi)^6}\int_0^{k_\mathrm{F}}k^2\,\mathrm{d}k\int_{k_\mathrm{F}}^{\infty}k'^2\,\mathrm{d}k'\frac{1}{\varepsilon_{\boldsymbol{k}'}-\varepsilon_{\boldsymbol{k}}}\\ &\times 2\pi\int_0^{\pi}\sin\theta\,\mathrm{d}\theta\,\mathrm{e}^{-\mathrm{i}kr\cos\theta}\,2\pi\int_0^{\pi}\sin\theta'\,\mathrm{d}\theta'\mathrm{e}^{\mathrm{i}k'r\cos\theta'}\\ = &-\frac{4}{(2\pi)^4}\int_0^{k_\mathrm{F}}k^2\,\mathrm{d}k\int_{k_\mathrm{F}}^{\infty}k'^2\,\mathrm{d}k'\frac{1}{\varepsilon_{\boldsymbol{k}'}-\varepsilon_{\boldsymbol{k}}}\frac{\sin kr}{kr}\,\frac{\sin k'r}{k'r}\,.\end{aligned}\tag{I.1.29}$$

Using the quadratic dispersion relation valid for free electrons and the notations $\kappa=kr$ and $\kappa'=k'r$, we have

$$I = -\frac{m_{\mathrm{e}}}{2\hbar^2\pi^4}\frac{1}{r^4}\int\limits_{0}^{k_{\mathrm{F}}r}\kappa^2\,\mathrm{d}\kappa\int\limits_{k_{\mathrm{F}}r}^{\infty}\kappa'^2\,\mathrm{d}\kappa'\frac{1}{\kappa'^2-\kappa^2}\frac{\sin\kappa}{\kappa}\frac{\sin\kappa'}{\kappa'}\,. \tag{I.1.30}$$

The κ'-integral is not affected significantly by shifting the lower limit of integration to $\kappa' = 0$ but, in order to avoid the singularity arising from $\kappa' = \kappa$, the principal value of the integral needs to be taken. By considering the integral

$$K = \mathrm{P}\int\limits_{0}^{\infty}\kappa'^2\,\mathrm{d}\kappa'\frac{1}{\kappa'^2-\kappa^2}\frac{\sin\kappa'}{\kappa'} \tag{I.1.31}$$

separately, the even character of the integrand implies

$$\begin{aligned} K &= \tfrac{1}{2}\mathrm{P}\int\limits_{-\infty}^{\infty}\kappa'^2\,\mathrm{d}\kappa'\frac{1}{\kappa'^2-\kappa^2}\frac{\sin\kappa'}{\kappa'} \\ &= \frac{1}{4\mathrm{i}}\mathrm{P}\int\limits_{-\infty}^{\infty}\mathrm{d}\kappa'\left[\frac{\kappa'\mathrm{e}^{\mathrm{i}\kappa'}}{\kappa'^2-\kappa^2}-\frac{\kappa'\mathrm{e}^{-\mathrm{i}\kappa'}}{\kappa'^2-\kappa^2}\right]. \end{aligned} \tag{I.1.32}$$

The principal-value integrals can be determined by using the complex variable $\kappa' \pm \mathrm{i}\eta$ instead of κ' (where η is an infinitesimal quantity), and performing the integral in the complex plane, along the contour shown in Fig. I.1. By using the variable $\kappa' + \mathrm{i}\eta$ in the first term, the poles are in the lower half-plane, and the integration contour is closed in the upper half-plane. The opposite is done in the second term.

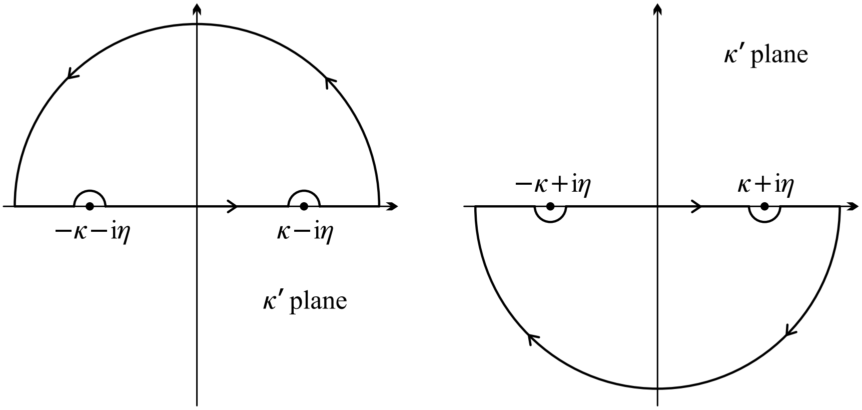


Fig. I.1. The integration contours used for the two terms in the integrand of K

Making use of the relation

$$\frac{1}{x \pm \mathrm{i}\eta} = \mathrm{P}\frac{1}{x} \mp \mathrm{i}\pi\delta(x)\,, \tag{I.1.33}$$

we have

$$K = \tfrac{1}{2}\pi\cos\kappa\,. \tag{I.1.34}$$

Substituting this back into (I.1.30),

$$\begin{aligned} I &= -\frac{m_\mathrm{e}}{4\hbar^2\pi^3}\frac{1}{r^4}\int\limits_0^{k_\mathrm{F}r} \mathrm{d}\kappa\,\kappa\sin\kappa\cos\kappa = -\frac{m}{16\hbar^2\pi^3}\frac{1}{r^4}\big(\sin 2\kappa - 2\kappa\cos 2\kappa\big)\Big|_0^{k_\mathrm{F}r} \\ &= -\frac{m_\mathrm{e}k_\mathrm{F}^4}{\hbar^2\pi^3}\frac{\sin 2k_\mathrm{F}r - 2k_\mathrm{F}r\cos 2k_\mathrm{F}r}{\big(2k_\mathrm{F}r\big)^4}\,. \end{aligned} \tag{I.1.35}$$

By collecting all factors, the effective interaction between two localized spins can be written as

$$\mathcal{H}_\mathrm{eff} = -2J(r)\boldsymbol{S}_1\cdot\boldsymbol{S}_2\,, \tag{I.1.36}$$

with an effective exchange constant

$$J(r) = \frac{m_\mathrm{e}J^2k_\mathrm{F}^4}{\hbar^2\pi^3}F(2k_\mathrm{F}r)\,, \tag{I.1.37}$$

where the function $F(x)$ is defined by

$$F(x) = \frac{x\cos x - \sin x}{x^4}\,. \tag{I.1.38}$$

This is the *RKKY interaction.*

The same result is obtained when the integral I is evaluated by another method. Using the variable $\boldsymbol{k}' = \boldsymbol{k}+\boldsymbol{q}$ but neglecting once again the restriction imposed on $\boldsymbol{q}$ by the Pauli exclusion principle,

$$I = \frac{2m_\mathrm{e}}{\hbar^2}\left(\frac{1}{V}\right)^2\sum_{\boldsymbol{q}}\sum_{|\boldsymbol{k}|<k_\mathrm{F}}\frac{1}{|\boldsymbol{k}+\boldsymbol{q}|^2-\boldsymbol{k}^2}\mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\,, \tag{I.1.39}$$

which is the Fourier transform of the formula given in (C.2.32). After performing the angular integral, we may change to a complex variable again. The previous result is then recovered through integration along the cuts of the logarithmic function.

It should be noted that the generator S of the transformation can be determined explicitly in this case. Since

$$\Big[c^\dagger_{\boldsymbol{k}'\alpha}c_{\boldsymbol{k}\beta}, \sum_{\boldsymbol{k}''\sigma}\varepsilon_{\boldsymbol{k}''}c^\dagger_{\boldsymbol{k}''\sigma}c_{\boldsymbol{k}''\sigma}\Big] = \big(\varepsilon_{\boldsymbol{k}'} - \varepsilon_{\boldsymbol{k}}\big)c^\dagger_{\boldsymbol{k}'\alpha}c_{\boldsymbol{k}\beta}\,, \tag{I.1.40}$$

it is straightforward to show that

$$S = -\frac{J}{2V}\sum_l \sum_{\boldsymbol{k}\boldsymbol{k}'}\sum_{\alpha\beta} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{R}_l}\frac{1}{\varepsilon_{\boldsymbol{k}'}-\varepsilon_{\boldsymbol{k}}}\boldsymbol{S}_l\cdot\boldsymbol{\sigma}_{\alpha\beta}c^{\dagger}_{\boldsymbol{k}'\alpha}c_{\boldsymbol{k}\beta}\,. \tag{I.1.41}$$

Using this formula in the first term of (I.1.9), making use of

$$(\boldsymbol{S}_l\cdot\boldsymbol{\sigma})(\boldsymbol{S}_{l'}\cdot\boldsymbol{\sigma}) = (\boldsymbol{S}_l\cdot\boldsymbol{S}_{l'}) + \mathrm{i}\boldsymbol{\sigma}\cdot(\boldsymbol{S}_l\times\boldsymbol{S}_{l'})\,, \tag{I.1.42}$$

and keeping only the state of the two localized spins from the entire Hilbert space, while integrating out the degrees of freedom of the s-electrons by taking the ground-state expectation value (as conduction electrons are treated as a reservoir), the previously derived effective Hamiltonian is recovered.

I.1.3 Derivation of the s–d Interaction

We shall now demonstrate that the s–d interaction given in (I.1.16) can also be viewed as an effective interaction. To this end, we shall start with the Anderson model (to be discussed in Chapter 35), in which d-electrons are not strictly bound to the atom but can become detached for short periods of time.

In the Anderson model s-electrons are described by the usual term

$$\mathcal{H}_{\mathrm{s}} = \sum_{\boldsymbol{k}\sigma}\varepsilon_{\boldsymbol{k}}c^{\dagger}_{\boldsymbol{k}\sigma}c_{\boldsymbol{k}\sigma}\,, \tag{I.1.43}$$

where the energies are referred to the chemical potential. For simplicity, we shall neglect the degeneracy of the d-states but take into account the Coulomb repulsion between the d-electrons when there are two of them (of opposite spins) on the same atom. Their Hamiltonian is then

$$\mathcal{H}_{\mathrm{d}} = \varepsilon_d\,(n_{d\uparrow}+n_{d\downarrow}) + Un_{d\uparrow}n_{d\downarrow}\,. \tag{I.1.44}$$

Finally, the term describing the hybridization of s- and d-electrons is

$$\mathcal{H}_{\mathrm{hybr}} = \frac{1}{\sqrt{V}}\sum_{\boldsymbol{k}\sigma}\Big(V_{d\boldsymbol{k}}c^{\dagger}_{\boldsymbol{k}\sigma}d_{\sigma} + V_{\boldsymbol{k}d}d^{\dagger}_{\sigma}c_{\boldsymbol{k}\sigma}\Big)\,, \tag{I.1.45}$$

where $d^{\dagger}_{\sigma}$ (d_{σ}) is the creation (annihilation) operator of d-electrons.

When $\varepsilon_d < 0$ but $\varepsilon_d + U > 0$, and there is no hybridization (that is, s- and d-states are not mixed), there is exactly one electron on the d-level, and therefore the atom has a localized moment. If $V_{\boldsymbol{k}d} \neq 0$, this electron can hop off the atom, and another electron can hop on the atom that has just become empty. If U is sufficiently large, there cannot be two electrons on the same atom simultaneously, whereas if $V_{\boldsymbol{k}d}$ is not too strong, there is always an electron on the atom – thus the atom continues to possess a spin and a magnetic moment, however their orientation may change because the spin of the electron that hops off the atom may be different from the spin of the electron that hops on it. At the same time, a spin flip also occurs in the electron system. This gives rise to the s–d interaction.

To obtain this from the Anderson model Hamiltonian by way of a canonical transformation, hybridization is chosen as the perturbation and all other terms are included in the Hamiltonian $\mathcal{H}_0$ of the unperturbed system. The generator S of the transformation that eliminates direct hybridization, the *Schrieffer–Wolff transformation*,[1] is

$$S = \frac{1}{\sqrt{V}} \sum_{\boldsymbol{k}\sigma} \left[\frac{V_{d\boldsymbol{k}}}{\varepsilon_{\boldsymbol{k}} - \varepsilon_d - U} n_{d,-\sigma} c^\dagger_{\boldsymbol{k}\sigma} d_\sigma + \frac{V_{d\boldsymbol{k}}}{\varepsilon_{\boldsymbol{k}} - \varepsilon_d} (1 - n_{d,-\sigma}) c^\dagger_{\boldsymbol{k}\sigma} d_\sigma \right.$$
$$\left. - \frac{V_{\boldsymbol{k}d}}{\varepsilon_{\boldsymbol{k}} - \varepsilon_d - U} n_{d,-\sigma} d^\dagger_\sigma c_{\boldsymbol{k}\sigma} - \frac{V_{\boldsymbol{k}d}}{\varepsilon_{\boldsymbol{k}} - \varepsilon_d} (1 - n_{d,-\sigma}) d^\dagger_\sigma c_{\boldsymbol{k}\sigma} \right] , \qquad \text{(I.1.46)}$$

where $V_{d\boldsymbol{k}} = V^*_{\boldsymbol{k}d}$. The canonical transformation leads to a term that describes the exchange between the localized spin on the d-level and the spin density of conduction electrons, as given in (I.1.17), with a coefficient

$$J_{\boldsymbol{k}\boldsymbol{k}'} = \tfrac{1}{2} V_{\boldsymbol{k}d} V_{d\boldsymbol{k}'} \left[\frac{1}{\varepsilon_{\boldsymbol{k}} - \varepsilon_d - U} + \frac{1}{\varepsilon_{\boldsymbol{k}'} - \varepsilon_d - U} - \frac{1}{\varepsilon_{\boldsymbol{k}} - \varepsilon_d} - \frac{1}{\varepsilon_{\boldsymbol{k}'} - \varepsilon_d} \right] . \qquad \text{(I.1.47)}$$

If both $\boldsymbol{k}$ and $\boldsymbol{k}'$ are close to the Fermi surface, the effective s–d exchange constant is

$$J = |V_{k_{\mathrm{F}}d}|^2 \frac{U}{\varepsilon_d(\varepsilon_d + U)} . \qquad \text{(I.1.48)}$$

Additional terms

$$\frac{1}{V} \sum_{\boldsymbol{k}\boldsymbol{k}'} \sum_{\sigma} V_{\boldsymbol{k}\boldsymbol{k}'} c^\dagger_{\boldsymbol{k}'\sigma} c_{\boldsymbol{k}\sigma} , \qquad \text{(I.1.49)}$$

which are independent of the spin of the d-level, also appear; they describe potential scattering. When $\boldsymbol{k}$ and $\boldsymbol{k}'$ are close to the Fermi surface, their strength is

$$V_{\mathrm{pot}} = -\frac{1}{2} |V_{k_{\mathrm{F}}d}|^2 \frac{U + 2\varepsilon_d}{\varepsilon_d(\varepsilon_d + U)} . \qquad \text{(I.1.50)}$$

I.2 Diagonalization of the Hamiltonian

Another application of unitary canonical transformations is the diagonalization of a Hamiltonian that is bilinear in the creation and annihilation operators.

As was found in connection with antiferromagnets, the spectrum of elementary excitations (magnons) can be obtained from the eigenvalues of the Hamiltonian given in (15.3.6),

$$\mathcal{H} = E_0 + 2|J|zS \sum_{\boldsymbol{k}} \left[a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}} + b^\dagger_{-\boldsymbol{k}} b_{-\boldsymbol{k}} + \gamma_{\boldsymbol{k}} \left(a_{\boldsymbol{k}} b_{-\boldsymbol{k}} + a^\dagger_{\boldsymbol{k}} b^\dagger_{-\boldsymbol{k}} \right) \right] , \qquad \text{(I.2.1)}$$

[1] J. R. Schrieffer and P. A. Wolff, 1966.

where the operators $a_{\boldsymbol{k}}$, $a^\dagger_{\boldsymbol{k}}$, $b_{-\boldsymbol{k}}$, and $b^\dagger_{-\boldsymbol{k}}$ satisfy bosonic commutation relations. A similar situation is encountered in the Bogoliubov treatment of superfluidity[2] (which we shall not discuss), and in Chapter 32, where the excitations of the one-dimensional Luttinger liquid is studied by means of bosonic density fluctuations.

A very similar but fermionic problem is encountered in Chapter 34 on the BCS theory of superconductivity, where the eigenstates of the Hamiltonian

$$\begin{aligned}\mathcal{H}_{\mathrm{BCS}} = E_0 &+ \sum_{\boldsymbol{k}} \xi_{\boldsymbol{k}} \left(c^\dagger_{\boldsymbol{k}\uparrow} c_{\boldsymbol{k}\uparrow} + c^\dagger_{-\boldsymbol{k}\downarrow} c_{-\boldsymbol{k}\downarrow} \right) \\ &- \sum_{\boldsymbol{k}} \left(\Delta_{\boldsymbol{k}} c^\dagger_{\boldsymbol{k}\uparrow} c^\dagger_{-\boldsymbol{k}\downarrow} + \Delta^*_{\boldsymbol{k}} c_{-\boldsymbol{k}\downarrow} c_{\boldsymbol{k}\uparrow} \right)\end{aligned} \tag{I.2.2}$$

are sought. For bosonic and fermionic systems alike, we shall use the Hamiltonian

$$\mathcal{H} = E_0 + \sum_{\boldsymbol{k}} \left[\varepsilon_{\boldsymbol{k}} \left(a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}} + b^\dagger_{-\boldsymbol{k}} b_{-\boldsymbol{k}} \right) + \gamma_{\boldsymbol{k}} \left(a_{\boldsymbol{k}} b_{-\boldsymbol{k}} + b^\dagger_{-\boldsymbol{k}} a^\dagger_{\boldsymbol{k}} \right) \right], \tag{I.2.3}$$

which is bilinear in the creation and annihilation operators, and demonstrate how it can be diagonalized by means of a canonical transformation

$$\widetilde{\mathcal{H}} = \mathrm{e}^{S} \mathcal{H} \mathrm{e}^{-S}. \tag{I.2.4}$$

Since the canonical transformation does not change the eigenvalues, the energy spectrum can be read off immediately from the diagonal form.

I.2.1 Bosonic Systems

We shall first consider a bosonic system, and show that diagonalization can be achieved by the choice

$$S = \sum_{\boldsymbol{k}} \theta_{\boldsymbol{k}} \left(b^\dagger_{-\boldsymbol{k}} a^\dagger_{\boldsymbol{k}} - a_{\boldsymbol{k}} b_{-\boldsymbol{k}} \right), \tag{I.2.5}$$

where $\theta_{\boldsymbol{k}}$ is real.

Performing the canonical transformation for each term of the Hamiltonian (I.2.3),

$$\widetilde{\mathcal{H}} = E_0 + \sum_{\boldsymbol{k}} \left[\varepsilon_{\boldsymbol{k}} \left(\widetilde{a}^\dagger_{\boldsymbol{k}} \widetilde{a}_{\boldsymbol{k}} + \widetilde{b}^\dagger_{-\boldsymbol{k}} \widetilde{b}_{-\boldsymbol{k}} \right) + \gamma_{\boldsymbol{k}} \left(\widetilde{a}_{\boldsymbol{k}} \widetilde{b}_{-\boldsymbol{k}} + \widetilde{a}^\dagger_{\boldsymbol{k}} \widetilde{b}^\dagger_{-\boldsymbol{k}} \right) \right], \tag{I.2.6}$$

where

$$\widetilde{a}^\dagger_{\boldsymbol{k}} = \mathrm{e}^{S} a^\dagger_{\boldsymbol{k}} \mathrm{e}^{-S}, \qquad \widetilde{a}_{\boldsymbol{k}} = \mathrm{e}^{S} a_{\boldsymbol{k}} \mathrm{e}^{-S}, \tag{I.2.7}$$

and $\widetilde{b}^\dagger_{-\boldsymbol{k}}$ and $\widetilde{b}_{-\boldsymbol{k}}$ are defined by similar formulas.

[2] N. N. Bogoliubov, 1947.

Applying the expansion (I.1.4) to $\widetilde{a}^\dagger_{\boldsymbol{k}}$, and using the relations

$$\begin{aligned} [S, a^\dagger_{\boldsymbol{k}}] &= -\theta_{\boldsymbol{k}} b_{-\boldsymbol{k}}\,, \qquad & [S, a_{\boldsymbol{k}}] &= -\theta_{\boldsymbol{k}} b^\dagger_{-\boldsymbol{k}}\,, \\ [S, b^\dagger_{-\boldsymbol{k}}] &= -\theta_{\boldsymbol{k}} a_{\boldsymbol{k}}\,, \qquad & [S, b_{-\boldsymbol{k}}] &= -\theta_{\boldsymbol{k}} a^\dagger_{\boldsymbol{k}} \end{aligned} \tag{I.2.8}$$

that follow from the explicit form of the generator S and the bosonic commutation relations, the repeated application of the commutators yields

$$\begin{aligned} \widetilde{a}^\dagger_{\boldsymbol{k}} &= a^\dagger_{\boldsymbol{k}} - \theta_{\boldsymbol{k}} b_{-\boldsymbol{k}} + \tfrac{1}{2}\theta^2_{\boldsymbol{k}} a^\dagger_{\boldsymbol{k}} - \tfrac{1}{3!}\theta^3_{\boldsymbol{k}} b_{-\boldsymbol{k}} + \dots \\ &= \cosh\theta_{\boldsymbol{k}}\, a^\dagger_{\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}}\,. \end{aligned} \tag{I.2.9}$$

Likewise, it can be proved that

$$\begin{aligned} \widetilde{a}_{\boldsymbol{k}} &= \cosh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, b^\dagger_{-\boldsymbol{k}}\,, \\ \widetilde{b}^\dagger_{-\boldsymbol{k}} &= \cosh\theta_{\boldsymbol{k}}\, b^\dagger_{-\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}}\,, \\ \widetilde{b}_{-\boldsymbol{k}} &= \cosh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, a^\dagger_{\boldsymbol{k}}\,. \end{aligned} \tag{I.2.10}$$

Inserting these formulas into the canonically transformed Hamiltonian,

$$\begin{aligned} \widetilde{\mathcal{H}} = E_0 + \sum_{\boldsymbol{k}} \Big\{ \varepsilon_{\boldsymbol{k}} \Big[&(\cosh\theta_{\boldsymbol{k}} a^\dagger_{\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}})(\cosh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, b^\dagger_{-\boldsymbol{k}}) \\ &+ (\cosh\theta_{\boldsymbol{k}}\, b^\dagger_{-\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}})(\cosh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, a^\dagger_{\boldsymbol{k}}) \Big] \\ &+ \gamma_{\boldsymbol{k}} \Big[(\cosh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, b^\dagger_{-\boldsymbol{k}})(\cosh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, a^\dagger_{\boldsymbol{k}}) \\ &+ (\cosh\theta_{\boldsymbol{k}}\, a^\dagger_{\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}})(\cosh\theta_{\boldsymbol{k}}\, b^\dagger_{-\boldsymbol{k}} - \sinh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}}) \Big] \Big\}\,. \end{aligned} \tag{I.2.11}$$

The off-diagonal terms vanish if

$$-\,2\varepsilon_{\boldsymbol{k}}\, \sinh\theta_{\boldsymbol{k}}\, \cosh\theta_{\boldsymbol{k}} + \gamma_{\boldsymbol{k}} \left(\cosh^2\theta_{\boldsymbol{k}} + \sinh^2\theta_{\boldsymbol{k}}\right) = 0\,. \tag{I.2.12}$$

The solution of this equation is

$$\cosh^2\theta_{\boldsymbol{k}} = \tfrac{1}{2}\left(\frac{\varepsilon_{\boldsymbol{k}}}{\sqrt{\varepsilon^2_{\boldsymbol{k}} - \gamma^2_{\boldsymbol{k}}}} + 1 \right), \qquad \sinh^2\theta_{\boldsymbol{k}} = \tfrac{1}{2}\left(\frac{\varepsilon_{\boldsymbol{k}}}{\sqrt{\varepsilon^2_{\boldsymbol{k}} - \gamma^2_{\boldsymbol{k}}}} - 1 \right). \tag{I.2.13}$$

The remaining diagonal Hamiltonian reads

$$\widetilde{\mathcal{H}} = E_0 + \sum_{\boldsymbol{k}} \hbar\omega_{\boldsymbol{k}} \left(a^\dagger_{\boldsymbol{k}} a_{\boldsymbol{k}} + b^\dagger_{-\boldsymbol{k}} b_{-\boldsymbol{k}} + 1 \right), \tag{I.2.14}$$

where

$$\hbar\omega_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}} \left(\cosh^2\theta_{\boldsymbol{k}} + \sinh^2\theta_{\boldsymbol{k}}\right) - 2\gamma_{\boldsymbol{k}}\, \sinh\theta_{\boldsymbol{k}}\, \cosh\theta_{\boldsymbol{k}} = \sqrt{\varepsilon^2_{\boldsymbol{k}} - \gamma^2_{\boldsymbol{k}}}\,. \tag{I.2.15}$$

Instead of this method, the reverse approach is usually applied. An inverse canonical transformation is performed on the operators, by introducing

$$\begin{aligned}\alpha_{\boldsymbol{k}} &= \mathrm{e}^{-S} a_{\boldsymbol{k}} \mathrm{e}^{S} = \cosh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}} + \sinh\theta_{\boldsymbol{k}}\, b^{\dagger}_{-\boldsymbol{k}}\,,\\ \alpha^{\dagger}_{\boldsymbol{k}} &= \mathrm{e}^{-S} a^{\dagger}_{\boldsymbol{k}} \mathrm{e}^{S} = \cosh\theta_{\boldsymbol{k}}\, a^{\dagger}_{\boldsymbol{k}} + \sinh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}}\,,\\ \beta_{\boldsymbol{k}} &= \mathrm{e}^{-S} b_{-\boldsymbol{k}} \mathrm{e}^{S} = \cosh\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}} + \sinh\theta_{\boldsymbol{k}}\, a^{\dagger}_{\boldsymbol{k}}\,,\\ \beta^{\dagger}_{\boldsymbol{k}} &= \mathrm{e}^{-S} b^{\dagger}_{-\boldsymbol{k}} \mathrm{e}^{S} = \cosh\theta_{\boldsymbol{k}}\, b^{\dagger}_{-\boldsymbol{k}} + \sinh\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}}\,.\end{aligned} \tag{I.2.16}$$

In terms of them, the original Hamiltonian becomes diagonal,

$$\mathcal{H} = E_0 + \sum_{\boldsymbol{k}} \hbar\omega_{\boldsymbol{k}} \left(\alpha^{\dagger}_{\boldsymbol{k}}\alpha_{\boldsymbol{k}} + \beta^{\dagger}_{-\boldsymbol{k}}\beta_{-\boldsymbol{k}} + 1\right). \tag{I.2.17}$$

I.2.2 Fermionic Systems

The same procedure can be applied to fermions – moreover, the formula (I.2.5) for the generator S of the transformation can be used without any modifications. The anticommutation relations for fermions then yield

$$\begin{aligned}[S, a^{\dagger}_{\boldsymbol{k}}] &= \theta_{\boldsymbol{k}} b_{-\boldsymbol{k}}\,, &\quad [S, a_{\boldsymbol{k}}] &= \theta_{\boldsymbol{k}} b^{\dagger}_{-\boldsymbol{k}}\,,\\ [S, b^{\dagger}_{-\boldsymbol{k}}] &= -\theta_{\boldsymbol{k}} a_{\boldsymbol{k}}\,, &\quad [S, b_{-\boldsymbol{k}}] &= -\theta_{\boldsymbol{k}} a^{\dagger}_{\boldsymbol{k}}\,,\end{aligned} \tag{I.2.18}$$

hence

$$\begin{aligned}\widetilde{a}^{\dagger}_{\boldsymbol{k}} &= a^{\dagger}_{\boldsymbol{k}} + \theta_{\boldsymbol{k}} b_{-\boldsymbol{k}} - \tfrac{1}{2}\theta^2_{\boldsymbol{k}} a^{\dagger}_{\boldsymbol{k}} - \tfrac{1}{3!}\theta^3_{\boldsymbol{k}} b_{-\boldsymbol{k}} + \dots\\ &= \cos\theta_{\boldsymbol{k}}\, a^{\dagger}_{\boldsymbol{k}} + \sin\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}}\,,\end{aligned} \tag{I.2.19}$$

and

$$\begin{aligned}\widetilde{a}_{\boldsymbol{k}} &= \cos\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}} + \sin\theta_{\boldsymbol{k}}\, b^{\dagger}_{-\boldsymbol{k}}\,,\\ \widetilde{b}^{\dagger}_{-\boldsymbol{k}} &= \cos\theta_{\boldsymbol{k}}\, b^{\dagger}_{-\boldsymbol{k}} - \sin\theta_{\boldsymbol{k}}\, a_{\boldsymbol{k}}\,,\\ \widetilde{b}_{-\boldsymbol{k}} &= \cos\theta_{\boldsymbol{k}}\, b_{-\boldsymbol{k}} - \sin\theta_{\boldsymbol{k}}\, a^{\dagger}_{\boldsymbol{k}}\,.\end{aligned} \tag{I.2.20}$$

The Hamiltonian can be diagonalized if

$$2\varepsilon_{\boldsymbol{k}} \sin\theta_{\boldsymbol{k}} \cos\theta_{\boldsymbol{k}} - \gamma_{\boldsymbol{k}} \left(\cos^2\theta_{\boldsymbol{k}} - \sin^2\theta_{\boldsymbol{k}}\right) = 0\,, \tag{I.2.21}$$

which implies

$$\cos^2\theta_{\boldsymbol{k}} = \tfrac{1}{2}\left(1 + \frac{\varepsilon_{\boldsymbol{k}}}{\sqrt{\varepsilon^2_{\boldsymbol{k}} + \gamma^2_{\boldsymbol{k}}}}\right), \qquad \sin^2\theta_{\boldsymbol{k}} = \tfrac{1}{2}\left(1 - \frac{\varepsilon_{\boldsymbol{k}}}{\sqrt{\varepsilon^2_{\boldsymbol{k}} + \gamma^2_{\boldsymbol{k}}}}\right), \tag{I.2.22}$$

and the new eigenvalues are given by

$$E_{\boldsymbol{k}} = \varepsilon_{\boldsymbol{k}} \left(\cos^2\theta_{\boldsymbol{k}} - \sin^2\theta_{\boldsymbol{k}}\right) + 2\gamma_{\boldsymbol{k}}\, \sin\theta_{\boldsymbol{k}} \cos\theta_{\boldsymbol{k}} = \sqrt{\varepsilon^2_{\boldsymbol{k}} + \gamma^2_{\boldsymbol{k}}}\,. \tag{I.2.23}$$

Just like for bosons, the inverse procedure is usually followed for fermions, too, as in Chapter 34 on superconductivity: the Hamiltonian is diagonalized in terms of the new creation and operation operators that are linear combinations of the original operators.

Figure Credits

Fig. 17.10(a)	W. Saslow, T. K. Bergstresser, and M. L. Cohen, *Phys. Rev. Lett.* **16**, 354 (1966), Fig. 1
Fig. 17.10(b)	J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**, 556 (1976), Fig. 1 (lower panel)
Fig. 18.10	W. A. Harrison, *Phys. Rev.* **118**, 1190 (1960), Fig. 2
Fig. 18.11	W. A. Harrison, *Phys. Rev.* **118**, 1190 (1960), Fig. 1
Fig. 19.2	W. Y. Ching and J. Callaway, *Phys. Rev. B* **11**, 1324 (1975), Fig. 1
Fig. 19.5	G. A. Burdick, *Phys. Rev.* **129**, 138 (1963), Fig. 3
Fig. 19.7	W. A. Harrison, *Phys. Rev.* **118**, 1190 (1960), Fig. 3
Fig. 19.8	W. A. Harrison, *Phys. Rev.* **116**, 555 (1959), Fig. 6
Fig. 19.9	J. R. Anderson and A. V. Gold, *Phys. Rev.* **139**, A1459 (1965), Figs. 6 and 11
Fig. 19.11	J. Callaway and C. S. Wang, *Phys. Rev. B* **16**, 2095 (1977), Fig. 1
Fig. 19.13	K. Fujiwara and O. Sueoka, *J. Phys. Soc. Japan* **21**, 1947 (1966)
Fig. 19.15	M. Lähdeniemi, E. Ojala, E. Suoninen, and I. Terakura, *J. Phys. F: Metal Phys.* **11**, 1531 (1981), Fig. 2
Fig. 19.16	N. V. Smith and M. M. Traum, *Phys. Rev. Lett.* **31**, 1247 (1973), Fig. 2
Fig. 19.17	P. Thiry, D. Chandesris, J. Lecante, C. Guillot, R. Pinchaux, and Y. Pétroff, *Phys. Rev. Lett.* **43**, 82 (1979), Figs. 1 and 2
Fig. 20.2(a)	J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **10**, 5095 (1974), Fig. 2
Fig. 20.5(b)	J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**, 556 (1976), Fig. 1 (middle panel)
Fig. 20.7	J. R. Chelikowsky and M. L. Cohen, *Phys. Rev. B* **14**, 556 (1976), Figs. 7 (middle panel) and 17 (upper panel)

Fig. 21.12	G. Dresselhaus, A. F. Kip, and C. Kittel, *Phys. Rev.* **98**, 368 (1955), Figs. 2 and 5
Fig. 21.14	A. F. Kip, D. N. Langenberg, and T. W. Moore, *Phys. Rev.* **124**, 359 (1961), Fig. 1
Fig. 21.17	J. B. Ketterson and R. W. Stark, *Phys. Rev.* **156**, 748 (1967), Fig. 17
Fig. 21.18	V. Shapira and B. Lax, *Phys. Rev.* **138**, A1191 (1965), Fig. 2
Fig. 21.19	M. S. Khaikin, *Soviet Physics JETP* **14**, 1260 (1962)
Fig. 21.21	V. F. Gantmakher, *Soviet Physics JETP* **16**, 247 (1963)
Fig. 22.10	D. R. Hofstadter, *Phys. Rev. B* **14**, 2239 (1976), Fig. 1
Fig. 22.11	C. Albrecht, J. H. Smet, K. von Klitzing, D. Weiss, V. Umansky, and H. Schweizer, *Phys. Rev. Lett.* **86**, 147 (2001), Figs. 2 and 4
Fig. 22.14	M. Tokumoto et al., *Solid State Commun.* **75**, 439 (1990)
Fig. 22.16	A. S. Joseph and W. L. Gordon, *Phys. Rev.* **126**, 489 (1962), Fig. 13
Fig. 22.17(a)	A. S. Joseph, A. C. Thorsen, E. Gertner, and L. E. Valby, *Phys. Rev.* **148**, 469 (1966), Fig. 1
Fig. 22.18	W. Kang, G. Montambaux, J. R. Cooper, D. Jérome, P. Batail, and C. Lenoir, *Phys. Rev. Lett.* **62**, 2559 (1989), Fig. 2
Fig. 24.4	N. Hanasaki, S. Kagoshima, and T. Hasegawa, *Phys. Rev. B* **57**, 1336 (1998), Figs. 1 and 2 (upper panels)
Fig. 24.5	M. Khoshenevisan, W. J. Pratt, Jr., P. A. Schroeder, and S. D. Steenwyk, *Phys. Rev. B* **19**, 3873 (1979), Fig. 3
Fig. 24.7	F. J. Blatt, *Physics of Electronic Conduction in Solids*, McGraw-Hill Book Co., New York (1968)
Fig. 24.8	G. K. White and S. B. Woods, *Philosophical Transactions of the Royal Society* **251** A, 273 (1959)
Fig. 24.9(a)	G. Grüner, *Advances in Physics* **23**, 941 (1974)
Fig. 24.9(b)	B. Knook, *Thesis*, Leiden (1962)
Fig. 24.11	J. Olsen, *Electron Transport in Metals*, Interscience Publishers, Inc., New York (1962)
Fig. 24.12(b)	H. E. Jackson, C. T. Walker, and T. F. McNelly, *Phys. Rev. Lett.* **25**, 26 (1970), Fig. 1
Fig. 24.13	K. v. Klitzing, G. Dorda, and M. Pepper, *Phys. Rev. Lett.* **45**, 494 (1980), Fig. 1
Fig. 25.2	F. Wooten, *Optical Properties of Solids*, Academic Press, New York (1972)
Fig. 25.3	J. R. Dixon and H. R. Riedl, *Phys. Rev.* **138**, A873 (1965), Fig. 1
Fig. 25.4	F. Wooten, *Optical Properties of Solids*, Academic Press, New York (1972)
Fig. 25.5	F. Wooten, *Optical Properties of Solids*, Academic Press, New York (1972)

Fig. 25.8	W. J. Turner and W. E. Reese, *Phys. Rev.* **127**, 126 (1962), Fig. 4
Fig. 25.10	J. C. Phillips, *Solid State Physics*, Vol. **18**, Academic Press, New York (1966)
Fig. 26.1	H. Kamerlingh Onnes, *Comm. Phys. Lab. Univ. Leiden*, No. 120b (1911)
Fig. 26.2	M. K. Wu et al., *Phys. Rev. Lett.* **58**, 908 (1987), Fig. 3
Fig. 26.3	L. H. Palmer and M. Tinkham, *Phys. Rev.* **165**, 588 (1968), Fig. 3
Fig. 26.6	B. S. Deaver, Jr. and W. M. Fairbank, *Phys. Rev. Lett.* **7**, 43 (1961), Fig. 2 (upper panel)
Fig. 26.12	N. E. Phillips, *Phys. Rev.* **114**, 676 (1959), Fig. 4
Fig. 26.14	I. Giaever and K. Megerle, *Phys. Rev.* **122**, 1101 (1961), Figs. 7 and 8
Fig. 26.24	H. F. Hess, R. B. Robinson, R. C. Dynes, J. M. Valles, Jr., and J. V. Waszczak, *Phys. Rev. Lett.* **62**, 214 (1989), Fig. 2
Fig. 26.27(b)	R. C. Jaklevic, J. Lambe, J. E. Mercereau, and A. H. Silver *Phys. Rev.* **140**, A1628 (1965), Fig. 1
Fig. 26.29	C. C. Grimes and S. Shapiro, *Phys. Rev.* **169**, 397 (1968), Fig. 1
Fig. 26.33	D. E. Langenberg et al., *Proc. IEEE* **54**, 560 (1966)
Fig. 26.36	R. C. Jaklevic, J. Lambe, J. E. Mercereau, and A. H. Silver, *Phys. Rev.* **140**, A1628 (1965), Fig. 7
Fig. 27.37(a)	B. J. van Wees et al., *Phys. Rev. Lett.* **60**, 848 (1988), Fig. 2
Fig. 27.37(b)	B. J. van Wees et al., *Phys. Rev. B* **43**, 12431 (1991), Fig. 6
Fig. 27.40	U. Meirav, M. A. Kastner, and S. J. Wind, *Phys. Rev. Lett.* **65**, 771 (1990), Figs. 2 and 5
Fig. 27.41(a)	J. H. F. Scott-Thomas, S. B. Field, M. A. Kastner, H. I. Smith, and D. A. Antoniadis, *Phys. Rev. Lett.* **62**, 583 (1989), Fig. 2 (top panel)
Fig. 27.41(b)	U. Meirav, M. A. Kastner, M. Heiblum, and S. J. Wind, *Phys. Rev. B* **40**, 5871 (1989), Fig. 1(a)
Fig. 27.42	S. Datta and B. Das, *Appl. Phys. Lett.* **56**, 665 (1990)

Name Index

Page numbers in italics refer to Volume 1: Structure and Dynamics.

Subject Index

Page numbers in italics refer to Volume 1: Structure and Dynamics.

Fundamental physical constants

Name	Symbol	Value
Bohr magneton	$\mu_\mathrm{B} = e\hbar/2m_\mathrm{e}$	$9.274\,009 \times 10^{-24}\,\mathrm{J\,T^{-1}}$
Bohr radius	$a_0 = 4\pi\epsilon_0\hbar^2/m_\mathrm{e}e^2$	$0.529\,177 \times 10^{-10}\,\mathrm{m}$
Boltzmann constant	k_B	$1.380\,650 \times 10^{-23}\,\mathrm{J\,K^{-1}}$
Conductance quantum	$G_0 = 2e^2/h$	$7.748\,092 \times 10^{-5}\,\mathrm{S}$
Electron g-factor	$g_\mathrm{e} = 2\mu_\mathrm{e}/\mu_\mathrm{B}$	$-2.002\,319$
Electron gyromagnetic ratio	$\gamma_\mathrm{e} = 2\lvert\mu_\mathrm{e}\rvert/\hbar$	$1.760\,860 \times 10^{11}\,\mathrm{s^{-1}\,T^{-1}}$
	$\gamma_\mathrm{e}/2\pi$	$28\,024.9540\,\mathrm{MHz\,T^{-1}}$
Electron magnetic moment	μ_e	$-9.284\,764 \times 10^{-24}\,\mathrm{J\,T^{-1}}$
		$-1.001\,160\,\mu_\mathrm{B}$
Electron mass	m_e	$9.109\,382 \times 10^{-31}\,\mathrm{kg}$
Electric constant	$\epsilon_0 = 1/\mu_0 c^2$	$8.854\,188 \times 10^{-12}\,\mathrm{F\,m^{-1}}$
Elementary charge	e	$1.602\,176 \times 10^{-19}\,\mathrm{C}$
Hartree energy	$E_\mathrm{h} = e^2/4\pi\epsilon_0 a_0$	$4.359\,744 \times 10^{-18}\,\mathrm{J}$
in eV		$27.211\,383\,\mathrm{eV}$
Josephson constant	$K_\mathrm{J} = 2e/h$	$483\,597.9 \times 10^{9}\,\mathrm{Hz\,V^{-1}}$
Magnetic constant	μ_0	$4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}$
Magnetic flux quantum	$\Phi_0 = h/2e$	$2.067\,834 \times 10^{-15}\,\mathrm{Wb}$
Nuclear magneton	$\mu_\mathrm{N} = e\hbar/2m_\mathrm{p}$	$5.050\,783 \times 10^{-27}\,\mathrm{J\,T^{-1}}$
Neutron mass	m_n	$1.674\,927 \times 10^{-27}\,\mathrm{kg}$
Neutron magnetic moment	μ_n	$-0.966\,236 \times 10^{-26}\,\mathrm{J\,T^{-1}}$
		$-1.913\,043\,\mu_\mathrm{N}$
Neutron g-factor	$g_\mathrm{n} = 2\mu_\mathrm{n}/\mu_\mathrm{N}$	$-3.826\,085$
Planck constant	h	$6.626\,069 \times 10^{-34}\,\mathrm{J\,s}$
in eV	$h/\{e\}$	$4.135\,667 \times 10^{-15}\,\mathrm{eV\,s}$
Proton g-factor	$g_\mathrm{p} = 2\mu_\mathrm{p}/\mu_\mathrm{N}$	$5.585\,695$
Proton gyromagnetic ratio	$\gamma_\mathrm{p} = 2\mu_\mathrm{p}/\hbar$	$2.675\,222 \times 10^{8}\,\mathrm{s^{-1}\,T^{-1}}$
	$\gamma_\mathrm{p}/2\pi$	$42.577\,482\,\mathrm{MHz\,T^{-1}}$
Proton magnetic moment	μ_p	$1.410\,607 \times 10^{-26}\,\mathrm{J\,T^{-1}}$
		$2.792\,847\,\mu_\mathrm{N}$
Proton mass	m_p	$1.672\,622 \times 10^{-27}\,\mathrm{kg}$
Reduced Planck constant	$\hbar = h/2\pi$	$1.054\,572 \times 10^{-34}\,\mathrm{J\,s}$
in eV	$\hbar/\{e\}$	$6.582\,119 \times 10^{-16}\,\mathrm{eV\,s}$
Rydberg constant	$R_\infty = \alpha^2 m_\mathrm{e} c/2h$	$10\,973\,731.569\,\mathrm{m^{-1}}$
Rydberg energy	$\mathrm{Ry} = R_\infty hc$	$2.179\,872 \times 10^{-18}\,\mathrm{J}$
in eV		$13.605\,692\,\mathrm{eV}$
Speed of light	c	$299\,792\,458\,\mathrm{m\,s^{-1}}$
Von Klitzing constant	$R_\mathrm{K} = h/e^2$	$25\,812.807\,572\,\Omega$