Electrons and Disorder in Solids

V. F. GANTMAKHER

THE
INTERNATIONAL SERIES
OF
MONOGRAPHS ON PHYSICS

SERIES EDITORS

International Series of Monographs on Physics

Electrons and Disorder in Solids

V.F. GANTMAKHER

Institute of Solid State Physics,
Russian Academy of Sciences,
Chernogolovka, 142432 Russia

Translated by
Lucia I. Man

PREFACE

What you hold in your hands is not a textbook but rather a guide. A guide cannot replace the trip itself. Nevertheless, it can give you a hint about the path to choose, where to turn or to stay, what is worth looking at and thinking about. If you do not know the local language, a guide may also be used as a phrase-book telling you how to ask the way or understand the legend on a direction sign.

Usually, a guide-book begins with a large-scale map. Its role in our book about the country of electrons and disorder in solids is played by a schematic list of contents (see the back endpaper) explaining the logic of the "administrative and territorial division of the country", the specification of its eleven chapters – provinces and links between them. Nine provinces depicted by white rectangles are well-known natural topics. The two remaining grey rectangles are specified by the form and the methods used in the appendices. The section entitled Percolation Theory summarizes the main concepts underlying this branch of mathematics. It is present in this book as a comparatively new discipline, nowadays widely used in physics, which has not been included into standard university courses for physicists. The section entitled Tunnel IV Characteristics describes experiments which are closely related simultaneously to several different phenomena such as, e.g., interelectron interference, the Coulomb gap, hopping conductivity, metal–insulator transitions, etc. By considering this topic in a individual section, we have tried to avoid unnecessary repetition. We hope that the reader, once acquainted with this section, would then repeatedly return to it when reading different topics in the book. Special symbols with a readily understood meaning indicate the chapters in which the material considered in the appendices is used.

As a rule, a large-scale map also indicates some contiguous territories indicated by grey ellipses in our schematic. We also indicate how to reach them: a list of necessary references and the most important review articles is given at the end of this preface. The reader is assumed to be well acquainted with such topics as electrons in an ideal lattice, transport in the τ-approximation, and scattering, or, at least, to be able to recollect these topics. In fact, no perfect knowledge of superconductivity or charge-density waves is necessary.

The suggested schematic represents not only a specific list of contents, but itself contains some unique information. However, one should keep in mind that since the country on the map is virtual, the map, as the ancient geographic maps, is far from being objective: its appearance depends on the author's knowledge and taste. However, the large contiguous territory named "Interacting Electrons" is not depicted on our map because, in fact, it is a country of tomorrow. Numerous discoveries in this land, its development, and road-building are all the deeds of

a not too distant future. Superconductivity with Cooper pairs and the Josephson effect and the fractional quantum Hall effect with composite fermions are the two peripheral provinces of this embryonic country. To fix the frontier, the book could have also been entitled "Noninteracting Electrons". However, this title would not have been quite correct either. It imposes a somewhat too strong constraint, because the Coulomb interaction between electrons is discussed in great detail when considering the Coulomb gap, Peierls and Mott transitions, and some other topics indicated by dotted arrows on the map. However, the problems of electron interaction are not reduced to Coulomb interactions alone. Thus, the book is only a starting beachhead for performing such studies.

However, the beachhead itself is far from being well developed. You are a tourist in a comparatively young still developing country. Some places have already acquired their final structure and the scenery has hardly changed. And you may also come across a building yard or a place which would change considerably to get updated, so that in a couple of years everything would look quite different in such places.

I do not know any other textbook covering similar topics altogether. Therefore, some chapters and even some sections are preceded by references, in which the respective topics are treated in detail and, at the same time, in an available form. These are books, reviews, and even original articles (of course, their choice is also somewhat subjective). As a whole, the system of references is not meant to reflect the priorities. Each figure of the book that deals with experimental results is supplied with references to the respective original paper, which would allow one to elucidate the details of the experiment. However, the experiments considered in the book were selected according to their eye-catching characteristics irrespective of the time of the study. Some pioneering theoretical papers often turn out to be unsuitable for primordial studies because of their too complicated mathematical apparatus. Such papers are not included in the lists of references. At the same time, there are some references to papers presenting vivid simple models, which were abandoned in later publications.

Now, have a nice trip! It is better to move following the thick arrows. Two-sided solid arrows indicate contacts with the outside world, whereas dashed and dotted arrows show inner ties. Remember that there exists no canonical universal route. So let everyone find his or her best path.

V. Gantmakher
Chernogolovka
1.06.2005

ACKNOWLEDGEMENTS

PRELIMINARY AND ADDITIONAL READING
on topics of metal physics not considered in this book

Electrons in an ideal lattice

A.A. Abrikosov. Fundamentals of the Theory of Metals. North-Holland, 1988.

N.W. Ashcroft and N.D. Mermin, Solid State Physics. Holt, Rinehart and Winston, 1969.

The Physics of Metals 1. Electrons (ed. J. Ziman). Cambridge University Press, 1969.

Transport in the τ-approximation and scattering

V.F. Gantmakher and Y.B. Levinson. Carrier Scattering in Metals and Semiconductors. North-Holland, 1987.

Spin and charge density waves

R.M. White and T.H. Geballe. Long-Range Order in Solids. Academic Press, 1979.

G. Grüner, Density Waves. Perseus Books, 2000.

G. Grüner, Rev. Mod. Phys. **60**, 1129 (1988); **66**, 1 (1994).

Superconductivity

A.A. Abrikosov. Fundamentals of the Theory of Metals. North-Holland, 1988.

V.V. Schmidt, The Physics of Superconductors. Springer, 1997.

Two-dimensional electron gas and fractional quantum Hall effect

T. Ando, A. Fowler, and F. Stern, Rev. Mod. Phys. **54**, #2 (1982).

Physics of Low-Dimensional Structures (ed. B. Butcher, N.H. March, M.P. Tosi) Plenum Press, 1993.

CONTENTS

1

METALS WITH STRONG DISORDER

The motion of itinerant electrons in crystals is determined by the long-range order. It is the long-range order that gives rise to the interference of electron waves scattered by individual atoms in a crystal, completely compensates electron scattering, and provides the stationary propagation of electron waves with practically arbitrary wave vectors $\boldsymbol{k}$ except for the waves with the wave vectors satisfying the condition

$$k^2 = (\boldsymbol{k} - \boldsymbol{K}_m)^2, \tag{1.1}$$

where $\boldsymbol{K}_m$ is an arbitrary reciprocal-lattice vector. Scattering of electrons with wave vectors (1.1) is of a resonance nature so that the respective electron waves cannot propagate in a crystal at all. All the other electrons are scattered only from the objects that break the crystal periodicity, which is taken into account by the mean free path l. In this context, the quantity l has the sense of a distance passed by an electron between two successive independent scattering events. Of course, in order to keep the length l no shorter than the electron wavelength $\lambda \simeq 2\pi/k_\mathrm{F}$ satisfying the condition

$$k_\mathrm{F} l > 1, \tag{1.2}$$

the scattering events should not occur too frequently. Here k_F is the radius of the Fermi sphere

$$k_\mathrm{F} = (3\pi^2 n)^{1/3}, \tag{1.3}$$

which is defined in terms of the free-carrier concentration n. The limitation imposed on the mean free path from below, $l \gtrsim 1/k_\mathrm{F}$, is called the Ioffe–Regel limit. In classical physics, inequality (1.2) may be described in the following way: a sinusoidal segment shorter than the wavelength cannot be considered as a sinusoid. The analogous quantum-mechanical argument is based on the uncertainty relation $\Delta k \Delta x \sim 1$. Since Δk should be much less than k_F, the minimum possible uncertainty of the electron trajectory is $\Delta x \gtrsim 1/k_\mathrm{F}$. It is natural that the distance l between the trajectory points where two successive scattering events take place should exceed the minimum uncertainty.

In this chapter, we consider "real metals" in which the concentration n is such that the mean intercarrier distance is of the order of the mean interatomic distance a,

$$n^{-1/3} \lesssim a \approx 3\,\text{Å}, \quad n \gtrsim n^* \approx 4 \cdot 10^{22}\,\text{cm}^{-3}. \tag{1.4}$$

In what follows, metals with $n \gtrsim n^*$ are called *standard*. The Fermi surfaces of materials with such a high electron density and long-range order have rather complicated shapes. However, all the necessary estimates can be made based on the model of a Fermi sphere described by eqn (1.3).

The mean free path l can be evaluated from the value of the conductivity σ. Substituting the Fermi radius given by eqn (1.3) into the expression for σ, we arrive at the Drude equation

$$\sigma = \frac{ne^2l}{\hbar k_{\mathrm{F}}} = (3\pi^2)^{-2/3}\frac{e^2}{\hbar}\,n^{1/3}(k_{\mathrm{F}}l). \tag{1.5}$$

In accordance with inequality (1.2), the minimum value of the dimensionless parameter $k_{\mathrm{F}}l$ equals unity. Then, it follows from eqn (1.5) that the resistivity cannot exceed $\rho^* \simeq 10(\hbar/e^2)a$. Of course, the numerical value of this quantity following from eqn (1.5), $\rho^* \approx 1000\,\mu\Omega\cdot\mathrm{cm}$, is only approximate and should be refined experimentally. The experiments discussed below yield the following ρ^* value for a standard metal, which will be used in what follows:

$$\rho^* \approx (200\text{–}300)\ \mu\Omega\cdot\mathrm{cm}. \tag{1.6}$$

The existence of the limiting ρ^* value allows us to formulate two questions:

1. What are the transport properties of a strongly disordered standard metal with a high ρ value obeying, nevertheless, the inequality $\rho < \rho^*$?
2. Is it possible to overcome the limit ρ^* and consider a standard metal with $\rho > \rho^*$?

In terms of the simplest model described by eqn (1.5), the existence of such a metal would mean that $l < a$, i.e., that an electron is blocked within one unit cell or is located in the vicinity of a single atom so that the metal is transformed into an insulator. The metal–insulator transition induced by disorder is called the Anderson transition; it is always observed in systems with electron density considerably less than the value given in eqn (1.4). Thus, we can reformulate the second question in this way: *Is an Anderson transition possible in systems with a high electron density?*

Below, we give the answer to the first question and also start discussing the second question and continue it in Chapter 7.

1.1. Diffraction theory of electron transport in liquid metals

The material of this paragraph is also discussed in the books by Faber (1969) and Ziman (1979).

If static defects in a periodic lattice are located far from one another, they scatter electrons independently. With a gradual increase in the defect concentration, the corresponding scattering potentials start overlapping. This makes it difficult to single out the space regions free from the scattering fields and unambiguously determine which of the static defects is really responsible for the scattering event.

A stronger disorder requires the development of a new approach to considering its possible consequences. This approach was suggested by Ziman in his theory of liquid metals.

It is well known that liquids possess only short-range order so that the nearest environment of each atom is almost the same as in a crystal. However, this "almost the same" gives rise to an uncertainty in the location of the atom with respect to the initial atom, which increases with the distance from the initial atom and, finally, leads to disappearance of the long-range order. It might seem that the absence of the long-range order (so that each atom scatters independently) should lead to $k_{\mathrm{F}}l \approx 1$ instead of the condition (1.2). However, in many instances this is not the case, and, in particular, in elemental liquid metals. This is seen from the resistivity values $\rho = 1/\sigma$. In order to estimate ρ, we express the carrier concentration n in eqn (1.5) in terms of the valence Z (i.e., the number of free electrons per atom) and the atomic concentration $N = 1/a^3$ determined from the specific weight of the melt, $n = Z/a^3$. The values of n and conductivity allow us to determine the ratio l/a. In most of the elemental liquid metals, this ratio exceeds 5, whereas in alkali metals it exceeds 100 (the only exception is Li for which $l/a \approx 13$). This signifies that the cross-section of scattering from individual atoms is not too high (several times less than a^2). A decrease in scattering is explained by a high electron density and resulting pronounced screening. Each electron "feels" not a true ion potential but only its small renormalized part preserved after screening by other electrons, the so-called pseudopotential. For the sake of brevity, we omit in this section the prefix "pseudo" and refer to a pseudopotential as simply a potential.

The theory stated here describes the electron scattering from a weak but rather extended random potential. In fact, the scattering is induced by the potential $V(\boldsymbol{r})$ extended over the whole sample volume. It is assumed that the energy spectrum is isotropic and that the electron energy depends only on the modulus of its wave vector, $\varepsilon = \varepsilon(k)$, whereas the wave functions have the form of unmodulated plane waves $\exp(i\boldsymbol{kr})$. The matrix element of the transition of the electron with such wave functions,

$$\int \psi_2^* V(\boldsymbol{r})\psi_1 \, d\boldsymbol{r} = \int e^{i(\boldsymbol{k}_1 - \boldsymbol{k}_2)\boldsymbol{r}} V(\boldsymbol{r}) \, d\boldsymbol{r} = V(\boldsymbol{q}), \qquad (1.7)$$

is the Fourier component of the scattering potential with argument equal to the wave-vector transfer during scattering, $\boldsymbol{q} = \boldsymbol{k}_1 - \boldsymbol{k}_2$. The potential $V(\boldsymbol{r})$ is a sum of the potentials $v(\boldsymbol{r} - \boldsymbol{R}_i)$ of individual atoms located at the points $\boldsymbol{R}_i$,

$$V(\boldsymbol{r}) = \sum_{\boldsymbol{R}_i} v(\boldsymbol{r} - \boldsymbol{R}_i). \qquad (1.8)$$

Therefore, the Fourier component of the potential $V(\boldsymbol{r})$ is expressed in terms of the Fourier components of the potentials $v(\boldsymbol{q})$ of individual atoms,

$$V(\boldsymbol{q}) = \sum_{\boldsymbol{R}_i} \int e^{i\boldsymbol{q}\boldsymbol{r}} v(\boldsymbol{r} - \boldsymbol{R}_i) \, d\boldsymbol{r} = \sum_{\boldsymbol{R}_i} e^{i\boldsymbol{q}\boldsymbol{R}_i} \int e^{i\boldsymbol{q}\boldsymbol{r}} v(\boldsymbol{r}) \, d\boldsymbol{r} = v(\boldsymbol{q}) \sum_{\boldsymbol{R}_i} e^{i\boldsymbol{q}\boldsymbol{R}_i}. \quad (1.9)$$

Note: Equation (1.8), which describes the potential $V(\boldsymbol{r})$, is derived under the assumption that the ions overlap only slightly, in other words, that there are no ions separated by too short $\boldsymbol{R}_i - \boldsymbol{R}_j$ distances.

Compare: Formally the same potential is also used in the model of the structural disorder discussed in Chapter 5 dedicated to Anderson transitions. In Chapter 5, we consider real and not pseudopotentials; the wells $v(\boldsymbol{r} - \boldsymbol{R}_i)$ are supposed to be rather deep, so that, in principle, each electron may be located in its own well.

Since the scattering probability is expressed in terms of a squared matrix element, one has to calculate the squared Fourier component $V^2(\boldsymbol{q})$. Let the volume occupied by a liquid metal be unity and the ion concentration be N. Then,

$$|V(\boldsymbol{q})|^2 = |v(\boldsymbol{q})|^2 \sum_{\boldsymbol{R}_i, \boldsymbol{R}_j} e^{i\boldsymbol{q}(\boldsymbol{R}_i - \boldsymbol{R}_j)} = |v(\boldsymbol{q})|^2 \left(N + \sum_{\boldsymbol{R}_i, \boldsymbol{R}_j, i \neq j} e^{i\boldsymbol{q}(\boldsymbol{R}_i - \boldsymbol{R}_j)} \right).$$
(1.10)

The latter equality is valid because all the N diagonal elements with $i = j$ of the double sum are equal to unity. Now, fix a certain ion $j = j_0$, translate the origin of the coordinate system to the point $\boldsymbol{R}_{j_0}$, and average the sum

$$\sum_{\boldsymbol{R}_i} e^{i\boldsymbol{q}(\boldsymbol{R}_i - \boldsymbol{R}_{j0})} \equiv \sum_{\boldsymbol{R}_i} e^{i\boldsymbol{q}\boldsymbol{R}_i'}, \qquad \boldsymbol{R}_i' = \boldsymbol{R}_i - \boldsymbol{R}_{j0},$$

over all the possible $\{\boldsymbol{R}_i\}$ configurations. Then, the summation over $\boldsymbol{R}_j$ can be changed to multiplication by N. Denoting averaging by a bar above the symbol, we obtain

$$\overline{|V(\boldsymbol{q})|^2} = |v(\boldsymbol{q})|^2 N \left(1 + \overline{\sum_{\boldsymbol{R}_i} e^{i\boldsymbol{q}\boldsymbol{R}_i}} \right) = |v(\boldsymbol{q})|^2 N S(q).$$
(1.11)

The expression in brackets denoted as $S(q)$ is called a structure factor. To rearrange $S(q)$ and clarify its physical meaning, we introduce the probability $NP(r)d^3r$ to find an ion in the volume d^3r under the condition that another ion is located at the origin $r = 0$. The integral $\int P(r)d^3r$ taken over the unit volume equals unity. If the positions of all the ions are statistically independent, then $P(r) \equiv 1$. The short-range order determines the P values at the r distances of the order of a, whereas as $r \to \infty$, the function P always tends to unity, $P(r) \to 1$, so that the integral $\int P(r)d^3r$ diverges. Therefore, the Fourier transform $P(q)$ of the function $P(r)$ has a singularity in the form of delta function at the origin $q = 0$: $P(q) = Q(q) + \delta(q)$. The function $Q(q)$ serves as the Fourier transform of the pair-correlation function $Q(r) = P(r) - 1$ related to $P(r)$ and tending to zero at large r. The function $Q(q)$ is regular at the point $q = 0$.

Now, replace the sum *averaged* over all the atomic positions $\boldsymbol{R}_i$ in the definition (1.11) of the structure factor $S(q)$ by the integral over d^3r where the probability $P(r)$ is a weighting factor. Ignoring the delta function (the difference

between the Fourier transforms $P(r)$ and $Q(r)$), we obtain

$$
\begin{aligned}
S(q) &= 1 + \overline{\sum_{\boldsymbol{R}_i} e^{i\boldsymbol{q}\boldsymbol{R}_i}} \\
&= 1 + N \int e^{i\boldsymbol{q}\boldsymbol{r}} Q(r)\, d^3 r \\
&= 1 + 2\pi N \int \int e^{iqr\cos\theta} \sin\theta\, d\theta Q(r) r^2\, dr \\
&= 1 + 4\pi N \int_0^\infty Q(r) \frac{\sin qr}{qr} r^2\, dr.
\end{aligned}
\tag{1.12}
$$

The function $S(q)$ appears in all the diffraction problems. It can be obtained from the relevant experiments and, in particular, from neutron and X-ray scattering. Similar to $P(q)$, this function contains exhaustive information on the correlations in ion positions. The appearance of this function in the case under consideration emphasizes that the approach used is based on the assumption that the wave functions of electrons are plane waves. Therefore, this theory is often referred to as the spectral or diffraction theory of transport in liquid metals.

The conventional form of the $S(q)$ function is shown in Fig. 1.1. After several attenuating oscillations, the function attains its asymptotic value, $S = 1$. The scale along the ordinate axis depends on the correlations: the weaker the correlations, the smaller the oscillation amplitude. For a system with statistically independent ion positions, $S \equiv 1$. The scale along the abscissa is set by the average interatomic distance, $a \approx N^{-1/3}$. The argument value in the first maximum is determined by the radius a_1 of the first coordination sphere, $1/q_1 \approx a_1 \approx a$. In a similar way, $1/q_2 \approx a_2 \approx 2a$, etc. Thus, it follows from eqn (1.5) that one may indicate certain points $q = 2k_{\mathrm{F}}$ at various Z (vertical dashed lines in Fig. 1.1) on the abscissa and also the positions of these points with respect to the maxima of the function $S(q)$. This leads to some nontrivial conclusions that can be verified experimentally.

The quantity l in eqn (1.5) "accumulates" the results of all the individual events of elastic scattering $|\boldsymbol{k}_1| = |\boldsymbol{k}_2| = k_{\mathrm{F}}$ with different scattering angles
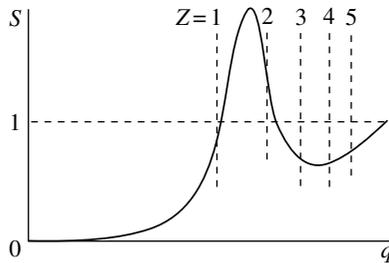
FIG. 1.1. Sketch of the $S(q)$ function showing the $2k_{\mathrm{F}}$ values for various number Z of free electrons per atom.
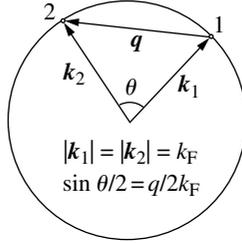
FIG. 1.2.   Illustrating eqn (1.13) and the relation between the angle of elastic
scattering, θ, and the modulus of the transferred momentum q.

$\theta = 2\arcsin(q/2k_F)$ (see Fig. 1.2) in accordance with the formula

$$1/l \propto \int_0^\pi \overline{|V(q)|^2}(1 - \cos\theta)\sin\theta\, d\theta \propto \int_0^1 S(2k_F x)v^2(2k_F x)x^3\, dx, \quad x = q/2k_F.$$

$$(1.13)$$

The presence of the factor x^3 in the integrand in eqn (1.13) indicates that the
main contribution to the integral comes from the region $x \approx 1$, i.e., we deal
here with almost backscattering within angles $\theta \approx \pi$. This signifies that in the
calculation of $1/l$, the important contribution comes not from the whole $S(q)$
function but only from its value in the vicinity of the point $q = 2k_F$ (Fig. 1.1).
Thus, the following elegant statement can be made.
    Obviously, with an increase in temperature, the correlations are weakened and
the ion system becomes more chaotic and approaches the state of an ideal gas.
Therefore, with an increase in temperature, the values of the function $S(q)$ at
fixed q tend to unity. As is seen from Fig. 1.1, in this case, the value $S(2k_F)$ in
metals with $Z = 1$ (liquid Na, K, Rb) and $Z = 3$ (liquid Al, Ga, In) increases,
whereas in metals with $Z = 2$ (liquid Zn, Cd, Hg) it decreases. Therefore, with
an increase in temperature, $1/l$ and the resistivity ρ (proportional to $1/l$) should
also behave differently. Thus, diffraction theory predicts a nontrivial dependence
of the sign of the temperature coefficient of resistivity on the valence of a liquid
metal – with an increase in temperature, the resistivity of alkali and trivalent
metals should increase, whereas the resistivity of divalent metals should decrease.
As is seen from Table 1.1 that lists the experimental values of the temperature
coefficient of resistivity for a number of metals, the prediction of diffraction
theory turned out to be quite correct. (All the metals indicated in the table have
comparatively low melting points – less than $200\,°\mathrm{C}$ for all the metals except for
Zn ($420\,°\mathrm{C}$) and Cd ($320\,°\mathrm{C}$).)
    It is worth considering one more aspect of the results obtained in diffrac-
tion theory. It turned out that a stronger disorder may result in a decrease
in resistivity. The prerequisite for occurrence of this phenomenon is the over-
lap of individual scatterers. If these scatterers were isolated, their effects would
have been additive. It is also important that the thermal phonons act against

TABLE 1.1.

Metal	Valence	$\rho(\mu\Omega \cdot \text{cm})$	$\dfrac{d\ln\rho}{d\ln T}$
Li	1	25	0.6
Na	1	10	0.85
K	1	13	0.76
Rb	1	22	0.70
Cs	1	37	0.69
Zn	2	37	−0.24
Cd	2	34	−0.22
Hg	2	91	−0.10
Ga	3	26	0.14
In	3	33	0.16

the background of the strong static disorder that has already destroyed the long-range order and anisotropy.

Note: An increase in resistivity with temperature lowering may also take place under the conditions of weak localization (see Chapter 2), but in this case, it is associated with the quantum corrections to conductivity that appear beyond the limits of the approximation of the kinetic equation. Diffraction theory demonstrates that the kinetic equation also may lead to a negative temperature coefficient of resistivity.

1.2. The Mooji rule

The success of diffraction theory shows that it is possible to use and refer to the parameter l even in those cases where l can hardly be called a distance passed by an electron between two successive scattering events. Calculations show that, in the case of strong disorder, scattering may be considered as a continuous process quantitatively described by the length l. Under these conditions, a decisive role is played by the dimensionless parameter $l/n^{-1/3} \approx k_F l$.

The brilliant confirmation of the conclusions that follow from diffraction theory applied to elemental metallic melts promoted the widespread use of this theory in processing of electrical conductivity data for disordered high-resistance alloys. First, the sign of the temperature coefficient of resistance is used to determine the average number of carriers per atom. This number is then used as an argument for determining the electron spectrum, the band overlap, the position of the Fermi level, etc. However, it turned out that, in many instances, the simplest estimates based on eqn (1.5) yielded $k_F l < 1$. In this case, plane waves are a poor approximation of the wave functions, and diffraction theory becomes invalid.

Under these conditions, one has to invoke experimental data. Figure 1.3 summarizes the experimental temperature coefficients of resistance, $\overline{\alpha} = (1/R)(dR/dT)$, for more than 100 high-resistance metal alloys ($\overline{\alpha}$ has the
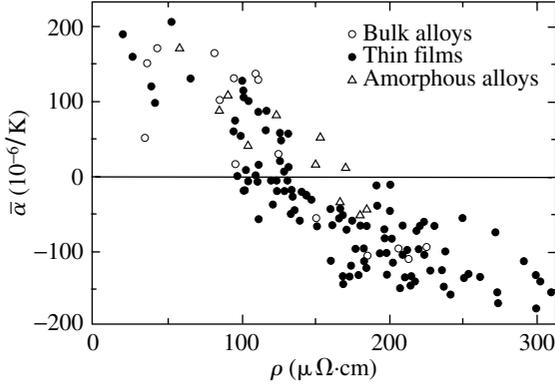
FIG. 1.3. The Mooji rule. Correlation between the resistivity and its temperature coefficient in high-resistance alloys (Mooji 1973).

dimensions of a reciprocal energy). These data point to the existence of a certain correlation between the resistivity and its temperature coefficient and lead to the empirical Mooji rule which reads that *alloys with resistivity lower than 100–150 $\mu\Omega \cdot cm$ usually have a positive temperature coefficient of resistance, whereas those with resistivity exceeding 100–150 $\mu\Omega \cdot cm$, a negative temperature coefficient of resistance.* In other words, if a static disorder gives rise to a "too high" resistivity, an increase in the temperature slightly reduces the resistivity.

Note: We consider here materials consisting only of metal atoms with the valence electrons weakly bound to the material ions and, therefore, delocalized. In other words, we consider only the metals which are standard in terms of eqn (1.4) and we do not consider metal oxides, Bi-type semimetals, MoGe-type compounds, high-temperature superconductors, and other materials with reduced electron density.
On the other hand: It is seen from Table 1.1 that almost all of the elemental liquid metals that can be considered within the diffraction theory have resistivity $\rho \lesssim 50\,\mu\Omega \cdot cm$ and, therefore, do not fall under the Mooji rule either.

It is no accident that Fig. 1.3 has no data on alloys with $\rho > 300\mu\Omega \cdot cm$. Hundreds of various metal alloys were prepared by different methods and they all had resistivity lower than $300\mu\Omega \cdot cm$. Although this value is three to four times lower than the maximum resistivity ρ^* evaluated using eqn (1.5), it is this value that seems to be the maximum possible resistivity value of a standard metal obtained in the limit $l \approx a$.

Both the existence of the empirical upper resistivity limit for a standard metal and the Mooji rule seem to be associated with important physical factors related to screening in the systems with high free-carrier densities. It is probable that screening (decreasing the apparent amplitude of a random potential) gives rise, in a self-consistent way, to the formation of such a critical level of randomness that the carriers still remain free.

1.3. Saturation of resistivity

Thus, making the static disorder stronger, we cannot set up the conditions necessary for an Anderson transition in a standard metal; the formation of a certain thermal disorder in addition to the strong static disorder can even result in a decrease in resistivity. However, there is one more chance to create such conditions – to take a metal with a weak static disorder but a strong electron–phonon interaction, which ensures a rapid increase in resistivity with temperature. This would allow us to study the effect of only thermal disorder. This attempt is based on the natural idea about the equivalence of the static (structural) and dynamic (thermal) disorder in terms of electron scattering. The velocity of the thermal motion of ions is of the order of the sound velocity, 10^5 cm/s, whereas the Fermi velocity of electrons is $v_F \approx 10^8$ cm/s. Therefore, moving electrons always "see" only the static pattern which includes the atomic thermal displacements.

The temperature dependence of resistance caused by electron scattering from phonons is described by the Grüneisen relation

$$\rho(T) = \rho_0 + \beta T^5, \quad T \ll T_D,$$
$$\rho(T) = \alpha T, \qquad T \gg T_D, \tag{1.14}$$

which is linear above the Debye temperature T_D. To reach high $\rho(T)$ values, one has to use the maximum possible α values and high temperatures. The coefficient α is the higher, the more pronounced the electron–phonon interaction, whereas the temperature is limited by melting. These factors dictate the choice of materials appropriate for experiments.

Figure 1.4 shows the $\rho(T)$ curves for single crystals of two intermetallic compounds, Nb_3Sn and Nb_3Sb. The former compound becomes superconducting at $T_c = 18$ K and, therefore, is widely used for manufacturing wires for superconducting magnets. The latter compound is also a superconductor with $T_c = 0.2$ K. We mentioned superconductivity deliberately. According to the classical theory of superconductivity, a high temperature of the superconducting transition indicates a strong electron–phonon interaction. This allows one to expect in eqn (1.14) rather high values of the β and α coefficients for Nb_3Sn. However, instead of linear behavior at high temperatures we see a gradual decrease of the slope of the $\rho(T)$ curve.

Now, consider the curve for Nb_3Sb. At $T < 200$ K, it is a typical Grüneisen curve – the portion of the residual resistivity at low temperatures, the portion proportional to $T^{3.6}$, and, finally, the linear portion described by eqn (1.14). The latter is so steep that one may expect the attainment of the critical values at admissible temperatures, $T \approx 500$ K. However, in actual fact, the $\rho(T)$ curve shows an obvious tendency to saturation toward about $150\,\mu\Omega\cdot$cm. It is remarkable that the saturation level is the same for both compounds, although, at low temperatures, the phonon-induced resistivity for Nb_3Sn is much higher than for Nb_3Sb, which correlates with the difference between their T_c temperatures.
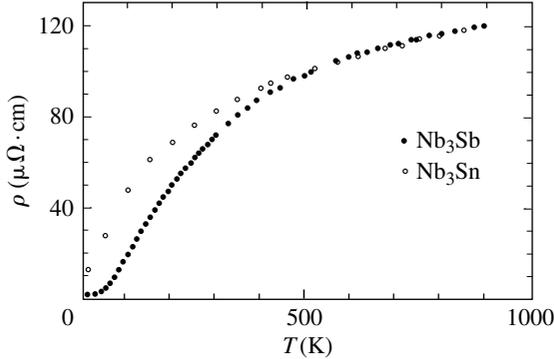
FIG. 1.4. $\rho(T)$ curves with the saturation of resistivity for single crystals of two intermetallic compounds (Fisk and Webb 1976).

Phenomenologically, the experimental curves in Fig. 1.4 are well described by the model of a shunting resistance

$$\rho^{-1} = \rho_{id}^{-1} + \rho_{sh}^{-1}. \tag{1.15}$$

The model assumes that when the resistivity ρ_{id} corresponding to the Grüneisen relation becomes too high, the current starts flowing through the shunting resistance ρ_{sh}. Equation (1.15) describes the experiments quite well, but its substantiation meets difficulties.

Usually, eqn (1.15) indicates the existence of parallel conduction channels, e.g., the existence of two independent groups of carriers described by different parameters and scattering laws. However, there are no grounds for the existence of such groups in these materials. One can also arrive at eqn (1.15) considering only one group of carriers by introducing certain correlations between the scattering events. If there were no correlations, the conductivity σ written in terms of the mean time between two collisions, τ, and the effective mass m is $\sigma = ne^2\tau/m$. Assume that two scattering events cannot take place in the same unit cell, i.e., that the scattering events are separated by a certain minimum time $\tau_0 = a/v_F$ during which an electron passes to the neighboring unit cell (Gurvitch 1981). After time τ_0, no limitations exist any more, and the scattering probability starts behaving in the conventional way. If the first scattering event takes place at the moment $t = 0$, the probability p of the following scattering event is

$$\rho = \begin{cases} 0, & t < \tau_0, \\ 1/\tau, & t > \tau_0. \end{cases}$$

Then, replacing the average time between the two collisions τ by $\tau + \tau_0$, we arrive at the eqn (1.15) with

$$\rho_{sh} \simeq (\hbar/e^2)a. \tag{1.16}$$

Although this result is consistent with experiment, the above arguments cannot be accepted as a formal proof of eqn (1.15). The main supposition that the scattering events occur instantly at a space point contradicts the more convincing assumption that the scattering events are lengthy either in space or in time. The latter assumption underlies the diffraction theory.

At the same time, it was experimentally established that saturation is attained at the resistivity given by eqn (1.16). This resistivity is characteristic of high-resistance alloys (Fig.1.3) and correspond to the smallest possible mean free path $l \approx a$, i.e., $\rho_{\mathrm{sh}} \approx \rho^*$. This confirms that the static and dynamic disorders produce almost the same effects on electrons. The equivalence of these types of disorder is also illustrated by the temperature curves of resistivity of TiAl alloys with different Al concentrations (Fig. 1.5). Pure titanium is characterized by a low residual resistivity, a pronounced increase in resistivity with temperature, and an obvious tendency to saturation. Qualitatively, the $\rho(T)$ curve for pure titanium behaves as those for Nb-based compounds in Fig. 1.4. With an increase in the Al concentration, the static disorder becomes stronger and the residual resistivity increases, whereas its increase with temperature becomes weaker (so that at 33% Al, the temperature coefficient of resistivity even becomes negative). At the same time, the limiting high-temperature resistivity varies comparatively slightly and does not go outside the range of values typical of high-resistance alloys.

In the light of the existence of the saturation of resistivity, it is interesting to consider the evolution of the resistivity anisotropy in single-crystal materials. This anisotropy is clearly seen from the temperature dependence of the electric
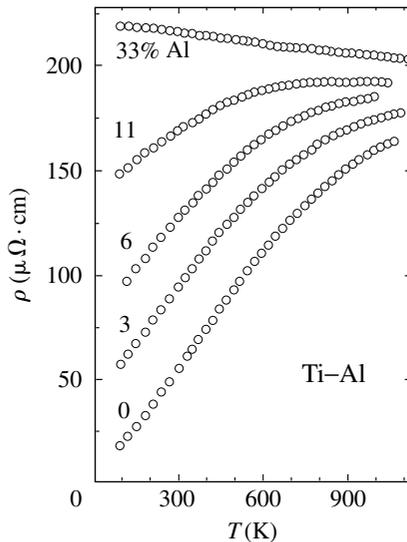
FIG. 1.5.   Resistivity of TiAl alloys with different Al concentrations as a function of temperature (Mooji 1973).
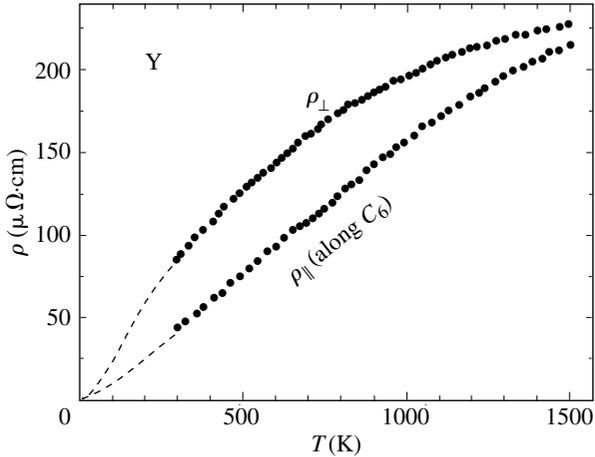
FIG. 1.6.   Anisotropy of the temperature dependence of resistivity for an yttrium single crystal (Zinov'ev et al. 1975).
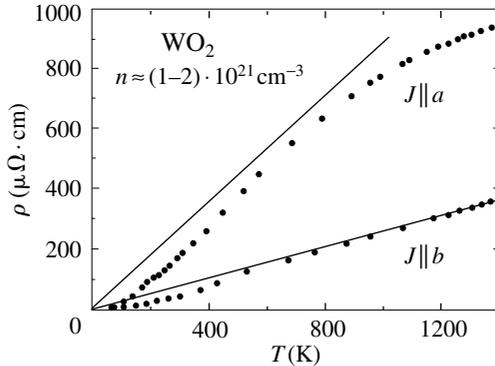
FIG. 1.7. $\rho(T)$ curves for a WO$_2$ single crystal measured along the crystallographic directions with the maximum and minimum resistivity (Gantmakher et al. 1986). Solid straight lines indicate the Grüneisen relation (1.14) in the high-temperature limit $\rho(T) = \alpha T$. The coefficient α for resistivity with the current $J \parallel \boldsymbol{a}$ is calculated under the assumption that eqn (1.15) is valid and that $\rho_{\mathrm{sh}} = 2000\,\mu\Omega\cdot\mathrm{cm}$.

conductivity of an yttrium single crystal shown in Fig. 1.6. In the temperature range 200–300 K, both resistivity and its temperature coefficient α measured in the plane normal to the sixfold axis are about twice higher than the values measured along this axis. Therefore, the in-plane resistivity approaches the critical value at lower temperatures and shows a tendency to saturation much earlier. As

a result, the difference between the resistivity measured in these two directions at $1400\,\mathrm{K}$ becomes less than 10%. The saturation resistivity ρ_{sh} measured along both directions is practically the same.

The saturation of resistivity is also observed in a "nonstandard" metal with a slightly lower concentration of the free electrons than in eqn (1.4). According to the data on the de Haas–van Alphen effect, the WO_2 oxide possessing a metallic conductivity has a carrier concentration of about $n \approx 2 \cdot 10^{21}\,\mathrm{cm^{-3}}$. This oxide has a monoclinic lattice and shows a pronounced (almost fourfold) anisotropy of resistivity at room temperature. Figure 1.7 shows the $\rho(T)$ curves measured along the directions with the minimum and maximum resistivity. Along the first direction (high resistivity), the curve has a tendency to saturation, $\rho_{\mathrm{sh}} \approx 2000\,\mu\Omega \cdot \mathrm{cm}$. Along the other direction, the resistivity is $\rho \ll \rho_{\mathrm{sh}}$ in the whole temperature range, follows the Grüneisen relation, and no signs of saturation are observed.

1.4. The Ioffe–Regel limit at high electron density

It is very useful to bring together all the facts and phenomena discussed above. Let us draw a diagram with the temperature T along one axis and the collision frequency (in energy units $\hbar/\tau$) along the other. As a characteristic scale along both axes we use the Fermi energy ε_{F}. Since for a standard metal $\varepsilon_{\mathrm{F}} \sim 10\,000\,\mathrm{K}$, the segment $(0\text{–}\varepsilon_{\mathrm{F}})$ on the temperature axis includes all the temperatures we are interested in. The frequency $1/\tau$ is assumed to be the sum of the frequencies of scattering from static defects, $1/\tau_0$, and phonons, $1/\tau_{\mathrm{ph}}$ (with due regard for the small-angle character of the phonon scattering at low temperatures), so that

$$\frac{1}{\tau} = \frac{1}{\tau_0} + \frac{1}{\tau_{\mathrm{ph}}}. \tag{1.17}$$

This frequency determines the resistivity ρ. The upper limit of the interval along the frequency axis, $\hbar/\tau = \varepsilon_{\mathrm{F}}$, corresponds to the Ioffe–Regel limit $l \approx k_{\mathrm{F}}^{-1}$. Higher values of collision frequencies would indicate the failure of the model of itinerant electrons with the wave functions in the form of plane waves. Therefore, all the transport phenomena in the gas of itinerant electrons should occur inside the square $0 \leqslant T$, $\hbar/\tau \lesssim \varepsilon_{\mathrm{F}}$ shown in Fig. 1.8.

All the topics considered in this chapter are associated with transport phenomena in the vicinity of the upper side of the square – in the region of pronounced scattering. First, we tried to approach the upper side of the square by moving along its left side where the temperatures are comparatively low. Here, the diffraction theory of liquid metals (developed mainly by Ziman) is applicable at the level of $(0.1\text{–}0.2)\varepsilon_{\mathrm{F}}$. An important element of this theory is the assumption that the wave functions of electrons are plane waves. In the vicinity of the upper left corner of the square, the dimensionless quantity $\hbar/\varepsilon_{\mathrm{F}}\tau$ stops being a small parameter, and the theory encounters serious difficulties. The experiments
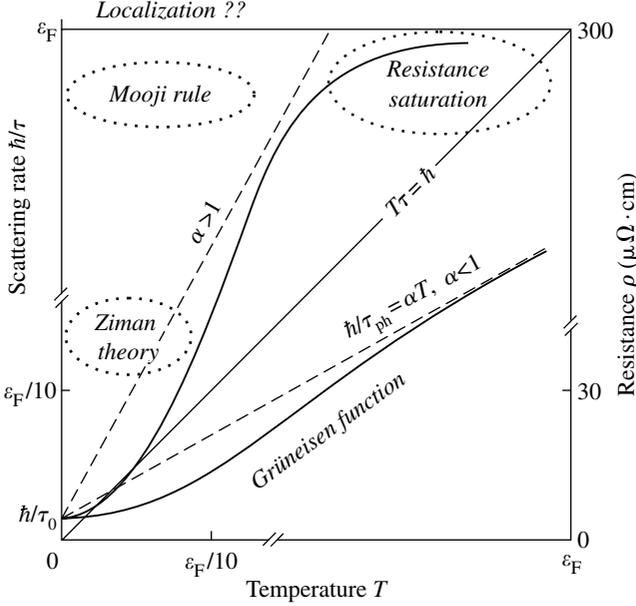
FIG. 1.8. The temperature–disorder plane $(T, \hbar/\tau)$. The dotted ellipses show the applicability regions of the Ziman theory and the Mooji rule and also the region where resistivity is saturated. The resistivity plots are drawn in the $(T, \hbar/\tau)$ coordinates for eqns (1.14) and (1.15). Dashed lines on both plots indicate the high-temperature asymptotic behavior of the Grüneisen function.

established two most important facts in this region, namely:

(i) The low-temperature resistivity of a standard metal does not exceed the value given by eqn (1.6),

$$\rho \lesssim \rho^* \approx 300 \, \mu\Omega \cdot \text{cm},$$

irrespective of how strong is the structural disorder.

(ii) In the upper left part of the square, the temperature curves of resistivity obey the Mooji rule, i.e., if the resistivity is close to that given by eqn (1.6), an increase in the temperature does not result in an increase in resistivity – it even slightly decreases it.

Now, in order to analyze the temperature dependence of resistivity of the materials with strong electron–phonon interaction, draw the diagonal of the square,

$$\hbar/\tau = T. \tag{1.18}$$

Remember this equation of the diagonal. Being written in the form $\tau = \hbar/T$, it recalls the expression for the dephasing time of two thermal electrons that determines

the quantum correction to conductivity induced by the interelectron interference. Therefore, the same square and its diagonal also appear in Chapter 2 dedicated to quantum corrections (see eqns (2.29) and (2.39) and Fig. 2.16 in Chapter 2).

Since $\rho \propto 1/\tau$, one can also draw the Grüneisen function (1.14) in the coordinate system $(T, \hbar/\tau)$ (for better understanding, we also show on the right the corresponding scale in resistivity units). In the $(T, \hbar/\tau)$ axes, the coefficient in the high-temperature asymptotic part of the Grüneisen function (1.14) becomes dimensionless; however, for the sake of simplicity, we preserve its notation α. Depending on the numerical values of the parameters of the metal under consideration, the asymptotic line αT can be located on either side of the diagonal (if numerical coefficients are ignored in calculations, all the algebraic factors are cancelled to yield $\alpha = 1$). If $\alpha < 1$, the asymptotic part of the curve is located in the lower triangle $T\tau > \hbar$. We are more interested in the case $\alpha > 1$, where the curve reaches the upper side of the square. However, the experiments described in the previous paragraph showed that in this case, one observes the saturation of resistivity at high temperatures not predicted by the Grüneisen relation.

Thus, neither an increase in the static or dynamic disorder nor their joint effect can ensure the intersection of the upper side of the square in Fig. 1.8, beyond which the Anderson localization is expected. This signifies that localization in a standard metal cannot be achieved by increasing the disorder. This is the experimental fact that should necessarily be taken into account. We shall return to the discussion of this problem in Chapter 5 when considering the concept of the metal–insulator transition and in Chapter 7 where we demonstrate how to find a way across the upper side of the square.

References

Faber, T. (1969). Electron transport phenomena in liquid metals, in: *The Physics of Metals 1. Electrons* (ed. J.M. Ziman). Cambridge University Press.

Fisk, Z. and Webb, G.W. (1976). Phys. Rev. Lett. **36**, 1084.

Gantmakher, V.F., Kulesko, G.I., and Teplinskii, V.M. (1986). JETP **63**, 832.

Gurvitch, M. (1981). Phys. Rev. B **24**, 7404.

Mooij, J.H. (1973). Phys. Stat. Sol. (a) **17**, 521.

Ziman, J.M. (1979). *Models of Disorder*. Cambridge University Press.

Zinov'ev, V.E., Sokolov, A.L., Gel'd, P.V., Chuprikov, G.E., and Epifanova, K.I. (1975). Sov. Phys. – Solid State **17**, 2353.

2

QUANTUM CORRECTIONS TO CONDUCTIVITY

A clear presentation of the main ideas associated with the quantum corrections to conductivity can be found in the book by Abrikosov (1988). The additional details, both experimental and theoretical, can be found in the reviews by Bergmann (1984) and Altshuler and Aronov (1985).

The creation of the theory of metals begins with the Bloch theorem on the behavior of electrons in an ideal periodic lattice. The next step is the allowance for deviations from periodicity in the τ-approximation. It is assumed that the violation of periodicity gives rise to electron scattering, i.e., transition of an electron from one stationary state in the ideal lattice to another. The sense of the parameter τ is associated with the electron response to various types of deviations from periodicity: impurities and other static defects, phonons, magnons, etc. At this stage, the additivity of scattering is postulated – it is assumed that the probability $1/\tau$ of the transition of an electron to a new state can be considered as the sum of the probabilities $1/\tau_i$ of electron scattering by various defects, impurities, phonons, etc. This assumption is quite natural if scattering events are rather rare as is the case, e.g., in pure metals.

The next step is the introduction of the so-called quantum corrections to conductivity. The necessary condition for their introduction is the occurrence of a series of scattering events.

Note: Here, conductivity is understood as specific conduction whose dimensions depend on the space dimension d:

$$\sigma_d[\Omega^{-1}\,\mathrm{cm}^{2-d}] \quad (d = 1, 2, 3). \tag{2.1}$$

There are two types of elementary scattering events. In the scattering events of the first type, an electron preserves its energy ε_j and, therefore, the temporal variation of the wave function, $\exp(i\varepsilon_j t/\hbar)$, is not changed either. The frequency of these elastic scattering acts is denoted as $1/\tau$. However, there are also some processes of inelastic scattering, e.g., collisions of electrons with phonons or other electrons. If an inelastic collision takes place at the moment t_0, then the electron changes its energy ε_j and "forgets" its phase that existed at time $t < t_0$ prior to this collision. We denote the probability of losing the phase memory as $1/\tau_\varphi$.

The quantum corrections we are going to consider may be introduced under the conditions that

$$\tau_\varphi \gg \tau. \tag{2.2}$$

Inequality (2.2) signifies that the events occurring with the highest frequency are the events of elastic scattering from static disorder. That is why when considering the quantum corrections to conductivity one often refers to dirty metals.

2.1. Weak localization

The quantum corrections to metal conductivity are associated with the electron wave properties that manifest themselves against the background of diffusion motion in the presence of a large number of elastic scatterers. The electron statistics is assumed to be degenerate, $T \ll \varepsilon_F$. An electron, which, at the moment $t = 0$, was at the origin of the coordinate system, $r = 0$, takes part in the diffusion motion with the Fermi velocity v_F and the mean free path $l = v_F \tau$. After time $t \gg \tau$, this electron may be found at the point $\boldsymbol{r}$ with probability

$$p(\boldsymbol{r}, t) = (4\pi Dt)^{-d/2} e^{-r^2/4Dt}, \quad r^2 = \sum_1^d x_i^2, \quad \int p(\boldsymbol{r}, t) \, d\boldsymbol{r} = 1. \tag{2.3}$$

Here d is the dimensionality of the space in which diffusion takes place and D is the diffusion coefficient defined as $D = l v_F/d$. The distribution width Δr gradually increases with time (Fig. 2.1) as

$$\Delta r \simeq \sqrt{Dt} \simeq l\sqrt{t/\tau} \simeq l\sqrt{N} \tag{2.4}$$

where $N = t/\tau$ is the number of steps in the diffusion process, i.e., the number of elastic collisions within the time t.

Equations (2.3) and (2.4) describe the diffusion of a classical particle. We will show now that the wave properties of an electron considerably change the function $p(r, t)$ at the particular point $r = 0$. To prove this statement, we select all the possible trajectories that have the form of a loop and return the electron to the point $r = 0$ and group them into pairs with equivalent sets of scatterers but with opposite directions of electron motion. For a classical particle, the probability $p(0, t)$ is the sum of the probabilities of its arrival at the point $r = 0$ along different trajectories. In quantum mechanics, this corresponds to the summation of the squared magnitudes of the corresponding wave functions. However, the quantum particle remembers its initial phase up to the moment $t \approx \tau_\varphi$. Hence, until $t < \tau_\varphi$ the wave functions should be summed up, and the probability $p(0, t)$ is the squared magnitude of this sum. The difference is noticeable:

classical summation $|A_1|^2 + |A_2|^2 = 2A^2,$
quantum summation $|A_1 + A_2|^2 = |A_1|^2 + |A_2|^2 + 2|A_1 A_2| = 4A^2.$ $\qquad$ (2.5)

Therefore, the probability of finding an electron at the origin, $p(0, t)$, should be doubled because of interference (Fig. 2.1), with the peak width being determined
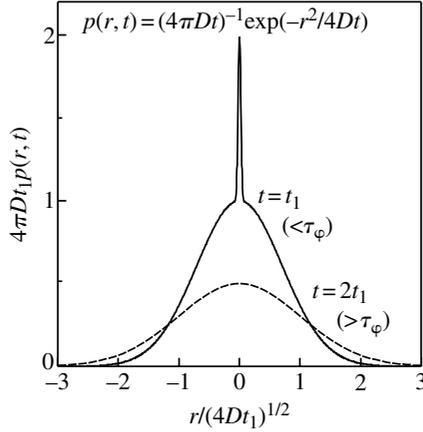
FIG. 2.1. The function $p(t)$ for two-dimensional diffusion at different moments, t_1 and $2t_1$. The additional peak on the curve $t < \tau_\varphi$ is due to quantum interference described by eqn (2.5).
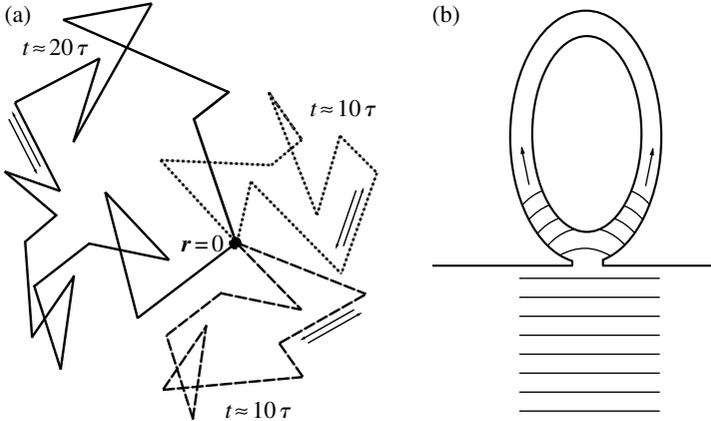
FIG. 2.2. (a) Closed diffusion trajectories which can be passed around by electron waves in opposite directions; (b) mechanical analogy: waves in a circular canal connected with a water pond.

by the uncertainty principle $\delta r \simeq \lambda \simeq 1/k_F$ (λ is the de Broglie wavelength and k_F is the Fermi wave vector). The division of the trajectories into interfering pairs is possible only for closed trajectories, i.e., the trajectories which end at the point $r = 0$ from which the electron started at the initial moment of time (Fig. 2.2a).

A higher probability of finding an electron at the point $r = 0$ means that it spends more time at this point. This leads to the corrections $\delta\sigma$ to conductivity and is called weak localization.

The physical processes underlying weak localization can be understood by invoking a mechanical analogy. Let a circular canal be connected with a large water pond (Fig. 2.2b). A wave approaching the canal from the pond enters both arms and forms two partial waves. If the waves do not damp in the canal (the damping time $\tau_\varphi > t$), both partial waves pass around the canal and meet again at the entrance.

The relative area under the additional peak of the function $p(r, t)$ depends on t. Therefore, in order to evaluate the relative correction to conductivity, $\delta\sigma/\sigma$, one has to calculate the addition $d\sigma$ to conductivity for the time dt formed because of the change in the function $p(r, t)$, and to integrate this addition over time. First, consider the three-dimensional case $d = 3$. The volume, at any point of which an electron can be found at the moment t, is of the order of $(Dt)^{3/2}$. The volume from which an electron can reach the origin within the time dt is of the order of $\lambda^2 v_{\mathrm{F}} dt$. The ratio of these two volumes determines the relative number of electrons that "revisit" the origin within the time dt. The minimum time necessary for an electron to return to the origin is the time of elastic scattering, τ. The returned electron can participate in interference if it has returned before the phase-breaking time, $t < \tau_\varphi$. As a result, we have for

$$d = 3: \quad \frac{\delta\sigma_3}{\sigma} \simeq -\int_\tau^{\tau_\varphi} \frac{v_{\mathrm{F}}\lambda^2 \, dt}{(Dt)^{3/2}} \simeq -\frac{v_{\mathrm{F}}\lambda^2}{D^{3/2}}\left(\frac{1}{\tau^{1/2}} - \frac{1}{\tau_\varphi^{1/2}}\right) = -(k_{\mathrm{F}}^2 l)^{-1}\left(\frac{1}{l} - \frac{1}{L_\varphi}\right).$$
$$(2.6)$$

The quantity introduced into eqn (2.6),

$$L_\varphi \simeq \sqrt{D\tau_\varphi} \simeq l\sqrt{\tau_\varphi/\tau} \gg l, \tag{2.7}$$

is called the *diffusion* or the *phase-breaking length*. Although the term with L_φ in eqn (2.6) is less than the term with l, it is this term that is responsible for the temperature dependence in the quantum correction to conductivity, because τ_φ depends on temperature and tends to infinity as $T \to 0$.

Since the normalization factor in eqn (2.3) for $p(\boldsymbol{r}, t)$ and, therefore, the denominator in the integrand of eqn (2.6), depends on the dimensionality d, the functions $\Delta\sigma/\sigma$ for various d turn out to be different. The characteristic length which is compared with the size of a particular sample is the diffusion length L_φ. In terms of diffusion, a film of thickness b and a wire with a diameter b are the objects of a reduced dimension if

$$b \ll L_\varphi. \tag{2.8}$$

For these objects, instead of eqn (2.6), we have

$$d = 2: \quad \frac{\delta\sigma_2}{\sigma} \simeq -\int_\tau^{\tau_\varphi} \frac{v_F \lambda^2 \, dt}{(Dt)b} \simeq \frac{v_F \lambda^2}{Db} \ln(\tau_\varphi/\tau) \simeq \frac{1}{(k_F l)(k_F b)} \ln(\tau_\varphi/\tau),$$

$$d = 1: \quad \frac{\delta\sigma_1}{\sigma} \simeq -\int_\tau^{\tau_\varphi} \frac{v_F \lambda^2 \, dt}{(Dt)^{1/2} b^2} \simeq \frac{2v_F \lambda^2}{Db^2}(L_\varphi - l) \simeq \frac{1}{(k_F b)^2}\left(\frac{L_\varphi}{l} - 1\right).$$

$$\text{(2.9)}$$

Condition (2.8) for dimension reduction is rather mild. It determines the configuration of the space operating with scale L_φ while the diffusion takes place in scale l. For a film with $l \ll b \ll L_\varphi$, the diffusion process remains three-dimensional; an electron in the film can move between two elastic-scattering events along an arbitrary direction, including the direction along the film normal. However, the quantum correction for such a film is determined by eqn (2.9) assuming that $d = 2$.

The corrections $\delta\sigma_i$ in expressions (2.6) and (2.9) are negative. Therefore, irrespective of dimension, beginning with a certain temperature, the conductivity $\delta\sigma_i$ decreases (resistivity increases) with lowering of the temperature. The interference effect is the more pronounced, the lower the dimension. In the three-dimensional case, a decrease in $\delta\sigma_3$ is limited and $\delta\sigma_3$ tends to a certain limit as $T \to 0$. At a reduced dimension, the corrections diverge because both τ_φ and L_φ tend to infinity at $T \to 0$. Since conductivity cannot be negative, there should exist some limits of the applicability of eqns (2.6) and (2.9). The limitation reduces to the requirement of the relative smallness of the corrections

$$\delta\sigma_i \ll \sigma \quad (i = 1, 2, 3). \tag{2.10}$$

It is useful to consider not the reduced values of corrections (2.6) and (2.9) to the 3D conductivity σ, eqn (1.5), but the absolute values of the changes of the conductivity $\Delta\sigma_d = \delta\sigma_d \, b^{3-d}$ which have the same dimensions as the conductivity σ_d, eqn (2.1):

$$d = 3: \quad \Delta\sigma_3 \approx -\text{const.} + \left(\frac{e^2}{\hbar}\right)L_\varphi^{-1},$$

$$d = 2: \quad \Delta\sigma_2 \approx -\left(\frac{e^2}{\hbar}\right)\ln\left(\frac{\tau_\varphi}{\tau}\right) \approx -2\left(\frac{e^2}{\hbar}\right)\ln\left(\frac{L_\varphi}{l}\right), \tag{2.11}$$

$$d = 1: \quad \Delta\sigma_1 \approx -\left(\frac{e^2}{\hbar}\right)L_\varphi.$$

All the corrections have the same scale, $e^2/\hbar$. This combination of the atomic constants with the dimensions of reciprocal resistance, $\hbar/e^2 = 4110\,\Omega$, is encountered in all the problems associated with localization. It is important that the quantum corrections (2.11) contain no carrier concentration, whereas the dependence of quantum corrections (2.11) on τ is weaker than the dependence of conductivity σ_d on τ. Therefore, the role of interference corrections increases with a decrease

in the initial conductivity of the material. This is another explanation of the fact that weak localization is usually considered as a dirty-metal effect.

Note: Since $\tau_\varphi \to \infty$ as $T \to 0$, then, at low temperatures, inequality (2.2) is fulfilled for any pure metal with an arbitrarily high τ. It is perfectly possible to grow a single crystal of pure metal, e.g., indium, with a mean free path of the order of $l \approx 0,1$ cm, so that the parameter $k_F l$ would be of the order of $k_F \approx 10^6$. At $T = 0,1$ K, the ratio τ_φ/τ is of the order of 10^3–10^4, i.e., inequality (2.2) is fulfilled. However, the contribution of the quantum correction to conductivity is negligible:

$$\frac{\delta\sigma_3}{\sigma} \sim \frac{1}{k_F^2 L_\varphi l} \sim 10^{-14}.$$

Note also: Formally, even a very thick film at a rather low temperature should be considered as two-dimensional. As $T \to 0$, the length L_φ tends to infinity, $L_\varphi \to \infty$, and, below a certain temperature, inequality (2.8) would be fulfilled. Then, the classical conductivity $\sigma_3 b$ should be corrected for $\Delta\sigma_2$.

Equations (2.6), (2.9) and (2.11) were repeatedly verified experimentally mainly on films. Fig. 2.3 summarizes the results of two experiments performed on Cu and Au films. Both materials are characterized by a logarithmic increase in conduction at temperatures below 10 K. Despite the fact that the corrections $\Delta R/R$ in the temperature range 1–10 K considerably differ, the absolute values of the corrections to conductivity, $\Delta\sigma$, differ much less, as was to be expected from eqn (2.11). The weak dependence of $\Delta\sigma$ on the material properties becomes obvious if one compares the measurements made on Cu and Au films with the measurements on amorphous In–O films (whose resistivity is 1000 times higher and a logarithmic increase in resistance starts with lowering of the temperature below 100 K). The $\Delta R/R$ values of the In–O films (Fig. 2.4) amount to 15, 6,
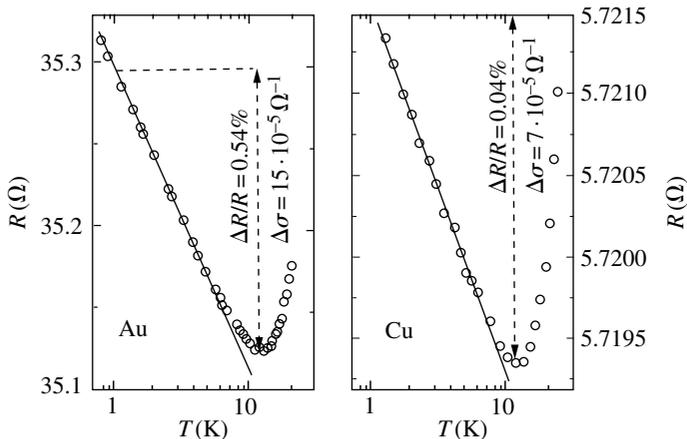
FIG. 2.3.   Temperature dependence of resistivity of thin Cu (Van der Dries 1981) and Au (Dorozhkin and Dolgopolov 1982) films. Dashed arrows indicate increase in resistivity at a tenfold decrease in temperature (from 10 to 1 K).
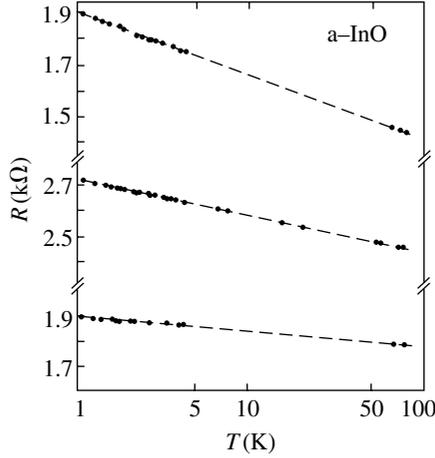
FIG. 2.4.   Resistivity of thin amorphous In–O films as a function of temperature (Ovadyahu and Imry 1981). The films have different oxygen content, which is reflected in the concentration of free carriers.
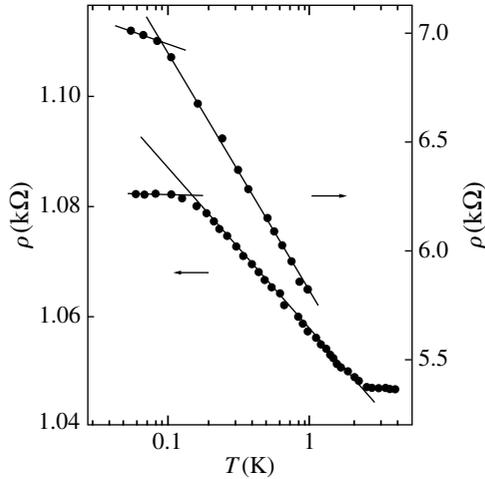
FIG. 2.5.   Logarithmic temperature increase in resistivity in the inverse Si layers for two samples (Bishop et al. 1980). Straight lines are drawn to guide the eye.

and 3% at a tenfold temperature change, whereas their $\Delta\sigma$ values range within $(1–2)\cdot10^{-5}\,\Omega^{-1}$, i.e., are only slightly lower than for Cu and Au films.

The logarithmic increase in resistance was also observed in true two-dimensional electronic systems, e.g., in inverse layers on Si surfaces in the temperature range 0.1–1.0 K (Fig. 2.5). The $\Delta R/R$ values for two samples were 16 and 3%, whereas the $\Delta\sigma$ values were of the order of $2.5\cdot10^{-5}\,\Omega^{-1}$.

Note: Irrespective of the acting mechanism of inelastic scattering of electrons, determining the phase-breaking time, the τ_φ and L_φ quantities are power functions of temperature, $L_\varphi \propto T^s$. Therefore, the quantum correction for films always manifests itself as the specific dependence $\Delta R \propto \ln T$ and s enters this relationship as the proportionality coefficient.

2.1.1. *Coherent backscattering of light*

Since weak localization is of the wave-interference nature, it should have an optical analogue. This analogue really exists – it is light scattering in disordered media. Illuminating the disordered medium from one side, we can see it at any angle, because the multiply scattered light exits from the illuminated volume in all directions with equal probability – the phenomenon of a blue sky. Since scattering is elastic, the modulus of the wave vector does not change and the scattering of a plane light wave $A \exp(i\boldsymbol{kr})$ incident on the medium may be described by the motion of the end of the vector $\boldsymbol{k}$ over the surface of a sphere $|\boldsymbol{k}| = $ const. In particular, part of the light beam is scattered in the backward direction. This part of the beam is described by the point $-\boldsymbol{k}$ on the sphere. In the classical description, there is nothing particular about this point, because if scattering is strong, the amplitude of the scattered wave is

$$A(\boldsymbol{k}) \rightarrow \text{const.} \quad \text{at } |\boldsymbol{k}| = \text{const.}$$

For plane waves, there exists an interference effect analogous to weak localization. The transitions from $\boldsymbol{k}$ to $-\boldsymbol{k}$ is the result of a random walk over the surface of the sphere with the step-by-step change of point $\boldsymbol{k}$ by $\Delta_i \boldsymbol{k} = \boldsymbol{q}_i$, i.e., $\boldsymbol{k} + \sum_i \boldsymbol{q}_i = -\boldsymbol{k}$. There exist an infinite number of sequences $\{\boldsymbol{q}_i\}$ that can be divided into interfering pairs of the same set of vectors but considered in the opposite order (Fig. 2.6) (partial permutation, e.g., of two random vectors
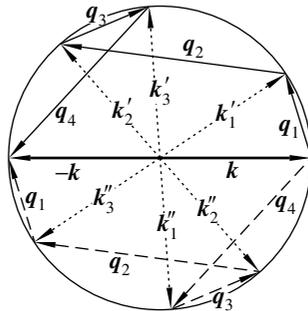
FIG. 2.6. Weak localization in the k-representation. Each random-walk trajectory on the surface of the Fermi sphere $\boldsymbol{k} \rightarrow \boldsymbol{k}'_1 \rightarrow \boldsymbol{k}'_2 \rightarrow \boldsymbol{k}'_3 \rightarrow -\boldsymbol{k}$ which results in a transition $\boldsymbol{k} \rightarrow -\boldsymbol{k}$ has the complementary trajectory $\boldsymbol{k} \rightarrow \boldsymbol{k}''_1 \rightarrow \boldsymbol{k}''_2 \rightarrow \boldsymbol{k}''_3 \rightarrow -\boldsymbol{k}$. The interference of waves scattered via these two series of states increases the backscattering.

in summation over i is inadmissible because the intermediate $\boldsymbol{k}_i$ values would not lie on the surface of the sphere). For each interfering pair of trajectories in k-space, the speculations made earlier in the derivation of (2.5) may be repeated. These considerations indicate a relative increase in backscattering.

The experimental setup for implementation of "weak localization in the k-representation" is schematically shown in Fig. 2.7. It records the angular dependence of the light intensity scattered by an aqueous suspension of polystyrene beads with diameters $0.46\,\mu$m occupying 10% of the total volume. The results of these experiments are shown in Fig. 2.8 (curve a) together with the analogous signals from the cell with pure water (curve b) and in the absence of any cell (curve c). The scales for the three curves are identical.
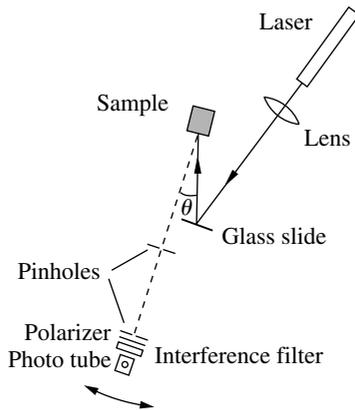
FIG. 2.7. Experimental setup for studying backscattering of light (Wolf and Maret 1985).
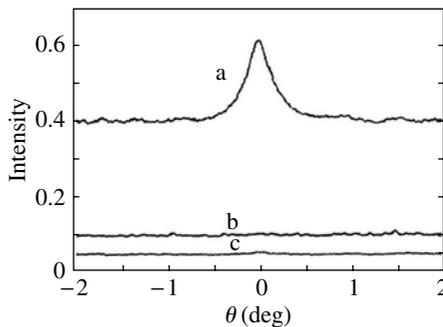
FIG. 2.8. Angular dependence of the scattered light intensity (a) by aqueous suspension of polystyrene beads (solid fraction 10%); (b) by the same cell filled with pure water, and (c) in the absence of any cell (Wolf and Maret 1985).

2.2. Effect of magnetic field on weak localization

A magnetic field B bends electron trajectories. Two electrons moving toward each other are deviated by the field in opposite directions. The angle of electron motion changes on the way between two impurities by a value of the order of $l/r = \Omega\tau$ (here r is the cyclotron radius and $\Omega = eB/mc$ is the cyclotron frequency). Usually, one considers dirty metals with a not too high τ value, therefore, in moderate fields $\Omega\tau \ll 1$, and the difference between the scattering angles can be ignored in comparison with other much more pronounced effects.

If an electron travels along the loop in a magnetic field, its wave function Ψ acquires an additional phase factor

$$\Psi \;\rightarrow\; \Psi \exp\left(i\frac{e}{\hbar c}\int \boldsymbol{A}\,d\boldsymbol{l}\right) = \Psi \exp\left(\pm\frac{i\pi BS}{\Phi_0}\right). \tag{2.12}$$

where $\boldsymbol{A}$ is the vector potential of the magnetic field, $\Phi_0 = \pi\hbar c/e$ is a magnetic-flux quantum, and $BS = \Phi$ is the total magnetic flux through the loop. The sign of the exponent depends on whether an electron moves clockwise or anticlockwise along the loop. Since an electron waves move along the loop in the opposite directions, then, upon their return to the origin, the two parts of the electron wave function would acquire the phase difference $\varphi = 2\pi(\Phi/\Phi_0)$.

The area S of the projection of a loop onto the plane perpendicular to the field is of the order of the squared mean distance $r(t)$ covered by the electron for the time t. Therefore,

$$\varphi = 2\pi\left(\frac{BS}{\Phi_0}\right) \simeq \frac{Br^2}{\Phi_0} \simeq \frac{BDt}{\Phi_0}. \tag{2.13}$$

The phase difference indicates that the second relationship in (2.5) should be replaced by

$$|A_1 + A_2|^2 = |A_1|^2 + |A_2|^2 + 2|A_1||A_2|\cos\varphi = 2A^2(1 + \cos\varphi). \tag{2.14}$$

If φ is small, this correction is inessential. However, if $\varphi > 1$, the scatter in the areas S of the loops becomes important because the phase (2.13) becomes uncertain. In Fig. 2.2, this scatter is illustrated by the difference between the areas of the dashed and dotted contours. The average value of the term with the phase in eqn (2.14) is $\overline{\cos\varphi} = 0$ and, therefore, the magnetic field destroys the interference. The longer the time t and the larger the average area $\overline{S} \simeq Dt$ of the closed trajectories formed within this time, the lower the field B that destroys the interference contribution. We are not interested in times $t > \tau_\varphi$, because the interference within these times is destroyed by inelastic collisions. The neighborhoods of the upper integration limit τ_φ in integrals (2.6) and (2.9) are important because it is these neighborhoods that enter eqn (2.11) via L_φ. Therefore, assuming that $\varphi = 1$ in the left-hand side of eqn (2.13) and substituting τ_φ instead of t in the right-hand side of eqn (2.13), we arrive at the field value B_φ sufficient to start breaking weak localization,

$$B_\varphi = \frac{c\hbar}{e}(D\tau_\varphi)^{-1}. \tag{2.15}$$

Writing the diffusion coefficient as $D \simeq \varepsilon_F \tau / m$ and introducing the cyclotron frequency $\Omega_\varphi = eB_\varphi / mc$ in the field B_φ, we obtain from eqn (2.15)

$$\Omega_\varphi \tau \frac{\varepsilon_F \tau_\varphi}{\hbar} \approx 1, \quad \text{whence } \Omega_\varphi \tau \ll 1. \tag{2.16}$$

Inequality (2.16) confirms that the bending of electron trajectories in the field B_φ can be ignored.

Now, introduce the magnetic time as

$$\tau_B = \frac{\Phi_0}{BD} \approx \frac{r_B^2}{D} \approx \Omega^{-1}(k_F l)^{-1}, \tag{2.17}$$

where $r_B = (\hbar c / 2eB)^{1/2}$ is the magnetic length. If $\tau_B \ll \tau_\varphi$, i.e., if $B \gg B_\varphi$, one has to change the upper integration limit in eqns (2.6) and (2.9), $\int_\tau^{\tau_\varphi} \rightarrow \int_\tau^{\tau_B}$. Then

$$0 < \Delta\sigma(B) - \Delta\sigma(0) \approx \begin{cases} 2\dfrac{e^2}{\hbar} \ln(L_\varphi / r_B), & d = 2, \\ & \qquad\qquad (l \ll r_B \leq L_\varphi). \\ \dfrac{e^2}{\hbar} \left[\dfrac{1}{r_B} - \dfrac{1}{L_\varphi} \right], & d = 3, \end{cases} \tag{2.18}$$

The conductivity $\sigma(B)$ in eqn (2.18) is compared with the conductivity in the zero field, $\sigma(0)$. In some instances, it is more convenient to insert into eqn (2.18) the conductivity $\sigma(\infty)$ in a very strong field instead of $\sigma(0)$.

Figure 2.9 illustrates the destruction of weak localization in Mg films by a magnetic field. The total decrease in conductivity is the more pronounced, the lower the temperature T. This experimental fact fits eqn (2.9) in which a decrease in T results in an increase of the upper integration limit. In this case, with a
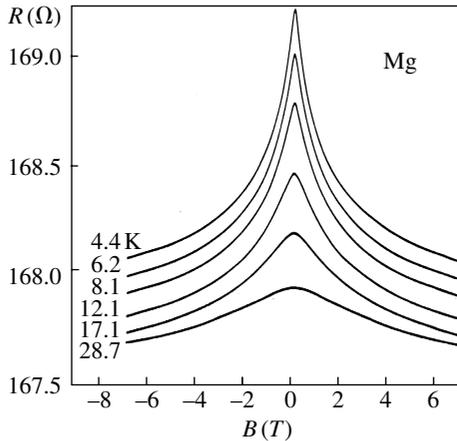
FIG. 2.9. Magnetoresistance of a thin Mg film for different temperatures (Bergmann 1984).

decrease in T, the region of the pronounced decrease in resistivity is displaced toward lower fields. This process is completed in the field

$$B_{\mathrm{el}} = \hbar c/el^2, \quad \text{at } \tau_B = \tau, \tag{2.19}$$

i.e., when the magnetic time is equal to the elastic time. Gradually decreasing the strong field, we observe, first, interference on the loops with the minimum areas S of the projections onto the plane normal to the magnetic field, $S \sim l^2$. This happens in the field $B \lesssim B_{\mathrm{el}}$. With a further decrease in the field, the interference is "switched on" at the trajectories with ever increasing S values, which is accompanied by an increase in resistivity. This process is finished in field (2.15)

$$B_{\varphi} = \hbar c/eL_{\varphi}^2, \quad \text{when } \tau_B = \tau_{\varphi}. \tag{2.20}$$

To conclude, the conductivity depends on the field B in the field interval $B_{\varphi} < B < B_{\mathrm{el}}$ and only the low-field edge of this interval is temperature dependent. The shift of the $R(B)$ curves in strong fields (Fig. 2.9) is explained by the temperature dependence of the classical part of the magnetoconductivity.

Thus, the magnetic field may be used to prove the existence of weak localization. What is really seen in the experiments is a negative magnetoresistance in weak fields. This kind of experimental observation of weak localization looks like a proof by contradiction, because we evidence the process of breaking the weak localization. However, in order to demonstrate weak localization, the resistivity dependence on the magnetic field in Fig. 2.9 is as objective as the resistivity dependence on temperature (Figs 2.3–2.5).

The consideration of eqn (2.14) shows that the wave correction is destroyed by a field of the order of B_{φ}, because of the scatter in the areas of the closed interfering trajectories. Designing the experiment in such a way that all the loop projections onto the plane normal to the magnetic field would have the same areas S, we avoid the destruction of the interference contribution. Instead, with an increase in φ, the contribution would start oscillating with the field with the period

$$\Delta B = \frac{\Phi_0}{S} = \frac{\pi \hbar c}{eS}. \tag{2.21}$$

This configuration was implemented experimentally. Evaporating a thin metal layer ($\delta r \ll r$) onto a quartz filament with diameter $2r \simeq 1$–$2\,\mu\mathrm{m}$, we obtain a cylindrical film. The area S of the projection of a closed diffusion trajectory onto the plane normal to the cylinder axis equals either zero or πr^2. The magnetic field directed along the cylinder axis does not affect the interference of the trajectories with $S = 0$. On the contrary, the contribution from the trajectories with $S = \pi r^2$ to the conductivity along the cylinder axis should oscillate with the period given by eqn (2.21). To make this contribution comparable with the conventional quantum correction, it is necessary to satisfy the condition $S < D\tau_{\varphi} \simeq L_{\varphi}^2$ (small cylinder diameter and low temperature). However, it seems that the oscillating addition may be observable also if $S \sim L_{\varphi}^2$.
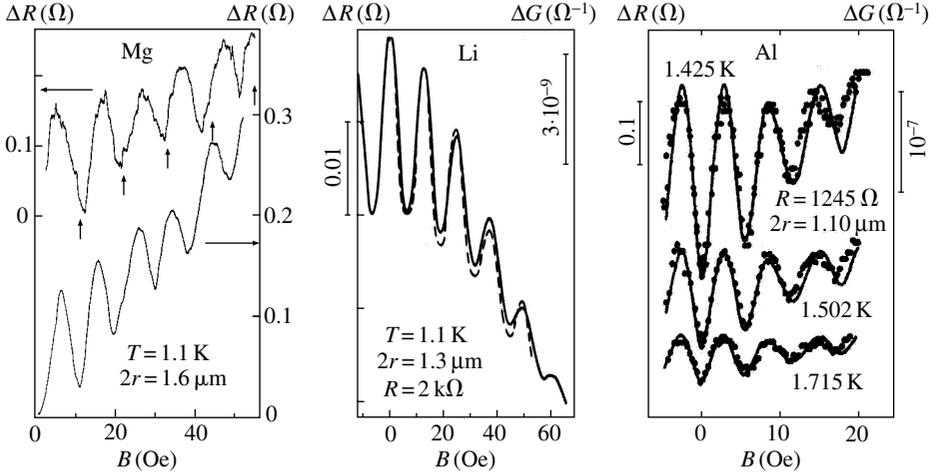
FIG. 2.10. Longitudinal magnetoresistance oscillations in cylindrical Mg, Al, and Li films. For Mg, solid lines show the experimental curves, the arrows indicate the calculated values of the field at which the resistivity extrema should be observed (Sharvin and Sharvin 1981). For Li, the solid line shows the experimental curves and the dashed line, the calculated ones (Altshuler et al. 1982). For Al, the points indicate the experimental values and the solid lines present the calculations (Gijs et al. 1984). For Li and Al curves, two various scales (of resistance variation, ΔR, and of conductance variation, $\Delta G = \Delta R/R^2$) are indicated.

Figure 2.10 shows the experimental $\Delta R/R$ curves obtained by two groups of experimentalists on Mg, Li, and Al cylindrical films. It is seen that the theoretical calculations may explain not only the oscillation periods but also the dependence of their amplitudes on the field and their monotonic run.

Our previous considerations show that, in the field $B = 0$, the resistivity should be maximal. However, this is observed only for Li films, whereas the oscillations in Mg and Al films have the opposite phases. We return to this fact in the next section after acquaintance with the effect of the spin–orbit interaction on interference.

2.3. Antilocalization

Electron waves propagating in opposite directions along a closed diffusion trajectory and interfering at the exit have the fixed spin projection, $+\frac{1}{2}$ or $-\frac{1}{2}$. In fact, two electrons with different spins coexist on each trajectory. Each of these electrons generates a pair of electron waves propagating in opposite directions. The pairs of waves meet at the initial point and each of these waves interferes with the complementary one. If no spin flip takes place, the two electrons are

absolutely independent and give their additive contributions into the quantum correction to conductivity. However, the situation becomes different if the spin–orbit interaction takes place. Then, the electron spin may flip in the course of elastic scattering and the two pairs of waves mix.

The spin–orbit interaction arises because the magnetic moment μ moving with velocity $\boldsymbol{v}$ gives rise to an electric field $\boldsymbol{e} \propto [\boldsymbol{\mu v}]$ in the stationary coordinate system. This field interacts with the field $\boldsymbol{E}$ of the ion charges, which results in the spin flip. The field intensity E and also the intensity of the spin–orbit scattering proportional to

$$\boldsymbol{E e} \propto \boldsymbol{E} \, [\, \boldsymbol{\mu v}\,], \tag{2.22}$$

are strongly dependent on the charge of the nucleus, i.e., on the atomic number Z. We have $\tau_{\mathrm{so}}^{-1} \propto (Z\alpha)^4$, where $\alpha = 1/137$ is the fine structure constant. Therefore, the effect of the spin–orbit interaction is more pronounced in materials containing heavy atoms.

Not every elastic-scattering event is accompanied by a spin flip, and, therefore, the spin-flip time is much longer than the elastic-scattering time, $\tau_{\mathrm{so}} \gg \tau$. The ratio τ_{so}/τ is the higher, the weaker the spin–orbit interaction, with the time τ_{so} being independent of the temperature. Since $\tau_\varphi \to \infty$ as $T \to 0$, then, at sufficiently low temperatures, we have

$$\tau \ll \tau_{\mathrm{so}} \ll \tau_\varphi. \tag{2.23}$$

To take into account the mixing of wave functions of electrons with different spins, consider a compound particle consisting of two electrons. The total spin of this pair can be in one of four possible states. These are the zero total spin (singlet) and unity total spin with three different projections (triplet). Therefore, the wave function Ψ of the pair has four components

$$\Psi = \begin{pmatrix} \Psi_0 \\ \Psi_{1,-1} \\ \Psi_{1,0} \\ \Psi_{1,1} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \left(\varphi_+^{(1)} \varphi_-^{(2)} - \varphi_-^{(1)} \varphi_+^{(2)} \right) \\ \varphi_-^{(1)} \varphi_-^{(2)} \\ \frac{1}{\sqrt{2}} \left(\varphi_+^{(1)} \varphi_-^{(2)} + \varphi_-^{(1)} \varphi_+^{(2)} \right) \\ \varphi_+^{(1)} \varphi_+^{(2)} \end{pmatrix}, \tag{2.24}$$

where $\varphi^{(1)}$ and $\varphi^{(2)}$ are the wave functions of the first and second electrons, respectively, and the subscripts $+$ and $-$ relate to the spin part of the wave function and indicate different spin projections. This electron pair interferes with itself if the corresponding waves meet after the bypass of a loop trajectory in opposite directions. The total interference is the sum of four terms – the terms due to the three components of the wave function Ψ_{1m} with spin 1 (and the projections $m = 0, \pm1$) and the component Ψ_0 with the zero spin.

Omitting the calculations, we indicate and comment here only on the final result

$$\frac{\delta\sigma_d}{\sigma} \simeq -\int_\tau^{\tau_\varphi} \frac{v\lambda^2\,dt}{b^{3-d}(Dt)^{d/2}} \left(\frac{3}{2}\,e^{-t/\tau_{\mathrm{so}}} - \frac{1}{2}\right), \quad (d=1,2,3). \tag{2.25}$$

Comparing this result with eqns (2.6) and (2.9), we see in the integrand a new factor in parentheses consisting of two terms. The first term is due to the three triplet components Ψ_{1m} with spin 1. The exponential attenuation of this term indicates that the interference of these functions is important only until the moment $t < \tau_{\mathrm{so}}$, i.e., until the moment at which the electrons still "remember" the initial spin. Within this time interval, we may assume that $\exp(-t/\tau_{\mathrm{so}}) \approx 1$, so that the expression in parentheses is of the order of unity, and the integrand coincides with that in eqns (2.6) and (2.9) written for a spinless electron. This is quite natural because up to this moment no spin flip has taken place.

The most interesting events occur within the time interval

$$\tau_{\mathrm{so}} < t < \tau_\varphi, \tag{2.26}$$

where the first term in parentheses may be ignored and then the expression is reduced to the negative second term associated with interference of the wave component Ψ_0. The minus sign indicates that the interference addition remaining after the spin flip does not decrease the conductivity; on the contrary, it even increases it. The final sign of the correction depends on the ratio $\tau_{\mathrm{so}}/\tau_\varphi$. If $\tau_{\mathrm{so}} \ll \tau_\varphi$, the interference contribution to conductivity is positive. Sometimes, this limit is called antilocalization.

One should not think that the *phases* of the two parts of the wave function of a singlet pair after the bypass of the loop in opposite directions would differ by a factor of π. The interference is observed if $t > \tau_{\mathrm{so}}$, i.e., after all the spins have been flipped. Each spin flip transforms the triplet pair into the singlet one and, vice versa, the singlet pair into the triplet one. Therefore, if $t > \tau_{\mathrm{so}}$, the singlet part of the wave function (three quarters of it) consist of ex-triplets. The singlet and triplet wave functions have different structures. It is seen from eqn (2.24) that the triplet components are symmetric with respect to particle permutations, whereas the singlet component is antisymmetric. This results in the appearance of the minus sign before the *amplitude* of the interference contribution from the singlet part of the wave function.

Note: All the above is valid only for the spin–orbital interaction and has no relation to scattering from a paramagnetic impurity accompanied by a spin flip. After the flip of the spin of the impurity center, an electron with spin μ turns out to be in the state with the spin $-\mu$, the state that initially was empty. This process is analogous to conventional scattering without spin flip and results not in antilocalization but in weak localization.

Of course, all the above speculations cannot replace direct calculations and, therefore, are somewhat unsatisfactory. But it is expedient to clarify which of

the antilocalization problems is only a technical one and which one is of principal importance. Interference may result nor only in an increase, but also in a decrease of the wave amplitude, with the sign of the changes often being unknown. Suffice it to recollect the discussion of the Fresnel formulas in the Paris Academy of Sciences. These formulas indicate the presence of a light spot on the screen along the axis of the beam diffracting from a nontransparent circular disk. This fact was first interpreted as an indication of the failure of the wave theory of light. Later, after the experimental observation of this spot, it was accepted as convincing proof of this theory. Logically, both weak localization and weak antilocalization are equally possible results of electron wave interference.

There is another aspect that is quite unexpected and very important – an electron wave continues taking part in the interference even after the spin–orbit interaction which resulted in the spin flip. Electrons are in contact with the surrounding quantum system, which may be called a thermostat. The effect of a scattering event on the electron-wave coherence (i.e., its ability to interfere) can be estimated from the state of the thermostat. Elastic scattering of a spinless electron leaves no traces in the thermostat. Thus, the coherence is preserved despite the fact that a scattered electron can be in any state on the sphere $|\boldsymbol{k}| = $ const. Inelastic electron scattering from the lattice results in a change of the thermostat state – the appearance or disappearance of a phonon. As a result, the electron becomes incoherent.

The spin flip may also occur in two different ways. If electron scattering from a magnetic impurity is accompanied by the spin flip, then, because of the conservation of the total spin, the impurity spin also flips. This signifies the change of the thermostat state although, generally speaking, both electron and thermostat energies may remain unchanged. Indeed, along with phonon scattering, the scattering from magnetic impurities makes the contribution to $1/\tau_\varphi$ and *does not lead* to antilocalization. Because of the spin–orbit interaction, the spin flip in scattering from nonmagnetic impurities leaves no traces in the thermostat and, thus, preserves the electron coherence. Nevertheless, the result of the interference is different because it took place under different interference conditions. Now, numerous coherent states are located on the surfaces of two spheres, $|\boldsymbol{k}_-| = $ const. and $|\boldsymbol{k}_+| = $ const., where the subscripts + and – indicate the direction of the spin projection.

Now consider the experimental study of antilocalization. It is most convenient to observe antilocalization by examining the dependence of conductivity on the magnetic field that destroys the interference. Figure 2.11 shows the corresponding $R(B)$ curves for copper films. Copper atoms are heavier than magnesium ones, and, therefore, the spin–orbit interaction in copper should be more pronounced. Indeed, comparing the curves for Cu and Mg (Figs 2.9 and 2.11), we see a small minimum on the curve for Cu in weak fields in the zero field. According to (2.17), the magnetic time is proportional to the field, $\tau_B \propto B^{-1}$. Therefore, increasing the field from the zero value, we see that, first, the τ_B value becomes comparable with τ_φ, which results in decreasing the upper integration limit and the gradual disappearance of that part of integral (2.25) which is responsible
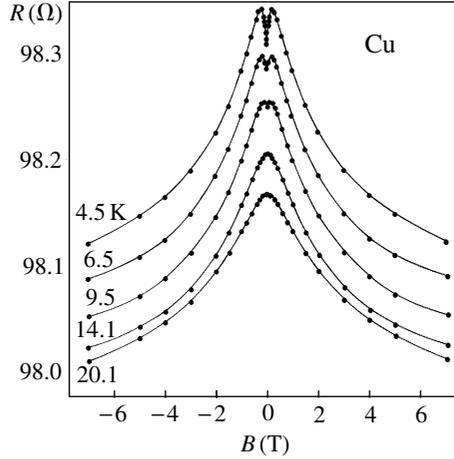
FIG. 2.11. Magnetoresistance of a thin Cu film at different temperatures (Bergmann 1984). The appropriate selection of the parameters allows one to make an ideal fit to the experimental points (solid lines).

for antilocalization. This should be accompanied by a decrease in conductivity and an increase in resistivity. Then, τ_B becomes comparable with τ_{so}, and the antilocalization region in the integrand completely disappears. A further increase in the field destroys the conventional weak localization. Since τ_{so} is temperature independent and τ_φ decreases with an increase in temperature, then the region (2.26) does not exist at higher temperatures and, therefore, the $R(B)$ curves have no minima.

It is interesting to follow the evolution of the $R(B)$ curves with a decrease in the time τ_{so}. This can be made by performing experiments on Mg films coated with a small amount of gold. Gold atoms play the part of the scattering centers that provide spin flip. The magnetoresistance of Mg films is shown in Fig. 2.12 for different coverage with Au. The τ_φ/τ_{so} ratio plays the role of the fitting parameter for comparing the experimental and calculated curves.

Now, return to the conductivity oscillations in cylindrical films (Fig. 2.10). It is natural to interpret the opposite phases of oscillations in the Mg and Al films, on the one hand, and of oscillations in Li, on the other hand, by the effect of spin–orbit interaction. Although Mg, Al, and Li are light elements, the atomic number of Li is much lower than the atomic numbers of the two other metals: $Z_{Li} = 3, Z_{Mg} = 12, Z_{Al} = 13$, so that the difference in τ_{so}, with all the other conditions being the same, amounts to two orders of magnitude.

Note: The inconsistency between the data for Mg in Figs 2.9 and 2.10 (opposite signs of the derivative $\partial R/\partial B$ at $B = 0$) is only apparent. The time necessary for an electron to pass around the cylinder is of the order of $\tau_{cyl} \simeq \tau(r/l)^2$. The interference along such trajectories occurs at times $t > \tau_{cyl}$. The magnetic field that destroys interference
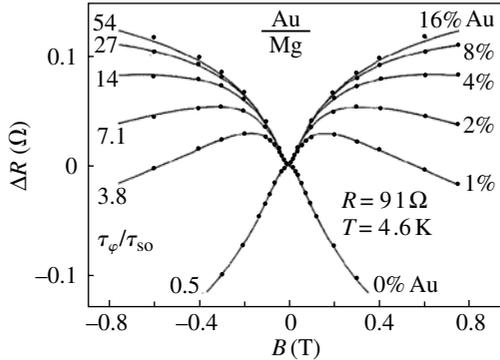
FIG. 2.12. Magnetoresistance of thin Mg films with different Au coverage (Bergmann 1984). The Au thickness as a percentage of an atomic layer is indicated on the right. On the left, the ratio $\tau_i/\tau_{\mathrm{so}}$ used for the theoretical fit is indicated. This ratio is practically proportional to the coating thickness.

within this time is determined by eqn (2.17). It is of the order of 10 Oe as well as the oscillation periods in Fig. 2.10. In fact, the oscillation phase for Mg and Al films show that the corresponding $R(B)$ curves should possess a 10 Oe-wide minimum similar to that of the curves for Cu. This minimum is indistinguishable in Fig. 2.9.

Equation (2.22) qualitatively describes the spin–orbit interaction. The crystal potential U enters this formula twice, via the electric field $\boldsymbol{E} = \nabla U$ and the velocity, which is a derivative in dispersion law, $\boldsymbol{v} = \hbar^{-1}\partial\varepsilon/\partial\boldsymbol{k}$, which is also determined by the crystal field. Therefore, the crystal structure of a material and, in particular, its symmetry is the second (after the atomic number Z) important factor determining the spin–orbit interaction and, thus, also spin–orbit scattering. The center of inversion of the crystal field is of key importance. If there is no center of inversion, eqn (2.22) determines the second contribution to the spin–orbit interaction, which, in a number of instances, may be the most important.

Most of the above classical examples of weak localization and antilocalization were observed for metal films. Today, studies of a two-dimensional electron gas are performed mainly on heterostructures in which electrons are localized in narrow quantum wells formed at the interfaces of two crystals. Usually, materials with noncentrosymmetric crystal structures are used. However, the spin–orbit interaction in such a two-dimensional gas also depends on the shape and asymmetry of the quantum well (the Rashba–Bychkov effect)

$$\nabla U = (\nabla U)_{\mathrm{cryst}} + (\nabla U)_{\mathrm{heter}}. \tag{2.27}$$

The well parameters are varied by applying a gate voltage V_{g} to the electrode on the top of the well. Thus, it is possible to pass from localization to antilocalization and back in a two-dimensional electron gas by varying the voltage V_{g} at the gate. Figures 2.13 and 2.14 illustrate the results of these experiments.
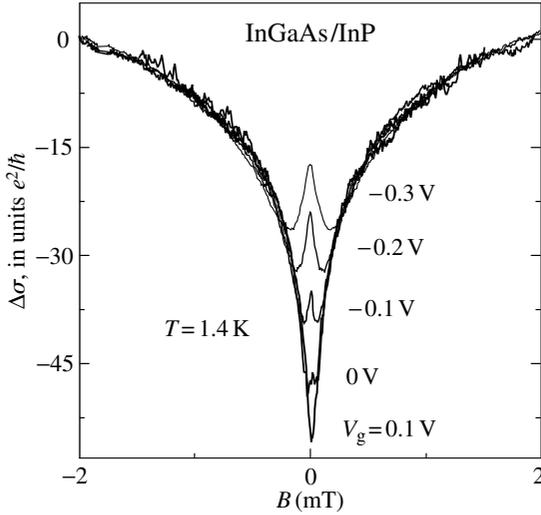
FIG. 2.13.   Antilocalization controlled by the gate voltage V_g in a InGaAs/InP quantum well with high electron mobility (Studenikin et al. 2003). The temperature is 1.4 K. The curves are shifted to superpose at the field 2 mT. It is assumed that this field completely destroys the interference contribution ($\Delta\sigma(B = 2\,\text{mT}) = 0$).
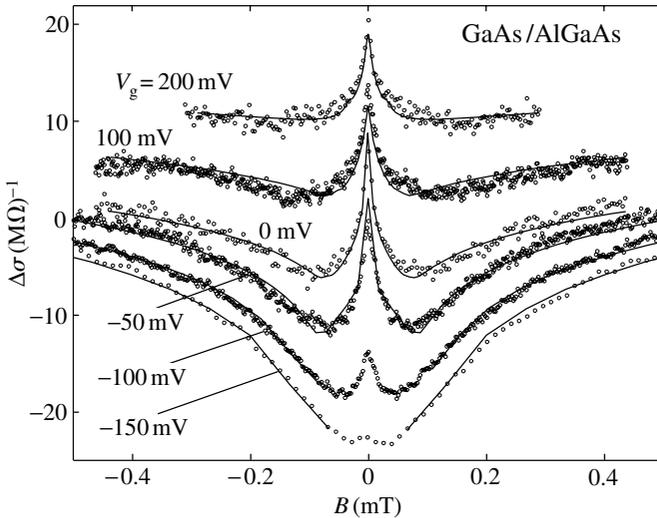
FIG. 2.14.   Antilocalization controlled by the gate voltage V_g in a GaAs/AlGaAs quantum well with high mobility (Miller et al. 2003). The temperature is 300 mK. Solid lines are the results of the theoretical fitting with the use of three parameters. For a more clear presentation, the curves are shifted along the vertical axis.

It is not accidental that we carefully follow eqns (2.11) and (2.25) by plotting conductivity and not resistivity in Figs 2.13 and 2.14. Otherwise, it would be difficult to compare the curves recorded at different gate voltages V_g because V_g considerably changes the carrier concentration and resistivity of the two-dimensional electron gas, and $\Delta R = \Delta\sigma/R^2$. In the experiments illustrated by Figs 2.9, 2.11 and 2.12, no pronounced difference in resistances was observed. Therefore, these figures show the resistance curves, which are measured in the experiments.

2.4. Interelectron interference (Aronov–Altshuler effect)

For a comprehensive treatment of the theory, see Altshuler and Aronov (1985).

On the terminology: The title of this section indicates that it is dedicated to the phenomena which, ultimately, follow from the wave properties of electrons. However, the same may also be said about weak localization. Therefore, the term electron interference is sometimes used instead of weak localization. In turn, the interference between waves from different electrons, we intend to discuss below, is often called the electron–electron interaction or the Aronov–Altshuler effect. However, following the optical terminology, one should classify weak localization as diffraction and the Aronov–Altshuler effect as interference.

Weak localization is the result of electron interference with itself provided by motion of the partial waves along closed trajectories. Interference between the wave functions of different electrons is also possible, e.g., if electrons meet twice. In the absence of scatterers, the electrons that happened to be close to one another at the moment $t = 0$ fly apart along ballistic trajectories with the Fermi velocity v_F. The distance r between these electrons increases linearly with time, $r \sim v_F t$. Under diffusion conditions, the mean distance between electrons increases considerably more slowly, $r \sim l(t/\tau)^{1/2} \sim v_F(t\tau)^{1/2}$. Diffusion keeps electrons close to one another and, thus, changes the conditions of their interaction.

In weak localization, the phase-breaking time τ_φ is determined by inelastic collisions. Under interference of two different electrons, it is not the inelastic collisions that are of primary importance, but the difference in initial energies ε_i the electrons may have. The change of phase of an electron with time is described by the relationship $\exp[i\varphi(t)] = \exp[i(\varepsilon_i/\hbar)t]$. Let two electrons with energies differing by $\Delta\varepsilon$ have the same phases at the moment $t = 0$. After time $\hbar/\Delta\varepsilon$, these phases would differ by a value of the order of unity.

Diffusion takes place only for electrons with energies ε ranging within

$$\varepsilon_F - T \lesssim \varepsilon \lesssim \varepsilon_F + T. \qquad (2.28)$$

For these electrons, the mean $\Delta\varepsilon \sim T$ value is proportional to T, and the characteristic dephasing time is

$$\tau_{ee} \simeq \hbar/T. \qquad (2.29)$$

The corresponding size of the interference region is

$$L_{\text{ee}} \simeq l \left(\frac{\tau_{\text{ee}}}{\tau} \right)^{1/2} \simeq v_{\text{F}} \left(\frac{\hbar\tau}{T} \right)^{1/2} \simeq \sqrt{\frac{\hbar D}{T}}. \qquad (2.30)$$

The τ_{ee} and L_{ee} quantities play the same role in interelectron interference as τ_φ and L_φ in weak localization, although the mechanisms of their influence on conductivity are quite different.

First of all, the diffusion character of motion changes the effective frequency of electron–electron collisions. In the ballistic regime, the energy and momentum of each of two charged particles after they fly apart differ from those before they close in. The event of their meeting is called a collision. The mean momentum transferred during the collision is of the order of k_{F}, the effective size of the interaction region is $1/k_{\text{F}}$, and the effective interaction time is $\hbar/\varepsilon_{\text{F}}$. Since each scattering event should satisfy the Pauli exclusion principle and the law of energy conservation, the initial and final states of the colliding electrons can lie only in the energy range given by eqn (2.28). Temperature limits the number of possible collisions – the number of electrons that may collide with the given electron is proportional to T and the number of finite states of this electron after each collision is also proportional to T. This results in the factor T^2 in the scattering probability $\hbar/\tau_{\text{e}}$, so that the collision frequency is described as

$$\frac{\hbar}{\tau_{\text{e}}} \sim \frac{T^2}{\varepsilon_{\text{F}}}. \qquad (2.31)$$

Above, we added the factor $1/\varepsilon_{\text{F}}$ to preserve the dimensions, because ε_{F} is the only independent parameter with the dimensions of energy.

In principle, eqn (2.31) for the collision frequency is changed if electrons do not move along ballistic trajectories, but diffuse in the vicinity of one another and are scattered by impurities. The effective size of the interaction region, $L_{\text{ee}} \gg 1/k_{\text{F}}$, becomes rather large, and the momentum transferred as a result of this interaction, $q \simeq 1/L_{\text{ee}}$, becomes small. It is this limitation on the value of the transferred momentum (and not the law of energy conservation) that determines the collision frequency τ_{e} in this case. Now, τ_{e} includes q^d and not T^2. In order to obtain the dimensions of energy for the variable $\hbar/\tau_{\text{e}}$, we have to multiply $q^d \propto 1/L_{\text{ee}}^d$ by the density of states at the Fermi level, also dependent on the space dimension d, $g_d \sim \varepsilon_{\text{F}}^{d/2-1} m^{1/2}$ (where $m = \hbar k_{\text{F}}/v_{\text{F}}$ is the electron mass). Thus, we obtain the following expression instead of eqn (2.31):

$$\frac{\hbar}{\tau_{\text{e}}} \sim \frac{1}{g_d L_{\text{ee}}^d}. \qquad (2.32)$$

Whence, we have

$$\hbar/\tau_{\text{e}} \sim T^{d/2} \varepsilon_F^{1-d} \tau^{-d/2} = \begin{cases} T^{1/2}\tau^{-1/2}, & d = 1, \\ T\varepsilon_F^{-1}\tau^{-1}, & d = 2, \ (T \ll 1/\tau). \\ T^{3/2}\varepsilon_F^{-2}\tau^{-3/2}, & d = 3, \end{cases} \qquad (2.33)$$

To distinguish between scattering in the ballistic and diffusion modes, we denote the collision frequency defined by (2.31) by $\hbar/\tau_{\mathrm{e}}^{\mathrm{ball}}$ and the frequency defined by (2.33), by $\hbar/\tau_{\mathrm{e}}^{\mathrm{diff}}$.

Note: The diffusion nature of electron motion within a time $\hbar/T$ does not exclude scattering with the transfer of a large momentum (2.31); it gives rise to additional scattering defined by eqn (2.33) in the diffusion channel. Scattering rates (2.31) and (2.33) should be summed.

The collisions described by eqn (2.33) introduce a direct contribution to the rate of energy relaxation and the formation of the phase-breaking time, τ_{φ}. However, the momentum exchange between electrons due to this process is unimportant and the collisions make no noticeable contribution to resistivity, because it was assumed at the very beginning that $1/\tau_{\mathrm{e}}^{\mathrm{diff}} \ll 1/\tau$. It is more important for electron transport that, in parallel with the switching-on of the diffusion channel of scattering, the electron interactions during diffusion also give rise to changes of the electron spectrum in the vicinity of the Fermi level.

Consider the interaction between two electrons with the energies measured from the Fermi level, ε and $-\varepsilon$. The Hamiltonian which describes their interaction contains the exchange term appearing because the electrons may be exchanged since they are equivalent. This term does not affect scattering, but introduces small additions to the electron energy and density of states. Usually, this is unimportant, because these additions are independent of energy. However, the correction to the exchange term that comes from the interactions during diffusion motion is noticeable against the total background, because the corresponding contribution to the density of states depends on energy and reaches the maximum value at the Fermi level. The effective time of interaction between the electrons with energies ε and $-\varepsilon$ is the more pronounced the lower the $|\varepsilon|$ value: $\tau_{\mathrm{ee}} = \hbar/\varepsilon$. At $T = 0$, the density of states in the vicinity of the Fermi level depends on energy ε as

$$\Delta g(T=0, \varepsilon) \sim (\hbar D)^{-d/2} \cdot \begin{cases} \sqrt{|\varepsilon|}, & d = 3, \\ \ln(\varepsilon \tau/\hbar), & d = 2, \\ 1/\sqrt{|\varepsilon|}, & d = 1. \end{cases} \qquad (2.34)$$

This is the most important result of the interelectron interaction in the diffusion channel. Diffusion that keeps the electrons close to one another in $\boldsymbol{r}$-space also increases the effective time of their interaction and, thus, pushes the levels away.

Generally speaking, the definition of the density of states also depends on interactions, because the energy distribution of the interacting electrons depends on the number of these electrons. Therefore, the density of states that enters, e.g., the formulas of heat capacity differs from the density of states determining the transport properties and the tunneling probability considered here. We return to this problem in Appendix B. In particular, it is indicated there that the Coulomb interaction decreases the effective tunneling density of states, because the space
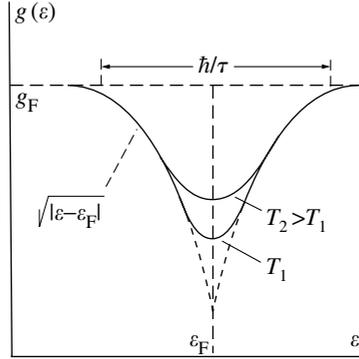
FIG. 2.15. The density-of-states minimum at the Fermi level formed due to interelectron interaction at different temperatures (three-dimensional case).

inhomogeneity of the charges formed by tunneling may be dissolved only within a certain finite time. This time is affected by the effectiveness of the diffusion.

The interaction results in the formation of a singularity in the $g(T = 0, \varepsilon)$ function at the Fermi level at $\varepsilon = \varepsilon_F$. The temperature determines the average difference between the energies of interacting electrons and removes the singularity in the function $g(0, \varepsilon)$ schematically depicted in Fig. 2.15 by the dashed line

$$g(T, \varepsilon) \simeq \begin{cases} g_F, & |\varepsilon_F - \varepsilon| > \hbar/\tau, \\ g(0, \varepsilon), & T < |\varepsilon_F - \varepsilon| < \hbar/\tau, \\ g(0, \varepsilon = T), & |\varepsilon_F - \varepsilon| < T. \end{cases} \qquad (2.35)$$

The dependence of the density of states on temperature in the direct vicinity of the Fermi level results in the temperature dependence of the corresponding correction to conductivity. This correction may be estimated in the same way as in the theory of weak localization.

The interelectron interaction takes place because the electrons that were close to one another at the moment $t = 0$ meet again within the time τ_{ee}. The probability η of this meeting is described by the same integrals as in the theory of weak localization

$$\eta(\Delta\varepsilon) \simeq \int_{\tau}^{\hbar/\Delta\varepsilon} \frac{v\lambda^2 \, dt}{b^{3-d}(Dt)^{d/2}}, \quad d = 1, 2, 3, \qquad (2.36)$$

where $\Delta\varepsilon$ is the difference in the energies of two interfering electrons. This predetermines the functional similarity of the expressions for the quantum corrections to conductivity in the cases of weak localization and interelectron

interference:

$$d = 3: \quad \Delta_{ee}\sigma_3 \approx -\text{const} + \left(\frac{e^2}{\hbar}\right) L_{ee}^{-1},$$

$$d = 2: \quad \Delta_{ee}\sigma_2 \approx -\left(\frac{e^2}{\hbar}\right) \ln\left(\frac{\tau_{ee}}{\tau}\right) \simeq -2\left(\frac{e^2}{\hbar}\right) \ln\left(\frac{L_{ee}}{l}\right), \quad\quad (2.37)$$

$$d = 1: \quad \Delta_{ee}\sigma_1 \approx -\left(\frac{e^2}{\hbar}\right) L_{ee}.$$

These corrections and the corrections described by (2.11) for weak localization depend on the dimensionality d in similar ways, have the same dimensionality criteria (2.8), etc. However, eqn (2.37) contains not the length L_φ determined by eqn (2.7), but the dephasing length determined by eqn (2.30).

It is natural to observe the Aronov–Altshuler effect experimentally by studying the transport properties of dirty metals. However, it is impossible to distinguish between interelectron interference and weak localization. We may use here as a hook the indifference of interelectron interference toward a magnetic field. However, it is far from being easy to use this hook. Therefore, the direct manifestations of this effect were observed in studies of the densities of states. The density-of-states minimum can be seen directly in special tunneling experiments described in Appendix B.

Thus, the interelectron interference is described with the use of three characteristic times – the dephasing time (2.29) of a diffusing electron and the relaxation times (2.31) and (2.33). To reveal the relations between these times, we constructed the "temperature–collision frequency" plane and used the Fermi energy ε_F as a scale along both axes. The squared unit area in this plane includes all the possible temperatures and collision frequencies (Fig. 2.16).

Note: the same square has been used on another occasion in Chapter 1 – see there Fig. 1.8.

In the effect of weak localization, collisions with static defects and collisions with phonons play different roles. The former provide diffusion, whereas the latter result in phase breaking. In interelectron interference, the two electrons are characterized from the very beginning by different rates of phase variation, which are proportional to the electron energies. Collisions of these electrons with phonons do not take the electrons from the thermal range described by eqn (2.28) and, therefore, do not increase the average difference in electron energies. Thus, the effective dephasing time τ_{ee} does not decrease. On the other hand, thermal displacements of atoms give rise to additional inhomogeneity and the scattering effects from phonons and static defects are summed up. The efficiency of the respective contributions of these two sources of scattering to diffusion and resistivity depends, via the factor $(1 - \cos\theta)$ in the collision integral, on the scattering angle θ. The frequencies of the electron–phonon, $1/\tau_{ph}$, and electron–impurity, $1/\tau_0$, collisions calculated with due regard for this efficiency yield the
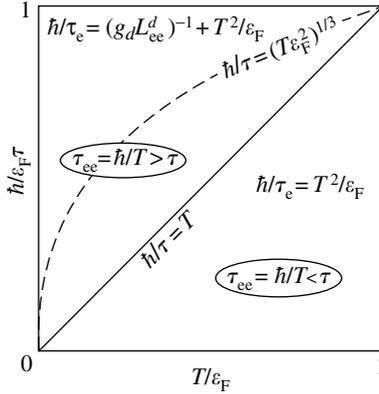
FIG. 2.16. Electron–electron interaction at relatively high (upper left side triangle) and low (lower right side triangle) elastic-scattering frequencies. In the upper triangle, no dephasing is observed for the time between two elastic collisions. In the lower triangle, the time between two collisions is sufficient for dephasing. The dashed line is drawn for the three-dimensional case. In three-dimensional systems, inequality (2.42) is valid only above this line (see the text).

so-called transport scattering frequency

$$\frac{1}{\tau_{tr}} = \frac{1}{\tau_0} + \frac{1}{\tau_{ph}}. \tag{2.38}$$

It is this frequency that is plotted along the vertical axis in Fig. 2.16. In metals, the transport time differs from the total time only in the case of phonon scattering at low temperatures. Therefore, the corresponding subscript of τ on the ordinate axis in Fig. 2.16 is omitted.

Now, draw the diagonal of the square in Fig. 2.16. The equation of the diagonal,

$$\tau = \hbar/T \tag{2.39}$$

is similar to eqn (2.29) defining the dephasing time τ_{ee}. However, it follows from the drawing that this equation includes not τ_{ee} but the transport scattering time τ. The diagonal divides the square into two triangles. The states in the lower triangle are the states of a metal in the pure limit, because the interactions between electrons take place during their motion along the ballistic trajectories and are described by the conventional scattering time, eqn (2.31). However, it is masked by phonon scattering and may be revealed only at low temperatures, i.e., where $\rho_{ph} \propto T^5$, and $\rho_e \propto T^2$.

Everywhere in the upper triangle, we have

$$\tau_{ee} > \tau, \tag{2.40}$$

in other words, the dephasing time is much longer than the time between the elastic (and quasielastic) collisions. Therefore, interelectron interference takes place under diffusion conditions. This triangle is the region of diffusion interaction, whereas the states in this triangle are the states of a metal in the dirty limit. In this region, there are two channels of interelectron scattering – the ballistic and diffusion channels, and

$$\frac{1}{\tau_{\mathrm{e}}} = \frac{1}{\tau_{\mathrm{e}}^{\mathrm{ball}}} + \frac{1}{\tau_{\mathrm{e}}^{\mathrm{diff}}}. \tag{2.41}$$

In systems with reduced dimension, $d = 1$ or $d = 2$, scattering in the diffusion channel in the upper triangle is more pronounced;

$$\frac{1}{\tau_{\mathrm{e}}^{\mathrm{diff}}} \gg \frac{1}{\tau_{\mathrm{e}}^{\mathrm{ball}}}. \tag{2.42}$$

It is this scattering that gives the main contribution to the formation of the phase-breaking rate, $1/\tau_{\varphi}$, that controls the weak-localization processes. For $d = 3$, inequality (2.42) takes place only above the curve

$$\hbar/\tau = (T\varepsilon_{\mathrm{F}}^2)^{1/3}, \tag{2.43}$$

shown by the dashed line in the upper triangle. Anyhow, irrespective of dimensionality, the interelectron interaction at $T > \hbar/\tau$ gives rise to the corrections to the density of states at the Fermi level described by eqn (2.34) and to the quantum corrections to conductivity described by eqn (2.37).

Note: No matter how low the residual resistivity $\rho_0 \propto 1/\tau_0$, i.e., no matter how pure the metal, at a low enough temperature, $T < \hbar/\tau_0$, this metal turns out to be formally in the dirty limit, so that its conductivity should necessarily contain the quantum correction. However, this correction can be formed only if the sample size exceeds L_{ee}. In any case, this correction is relatively small.

References

Abrikosov, A.A. (1988). Fundamentals of the Theory of Metals. North-Holland.

Altshuler, B.I. and Aronov, A.G. (1985). Electron–electron interaction in disordered conductors, in *Electron–electron Interactions in Disordered Systems*. (eds. A.L. Efros and M. Pollak). North-Holland.

Altshuler, B.I., Aronov, A.G., Spivak, B.Z., Sharvin, D.Yu., and Sharvin, Yu.V. (1982). JETP Letters **35**, 588.

Bergmann, G. (1984). Phys. Rep. **107**, 1.

Bishop, D.J., Tsui, D.C., and Dynes, R.C. (1980). Phys. Rev. Lett. **44**, 1153.

Dorozhkin, S.I. and Dolgopolov, V.T. (1982). JETP Letters **36**, 18.

Gijs, M., Van Haesendonck, C., and Bruynseraede, Y. (1984). Phys. Rev. B **30**, 2964.

Miller, J.B., Zumbuhl, D.M., Marcus, C.M., Lyanda-Geller, Y.B., Goldhaber-Gordon, D., Campman,K., and Gossard,A.C. (2003). Phys. Rev. Lett. **90**, 076807.

Ovadyahu, Z. and Imry, Y. (1981). Phys. Rev. B **24**, 7440.

Sharvin, D.Yu. and Sharvin, Yu.V. (1981). JETP Letters **34**, 272.

Studenikin, S.A., Coleridge, P.T., Ahmed, N., Poole, P., and Sachrajda, A. (2003). JETP Letters **77**, 311.

Van der Dries, L., Van Haesendonck, C., Bruynseraede, Y., and Deutscher, G. (1981). Phys. Rev. Lett. **46**, 565.

Wolf, P. and Maret,G. (1985). Phys. Rev. Lett. **55**, 2696.

EFFECT OF THE COULOMB INTERACTION ON THE ELECTRON ENERGY SPECTRUM

Most of the problems considered in this chapter are also considered in the book by Shklovskii and Efros (1984).

3.1. The Peierls transition

The details of the problems considered in this paragraph can be found in the books by Peierls (1955) and Grüner (2000).

All the electrons of a solid exist in the initial potential created by the ions. However, an inhomogeneous distribution of the electrons in space gives rise to the formation of additional electric fields and renormalization of the bare potential. Therefore, each electron "feels" not only the potential created by ions but also the potential created by other electrons. In particular, the screening is a kind of such renormalization. In metals, the concentration of renormalized electrons is very high and screening is pronounced, so the initial potential retains only its small part, the so-called pseudopotential. The pseudopotential affects the single-electron spectrum mainly due to the preserved translation symmetry of the initial potential.

However, in many instances the effect of electrons on their own energy spectrum is not reduced only to screening. Electrons may considerably change the bare potential and lower its symmetry. A striking example here is the Peierls instability. Consider a regular one-dimensional atomic chain with period a. The corresponding Brillouin zone for this chain is the segment between the $-\pi/a$ and π/a values along the k-axis (Fig. 3.1). This zone has $1/a$ states for electrons per a unit length of the chain. Each state may contain two electrons – altogether $2/a$ electron sites. Let one electron of each atom be collectivized so that the concentration of delocalized electrons is $n = 1/a$. Let the temperature be very low, for simplicity, $T = 0$. Then n electrons would occupy $n/2$ states in the central part of the segment, from $-\pi/2a$ to $\pi/2a$, in which the energy is minimal. Now, imagine that the period of the atomic chain is doubled by a small displacement of each second atom. Then the period in the reciprocal space would be twice reduced, so that an additional boundary of the Brillouin zone would be formed exactly between the occupied and unoccupied states. At the Brillouin-zone boundary, the $\varepsilon(k)$ spectrum has a gap, with the derivative $d\varepsilon/dk$ having zero values on both sides. As is seen from Fig. 3.1, the formation of the additional boundary increases the energy of the unoccupied states $|k| > \pi/2a$
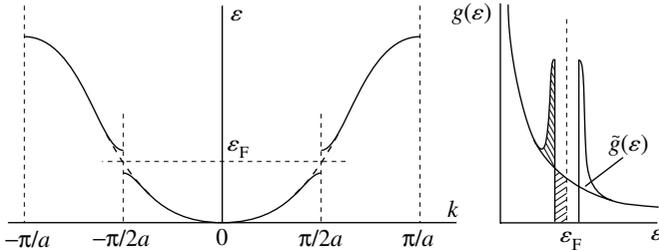
FIG. 3.1. Appearance of a gap around the Fermi level in a one-dimensional
  Brillouin zone for a chain of metal atoms with one valence electron per
  atom. Right: the density of states $g(\varepsilon)$; the hatched areas are equal. The
  dashed line shows the density of states $\tilde{g}(\varepsilon)$ in an unperturbed one-dimensional
  electron gas.

but decreases the energy of the occupied states $|k| < \pi/2a$. This is accompanied
by a decrease of the energy of the electron system as a whole. The Fermi level ε_F
with respect to the band bottom at the point $k = 0$ remains unchanged.

Thus, doubling of the chain period is favorable for an electron system with
one electron per atom because it decreases its energy. The problem is how to
attain period doubling. The simplest way is to provide a small displacement
$(\Delta a)_2$ of each second atom, although this would result in a certain loss of the
elastic energy that determined the period a. Nevertheless, a small displacement
$(\Delta a)_2$ is still somewhat advantageous in most cases, and its value is determined
by the minimum of the total energy.

The assumption that the chain has one electron per atom is not very import-
ant. The appropriate period can be selected at any carrier concentration. Let
the electron concentration n be a fraction α of the atomic concentration $1/a$,
so that $n = \alpha/a$. If a new period in the chain is formed, $\tilde{a} = 2a/\alpha$, then the
Brillouin zone acquires an additional boundary between the free and occupied
states. Such a period may appear even without displacements of the chain atoms
because of a modulation due to a charge-density wave in the electron system.
Since the formation of the new period gives rise to the formation of a gap between
the occupied and the unoccupied states in the spectrum and the transformation
of a metal into an insulator, we may state that no one-dimensional metal can
exist at the absolute zero of temperature. With an increase in the temperature,
the gain in the electron energy gradually diminishes because of the appearance
of empty states under the gap (where the energy of states is lower than in the
unperturbed spectrum) and occupied states above the gap (where the energy
of states is higher than in the unperturbed spectrum). Therefore, at a certain
critical temperature, the material undergoes a transition to the metal state.

In the Peierls consideration, the assumption that the system is one-
dimensional is essential because the Fermi surface of systems with higher
dimension always has a certain curvature, whereas the Brillouin zone bound-
aries, eqn (1.1), where the energy is discontinuous are always flat. Therefore, it

is impossible to transform a three-dimensional metal into an insulator by reducing its symmetry. However, a decrease in the total electron energy with the appearance of new discontinuity planes of the $\varepsilon(k)$ function in k-space may often be important. The very existence of Bi, Sb, and As semimetals and metal alloys with complicated structures containing a large number of atoms in their unit cells is due to the tendency to diminish the electron energy by means of new Brillouin planes. Moreover, it was revealed that anisotropic media may be unstable against the charge- and spin-density waves whose appearance is accompanied by phase transitions. All these phenomena gave rise to a new branch of solid state physics not considered in the present book. Therefore, we only mention here the Peierls transition but leave aside such problems as the scale of atomic displacements at $T = 0$, the critical temperature of the transition where the gap becomes zero, etc. (see, e.g., Peierls 1955 or Grüner 2000).

Nevertheless, we mention here the Peierls transition by two reasons.

First, the Peierls transition is a very convenient example for the demonstration of principle of self-action of an electron gas onto its own energy spectrum. Below, we consider the manifestation of this principle in a disordered medium. The analogy becomes obvious if one compares Fig. 3.1 with Fig. 3.7 presented below.

Second, in the discussion at the beginning of Chapter 5 on the classification of electron phase transitions, the Peierls transition is considered as a clear example: despite the fact that a decrease in the electron energy is the key factor, the transition point is determined not by the electron energy only, but by the energy balance that includes other types of energies, e.g., the elastic energy.

3.2. Structure of the impurity band in the case of low doping

The band theory successfully explains the division of solids with ideal crystal lattices into metals and insulators. The Fermi level in metals lies inside one band or several overlapping bands, i.e., in the range of allowed energies. In an insulator, the Fermi level lies in the range of forbidden energies, with the number of free carriers at the temperature T being determined by the exponent, where the distance $|E_b - \mu|$ from the Fermi level μ to the nearest edge of the allowed energies, E_b, is compared with T. If $|E_b - \mu|$ is of the order of the atomic energy ($1\,\mathrm{eV}$, i.e., $10\,000\,\mathrm{K}$), so that at room temperatures $T_r \approx 300\,\mathrm{K}$

$$|E_b - \mu| \gg T_r, \qquad (3.1)$$

then the conductivity at room temperature is negligible, and the material is a true insulator. If inequality (3.1) is not too strong, the material possesses a finite intrinsic conductivity, which is frozen out with a decrease in temperature.

On terminology: In Chapters 3 and 4, we denote the Fermi level by μ in accordance with the tradition of physics of semiconductors, whereas in all the other chapters of the book we denote the Fermi energy by ε_F. Both μ and ε_F signify the energy level which, at the absolute zero of temperature, $T = 0$, separates the occupied electron

states from unoccupied states. In metals, the states in the vicinity of the Fermi level are delocalized. In insulators, these levels are localized. They may even be absent if the Fermi level is located in the energy gap.

Impurities may create electron levels E_0 in the forbidden band. We are interested in those energy levels E_0 which are located in the vicinity of the band edge E_b. These impurities are called donors and acceptors. The presence of donors or acceptors displaces the Fermi level close to E_0, whereas the number of free carriers and the conductivity are controlled by the impurity concentration N and the ionization energy of the impurities, $|E_b - E_0|$, compared with T. The corresponding conductivity is called the impurity conductivity.

The materials possessing intrinsic or impurity conductivity are called semiconductors. The above consideration makes it clear that placing a particular materials among semiconductors is rather conditional and, in particular, depends on temperature. At low temperatures, semiconductors stop existing as a class of materials because, with the freezing out of conductivity, all the materials become insulators. However, a number of materials, such as Ge, Si, GaAs, InSb, InP, etc., are traditionally called semiconductors.

For definiteness, consider an n-type semiconductor with band gap Δ. Let the main impurities in this semiconductor be shallow hydrogen-like donors, i.e., donors with ionization energy $E_0 \ll \Delta$ and the electron wave function vanishing at large distances as $\exp(-r/a_B)$ (Bohr radius $a_B = \hbar(2m^* E_0)^{-1/2} = \hbar^2 \kappa/m^* e^2$, where κ is the permittivity and m^* is the effective mass). It is impossible to purify and dope a semiconductor to such a degree that it would contain solely donors and no acceptors. The presence of a number of acceptors in an n-type semiconductor becomes crucial. Since the electron energy at the acceptor is considerably lower than its energy at a donor, all the N_A acceptors would capture electrons from the donors and acquire a negative charge, whereas a certain number N_A from the total number of N_D donors, $N_D > N_A$, would loose their electrons and acquire a positive charge. This results in the formation of a random electric field created by $2N_A$ randomly distributed charges. The energies of all the donors are modified by this random field:

$$\varepsilon_j = E_0 + \Delta\varepsilon_j, \quad \Delta\varepsilon_j = \frac{e^2}{\kappa} \left[\sum_a \frac{1}{|r_j - r_a|} - \sum_{id}' \frac{1}{|r_j - r_{id}|} \right]. \qquad (3.2)$$

The first sum is taken over all the acceptors (variable subscript "a") and the second one, over all the ionized donors (subscript "id"). The subscript j indicates an arbitrary donor, the prime on the second sum indicates that if the subscript j belongs to an ionized donor, the term $j = $ id with zero in the denominator should be excluded from the summation.

The energy shifts $\Delta\varepsilon_i$ indicate that the existence of a random electric field even in a low-doped semiconductor ($Na_B^3 \ll 1$) gives rise to spreading of an N_D-tuply degenerate donor level E_0 into the so-called impurity band. As we are interested in the energy shifts, we assume below in this chapter that $E_0 = 0$. The
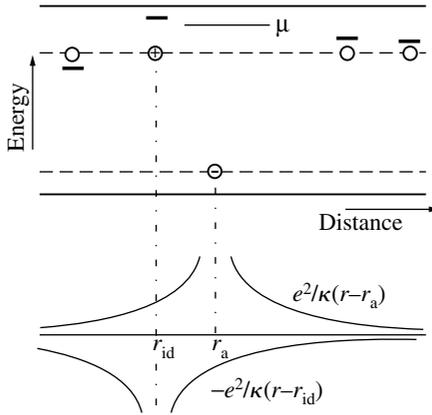
FIG. 3.2. Electric field of charged centers. The circles on the energy sketch (above) show the initial values of the donor and acceptor energies. The horizontal bars show the values of the donor energies corrected for the Coulomb potential of charged impurities. The contributions to the potential that come from individual charged impurities are shown below.

Fermi level μ separates the occupied levels of neutral donors and the unoccupied levels of ionized donors, i.e., lies inside this impurity band (Fig. 3.2).

Note: This transformation of a level into an impurity band at $N a_{\mathrm{B}}^3 \ll 1$ is not associated with the overlap of the wave functions and is of pure classical origin. The overlap and delocalization (at a certain degree of overlap) arise under higher doping levels. Until this moment, all the states in this impurity band are localized and the term "band" signifies the range of energies with nonzero density of states.

3.2.1. *Weak compensation*

Consider the case of a relatively small number of acceptors, so that the compensation coefficient is small, $K = N_{\mathrm{A}}/N_{\mathrm{D}} \ll 1$. Since the opposite charges are attracted, then in equilibrium, only the donors located close to acceptors are ionized. The inequality $N_{\mathrm{A}} \ll N_{\mathrm{D}}$ signifies that the average interacceptor distance is much larger than the average distance between donors. Therefore, each ionized donor can be attributed to a certain acceptor and, thus, the problem is reduced to the consideration of possible *local configurations* of charged donors around a charged acceptor.

If two donors are in the vicinity of an acceptor, they both may be ionized. Let the negatively charged acceptor be located in the middle between two donors at distances r from these donors (Fig. 3.3a). Now, let one of the donors be ionized, i.e., be positively charged. The arrival of a positive charge from infinity to the second donor would result in an energy gain $e^2/\kappa r$ because of the donor attraction to the acceptor, whereas the loss in energy caused by the repulsion
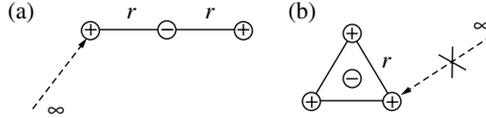
FIG. 3.3.   (a) Possible and (b) impossible configurations of charged centers.

from the first donor, $e^2/2\kappa r$, is less by half. Therefore, such a configuration is possible and one acceptor may keep in its possession two ionized donors.

However, an acceptor cannot keep in its possession three ionized donors. In the most favorable configuration, donors are located at the apices of the equilateral triangle with sides r, with the acceptor being located in the triangle center (Fig. 3.3b). If two donors have already been ionized, then the ionization of the third donor (described as the arrival of a positive charge from infinity to this donor) would result in energy loss, $(e^2/\kappa r)(\sqrt{3} - 2) < 0$.

All this allows one to divide all the acceptors into three groups: N_0 acceptors having no ionized donors in their vicinities; N_1 acceptors having one donor in their vicinities; and N_2 acceptors with two ionized donors in their vicinities. The condition of electrical neutrality $N_0 + N_1 + N_2 = N_A$ yields for the Fermi level

$$N_0(\mu) = N_2(\mu). \tag{3.3}$$

Now, write the expression for $N_0(\mu)$. Let the origin of the energy scale correspond to the energy of an isolated donor. As is seen from Fig. 3.2, an acceptor belongs to the group N_0 if it has no donors within a radius $r_\mu = e^2/\kappa\mu$. Therefore, N_0 is equal to the acceptor concentration N_A multiplied by the probability that the volume $(4\pi/3)r_\mu^3$ does not contain a donor:

$$N_0(\mu) = N_A \exp\left(-\frac{4\pi}{3}r_\mu^3 N_D\right) = N_A \exp\left[-\frac{4\pi}{3}\left(\frac{\varepsilon_D}{\mu}\right)^3\right], \tag{3.4}$$

where $\varepsilon_D = (e^2/\kappa)N_D^{1/3}$ is the energy of the Coulomb interaction at the average interdonor distance $N_D^{-1/3}$.

The calculation of N_2 is somewhat more complicated. The probability that two donors closest to the acceptor are located in the volumes elements $d\boldsymbol{r}_1$ and $d\boldsymbol{r}_2$ at distances $r_1 = |\boldsymbol{r}_1|$ and $r_2 = |\boldsymbol{r}_2| \geqslant r_1$ from the acceptor is

$$N_D^2 \exp\left(-\frac{4\pi}{3}r_1^3 N_D\right) \exp\left(-\frac{4\pi}{3}(r_2^3 - r_1^3)N_D\right) d\boldsymbol{r}_1\, d\boldsymbol{r}_2$$

$$= N_D^2 \exp\left(-\frac{4\pi}{3}r_2^3 N_D\right) d\boldsymbol{r}_1\, d\boldsymbol{r}_2. \tag{3.5}$$

Both donors will be ionized if their energy levels exceed μ,

$$\varepsilon_1 = \frac{e^2}{\kappa|\boldsymbol{r}_1|} - \frac{e^2}{\kappa|\boldsymbol{r}_1 - \boldsymbol{r}_2|} > \mu, \quad \varepsilon_2 = \frac{e^2}{\kappa|\boldsymbol{r}_2|} - \frac{e^2}{\kappa|\boldsymbol{r}_1 - \boldsymbol{r}_2|} > \mu. \tag{3.6}$$

Then

$$N_2(\mu) = N_A N_D^2 \int d\boldsymbol{r}_1 \int_{r_2 > r_1} d\boldsymbol{r}_2 \exp\left(-\frac{4\pi}{3} r_2^3 N_D\right) \Theta(\varepsilon_1 - \mu)\Theta(\varepsilon_2 - \mu),$$

$$\Theta(x) = \begin{cases} 1 & \text{at} \quad x > 0, \\ 0 & \text{at} \quad x < 0. \end{cases} \tag{3.7}$$

The $N_2(\mu)$ value calculated from eqn (3.7) is somewhat underestimated. It may happen that two donors closest to the acceptor cannot be ionized simultaneously, because they are too close to one another. Such an acceptor is not taken into account in eqn (3.7). However, one more donor may be located on the other side of this acceptor at a distance of $r_3 > r_2$ from it and may form an ionized pair with one of the closest donors. However, the respective correction is negligible – less than one percent. Substituting the data calculated by eqn (3.7) with the necessary correction and eqn (3.4) into eqn (3.3), we obtain $N_0 = N_2$ and N_1, and also the position of the Fermi level

$$N_0 = N_2 = 0.013 N_A, \quad N_1 = 0.974 N_A, \tag{3.8}$$

$$\mu \approx 0.99 \varepsilon_D. \tag{3.9}$$

The plot of the density of states is sketched in the left-hand side of Fig. 3.4.

Note: The area under the $g(\varepsilon)$ curve equals N_D and the effective width of the distribution is of the order of ε_D. The sense of the coefficient 0.99 in eqn (3.9) is symbolic. Despite the fact that it is almost unity, we write this coefficient here to emphasize that it was calculated. At the same time, this calculation gives no information on the symmetry of the $g(\varepsilon)$ curve with respect to E_0.

The dependence of $\mu - E_0$ on N_D can be verified experimentally because this difference is, in fact, the activation energy of the nearest-neighbor hopping
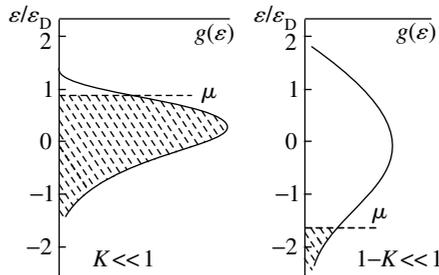
FIG. 3.4.   Sketches of the density of states in the impurity band at weak ($K \ll 1$) and strong ($1 - K \ll 1$) compensation. The origin of the energy scale $\varepsilon/\varepsilon_D = 0$ corresponds to the level E_0. The occupied states are hatched.
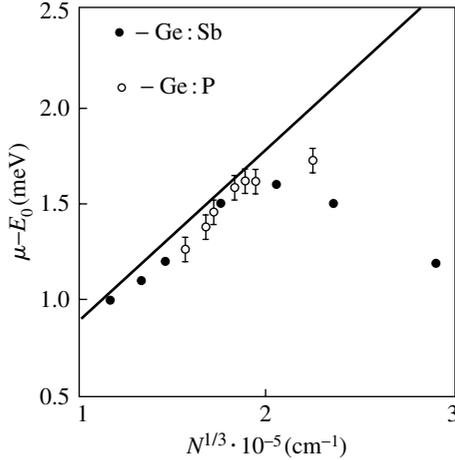
FIG. 3.5. Activation energy of the nearest-neighbor hopping conductivity as a
function of the impurity concentration in Ge : P (from Shklovskii and Shlimak
1972) and Ge : Sb samples (from Fritzsche 1958). The straight line shows the
theoretical dependence described by (3.9).

conductivity (see Chapter 4). Figure 3.5 shows the data of two experiments on
n-type germanium. It is seen that, at least at low N_D, the relation $\mu - E_0 \propto N^{1/3}$
is confirmed, in accordance with eqn (3.9). At high donor densities N_D, the
overlap of the electron wave functions localized at the impurity centers (not taken
into account by the above classical calculation) seems to become important.

Note: Both the density of states and the position of the Fermi level were formed in
the potential relief induced by electrons that, in principle, could be distributed over
donors in $C_{N_D}^{N_A}$ ways. The solution in the limit $K \ll 1$ obtained from eqn (3.3) describes
the function $g(\varepsilon)$ on the scale of ε_D. With an increase in K, the number of ways $C_{N_D}^{N_A}$
of possible arrangement of electrons at the donors also increases, which allows one to
expect that the $g(\varepsilon)$ function would become more complicated. Indeed, as will be shown
in the last section of this chapter, the $g(\varepsilon)$ function forms an additional structure on a
lower scale in the vicinity of μ.

3.2.2. *Strong compensation*

Consider the other limiting case

$$1 - K \ll 1, \qquad n = N_D - N_A \ll N_D. \qquad (3.10)$$

On the terminology: The letter n denotes the number of carriers and, if the electrons are
localized, the number of "possible carriers". Therefore, here $n = N_D - N_A$ determines
the number of electrons remaining in the donor-impurity band. On the other hand, N_A
electrons departed to the acceptors have obviously left the game (both in weak and

strong compensation). Here, we call the game (or a future game) the participation of electrons in the hopping conductivity at a low but not zero temperature, the metal–insulator transition with an increase in n, etc.

Under the conditions defined by eqn (3.10), the larger part of the donors are ionized. Only a donor too close to another ionized donor may remain neutral; in pairs of closely located donors, one donor remains neutral because its energy is reduced by $\varepsilon = e^2/\kappa r$ in the field of a positively charged neighbor (here r is the distance between the donors in the pair).

The probability of having a nearest neighbor at a distance from r to $r + dr$ is

$$\nu_r \, dr = N_{\mathrm{D}} \cdot 4\pi r^2 \, dr, \tag{3.11}$$

and the probability of decreasing the energy by $\varepsilon = e^2/\kappa r$ at the expense of the neighbor is

$$\nu_\varepsilon \, d\varepsilon = N_{\mathrm{D}} \cdot 4\pi r^2 \frac{dr}{d\varepsilon} \, d\varepsilon = -N_{\mathrm{D}} \cdot 4\pi \frac{(e^2/\kappa)^3}{\varepsilon^4} \, d\varepsilon. \tag{3.12}$$

Multiplying both sides of the above equation by N_{D} and dividing by 2 (in order not to take into account a donor twice), we arrive at the density of states for pairs

$$g_2(\varepsilon) = 2\pi N_{\mathrm{D}} \frac{\varepsilon_{\mathrm{D}}^3}{\varepsilon^4}, \qquad \varepsilon_{\mathrm{D}} = \frac{e^2 N_{\mathrm{D}}^{1/3}}{\kappa}. \tag{3.13}$$

The Fermi level is determined from the condition

$$\int_{-\infty}^{\mu} g_2(\varepsilon) \, d\varepsilon = n, \tag{3.14}$$

whence

$$\mu = -\left(\frac{2\pi}{3}\right)^{1/3} \varepsilon_{\mathrm{D}} \left(\frac{N_{\mathrm{D}}}{n}\right)^{1/3} = -\left(\frac{2\pi}{3}\right)^{1/3} \frac{\varepsilon_{\mathrm{D}}}{(1-K)^{1/3}} \tag{3.15}$$

(see the sketch on the right-hand side of Fig. 3.4).

In addition to N_{D}, eqn (3.15) contains one more variable parameter, $(1 - K)$. Experimentally, the effect of this parameter on conductivity is confirmed by measuring the activation energy of the hopping conductivity over the nearest neighbors (Fig. 3.6).
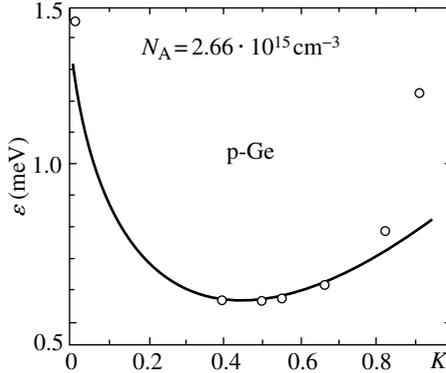
FIG. 3.6. Activation energy of the conductivity by nearest-neighbor hopping as a function of the compensation coefficient (Mott and Twose 1963). On the right, for $K > 0.5$, the theoretical curve attains the dependence described by eqn (3.15).

3.3. The Coulomb gap

Within both weak, $K \ll 1$, and strong, $1 - K \ll 1$, compensations, we obtained a bell-shaped density of states $g(\varepsilon)$ with the maximum close to E_0 and the Fermi level at one of the wings of the distribution (Fig. 3.4). It could be expected that with a gradual change in K from one limit to another, the Fermi level would smoothly shift from one edge of a weakly varying function $g(\varepsilon)$ to the other, being located near the maximum at $K \approx 0.5$ (when the number of electrons is half as great as the number of centers in the impurity band).

However, the results of a computer simulation at $K \approx 0.5$ (Fig. 3.7) show that this is not so. The density of states in Fig. 3.7 is plotted in units of g_0 and the energy in units of ε_D:

$$g_0 = \frac{N_D}{\varepsilon_D}, \qquad \varepsilon_D = \left(\frac{e^2}{\kappa} \right) N_D^{1/3}, \tag{3.16}$$

so that the area under the $g(\varepsilon)$ curve equals unity. As was to be expected, the Fermi level μ is really close to $E_0 = 0$, but at $\varepsilon = \mu$, the function $g(\varepsilon)$ is *minimal* and not maximal, $g(\mu) = 0$.

This result should be considered and discussed in more detail.

The *prerequisite* for obtaining such a result is a large number $C_{N_D}^{N_A}$ of possible electron distributions over the centers. This signifies the possible existence of a large number of different density-of-states functions from which the optimum function with a Coulomb gap should be selected.

One can readily explain *why* the selected distribution is preferable. Compare this distribution with the "expected" one $\tilde{g}(\varepsilon)$ (dashed line in Fig. 3.7) having

FIG. 3.7. Density of states in the impurity band for the compensation coefficient $K = 0.5$ according to the data of a computer simulation (Efros et al. 1979). The hatched areas are equal (compare with Fig. 3.1).

the same wings and area

$$\int_{-\infty}^{\infty} \tilde{g}(\varepsilon) \, d\varepsilon = \int_{-\infty}^{\infty} g(\varepsilon) \, d\varepsilon = N_{\mathrm{D}}, \tag{3.17}$$

but without a minimum. One can readily see that the average energy $\bar{\varepsilon}_-$ of states below the Fermi level is lower

$$\bar{\varepsilon}_- = \frac{1}{N_{\mathrm{D}} - N_{\mathrm{A}}} \int_{-\infty}^{\mu} \varepsilon g(\varepsilon) \, d\varepsilon < \frac{1}{N_{\mathrm{D}} - N_{\mathrm{A}}} \int_{-\infty}^{\mu} \varepsilon \tilde{g}(\varepsilon) \, d\varepsilon, \tag{3.18}$$

because the center area of this part of the distribution is shifted to lower energies. At low temperatures, the occupied states are located mainly in this part of the spectrum. Therefore, the total energy of the electron gas,

$$\widehat{E} = \int_{-\infty}^{\infty} \varepsilon g(\varepsilon) f(\varepsilon/T) \, d\varepsilon, \qquad f(x) = (\exp x + 1)^{-1} \tag{3.19}$$

decreases despite an increase in the average energy $\bar{\varepsilon}_+$ of the states in the spectrum range $\varepsilon > \mu$. These speculations are similar to those explaining the Peierls transition in a one-dimensional system. Comparing the changes $\tilde{g}(\varepsilon) \to g(\varepsilon)$ in the density of states because of the Peierls instability (Fig. 3.1) and due to the formation of the Coulomb gap (Fig. 3.7), we see that the equations of type eqns (3.17)–(3.19) are applicable to both cases. However, it follows from this analogue that the *temperature may change the density-of-states function $g(\varepsilon)$ and affect the Coulomb gap.*

To show *how* the $g(\varepsilon)$ function changes in the vicinity of μ, consider the following situation. Let two states from the energy range $(\mu - \varepsilon/2, \mu + \varepsilon/2)$ be

in equilibrium at $T = 0$. One of these states with energy $\varepsilon_i < \mu$ is occupied, whereas the other with energy $\varepsilon_j > \mu$ is empty. Let these states be spaced by the distance r_{ij}. Now, take the electron from the center i away to infinity. Then, the energy of the level j decreases down to $\varepsilon_j - e^2/\kappa r_{ij}$. Nevertheless, this energy should exceed ε_i, otherwise the alternative configuration in which the center j is occupied and the center i is empty would be preferable. The inequalities

$$\varepsilon_j - \varepsilon_i - e^2/\kappa r_{ij} > 0, \qquad \varepsilon_j - \varepsilon_i < \varepsilon \tag{3.20}$$

yield the ε-dependent constraint for r_{ij} from below and constraint for N from above:

$$\varepsilon > \frac{e^2}{\kappa r_{ij}}, \qquad N(\varepsilon) = r_{ij}^{-3} < \left(\frac{\kappa\varepsilon}{e^2}\right)^3. \tag{3.21}$$

Since the energy ε may take arbitrarily small values, the inequality (3.21) limits the rate of $N(\varepsilon)$ variations near $\varepsilon = 0$. We may replace the inequality sign by equality and differentiate the equality obtained:

$$g(\varepsilon) = \frac{\partial N}{\partial \varepsilon} \propto \frac{\kappa^3}{e^6}(\varepsilon - \mu)^2. \tag{3.22}$$

The equality sign indicates that $(\varepsilon - \mu)^2$ is the lowest possible power satisfying eqn (3.21).

This shape of the $g(\varepsilon)$ function in the vicinity of μ is referred to as a *soft Coulomb gap*. It is called soft because $g(\varepsilon)$ goes to zero only at one point and it is called Coulomb because it is caused by the Coulomb interaction.

Equations (3.21) and (3.22) are written for three-dimensional space. In the two-dimensional case, a similar consideration yields

$$N(\varepsilon) = r_{ij}^{-2} < \left(\frac{\kappa\varepsilon}{e^2}\right)^2, \qquad g(\varepsilon) = \frac{\partial N}{\partial \varepsilon} \propto \frac{\kappa^2}{e^4}|\varepsilon - \mu|. \tag{3.23}$$

The compensation coefficient K was not used in explicit form for considering the Coulomb gap. It should be assumed that the function $g(\varepsilon)$ goes to zero at the Fermi level not only at $K \approx 0.5$ but also at the K values closer to the end of the (0, 1) interval. This is confirmed by the computer simulation illustrated by Fig. 3.8 on the same scale determined by eqn (3.16) as in Fig. 3.7.

Note: The distribution for $K = 0.9$ is broader than for $K = 0.1$. This was to be expected from the comparison of eqns (3.9) and (3.15). Equation (3.15) has an additional numerical factor of about 1.3 and the factor $(1-K)^{1/3}$ in the denominator. Although eqns (3.9) and (3.15) do not include the Coulomb gap, they give the appropriate curve width and the position of the Fermi level.

The Coulomb gap directly affects the temperature dependence of the hopping conductivity in the vicinity of the Fermi level (variable-range hopping). This allows one to study the Coulomb gap experimentally. The problem is considered in detail in Chapter 4 dedicated to hopping conductivity. The experimental curves presented there allow one, in particular, to determine to what extent
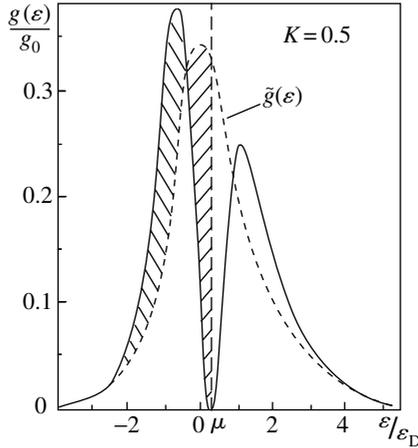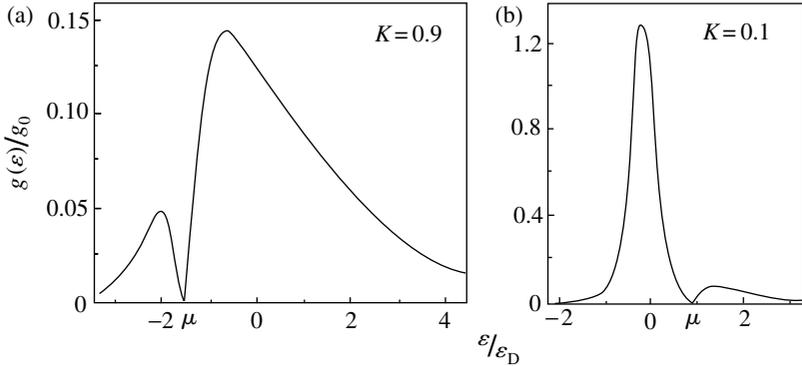
FIG. 3.8. Density of states in the impurity band for the compensation coeffi-
cients $K = 0.1$ and $K = 0.9$ according to the data of a computer simulation
(Efros et al. 1979).

the Coulomb gap may be considered as a universal phenomenon and answer the
question whether insulators with a nonzero density of states in the vicinity of
the Fermi level may exist at all.

The hopping conductivity is indirect proof of the existence of the Coulomb
gap. This can also be proved by more direct methods, i.e., by measuring the
current–voltage characteristics of tunnel junctions. These characteristics allow
one to directly determine the density-of-state function in the vicinity of the
Fermi level. Usually, one gradually varies the material properties from those of
metal to insulator. In the course of this transformation, the minimum density
of states at the Fermi level is gradually broadened and deepened and, finally,
attains the value $g(\mu) = 0$. As a rule, one simultaneously observes dramatic
changes in the character of the temperature dependence of the conductivity.
The interpretation of these experiments requires the invocation not only of the
Coulomb gap but also of the specific interelectron interaction in dirty metals,
metal–insulator transitions, hopping conductivity, etc. (see Appendix B). The
Coulomb gap plays an important part in the consideration of all the above
problems.

It is useful to trace the studies of the Coulomb gap. They started as a reaction
to experimental studies of hopping conductivity, which revealed that sometimes
the temperature dependence of the conductivity cannot be described either by an
activation law (4.10) or the law (4.15) for variable-range hopping at a constant
density of states in the vicinity of the Fermi level. The introduction of a Coulomb
gap allows one to describe both consistently and fully the experimental data on
the hopping conductivity (Chapter 4). Moreover, the shape of the Coulomb gap
predicted based on eqns (3.22) and (3.23) was confirmed experimentally by the
tunneling experiments.

Figures 3.9 and 3.10 show the characteristics obtained for tunnel metal–
insulator junctions. As is shown in Appendix B, the dI/dV derivative plotted
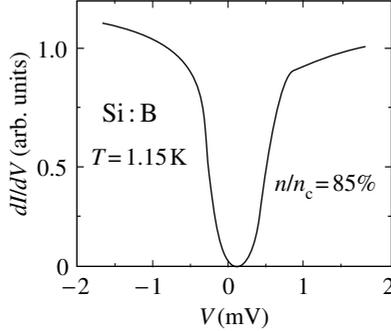
FIG. 3.9. Differential conductivity of the tunnel junction formed on an insulating Si : B sample with a low carrier density (Massey and Lee 1996). The complete series of curves is given in Fig. B.5 of Appendix B. The meaning of the critical concentration n_c is also explained in Appendix B.
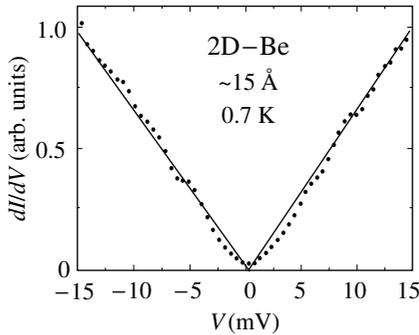
FIG. 3.10. Differential conductivity of the tunnel junction formed on an ultrathin high-resistance Be film at $T = 0.7\,\mathrm{K}$ (Butko et al. 2000). The complete series of the curves is given in Fig. B.6 of Appendix B.

along the ordinate axis is proportional to the density of states in the insulator, whereas the voltage V over the junction determines the energy (measured from the Fermi level) related to this density of states. In both cases, the Coulomb gaps are revealed. In the bulk Si : B sample, the gap shows parabolic behavior, in accordance with eqn (3.22). In ultrathin Be films, the gap varies linearly with the energy, in full accordance with the prediction based on eqn (3.23). The measurements made on both materials are discussed in more detail in Appendix B.

At the same time, it was clear from the very beginning that the theory is far from being perfect. Expressions (3.20)–(3.22) are equivalent to the allowance of solely the pair interelectron interaction, because it was assumed that the electron displacement from the center i to the center j leaves the arrangement of

all the other electrons intact. However, numerous simultaneous displacements of other electrons are also possible. In this case, one should compare the energies of the configurations where the occupancies not of two but of a larger number of centers differ. In turn, this may influence the electron spectrum. Therefore, decisive arguments in finding the shape of a Coulomb gap should be expected from experiments. We return to this problem in the next chapter.

References

Butko, V.Yu., DiTusa, J.F., and Adams, P.W. (2000). Phys. Rev. Lett. **84**, 1543.

Efros, A.L., Lien, N.V., and Shklovskii, B.I. (1979). J. Phys. C**12**, 1023.

Fritzsche, H. (1958). J. Phys. Chem. Sol. **6**, 69.

Grüner, G. (2000). *Density Waves*. Perseus Books.

Massey, J.G. and Lee, M. (1996). Phys. Rev. Lett. **77**, 3399.

Mott, N.F. and Twose, W.D. (1963). Adv. Phys. **10**, 707.

Peierls, R.E. (1955). *Quantum Theory of Solids*. Clarendon Press, Oxford.

Shklovskii, B.I. and Efros, A.L. (1984). *Electronic Properties of Doped Semiconductors*. Springer.

Shklovskii, B.I. and Shlimak, I.S. (1972). Sov. Phys. – Semicond. **6**, 104.

4

HOPPING CONDUCTIVITY

A somewhat different presentation of this problem may be found in the book by Shklovskii and Efros (1984); the words "hopping transport" are in the title of the large book edited by Pollak, M. and Shklovskii, B. (1991).

4.1. Localized states and transitions between these states

This chapter is dedicated solely to the localized states which are defined as the states whose wave functions are concentrated mainly in a confined region and exponentially decay outside this region,

$$\psi \to f(r)e^{-r/\xi} \quad \text{as } r \to \infty.$$

The coefficient ξ in the exponent is called the *localization radius* or the *localization length*. Thus, in a three-dimensional isotropic potential well

$$U(r) = \begin{cases} 0, & r > a \\ -U_0, & r \leqslant a \end{cases}$$

the wave function at infinity decreases as

$$\psi \propto \frac{1}{r}e^{-r/\xi}, \quad \xi = \frac{\hbar}{\sqrt{2m|E|}}, \tag{4.1}$$

where E is the position of the electron level measured from the upper edge of the well.

Electron wave functions at the levels of a one-dimensional rectangular well attenuate according to the law

$$\psi(x) \propto \exp(-x/\xi), \quad \text{i.e., } f(x) = \text{const.} \tag{4.2}$$

Again, the localization radius ξ is defined by expression (4.1). In particular, the wave function in a shallow well of width a_w and depth $U_0 \ll \hbar^2/ma_\text{w}^2$ with only one level attenuates within the length $\xi = \hbar^2/ma_\text{w}U_0$.

Note: the difference in the latter example between the well size a_w and the localization length ξ. We assume everywhere that $a_\text{w} \ll \xi$.

The localization length in an attractive Coulomb potential $U = -(e^2/r)$ created by a unit charge e is called the Bohr radius $a_\text{B} = \hbar^2/me^2$. The energy levels

E_n and the asymptotic behavior of isotropic electron wave functions ψ_n in a hydrogen atom depend on the principal quantum number n and are equal to

$$E_n = E_1/n^2, \quad \psi_n(r) \to C(n)r^{n-1}\exp(-r/na_{\mathrm{B}}) \text{ as } r \to \infty, \ (n = 1, 2, 3, \ldots),$$
$$(4.3)$$

where $C(n)$ are constants. It is seen from eqn (4.3) that the localization length of an electron depends on the principal quantum number as $\xi_n = na_{\mathrm{B}}$. In the ground state, $\xi_1 = a_{\mathrm{B}}$.

In order to describe impurity centers in solids, in particular, shallow donors and acceptors introduced in the previous chapter, one usually considers the model of hydrogen-like centers with the Bohr radius

$$a_{\mathrm{B}} = \kappa\hbar^2/m^*e^2, \tag{4.4}$$

whose numerator includes an additional factor, the permittivity κ, and the effective electron mass m^* replaces the conventional electron mass m in the denominator.

At finite temperatures, the localized states make their contribution to transport processes. Transport between the localized states arises as a result of carrier hopping from the occupied to free states and, therefore, is called hopping conductivity. An important postulate of the concept of hopping conductivity is the assumption that practically all the states have different energies: two states with equal energies are located at infinite distance from one another. Hence, the hopping is accompanied by emission or absorption of a phonon. When considering hopping conductivity, we shall proceed from the impurity-band model in lightly doped material (Chapter 3).

Let impurity centers with coordinates r_i have the concentration N so low that $Na_{\mathrm{B}}^3 \ll 1$ and electron wave functions are localized. Nevertheless, since the wave functions have exponential tails, the neighboring impurity centers overlap, which results in a certain finite probability of electron transition (hopping) from one center to another:

$$1/\tau_{ij} \propto F(\varphi_{ij}, f_i, f_j) \int |M_q|^2 \delta(\hbar qs - \Delta_{ij})\, d^3q$$

$$\propto F(\varphi_{ij}, f_i, f_j) \left| \int \psi_j^* e^{iqr} \psi_i\, d^3r \right|^2 \tag{4.5}$$

where s is the sound velocity. The factor $\exp(iqr)$ in the integrand in eqn (4.5) is a phonon wave function. A delta function ensures the fulfillment of the law of energy conservation and selection of phonons that should be either emitted or absorbed in order to compensate the energy difference between the initial and finite states, $\Delta_{ij} = \varepsilon_i - \varepsilon_j$. In the first approximation, the squared integral reduces to the factor $\exp(-2r_{ij}/a_{\mathrm{B}})$. The function F accumulates all the statistical factors affecting a transition: according to the Pauli principle, only a transition to an empty level ε_j may take place, a phonon may be absorbed only if the thermostat has such a phonon, etc. The arguments of the function

F are the Fermi and Bose distribution functions, $f_i = [\exp(\varepsilon_i - \mu)/T + 1]^{-1}$ and $\varphi_{ij} = [\exp(\Delta_{ij}/T) - 1]^{-1}$, respectively. Performing certain transformations and ignoring unimportant terms, one may reduce the function F to the factor $\exp(-\varepsilon_{ij}/T)$, where ε_{ij} is a certain characteristic energy not necessarily equal to $\Delta_{ij} = \varepsilon_i - \varepsilon_j$. One can write this factor proceeding from simple physical considerations.

Note: Although the hopping conductivity may arise only at a nonzero temperature, this temperature is assumed to be rather low, so that no excited impurity centers appear. According to eqn (4.3), the energy of a lower excitation attains 3/4 of the ionization energy. Therefore, with an increase in temperature, excited states appear practically simultaneously with ionized carriers, so that hopping conductivity is masked with band conductivity and becomes unimportant.

Thus, the probability of transitions between the centers i and j, $1/\tau_{ij}$, is proportional to the product of the two exponents. Now, connect each pair of impurity centers by a fictitious resistance R_{ij} inversely proportional to the transition probability,

$$R_{ij} = R_0 e^{u_{ij}}, \quad u_{ij} = \frac{2r_{ij}}{a_{\rm B}} + \frac{\varepsilon_{ij}}{T}. \tag{4.6}$$

Thus, we arrive at an Abrahams–Miller network of random resistances modelling an insulator. In equilibrium, the electron transitions between all the net points in both directions are equally probable (the principle of detailed equilibrium). The application of an external electric field to the system gives rise to a directed flow of electrons, i.e., to finite conductivity. To calculate this conductivity one has to apply Kirchhoff's laws to such a net.

At first sight, an Abrahams–Miller net looks frightening because, formally, each site of this net is connected with all the remaining sites. However, the resistances between the net sites lying at large distances from one another are exponentially high, and may be discarded because they are shunted by considerably lower resistances. This underlies the general method of the problem solution. One gradually removes high resistances from the net unless it remains connected. The resistance of the net depends on the highest resistances that one was forced to leave in the net to preserve its connectedness. Therefore, for better understanding of the material considered in this chapter, one has to be acquainted with Appendix A, and, in particular, with Section A.4. Below, we consider several variants of hopping conductivity. In each variant, we have to single out important sites, calculate R_{ij} values, and solve the corresponding percolation problem.

Note: The resistivity ρ of a regular orthogonal net with cell size a and bond resistance R depends on the net dimension d,

$$\begin{aligned} d = 1: \quad &\rho = R/a, \\ d = 2: \quad &\rho = R, \\ d = 3: \quad &\rho = Ra. \end{aligned} \tag{4.7}$$

The same relationships are also valid for an arbitrary net, but, in this case, a and R are assumed to be average values. (see Section A.1)

4.2. Nearest-neighbor hopping

The simplest form of hopping conductivity is presented by transitions between the nearest neighbors. The density of states in the donor impurity band at low donor concentration is maximum at energy value on the order of the ionization energy of an isolated donor, E_D. When the initial and finite points of such a hop are among the nearest neighbors, it is most probable that the energy levels of these points are in the vicinity of the maximum density of states. A hopping event takes place only if the terminating site is free. The probability that it is free depends on its energy with respect to the Fermi level μ and is proportional to

$$\exp(-|\mu - E_D|/T). \tag{4.8}$$

This is the smallest factor entering the function F in eqn (4.5). The term $\exp(-\Delta_{ij}/T)$ determining the number of phonons which may participate in transitions is larger because $\Delta_{ij} \ll |\mu - E_D|$. Therefore, u_{ij} in eqn (4.6) for all the hops is always equal to the distance from the maximum density of states to the Fermi level, $\varepsilon_{ij} = |\mu - E_D|$.

Note: This is the case, where $\varepsilon_{ij} \neq \Delta_{ij}$. Above, we have mentioned such possibility.

Since factor (4.8) enters all the u_{ij}, it makes no contribution to scattering in R_{ij} and, thus, may be "factored out", i.e., ignored in the analysis of net properties,

$$R_{ij} = R_0 \exp(2r_{ij}/a_B)$$

(the closer the sites to one another, the lower the resistance between these sites in the model net).

Thus, the problem is reduced to the determination of the percolation radius r_c in a system of randomly distributed sites with concentration N. Percolation theory states (eqn (A.10) in Appendix A) that

$$(4\pi/3)r_c^3 N = B_c = 2.7, \tag{4.9}$$

whence

$$r_c = 0.865 N^{-1/3} \quad \text{and} \quad \rho = \rho_0 \exp(1.73/N^{1/3}a_B). \tag{4.10}$$

Now, consider the experiment. Figure 4.1 shows the temperature dependence of the resistance of germanium samples with different dopant concentrations. This experiment has two characteristic features. First, the enormous range of the measured resistivity (12 orders of magnitude!) and dopant concentrations (almost three orders of magnitude). Second, a remarkable doping method ensuring the absence of any correlations in dopant distribution and rigorous maintenance of the compensation coefficient. A sample of pure Ge was irradiated with a neutron flux in a reactor. As a result of nuclear reactions with neutrons, nuclei of one of the Ge isotopes turn into Ga nuclei and create shallow acceptors, and nuclei of another Ge isotope turn into As nuclei and create shallow donors. The relative numbers of shallow donors and acceptors are determined by the cross-sections
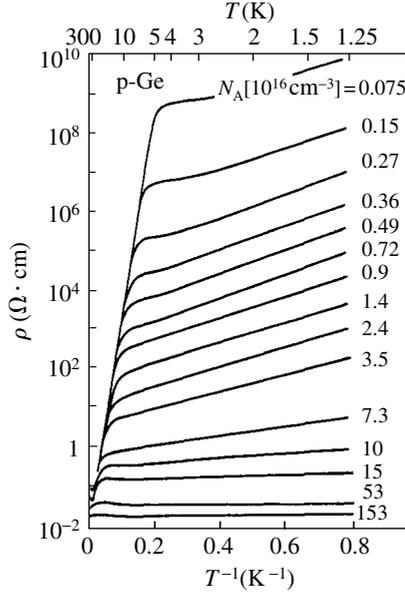
FIG. 4.1.   Resistivity of p-Ge samples with different dopant concentrations in
$(T^{-1}, \log \rho)$ coordinates (Fritzsche and Guevas 1960).

of the corresponding nuclear reactions and relative isotope concentrations in
the irradiated samples. Irradiation of natural germanium yields p-Ge with the
compensation coefficient $K = N_{\mathrm{A}}/N_{\mathrm{D}} = 0.4$. The N_{A} and N_{D} concentrations and
the carrier concentration $n = N_{\mathrm{A}} - N_{\mathrm{D}}$ depend only on the irradiation time.

To describe the set of curves thus obtained, it is assumed that there are two
conducting channels acting in parallel with two different exponentially varying
temperature factors

$$\sigma = \sigma_{\mathrm{b}} + \sigma_{\mathrm{h}} = \sigma_{\mathrm{b0}} \exp(-\varepsilon_{\mathrm{b}}/T) + \sigma_{\mathrm{h0}} \exp(-\varepsilon_{\mathrm{h}}/T); \quad \varepsilon_{\mathrm{b}} \gg \varepsilon_{\mathrm{h}}, \; \sigma_{\mathrm{b0}} \gg \sigma_{\mathrm{h0}}.$$
(4.11)

The channel σ_{b} is the band conductivity conventional for semiconductors with
shallow impurities, which is provided by thermal excitation of electrons from
the impurity levels to the conduction band and their transformation there into
free carriers. This channel is dominating at higher temperatures. The resistivity
curves in this region attain the limiting line in the left-hand side of the plot in
Fig. 4.1. An exponential increase in resistivity along the straight line

$$\rho = \rho_{\mathrm{b0}} \exp(\varepsilon_{\mathrm{b}}/T)$$
(4.12)

(where ε_{b} is the approximate ionization energy of the impurity) is explained by
freezing out of carriers in the conduction band at low temperatures. Resistivity
ceases increasing below the temperature when another conductivity mechanism
becomes prevalent – electron hopping from one impurity to another without

involving the conduction band. The temperature of cross-over from one conductivity mechanism to another (the channels σ_b and σ_h become comparable) is the higher, the higher the dopant concentration (N_A in Fig. 4.1).

We start the analysis of Fig. 4.1 noting the strong dependence of the σ_{h0} values on the impurity concentration. Extrapolating the functions $\rho(T^{-1})$ to $T = \infty$, we obtain Table 4.1.

Table 4.1 indicates that σ_{h0} includes an additional exponential factor, so that the experimental data are described by the expression

$$\sigma_h(T) = \sigma_{h0} \exp\left(\frac{-\varepsilon_h}{T}\right) = \sigma_{h0}^* \exp(-f(N)) \exp\left(\frac{-\varepsilon_h}{T}\right). \qquad (4.13)$$

Comparing eqn (4.13) with eqns (4.8) and (4.10), we obtain

$$\sigma_h(T) = \sigma_{h0}^* \exp\left(\frac{-1.73}{N^{1/3} a_B}\right) \exp\left(\frac{-|\mu - E_D|}{T}\right). \qquad (4.14)$$

The functional dependence following from eqn (4.14), $\ln \rho \propto N^{-1/3}$, is illustrated by the two experimental straight lines in Fig. 4.2. The slopes of these lines yield the numerical coefficient in the first exponent in expression (4.14). The values of this coefficient obtained from the two experiments illustrated by Fig. 4.2 and other analogous experimental data are listed in Table 4.2.

Although Table 4.2 indicates coefficients for different materials, they are all close to the coefficient value 1.73 in eqn (4.10) following from percolation theory.

TABLE 4.1.

$10^{-15} N$ (cm^{-3})	0.15	1.5	3.5	35
σ_{h0} ($\Omega \cdot$ cm)$^{-1}$	10^{-8}	10^{-6}	10^{-4}	10^{-1}
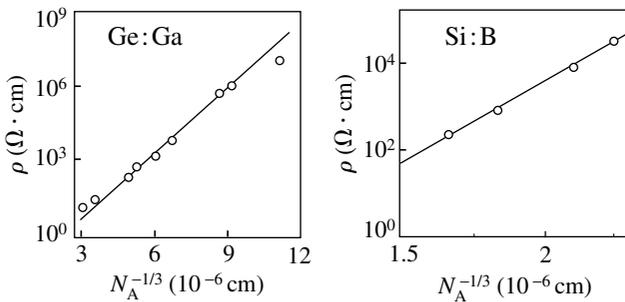
FIG. 4.2. Resistivity as a function of the average distance between acceptors, $N_A^{-1/3}$, in Ge : Ga (Fritzsche and Guevas 1960) and Si : B (Ray and Fan 1961).

TABLE 4.2.

n-GaAs	n-InP	p-Ge	p-Si
1.7	1.9	1.9	1.8
1.88		1.75	
1.9		2.0	

It is very important that the temperature dependence in Fig. 4.1 experimentally confirms the functional dependence $\ln(\rho/\rho_0) \propto 1/T$: the right-hand sides of the experimental curves in Fig. 4.1 are straight lines.

4.3. Variable-range hopping

4.3.1. *The Mott law*

The condition necessary for realization of nearest-neighbor hopping is the existence of a large number of pairs of close neighbors with one of them being free. A decrease in the temperature, such that

$$T \ll |\mu - E_{\mathrm{D}}|, \tag{4.15}$$

makes the number of empty sites among the nearest neighbors (the majority of which have energy E_{D}) too small, and hopping to the nearest centers freezes out. In turn, this increases the importance of hopping between the centers with energies lying in some ε-vicinity of the Fermi level where empty sites certainly exist. The problem now is to estimate how close these centers are to one another.

Consider the vicinity of the Fermi level $\mu \pm \varepsilon$ and assume that the density of states in this vicinity is constant, $g = g_\mu$. Then the number of states in this vicinity is $N(\varepsilon) = g_\mu \varepsilon$, the average distance between such states is $\bar{r}_{ij} \approx [N(\varepsilon)]^{-1/3}$, and the difference in their energies is on the order of ε. Now, the main statistical factor F in eqn (4.5) is determined by the existence of phonons φ_{ij} which are necessary for the fulfillment of the law of energy conservation. Since the energy of a necessary phonon is $\Delta_{ij} \sim \varepsilon$, the main factor in F is on the order of $\exp(-\varepsilon/T)$.

Now, retain in the random Abrahams–Miller net only the sites with energies lying within the range $\mu \pm \varepsilon$. The density $N(\varepsilon)$ and the average distance $\bar{r}_{ij}$ between the sites of the subnet thus obtained depend on the still unknown ε value. The neighbors in this subnet are connected by resistances (4.6) in which u_{ij} equals

$$u_{ij} = \frac{2}{a_{\mathrm{B}}[N(\varepsilon)]^{1/3}} + \frac{\varepsilon}{T} = \frac{2}{g_\mu^{1/3} a_{\mathrm{B}} \varepsilon^{1/3}} + \frac{\varepsilon}{T} . \tag{4.16}$$

Now determine ε from the condition that the u_{ij} value is minimal:

$$\frac{d}{d\varepsilon} u_{ij}(\varepsilon) = 0, \quad \varepsilon = \varepsilon_{\min} = \left(\frac{2T}{3g_\mu^{1/3} a_{\mathrm{B}}}\right)^{3/4} = (T^3 T_{\mathrm{M}})^{1/4}, \quad T_{\mathrm{M}} \approx (g_\mu a_{\mathrm{B}}^3)^{-1} > T. \tag{4.17}$$

The average energy change in hopping inside the subnet thus defined is of the order of $\varepsilon_{\min} \approx (T^3 T_{\mathrm{M}})^{1/4}$, whereas the average hopping length is

$$\bar{r} \approx \left(\frac{a_{\mathrm{B}}}{g_\mu T}\right)^{1/4} \approx a_{\mathrm{B}} \left(\frac{T_{\mathrm{M}}}{T}\right)^{1/4}. \tag{4.18}$$

The probability of hopping for a shorter distance is lowered because of the factor $\exp(-\Delta_{ij}/T)$, and the probability of hopping for a larger distance is lowered because of the factor $\exp(-2r_{ij}/a_{\mathrm{B}})$. The average hopping length $\bar{r}$ depends on temperature T; this justifies the term "variable-range hopping."

In terms of percolation theory, the hopping length $\bar{r}$ has the meaning of the interaction radius (A.10). It is easy to be certain that it exceeds the average distance $\bar{r}_{ij}$ between the sites of the subnet:

$$\bar{r}_{ij} = a_{\mathrm{B}} \left(\frac{T_{\mathrm{M}}}{T}\right)^{1/12} \ll \bar{r}.$$

The subnet itself is *temperature dependent*, but at any temperature, its centers connected by a finite hopping probability,

$$u_{ij} \approx 2 \left(\frac{T_{\mathrm{M}}}{T}\right)^{1/4},$$

form an infinite cluster. The resistivity of the subnet is

$$\rho = \rho_0 \exp \left(\frac{T_{\mathrm{M}}}{T}\right)^{1/4}. \tag{4.19}$$

The factor ρ_0 is a power function of temperature but this dependence is usually neglected.

The power $1/4$ in eqn (4.19) is inherent only in a three-dimensional insulator. For a thin film, we have

$$\bar{r}_{ij} \approx [N(\varepsilon)]^{-1/2}, \quad u_{ij} = \frac{2}{g_\mu^{1/2} a_{\mathrm{B}} \varepsilon^{1/2}} + \frac{\varepsilon}{T}, \tag{4.20}$$

$$\varepsilon_{\min} = \left(\frac{T}{g_\mu^{1/2} a_{\mathrm{B}}}\right)^{2/3} = (T^2 T_{\mathrm{M}})^{1/3}, \quad T_{\mathrm{M}} \approx (g_\mu a_{\mathrm{B}}^2)^{-1}. \tag{4.21}$$

Then, instead of eqns (4.18) and (4.19), we obtain

$$\bar{r} \approx \left(\frac{a_{\mathrm{B}}}{g_\mu T}\right)^{1/3} \approx a_{\mathrm{B}} \left(\frac{T_{\mathrm{M}}}{T}\right)^{1/3}. \tag{4.22}$$

and

$$\rho = \rho_0 \exp \left(\frac{T_{\mathrm{M}}}{T}\right)^{1/3}. \tag{4.23}$$

Using the notation d for the system dimension, we may combine eqns (4.19) and (4.23) to obtain one general expression

$$\rho = \rho_0 \exp \left(\frac{T_{\mathrm{M}}}{T} \right)^{1/(d+1)}, \quad T_M \approx (g_\mu a_{\mathrm{B}}^d)^{-1}. \tag{4.24}$$

4.3.2. The Efros–Shklovskii law

When deriving the Mott law, eqn (4.24), we assumed the density of states, $g(\varepsilon)$, in the vicinity of the Fermi level to be constant. However, the situation becomes different in the presence of a Coulomb gap. Then

$$g(\varepsilon) \propto \left(\frac{\kappa}{e^2} \right)^d |\varepsilon|^{d-1}, \quad g(0) = 0 \tag{4.25}$$

(d is the space dimension and the energy ε is measured from the Fermi level, see eqns (3.22) and (3.23) in Chapter 3). As in the derivation of the Mott law, consider again a certain vicinity of the Fermi level $\mu \pm \varepsilon$ symmetric with respect to μ. The number of states in this vicinity depends on the dimension d and is equal to

$$N(\varepsilon) \propto \left(\frac{\kappa \varepsilon}{e^2} \right)^d.$$

However, the average distance between the centers in this vicinity, $\overline{r}_{ij}$, does not depend on the dimension d,

$$\overline{r}_{ij} \approx [N(\varepsilon)]^{-1/d} \approx \frac{e^2}{\kappa \varepsilon}.$$

Now, repeating the reasoning of the previous section, we see that the resistivity exponent in the Abrahams–Miller subnet within the energy range ε

$$u_{ij} = \frac{2}{a_{\mathrm{B}}[N(\varepsilon)]^{1/d}} + \frac{\varepsilon}{T} = \frac{2e^2}{\kappa a_{\mathrm{B}} \varepsilon} + \frac{\varepsilon}{T} \tag{4.26}$$

is minimum at the values

$$\varepsilon_{\mathrm{min}} = (T T_{\mathrm{ES}})^{1/2}, \quad T_{\mathrm{ES}} \approx \frac{e^2}{\kappa a_{\mathrm{B}}} \tag{4.27}$$

and, irrespective of dimension, the average hopping length is

$$\overline{r} \approx \left(\frac{e^2 a_{\mathrm{B}}}{\kappa T} \right)^{1/2} \approx a_{\mathrm{B}} \left(\frac{T_{\mathrm{ES}}}{T} \right)^{1/2}, \tag{4.28}$$

whereas the conductivity obeys the law

$$\rho = \rho_0 \exp \left(\frac{T_{\mathrm{ES}}}{T} \right)^{1/2}. \tag{4.29}$$

Now, we have eqns (4.26)–(4.29) instead of eqns (4.16)–(4.19) for three-dimensional space and eqns (4.20)–(4.23) for two-dimensional space.

Note: The quantity T_{ES} having the dimensions of energy is *not* the width $\delta\varepsilon$ of the Coulomb gap. It is possible to evaluate $\delta\varepsilon$ by equating the quantity given by eqn (4.25) to the density of states at the Fermi level $g(\mu)$ calculated without the allowance for such a gap.

4.4. Experimental observation of hopping conductivity

Thus, the theory operates with several variants of the hopping conductivity. Experiment has to establish where and why each variant is implemented in practice. It should be kept in mind that a variant of hopping conductivity implemented under certain specific conditions depends not only on the material parameters but on temperature as well. Figure 4.3 shows a natural sequence of conductivity mechanisms replacing one another with a decrease in temperature. This sequence was drawn under the assumption that the Fermi level lies in the impurity band of the localized states of a lightly doped semiconductor. Each following sequence stage corresponds to a lower temperature T. However, each time, a lower T value in Fig. 4.3 is compensated with a larger scale, to keep the same delineation of the energy interval T.

The cross-over from the band conductivity of thermally excited carriers to nearest-neighbour hopping is well seen on the example of p-Ge in Fig. 4.1. One
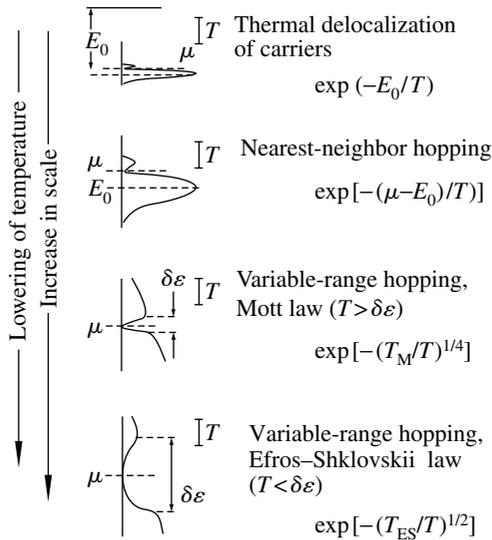
FIG. 4.3.   Changes of the hopping-conductivity mechanisms with lowering of temperature. The density of states at each temperature is indicated on an appropriate scale.
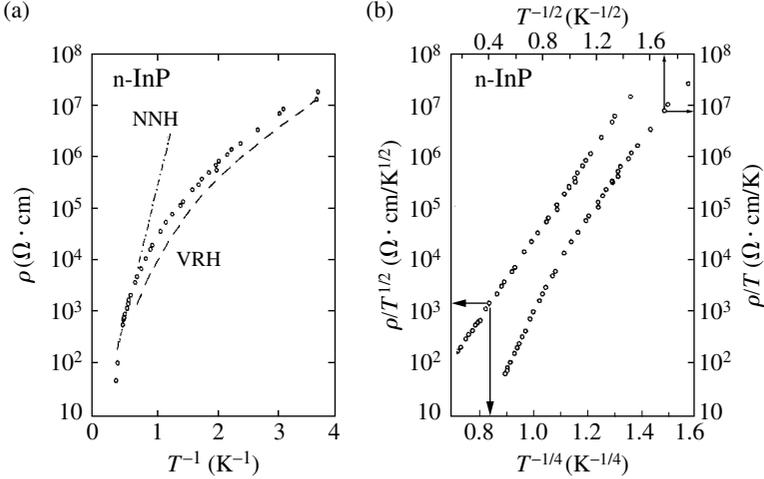
FIG. 4.4.   The determination of the functional temperature dependence of resist-
ivity of n-InP by using different coordinates (Mansfield et al. 1988).The dash-
dot line NNH in (a) shows the expected disposition of the experimental points
if the nearest-neighbor hopping were the leading transport mechanism; the
dash line VRH shows the same for variable-range hopping.

can see a kink between the straight line $\ln \rho \propto (1/T)$ common for all N_A con-
centrations in the left-hand part of Fig. 4.1 and the straight lines on the right
coming to it from the low-temperature region.

Some variants of the change of the prevailing hopping-conductivity mechan-
ism were also observed but they were less obvious. This is explained by the
procedure of the experimental determination of the temperature exponent. Plot-
ting the experimental plots in $(T^{-1/\nu}, \ln \rho)$ coordinates, where $\nu = 1, 2, 3,$ or 4,
we obtain straight lines only for an appropriate choice of the power ν, otherwise
we obtain certain bent curves. To determine with sufficient accuracy whether
the dependence plotted at a certain power of T along the abscissa fits a straight
line or not, we have to consider a rather large range of function variation. This is
illustrated by the plots of the resistivity data for n-InP. One may readily distin-
guish nearest-neighbor hopping from variable-range hopping ($\nu = 1$ from $\nu \neq 1$).
Resistivity in the case of nearest-neighbor hopping in the coordinates shown in
Fig. 4.4a should be depicted by a straight line, whereas the dashed curve in
Fig. 4.4a corresponds to the Mott law, eqn (4.19). There is no doubt that the
experimental points are considerably closer to the curve than to the straight line.
However, to be able to distinguish $\nu = 1/2$ from $\nu = 1/4$ in n-InP (Fig. 4.4b),
we have to incorporate the range of ρ variation by five orders of magnitude.

The most often used natural temperature range in experiments on hopping
conductivity is 4–0.04 K. Which of the mechanisms really acts in this range
depends on specific parameters of each material. As is seen from Fig. 4.5,
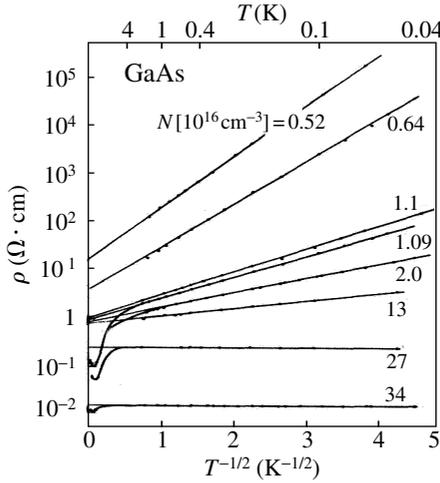
FIG. 4.5. Resistivity of GaAs samples with different dopant concentrations in $(T^{-1/2}, \log \rho)$ coordinates (Rentzsch et al. 1986).
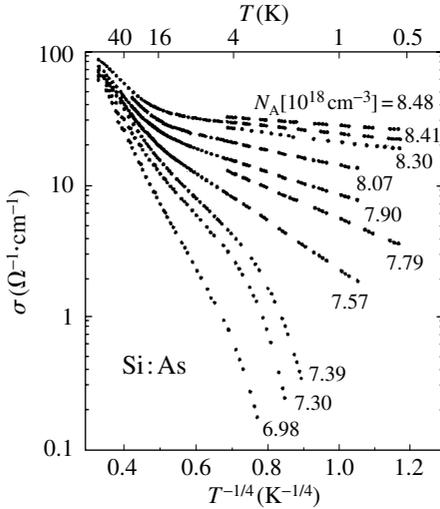
FIG. 4.6. Resistivity of Si : As samples with different dopant concentrations in $(T^{-1/4}, \log \rho)$ coordinates (Shafarman et al. 1989).

the resistivity of GaAs obeys the law $T^{-1/2}$ in this temperature range without any deviations. Therefore, this material has a Coulomb gap with a width of several degrees. The resistivity of a Si : As sample (Fig. 4.6) at some range of As concentrations behaves according to the law $T^{-1/4}$ in the temperature range
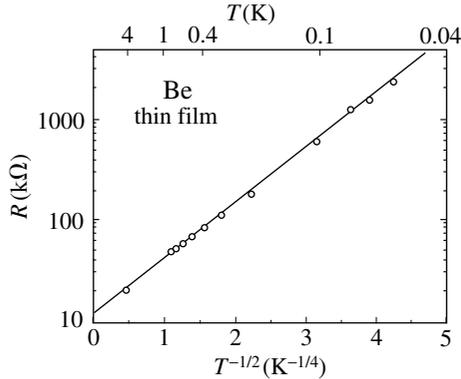
FIG. 4.7. Temperature dependence of the resistance of an ultrathin Be film
   (Butko et al. 2000).

from about $10\,\mathrm{K}$ to, at least, $0.5\,\mathrm{K}$. This signifies that in $\mathrm{Si:As}$ samples, the
Coulomb gap certainly has a width less than $0.5\,\mathrm{K}$ or that the spectrum has no
gap at all.

   All the materials considered up to now in this chapter are classical semicon-
ductors with shallow donors or acceptors. With lowering of the temperature, all
these materials become insulators. One should not think that hopping conductiv-
ity is a property inherent only in this rather limited class of materials. Consider
the conductivity in a material of quite a different nature – in a metal film so
thin that its conductivity is blocked by surface scattering. Tunnel experiments
on an ultrathin Be film performed in the temperature range 1–$0.5\,\mathrm{K}$ prove the
existence of a Coulomb gap (Fig. 3.10 in Chapter 3 and Fig. B.5 in Appendix B).
Figure 4.7 shows that the film is insulating and that transport in it proceeds by
the mechanism of hopping conductivity. As should be expected in the case of
the existence of a Coulomb gap, the film resistivity varies according to the law
$\ln \rho \propto T^{-1/2}$.

   Thus, all the hopping-conductivity mechanisms predicted theoretically were
also observed experimentally. Nevertheless, a number of fundamental questions
still received no experimental answers. First, it is still unclear whether a Coulomb
gap arises in all cases or an insulator with a finite density of states at the Fermi
level is also possible. The fact that the experiments on n-InP (Fig. 4.4) and
$\mathrm{Si:As}$ (Fig. 4.6) samples showed that conductivity varies according to the Mott
law does not necessarily signify the absence of a Coulomb gap in these materials.
It allows us only to establish the upper boundary $(0.3$–$5\,\mathrm{K})$ for the width of the
possible Coulomb gap.

   In principle, the existence of an insulator with localized states at the Fermi
level without a Coulomb gap is quite possible. As we saw above, the gap may be
formed if electrons have enough space for a maneuver: the number of electron
distributions over the centers should be rather large. However, this condition is
not always fulfilled because of some specific structural features associated with

the defect arrangement in a material. The experimental study of this problem requires low-temperature measurements. At the same time, with the extension of the measurement range to lower temperatures, conductivity becomes so low that it is difficult to perform reliable measurements.

There is another fundamental question: once a Coulomb gap is formed, to what degree of accuracy it is soft and eqns (3.22) and (3.23) are exact? According to theoretical estimates, due to higher orders of Coulomb interaction, the gap should become hard. The experimental answer is hampered by difficulties in low-temperature measurements of the conductivity. Thus, an important experimental problem is to descend for one step lower in the scheme shown in Fig. 4.3 and to determine the structure of the function $g(\varepsilon)$ at distances from the Fermi level closer than the gap width.

One should bear in mind that there is one more possibility. The scheme in Fig. 4.3 is drawn under the assumption that the density of states $g(\varepsilon)$ does not depend on temperature and that the change of a conductivity mechanism is provided by the scale of energies that may be transferred to a thermostat, i.e., by the temperature. However, it might happen that $g(\varepsilon)$ may depend on temperature by itself; some arguments in favor of such a possibility were presented in section 3.3, see eqn (3.19) and the text after it. For instance, the soft gap may transform into a hard gap with a decrease in temperature.

All these problems and difficulties are well illustrated on a Si:B sample. Figure 4.8 shows the resistivity data for a Si:B sample in the temperature range from 4 to 0.1 K illustrating cross-over from 1/4 to 1/2 in the exponent of the activation process. The deviation from the straight line in the lower part of Fig. 4.8 with $T^{-1/2}$ designated along the abscissa indicates that at temperatures $T > 1\,\mathrm{K}$ the 1/2 law becomes invalid. However, the inset shows that this part of the curve becomes a straight line being plotted as a function of $T^{-1/4}$. Thus,
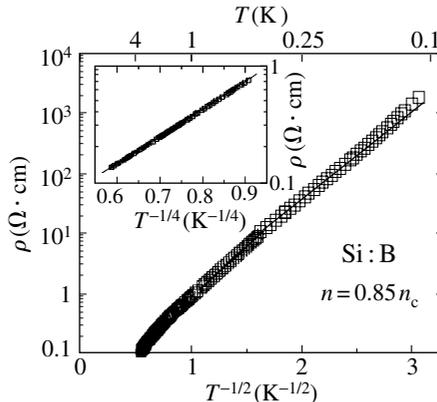
FIG. 4.8.   Temperature dependence of resistivity of Si:B samples plotted in $(T^{-1/2}, \log \rho)$ and $(T^{-1/4}, \log \rho)$ coordinates (Massey and Lee 1995).
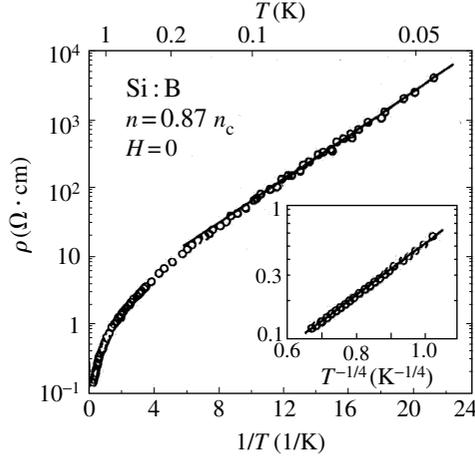
FIG. 4.9. Cross-over of the temperature dependence of a Si : B sample from the Mott ($\ln \rho \propto T^{-1/4}$) to Arrhenius ($\ln \rho \propto T^{-1}$) law (Dai et al. 1995). The temperature range is extended toward low temperatures as compared to the range presented in Fig. 4.8.

above 1 K, the Mott law is valid, whereas below this temperature, the cross-over to the Efros–Shklovskii law takes place. This agrees with the electron spectrum obtained with the use of a tunnel junction, which demonstrates the presence of a Coulomb gap in a Si : B sample at 1.15 K (see also Fig. 3.9 in Chapter 3 and Fig. B.5 in Appendix B).

However, when a cross-over takes place the involved temperature range is cut into two subranges. One of them, or even both, may turnout to be not large enough to determine the exponent with confidence. Although the points of the lowest temperatures in Fig. 4.8 deviate from the straight line in ($T^{-1/2}$, $\ln \rho$) coordinates, these deviations do not exceed the admissible errors. Only extending the range of measurements in Fig. 4.9 toward lower temperatures makes the conclusions drawn more accurate. The Mott law for $T > 1$ K is in fact valid. However, at lower temperatures, the resistivity of a Si : B sample in a comparatively weak magnetic field and without such a field behaves quite differently. In the zero magnetic field, a cross-over from the power $1/4$ is observed not to the $1/2$, but to 1, i.e., the resistivity at lowest temperatures follows an Arrhenius law, $\rho \sim \exp(T_0/T)$ (Fig. 4.9). This signifies that a soft Coulomb gap in the vicinity of the Fermi level with lowering of temperature is changed for a hard gap: the density of states $g(\varepsilon)$ goes to zero not at one point $\varepsilon = \varepsilon_F$, but within a certain energy region T_0 in the vicinity of the Fermi level. It is seen from Fig. 4.9 that $T_0 \simeq 0.37$ K (according to tunnel measurements indicated above, the total gap width is on the order of 10 K).

It is also worth noting that in a magnetic field, the Coulomb gap remains soft. As is seen from Fig. 4.10, the points in fields $H > 2$ T fit straight lines even
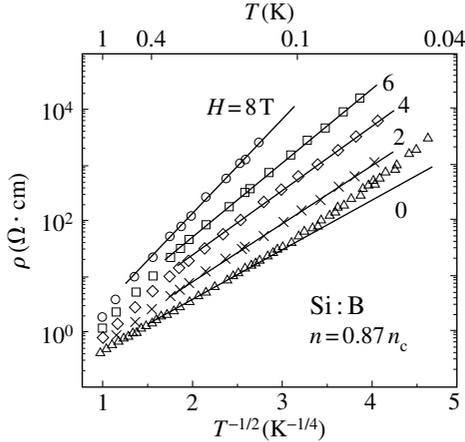
FIG. 4.10.   Change of the functional dependence of resistivity of a Si : B sample in an applied magnetic field (Dai et al. 1995).

at the lowest temperatures. It is unclear whether the gap becomes hard due to Coulomb or some other interaction. Its dependence on the magnetic field gives some grounds to believe that the gap becomes hard because of spin correlations, but this problem is beyond our scope.

References

Butko, V.Yu., DiTusa, J.F., and Adams, P.W. (2000). Phys. Rev. Lett. **84**, 1543.

Dai, P., Zhang, Y., and Sarachik, M.P. (1992). Phys. Rev. Lett. **69**, 1804.

Fritzsche, H. and Guevas, M. (1960). Phys. Rev. **119**, 1238.

Mansfield, R., Abboudy, S., and Foozoni, F. (1988). Phil. Mag. B**57**, 777.

Massey, J.G. and Lee, M. (1995). Phys. Rev. Lett. **75**, 4266.

Pollak, M. and Shklovskii, B. (eds), (1991). *Hopping Transport in Solids.* North-Holland.

Ray, R. and Fan, H. (1961). Phys. Rev. **121**, 768.

Rentzsch, R., Friedland, K.J., Ionov, A.N., et al. (1986). Phys. Stat. Sol. B **137**, 691.

Shafarman, W.N., Koon, D.W., and Castner, T.G. (1989). Phys. Rev. B **40**, 1216.

Shklovskii, B.I. and Efros, A.L. (1984). *Electronic Properties of Doped Semiconductors.* Springer.

5

METAL–INSULATOR TRANSITIONS

The same problems are also discussed, although in a somewhat different style, in the book by Mott (1990) and the review articles by Lee and Ramakrishnan (1985) and Kramer and MacKinnon (1993).

The fundamental difference between insulators and metals consists in that the electron states at the Fermi level in insulators are localized, whereas in metals, they are itinerant. If, successively varying a certain parameter, one manages to bring an insulator to a metal state, the symmetry of the wave functions at the Fermi level changes. It is this symmetry that distinguishes a metal from an insulator. Defining a metal and an insulator in terms of the wave functions of electrons, we should like to indicate that the main physical property that distinguishes the materials of these two types is the conductivity, i.e., the possibility of materials to conduct currents in an infinitely weak electric field. This characteristic is of the "yes–no type" – the conductivity has either a zero value, $\sigma = 0$, or any, even infinitesimal, nonzero value. However, at a finite temperature $T \neq 0$, an insulator can also conduct a current by the mechanism of hopping conductivity. Therefore, the above definition of an insulator is valid only at the temperature $T = 0$. To answer the question whether the material studied is a metal or an insulator, it is necessary to extrapolate the experimental $\sigma(T)$ dependence to the temperature $T = 0$ despite the fact that the corresponding procedure is rather inconvenient and cumbersome and, in many instances, may yield ambiguous results.

Note: For quite a long time, it has been believed that it is possible to distinguish a metal from an insulator by the sign of the conductivity derivative at low temperatures. Materials with $\partial \sigma / \partial T \leqslant 0$ were believed to be metals, whereas those with $\partial \sigma / \partial T > 0$, insulators. The study of quantum corrections to the conductivity showed that the conductivity derivative may be positive, $\partial \sigma / \partial T > 0$, for metals as well, so that the sign of the derivative $\partial \sigma / \partial T$ cannot be considered as a criterion for distinguishing metals from insulators.

Since the existence of conductivity allows one to distinguish a metal from an insulator only at $T = 0$ and the wave functions (whose symmetries are compared) are assumed to be functions of the ground state, the definition of a metal–insulator transition makes sense only at $T = 0$. There are two main factors that affect the wave functions of the ground state and whose variation can give rise to a metal–insulator transition. These are disorder and the

electron–electron interaction. A transition induced by disorder in a system of noninteracting electrons is called an Anderson transition. A transition induced by the interelectron interaction is called a Mott transition. As a rule, both factors vary simultaneously, so that real transitions may be called Anderson–Mott transitions. The real control parameter that affects either one or both main factors may be an impurity concentration, pressure, magnetic field, etc. We denote the control parameter by x. On the phase plane (x, T), the transition is depicted as an isolated point on the line $T = 0$.

It is very important that hereafter we consider those electron phase transitions which involve only electrons. There also exist *structural* transitions in which the change of the atomic-system state is accompanied by electron delocalization but we do not consider these transitions here. However, a decrease in electron energy is often the main cause of a structural phase transition. An example is a Peierls transition. In the model considered in Section 3.1, the transition to the state of an insulator occurs at a finite temperature. At this temperature, the gain in the temperature-dependent electron energy associated with the appearance of a new period becomes equal to the loss in the elastic energy caused by ionic displacements.

The structural transitions are diverse. The states of an electron subsystem on both sides of this transitions are usually evaluated based on the band theory proceeding from the position of the Fermi level with respect to the energy bands. For example, at the temperature $T = 18\,°C$, the transition from white to gray tin takes place. The high-temperature phase (white tin) is a good metal in which the Fermi level intersects several energy bands and forms Fermi surfaces in these bands. In gray tin (thermodynamically stable at low temperatures), the Fermi level lies in the energy gap and, therefore, gray tin is an insulator. We call such an insulator a band insulator in order to distinguish it from an Anderson insulator, where, in the close vicinity of the Fermi level, there exist electron states but only with localized wave functions.

Considering an arbitrary insulator, e.g., Ge or Si or gray tin, as a matrix and doping it with impurities, it is possible to give rise to an electron transition to the metal state. In this case, the critical impurity concentration may be determined experimentally in low-temperature transport measurements.

Thus, a metal–insulator transition is understood here as an electron transition and, therefore, we consider the occurrence of these transitions with the change of the electron density and the degree of disorder.

5.1. The Anderson transition

Following Anderson, consider a periodic lattice of rectangular wells having different depths and concentration $N = a^{-d}$ (here a is the lattice period and d is the lattice dimensions). Let the levels in these wells exist in the energy interval W and the density of states in this range be constant (Fig. 5.1). The energy is measured from the point $\varepsilon = 0$ in the center of the interval W.
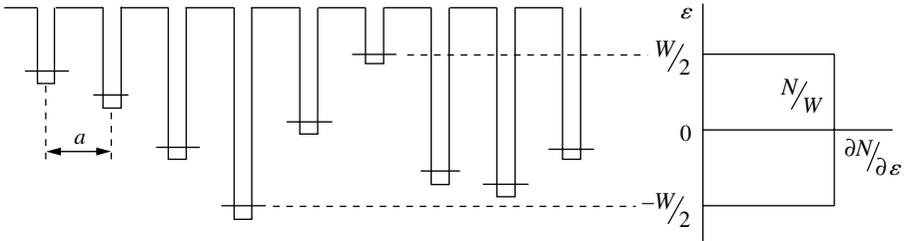
FIG. 5.1. Anderson model: periodically arranged wells of different depths.

Then, we have

$$\frac{\partial N}{\partial \varepsilon} = \begin{cases} N/W & \text{at } |\varepsilon| \leqslant W/2, \\ 0 & \text{at } |\varepsilon| > W/2. \end{cases} \tag{5.1}$$

Note: In this case, the level $\varepsilon = 0$ is not the Fermi level. The number of electrons in the system depends on various external factors.

The wave functions have tails $\exp(-r/a_\mathrm{B})$ and, therefore, the wave functions of the electrons localized in the neighboring wells (see eqns (4.1) and (4.2) in Section 4.1) overlap. If the distance r_{12} between two neighboring wells obeys the inequality $r_{12} \gg a_\mathrm{B}$, then the *transfer integral*

$$J = \int \psi_1^* \hat{H} \psi_2 \, d^3 r \propto \exp(-r_{12}/a_\mathrm{B}) \tag{5.2}$$

has a low value determined by the factor $\exp(-r_{12}/a_\mathrm{B})$.

In principle, two limiting cases are possible. Each electron may be localized in its own well (this is the case, e.g., if wells are very deep and differ from one another). On the other hand, all the electrons may be delocalized, so that any electron is spread over all the wells. For instance, if all the wells have the same depths, the wave functions of electrons are the Bloch waves.

Anderson considered the case of a three-dimensional set $(d = 3)$ of periodically located wells and came to the conclusion that the symmetry of the wave functions of the ground state depends on the relative degree of disorder. The parameter in the Anderson problem is the energy ratio J/W. If the condition

$$\frac{J}{W} \geqslant \left(\frac{J}{W} \right)_{\mathrm{crit}}. \tag{5.3}$$

is fulfilled, then at least some states are delocalized, and metal conductivity exists. At the critical value of the ratio J/W, the delocalized states appear in the band center at $\varepsilon = 0$. A further increase in the ratio J/W results in a gradual broadening of the layer of delocalized states.

The sense of the ratio J/W and its role may be explained by considering the simplest problems of quantum mechanics. In first-order perturbation theory,
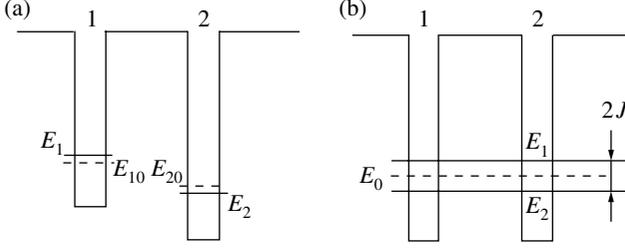
FIG. 5.2. (a) Shifts of the levels in two rectangular wells of different depths and (b) splitting of the levels in two equivalent wells with due regard for the overlap of the wave functions in the first order of perturbation theory.

a finite overlap of the φ_1 and φ_2 wave functions in two wells of different depths with levels E_{10} and E_{20} results in small corrections to the wave functions

$$\psi_1 = c_1\varphi_1 + c_2\varphi_2, \quad \psi_2 = -c_2\varphi_1 + c_1\varphi_2, \quad c_2 = J/(E_{10} - E_{20}). \tag{5.4}$$

As long as $c_2 \ll c_1 \approx 1$, all the electrons are predominantly located in their own wells.

If the wells are equivalent ("resonant"), $E_{10} = E_{20} = E_0$, the answer is quite different. The level E_0 is split into $E_{1,2} \simeq E_0 \pm J$ (Fig. 5.2b), and the wave functions of both states are uniformly smeared over the two wells:

$$\psi_{1,2} = \frac{1}{\sqrt{2}}(\varphi_1 \pm \varphi_2). \tag{5.5}$$

Not only are the structures of the wave functions (5.4) and (5.5) different, but the orders of magnitude of the level shifts are also different (Fig. 5.2). For definiteness, consider rectangular one-dimensional wells. For wells with different depths, each of the two wells is a perturbation for the electron located in the other well. Since the unperturbed wave function φ_1 in the vicinity of well 2 is of the order of $\exp(-r_{12}/a_B)$, then the shift of the level E_{10} is of the order of

$$\Delta_1 E \equiv E_1 - E_{10} \simeq \int \varphi_1^* \hat{H}_2 \varphi_1 \, d^3r \propto \exp(-2r_{12}/a_B), \tag{5.6}$$

In other words, the squared small factor $\exp(-r_{12}/a_B)$ of the transfer integral J enters $\Delta_1 E$ (Fig. 5.2a). For resonant wells (Fig. 5.2b), we have

$$\Delta E \sim J, \quad \text{i.e., } \Delta E \propto \exp(-r_{12}/a_B). \tag{5.7}$$

The wells behave as resonant wells as long as the difference between their unperturbed energies obeys the inequality $|E_{10} - E_{20}| < J$. Therefore, J/W is the fraction of resonant wells. Then the critical value $(J/W)_{\text{crit}}$ may be interpreted as a percolation threshold above which the spectrum acquires the states with delocalized wave functions.

If the value of the parameter J/W is lower than the critical value and the wave functions at the Fermi level are localized, the material is called an Anderson

insulator. Its structure, in principle, differs from that of a band insulator. In a band insulator, the Fermi level is located in the forbidden band, where the density of states equals zero and conductivity is provided by electrons thermally excited to the conduction band or by holes thermally excited to the valence band. The density of states at the Fermi level in an Anderson insulator is finite. The energy level beginning with which the states become delocalized is spaced by a certain distance from the Fermi level. This level is called the *mobility edge*. It plays the role of the bottom of the conduction band. The electrons or holes thermally excited above the mobility edge participate in conductivity via diffusion; the holes or electrons below the mobility edge participate in the hopping conductivity. The transition to the metal state takes place if the Fermi level is brought into coincidence with the mobility edge. This may be attained either by varying the parameter J/W or by shifting the Fermi level.

The Anderson model is also studied for systems with reduced dimensionality. In phase transitions, the dimensionality is usually an important parameter. In Chapter 2, its role was discussed in connection with weak localization. The criterion of dimensionality reduction introduced in Chapter 2 is very soft, because the characteristic size of an object b was compared in eqn (2.8) with a relatively large diffusion length L_φ. Here, we consider strong localization and, correspondingly, the criterion should be much more rigid. It is associated with the structure of the spectrum of confined electrons. The spectrum takes the form

$$\varepsilon = \hbar^2 k_\parallel^2 / 2m + \varepsilon_\perp(i), \quad i = 1, 2, \dots . \tag{5.8}$$

Here $k_\parallel$ is the wave vector along the directions of unlimited electron motion, $\varepsilon_\perp$ is the part of the energy quantized due to the confinement, and i is the number of the quantized subband. In the film, $k_\parallel^2 = k_x^2 + k_y^2$; a standing wave is formed along the normal Oz and the corresponding energy part is quantized. In a wire, $k_\parallel$ is the wave vector directed along the wire axis, and $\varepsilon_\perp$ is determined by the quantization in two transverse directions.

If all the electrons find room in the lower of the quantized subbands, the system has a reduced dimensionality. For a degenerate electron system, the criterion is

$$\varepsilon_F < \Delta_s, \quad \Delta_s \equiv \varepsilon_\perp(i=2) - \varepsilon_\perp(i=1) \propto b^{-2}. \tag{5.9}$$

The properties of the Anderson model in systems with reduced dimensionality may be briefly formulated in the following way. In one-dimensional systems, $d = 1$, an infinitesimal disorder gives rise to localization, whereas two-dimensional systems, $d = 2$, should be considered as a boundary case with respect to possible appearance of delocalized states.

5.2. The Landauer formula for one-dimensional (1D) systems

The theoretical aspects of this problem are consistently described by Imry (1997).

Connect two reservoirs with an *ideal* wire of length Λ and apply to these reservoirs the potential difference V. The word "ideal" is used to show that there

is no scattering (even elastic) in the wire. Then, any electron entering one end would depart from the other end of the wire with probability equal to unity. Let diameter b of a wire be so small that under the Fermi level ε_F there is only a limited number of quantized subbands $\nu = 2N_s$ of its spectrum (5.8),

$$\varepsilon_\perp(i) < \varepsilon_F \quad \text{for } i = 1, 2, \ldots, N_s. \tag{5.10}$$

These subbands are also called channels; in the absence of a magnetic field, each $i \leqslant N_s$ gives rise to two channels with opposite spins.

If $N_s = 1$, then the 1D system is called a one-channel system (taking into account the spin, it could also be called a two-channel system). Systems with $N_s > 1$ are called multichannelled. Since we consider an ideal wire, the wire channels are independent and cannot exchange electrons. The electron density n_i in channel i, the longitudinal electron velocity v_i, and the density of states g_i at the Fermi level are related as

$$v_i = \hbar^{-1}\left(\frac{\partial\varepsilon}{\partial k}\right)_{\varepsilon=\varepsilon_i}, \quad g_i = \left(\frac{\partial n_i}{\partial\varepsilon}\right)_{\varepsilon=\varepsilon_i} = \frac{1}{2\pi\hbar v_i}, \quad \varepsilon_i = \varepsilon_F - \varepsilon_\perp(i), \ 2\sum_{i=1}^{N_s} n_i = n. \tag{5.11}$$

The existence of the potential difference V between the reservoirs signifies that the difference in the electron density $\delta n_i = g_i eV$ induces the difference between the electron flows entering the channel i from the right- and left-hand sides. The specific channel parameters that enter eqn (5.11) are cancelled in the equation for the current $J_i = ev_i\delta n_i$ in the channel, so that it does not depend on the subscript i and is equal to

$$J_i = ev_i g_i eV = \left(\frac{e^2}{2\pi\hbar}\right)V. \tag{5.12}$$

The conductance $y_{id} = J/V$ and resistance $\varrho_{id} = 1/y_{id}$ of the wire are determined by the total current, $J = \sum_1^\nu J_i$, over all ν operating channels and are equal to

$$y_{id} = \left(\frac{e^2}{2\pi\hbar}\right)\nu, \qquad \varrho_{id} = \left(\frac{2\pi\hbar}{e^2}\right)(1/\nu). \tag{5.13}$$

The subscript emphasizes that eqn (5.13) is written for an ideal wire.

The result described by eqn (5.13) is remarkable from several standpoints. First, it turned out that an ideal 1D system (even if it is multichannel) is always characterized by a certain dissipation though there is no scattering in it. This is a manifestation of the *nonlocality principle*. Electrons take the energy from the field passing the wire and give it away somewhere outside the wire, i.e., being thermalized in the reservoir. Second, amazing as it is, the wire resistance ϱ_{id} is independent of its length and is determined solely by the quantization of the electron spectrum.

It might seem that the statement that the wire resistivity is independent of its length is inconsistent with the following simple consideration. Imagine that the wire is cut into two parts connected in series. If each part has resistance ϱ_{id}, then the total resistance
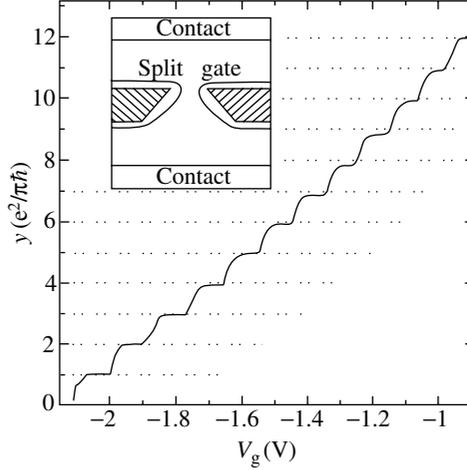
FIG. 5.3. Conductance y of the ballistic contact between two 2D regions of the GaAs–Al$_x$Ga$_{1-x}$As heterostructure as a function of the gate voltage controlling the contact width (van Wees et al. 1988). Inset: schematic of a measuring cell.

would have been $2\varrho_{\mathrm{id}}$. However, it is not sufficient only to cut the wire into two parts. To make independent resistors from these two parts, one has to insert between them an additional reservoir–thermostat to make the propagating electron waves incoherent. If the temperature of the wire differs from absolute zero, $T \neq 0$, so that there exists a finite length $L_\varphi < \infty$ within which phase breaking takes place, then such "thermostats" would appear automatically spaced by distances L_φ from one another.

So, the temperature limits the possible length of the ideal wire from above, $\Lambda < L_\varphi$. The limit from below comes from the wire diameter: $\Lambda > b$, i.e. the wire may be very short. This gives a way to verify eqn (5.13), since it is simpler to prepare a small region without defects.

Figure 5.3 illustrates conductivity measurements in a narrow channel connecting two regions of the 2D electron gas under the split gate in the GaAs–Al$_x$Ga$_{1-x}$As heterostructure (van Wees et al. 1988). With an increase in the gate voltage V_{g}, the nonconducting region becomes somewhat broadened because it slightly protrudes outside the gate edges. As is seen from the inset in Fig. 5.3, this narrows the conducting channel, which reduces the number of channels, N_{s}. The small length of these channels makes it possible to get rid of scattering but does not affect applicability of eqn (5.13). In the structure shown in Fig. 5.3, the electron density is $3.56 \cdot 10^{11}\,\mathrm{cm}^{-2}$, the mean free path at $0.6\,\mathrm{K}$ is $8.5\,\mu\mathrm{m}$, and the characteristic wire sizes are of the order of $0.25\,\mu\mathrm{m}$.

It is seen from the inset in Fig. 5.3 that measurements were made using a two-contact scheme, so that the measured resistance R_{mea} includes the resistance R_{cont} of the contacts and the adjacent broad regions of the 2D layer.
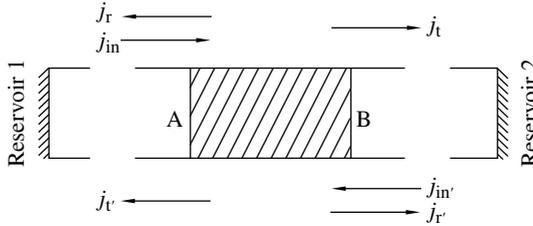
FIG. 5.4. A one-dimensional conductor connecting two reservoirs and consisting of two ideal end portions and the scattering portion AB in the middle.

The conductance y we are interested in is obtained by subtraction of R_{cont} from R_{mea}, namely $y \equiv \varrho^{-1} = (R_{\text{mea}} - R_{\text{cont}})^{-1}$. We used as R_{cont} the value of $4.35 \, \text{k}\Omega$. This value approximately corresponds to the data of independent measurements. In full accordance with eqn (5.13), the function $y(V_{\text{g}})$ after the subtraction of this value is transformed into a sequence of steps with equal heights

$$\Delta y_{\text{id}} = \left(\frac{e^2}{2\pi\hbar}\right) \Delta\nu = \frac{e^2}{\pi\hbar}, \quad \Delta\nu = 2. \tag{5.14}$$

Note: The considerations that result in eqns (5.12) and (5.13) do not predetermine the electric-field distribution along the wire. Thus, in the ideal one-dimensional edge channels formed due to the quantum Hall effect along the sample edge between the contacts, the voltage drop V is concentrated at the boundary with one of the contacts (see Fig. 9.16 in Chapter 9).

Now, consider a nonideal and, for simplicity, one-channel wire $\nu = 1$ (Fig. 5.4). Let some elastic scatterers be located in the hatched part of the wire. At the moment, we do not specify their mutual arrangement and shall consider the whole region as one scattering object. In the quantum mechanics of a one-dimensional system, such an object is characterized by complex reflection and transmission coefficients, r and t, relating the amplitudes of the reflected (backscattered) and transmitted waves with the amplitude of the incident wave. The electron flows j_{in}/e and $j_{\text{in}'}/e$ are incident onto the hatched region from the left and the right, with each flow being reflected with probability $\mathcal{R} = |r|^2$ and transmitted with probability $\mathcal{T} = |t|^2$. From the symmetry of the quantum mechanical relations it follows that

$$\mathcal{R} = \frac{j_r}{j_{\text{in}}} = \frac{j_{r'}}{j_{\text{in}'}}, \quad \mathcal{T} = \frac{j_t}{j_{\text{in}}} = \frac{j_{t'}}{j_{\text{in}'}}, \quad \mathcal{R} + \mathcal{T} = 1. \tag{5.15}$$

If the voltage drop across the hatched region is zero, then the total electron current in the wire is also zero. The potential difference at the region boundaries, δV, gives rise to the difference of densities at these boundaries, $\delta n = ge\delta V$. In one-dimensional systems, all the electrons move along the conductor and, therefore, they all belong to one of the flows entering eqn (5.15). This allows one to relate δn to the electron densities in the flows and express δV in terms of

currents as

$$\delta V = \frac{\delta n}{ge} = \frac{j_{\text{in}} + j_r + j_{t'}}{e^2 gv} - \frac{j_{\text{in}'} + j_{r'} + j_t}{e^2 gv} = \frac{2\mathcal{R}(j_{\text{in}} - j_{\text{in}'})}{e^2 gv}. \tag{5.16}$$

Here g and v are the density of states and the magnitude of the electron velocity at the Fermi level. Since the total current J equals

$$J = j_{\text{in}} - j_r - j_{t'} = j_{\text{in}'} - j_{r'} - j_t = \mathcal{T}(j_{\text{in}} - j_{\text{in}'}), \tag{5.17}$$

the ratio $J/\delta V$ allows one to write the conductance $y_{\text{imp}} = J/\delta V$ and resistance $\varrho_{\text{imp}} = y_{\text{imp}}^{-1}$ of the hatched region as

$$y_{\text{imp}} = \left(\frac{e^2}{2\pi\hbar}\right)\left(\frac{\mathcal{T}}{\mathcal{R}}\right) = \left(\frac{e^2}{2\pi\hbar}\right)\frac{\mathcal{T}}{1 - \mathcal{T}}, \quad \varrho_{\text{imp}} = \left(\frac{2\pi\hbar}{e^2}\right)\left(\frac{\mathcal{R}}{\mathcal{T}}\right) = \left(\frac{2\pi\hbar}{e^2}\right)\frac{\mathcal{R}}{1 - \mathcal{R}} \tag{5.18}$$

Landauer suggested representing elastic scattering centers in the form of potential barriers. Then the transport characteristics of the system are expressed in terms of the reflection and transmission coefficients of the wave propagating across these barriers. Therefore, the corresponding formulas and, in particular, eqn (5.18) for conductance are called the Landauer formulas. In principle, the method suggested by Landauer is applicable to systems of any dimensionality, but it is especially convenient for one-dimensional systems.

Note: The Landauer formula in the form (5.18) was derived under the assumption that the potential difference is applied directly to the scattering region between the points A and B (Fig. 5.4). Therefore, the conductance (5.18) for weak scattering ($\mathcal{T} \sim 1$, $\mathcal{R} \ll 1$) may be higher than the conductance (5.13) of the system having no scatterers at all.

If the potential difference in the system shown in Fig. 5.4 is applied to the reservoirs, the resistances of both an ideal wire and the scattering region are switched in series and the system conductance is

$$y^{-1} = y_{\text{id}}^{-1} + y_{\text{imp}}^{-1} \equiv \varrho_{\text{id}} + \varrho_{\text{imp}} = \left(\frac{2\pi\hbar}{e^2}\right)\left(1 + \frac{\mathcal{R}}{\mathcal{T}}\right), \quad y = \left(\frac{e^2}{2\pi\hbar}\right)\mathcal{T}. \tag{5.19}$$

Now, if $\mathcal{T} \to 1$, the conductance $y \to y_{\text{id}}$, as was to be expected. Equation (5.19) for y may also be obtained directly by applying a potential difference to the reservoirs and writing the electron flow from one reservoir to another with due regard for single scattering (cf. the derivation of eqn (5.13)). This signifies that the summation of resistances in eqn (5.19) in accordance with Ohm's law is quite justified. As is shown below, it is not always justified in one-dimensional systems because of the interference of the incident and reflected waves.

5.3. Localization and role of correlations in 1D systems

Consider two successive barriers in a one-channel one-dimensional conductor (Fig. 5.5) and write the parameters $\mathcal{T}$ and $\mathcal{R} = 1 - \mathcal{T}$ of a compound scattering object in terms of the parameters $\mathcal{T}_1$, $\mathcal{R}_1$, $\mathcal{T}_2$ and $\mathcal{R}_2$ of the prime defects. If a
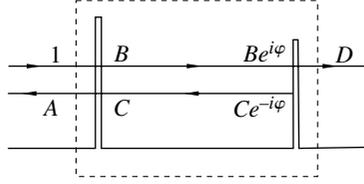
FIG. 5.5. The scattering portion of a one-dimensional conductor consisting of two barriers. The complex amplitudes $A, \ldots, D$ of the waves that arrive at and depart from the barriers are normalized to the amplitude of the initial arriving wave 1.

wave with amplitude 1 is incident onto barrier 1 from the left, then the stationary wave field thus formed contains four more waves: the reflected wave A, the transmitted wave D, and two waves B and C propagating between the barriers in opposite directions (where A, B, C, and D are the complex wave amplitudes). Expressing the amplitudes of the waves propagating to the right and the left from each barrier in terms of the amplitudes of the incident waves, we arrive at four equations

$$A = r_1 + C t_1, \quad B = t_1 + C r_1, \quad C e^{-i\varphi} = B e^{i\varphi} r_2, \quad D = B e^{i\varphi} t_2. \qquad (5.20)$$

Above, we used the fact that the reflection coefficients are the same for the waves incident on the right and on the left, $r_1 = r_1'$. The factors $\exp(\pm i\varphi)$ take into account the phase incursion of the wave within the interbarrier distance. It follows from eqn (5.20) that

$$D = \frac{e^{i\varphi} t_1 t_2}{1 - e^{2i\varphi} r_1 r_2}, \quad \mathcal{T} = |D|^2 = \frac{\mathcal{T}_1 \mathcal{T}_2}{1 + \mathcal{R}_1 \mathcal{R}_2 - 2\sqrt{\mathcal{R}_1 \mathcal{R}_2} \cos \vartheta}, \qquad (5.21)$$

where $\vartheta = 2\varphi + \arg(r_1 r_2)$. The conductance Y_2 of the compound "two-barrier" scatterer shown by the dashed line in Fig. 5.5 is

$$Y_2 = \frac{e^2}{2\pi\hbar} \frac{\mathcal{T}}{1 - \mathcal{T}} = \frac{e^2}{2\pi\hbar} \frac{\mathcal{T}_1 \mathcal{T}_2}{\mathcal{R}_1 + \mathcal{R}_2 - 2\sqrt{\mathcal{R}_1 \mathcal{R}_2} \cos \vartheta}. \qquad (5.22)$$

If the barriers in the compound scatterer are equal, $r_1 = r_2 = r', t_1 = t_2 = t'$, $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}'$, etc., then

$$Y_2 = \frac{e^2}{2\pi\hbar} \frac{(\mathcal{T}')^2}{4\mathcal{R}' \sin^2 \vartheta/2}, \quad \vartheta/2 = \varphi + \arg(r') = kl + \arg(r'), \qquad (5.23)$$

where k is the wave vector, and l is the distance between the barriers.

The conductance (5.22) depends not only on the parameters of the two prime barriers but, via the angle ϑ, also on the distance between these barriers. Since we are interested in a 1D conductor with a large number of randomly located

barriers, we may average all the possible distances between the barriers by assuming that the angle ϑ may take any value from 0 to 2π with equal probabilities. This assumption is not quite justified, but it allows one to follow the tendencies arising for long chains of one-dimensional barriers; for details, see Imry (1997) and the original publication by Anderson et al. (1980). Using the average value $\overline{\cos\vartheta} = 0$, we obtain the average conductance $\overline{Y_2}$ of a system of two barriers as

$$\overline{Y_2} = \frac{e^2}{2\pi\hbar} \frac{\mathcal{T}_1\mathcal{T}_2}{\mathcal{R}_1 + \mathcal{R}_2} = \frac{e^2}{2\pi\hbar} \frac{(1 - \mathcal{R}_1)(1 - \mathcal{R}_2)}{\mathcal{R}_1 + \mathcal{R}_2}. \quad (5.24)$$

For comparison, we also write the classical expression for the sum of the resistances $\varrho_1 = y_1^{-1}$ and $\varrho_2 = y_2^{-1}$ connected in series:

$$Y_2^{(\text{cl})} = (y_1^{-1} + y_2^{-1})^{-1} \equiv (\varrho_1 + \varrho_2)^{-1} = \frac{e^2}{2\pi\hbar} \left(\frac{\mathcal{R}_1}{1 - \mathcal{R}_1} + \frac{\mathcal{R}_2}{1 - \mathcal{R}_2} \right)^{-1}$$

$$= \frac{e^2}{2\pi\hbar} \frac{(1 - \mathcal{R}_1)(1 - \mathcal{R}_2)}{\mathcal{R}_1 + \mathcal{R}_2 - 2\mathcal{R}_1\mathcal{R}_2} \quad (5.25)$$

It is seen that, comparing with eqn (5.24), eqn (5.25) has an additional term in the denominator which is proportional to the product of the transmission coefficients of two barriers, $\mathcal{R}_1\mathcal{R}_2$.

Consider a long chain of equivalent weakly scattering barriers $\mathcal{R}' \ll 1$, $\mathcal{T}' \sim 1$, located at random distances l_i. The average distance between the barriers $l = \overline{l_i}$ has the sense of an elastic mean free path. Each barrier introduces the same low resistance $\varrho' = (2\pi\hbar/e^2)(\mathcal{R}'/\mathcal{T}') \ll 2\pi\hbar/e^2$. Let us calculate the resistance $R_N = Y_N^{-1} = (2\pi\hbar/e^2)(\mathcal{R}_N/\mathcal{T}_N)$ of a compound scattering object consisting of N barriers by the recurrence formula following from eqn (5.24),

$$\frac{\mathcal{R}_N}{\mathcal{T}_N} = \frac{\mathcal{R}_{N-1} + \mathcal{R}}{\mathcal{T}_{N-1}\mathcal{T}'}. \quad (5.26)$$

If the number of barriers N is small so that $N\varrho' \ll 2\pi\hbar/e^2$, the R_N value increases linearly, $R_N \approx N\varrho \propto N$. The probability of reflection, $\mathcal{R}_N$, at first also increases linearly. However, as it cannot exceed unity, its growth is limited. Therefore, beginning with a certain N value, it is possible to assume that $\mathcal{R}_N \approx \mathcal{R}_{N-1} \approx 1$ in eqn (5.26) and, therefore, to write

$$\mathcal{T}_N \approx \mathcal{T}_{N-1}\mathcal{T}', \quad \mathcal{T}_N \to s(\mathcal{T}')^N = se^{aN} \quad \text{as } N \to \infty (s = \text{const}, a = \ln\mathcal{T}' < 0). \quad (5.27)$$

An exponential decrease in the intensity of the transmitted wave $\mathcal{T}_N$ with an increase in the length-scale N is a manifestation of the 1D localization.

Consider one specific feature of the transport in one-dimensional systems. Figure 5.6 shows the transport characteristics of a quasi-one-dimensional system consisting of an accumulating layer of the field transistor on an n-Si surface (Fowler et al. 1982). At low temperatures, the dependence of the conductance y on gate voltage V_{g} acquires a noise-like component with a very large amplitude. It is not real noise. The signal does not depend on time. If the sample is not heated
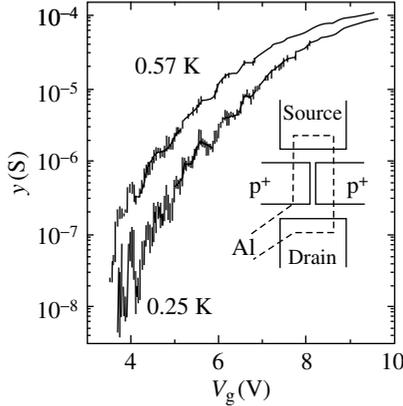
FIG. 5.6.   Conductance y of a long quasi-one-dimensional channel in a field
transistor prepared on an n-Si surface and operating in the mode of an accu-
mulating layer as a function of the gate voltage V_g (Fowler et al. 1982). The
channel width may vary from zero to the maximum value $\sim 1\,\mu$ set by the
design (see the schematic in the inset) by applying voltages to the control
electrodes p$^+$ and to the gate.

to room temperature, the $y(V_{\mathrm{g}})$ curve is reproduced in repeated experiments up
to the smallest details. It is seen that at low temperatures and gate voltages V_{g},
which provide a narrow channel and a low carrier concentration, the conductance
shows random oscillations with the variation of V_{g}, whose amplitude increases
with lowering of the temperature. Another sample shows similar general features
of the signal variation with voltage V_{g} and temperature, but the details of the
oscillation structure become quite different. The same is true for the sample
cooled for the second time from room temperature

The random oscillations are explained mainly by their one-dimensional
character. All the wire defects are "switched in series" and the current lines
cannot avoid any defect. The elimination of only one strongly scattering defect
may dramatically influence the total resistance. We have to explain how the
variation of V_{g} (which changes the concentration of the carriers and their Fermi
energy ε_{F}) may switch on, switch off, or change the efficiency of individual
defects.

Above, we averaged eqn (5.22) over $\cos\vartheta$ based on the fact that interbarrier
distances l_i have a certain dispersion. However, the angle $\varphi = k l_i$, which enters ϑ,
depends not only on l_i but also on the wave vector k, i.e., on the electron energy
ε_{F}. It is seen from eqn (5.23) that $R_2 = Y_2^{-1}$ of one pair of scattering barriers
with resistances ϱ each and fixed distance l between them varies from zero to 4ϱ:

$$0 \leqslant R_2 \leqslant 4\varrho, \tag{5.28}$$

depending on energy of the incident electron.

It is timely to recall that the transport properties of the 1D system are determined by the electrons from the vicinity of ε_F; at lower energies, there are always two electron flows that compensate one another. Also, we have mentioned above the difficulties in averaging eqn (5.22) for resistance over the angle ϑ; they are associated just with the large range (5.28) of the resistance variation R_2 of a compound scatterer.

The space between two barriers is a potential well. Generally speaking, this well has a set of levels ε_i whose width depends on the transparency of the barriers t_1 and t_2. With the displacement of electron energy ε_F with respect to the levels in the well, the probability of tunneling starts oscillating and attains a maximum under the resonance conditions $\varepsilon_F = \varepsilon_i$. Therefore, enormous random oscillations of resistance may be theoretically described in terms of resonant tunneling.

The model of localized states in 1D systems uses ideas about electron levels inside compound scatterers. At sufficiently low temperatures, reflections from far barriers

$$1 \ll N \ll L_\varphi / l \qquad (5.29)$$

remain coherent. Therefore, according to eqn (5.27), these reflections at sufficiently high L_φ values compensate the transparency of the barriers t_1 and t_2 and make the state really localized between these barriers. Under these conditions, one can expect the manifestation of the hopping conductivity. And, indeed, Fig. 5.7 shows the temperature dependence of conductance measured at several minima of the curve shown in Fig. 5.6. In the left-hand side of Fig. 5.6 (at low V_g and low conductance) the oscillations are quite pronounced, and there are
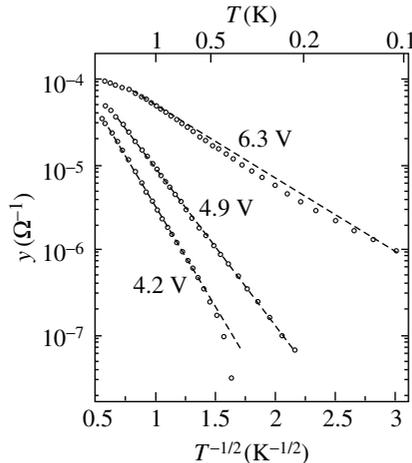
FIG. 5.7.   Logarithm of conductance as a function of $T^{-1/2}$ of a field-transistor channel at three gate voltages which correspond to different conductance minima (Fowler et al. 1982).

grounds to believe that the channel would be one-dimensional. The changes of this part of the $y(V_\mathrm{g})$ curve with temperature fit the functional dependence

$$y = y_0 \exp[-(T_\mathrm{M}/T)^{1/2}], \qquad (5.30)$$

in full accordance with the Mott law, eqn (4.24). At high V_g values, the channel is broadened and gradually transformed into a two-dimensional system. The conductance increases, whereas the amplitude of random oscillations decreases. The experimental points of the dependence $\log y(T^{-1/2})$ measured at the gate voltage $V_\mathrm{g} = 6.3\,\mathrm{V}$ deviate from the straight line in Fig. 5.7, but are linearized in the $(\log y, T^{-1/3})$ axes, also in full accordance with eqn (4.24).

Let us revisit the general statement that a 1D random potential unavoidably leads to localization. Equation (5.22), which is the starting point in constructing the concrete model (5.27) aimed to demonstrate localization, contains a hint that correlations can help to avoid localization. To show this, consider a simplified version of eqn (5.22), namely eqn (5.23) for the conductance Y_2 of a symmetrical compound scatterer. It follows from eqn (5.23) that for the wave with a special wave vector $k_0 = -\arg(r')/l$ the compound scatterer is completely transparent and there is no reflected wave, $\mathcal{R}'_2 = 0$. If all the single barriers in the model (5.26)–(5.27) are substituted for identical compound barriers (5.23), then the electron with energy $\varepsilon_0 = \hbar^2 k_0^2/2m$ turns out to be delocalized, in spite of the randomness in the disposition of the scatterers.

This idea is elaborated in the so-called dimer model (Dunlap et al. 1990). The model uses a 1D periodic lattice of potential wells instead of a random set of barriers. The lattice consists of two types of wells with energy levels E_a and E_b. These types of wells are randomly distributed over the odd lattice sites without any correlations and each even site has a well of the same kind as the odd one on the left. This signifies that the wells of the same type form pairs, whence the name of the dimer model (Fig. 5.8a). If the distance between the wells is a, the lattice thus formed may be represented as a sum of two random but equivalent
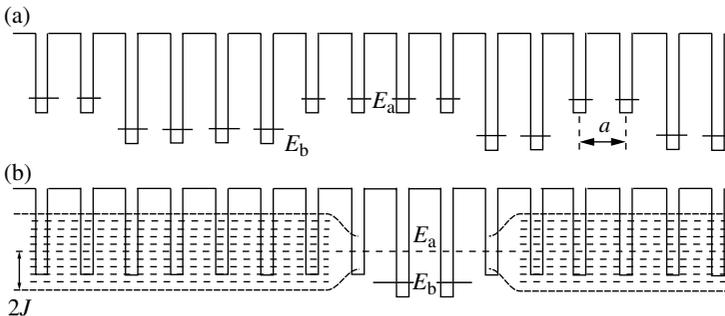
FIG. 5.8.   (a) The dimer model of a one-dimensional random potential. The changes in the level positions due to the overlap of wave functions are not shown. (b) Electron levels in a one-dimensional lattice with one dimer defect.

sublattices displaced relative to one another by a distance a. The period of both sublattices equals $2a$.

Assume that the pairs E_a belong to the basic lattice and pairs E_b are defects. We have already seen that the model of paired defects contains delocalized states at some separate energies determined by the structure of the defect. The condition that a free electron of the basic lattice with such energy can pass freely through the defects may readily be understood from Fig. 5.8b, where one dimer-type defect is immersed into the basic lattice formed by E_a wells. Let the transfer integral for two neighboring wells be J. Then, both on the right and on the left of the defect, bands with a quasi-continuous level distribution are formed, $\varepsilon = E_a - 2J\cos ka$. If the relation

$$|E_a - E_b| < 2J, \tag{5.31}$$

is fulfilled, then the level E_b turns out to be inside the bands and the reflection probability becomes zero, $\mathcal{R}' = 0$, for the wave vector k_0 defined from the relation $\cos k_0 = -2J/(E_a - E_b)$.

In the dimer model, the correlations were introduced only between the nearest neighbors. With such correlations, the delocalized states arose only at discrete energy values. To obtain the band of delocalized states, it is necessary to consider long-range correlations in the potential which should nevertheless remain random. The routine for construction of such potential was proposed by Izrailev and Krokhin (1999). In the next section, we shall give an example of this potential which was written to be tested by microwave modelling.

5.4. Microwave modelling

A comprehensive discussion of these problems is given by Stöckmann (1999).

Both the time-dependent Schrödinger equation,

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\triangle\Psi + U\Psi, \tag{5.32}$$

and the classical wave equation,

$$\frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2} = \triangle\Psi - U\Psi \tag{5.33}$$

(c is the light velocity), after the substitution $\Psi = e^{-i\omega t}\psi$, are reduced to the same equation

$$(\triangle - U + k^2)\psi = 0 \tag{5.34}$$

with the only difference that for the Schrödinger equation we have

$$\omega = (\hbar/2m)k^2, \tag{5.35}$$
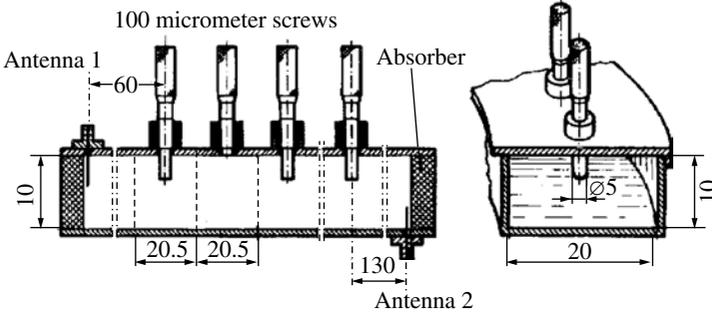
FIG. 5.9.   Sketch of a one-mode waveguide with an array of 100 scatterers used to measure the transmission coefficient t of an electromagnetic wave as a function of frequency (Kuhl et al. 2000). All dimensions are given in millimeters.

whereas for the wave equation, the relation between ω and k is linear:

$$\omega = ck. \tag{5.36}$$

This makes it possible to study localization processes by considering the distribution of electromagnetic fields in microwave devices. Two examples are presented below, one with a 1D system, the other with a 2D system.

In the paper by Kuhl et al. (2000), the 1D random potential was imitated by a long waveguide with scatterers inside and the transmission coefficient of an electromagnetic wave was measured as a function of frequency. A sketch of this waveguide is depicted in Fig. 5.9. The working range of frequencies ν was selected in the frequency range of waveguide operation in the single-mode regime: $7.5\,\mathrm{GHz} = c/2a < \nu < c/a = 15\,\mathrm{GHz}$, where a is the maximum size of the transverse cross-section of the waveguide.

Then, $N = 100$ screw scatterers (modelling a random potential) were screwed into the waveguide at equal distances from one another. The screw lengths u_n (where $1 \leq n \leq N$) in the waveguide can be varied thus changing the effectiveness of the scatterers. The screw lengths were determined from the expression

$$u_n = \sqrt{\overline{u_n^2}} \sum_{m=-\infty}^{\infty} \beta_m Z_{n+m}, \quad \beta_m = \frac{2}{\pi} \int_0^{\pi/2} \sqrt{\varphi(\mu)} \cos(2m\mu)\, d\mu. \tag{5.37}$$

Here, Z_{n+m} are random numbers ranging from -1 to $+1$. It is these numbers that introduce randomness into the potential. The correlation between all the u_n depths is provided by the factor β_m defined in terms of the function $\varphi(\mu)$. The latter function is selected with the invocation of a specified mathematical algorithm for the necessary transmission spectrum of a one-dimensional system. The results obtained with the use of this algorithm are shown in Fig. 5.10. The $\varphi(\mu)$ function was selected in such a way that two transmission bands were formed inside the working range. The transmission coefficient of the waveguide
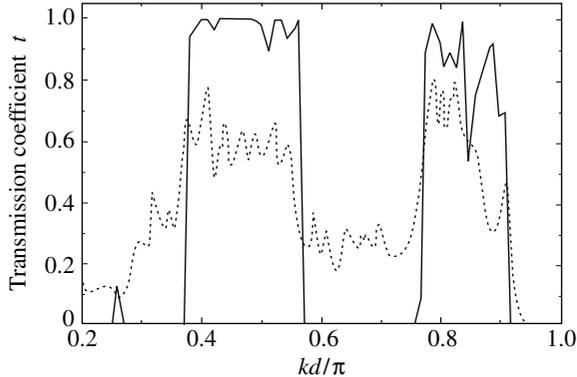
FIG. 5.10. Transmission t through a one-dimensional channel with correlated
  periodically arranged random scatterers; the correlation is governed by a spe-
  cially selected correlation function. The solid line relates to the numerical
  experiment with $N = 10^4$ scatterers; the dashed line relates to microwave
  transmission through an array of $N = 100$ scatterers averaged over five inde-
  pendent measurements. The correlation function is the same in both cases
  (Kuhl et al. 2000).

with the one-dimensional sequence of $N = 10^4$ scatterers calculated in the com-
puter simulation is shown by the solid line in Fig. 5.10. The result of the real
microwave experiment with the sequence of $N = 100$ scatterers averaged over the
five measurements with different sets of Z_n numbers is shown by the dashed line.

In solid state physics, the existence of a transmission band with a finite width
signifies the existence of a mobility edge and, at a selected value of a control
parameter, also of a metal–insulator transition. Thus, one-dimensional models,
which allow one to achieve considerable progress in the solution of the problem
not only numerically but also analytically, demonstrate the important role of
correlations in a random potential in the localization problems.

The microwave method may also be used for simulating two-dimensional loc-
alization. The corresponding experiments are called *microwave billiards*. Their
main elements are plane resonators with feeding antennas. The transverse dimen-
sion of such a resonator along the z-axis is of the order of the wavelength or less,
whereas the longitudinal dimension along the x- and y-axes are much larger
than the wavelength. A random potential U is modelled by metal scatterers
randomly located inside the resonator. The stationary distribution of an elec-
tromagnetic field inside the resonator, e.g., of an electric field $E_z(x, y)$, at the
corresponding wave mode satisfies not only the Maxwell equations, but also the
Schrödinger eqns (5.32) and (5.34) for the eigenfunction ψ. This distribution may
be measured, e.g., by moving a small metal sphere inside the resonator. The per-
turbation introduced by this sphere results in a change $\Delta\omega$ of the resonator
frequency proportional to the squared electrostatic field at the given point,

<div align="center">1.73 GHz   2.43 GHz   4.89 GHz</div>
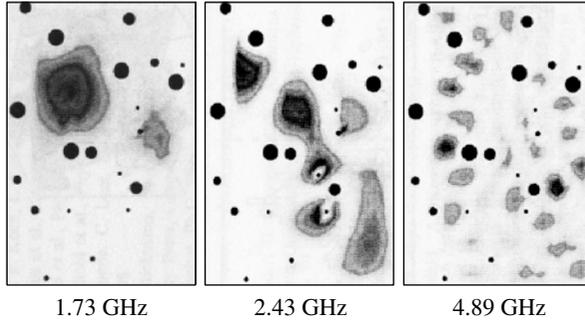
FIG. 5.11. Eigenfunctions obtained in a $240 \times 340\,\text{mm}^2$ rectangular billiard with randomly distributed scatterers (black circles) at different frequencies (Stöckmann et al. 2001). The values of $|\psi|^2$ are given on a gray scale.

$\Delta\omega \propto E_z^2(x,y)$, and, therefore, also the squared eigenfunction of an electron at this point, $\Delta\omega \propto \psi^2(x,y)$, at the energy value $\hbar\omega$.

As an example, Fig. 5.11 borrowed from Stöckmann et al. (2001), shows three eigenfunctions in a rectangular billiard with random scatterers obtained at three different frequencies. The values of the squared amplitude $\psi^2(x,y)$ are converted into a gray scale. This example illustrates delocalization of an electron with an increase in its energy.

5.5. Model of structural disorder

Named after I.M. Lifshitz, this model was applied to the metal–insulator transitions in the book by Shklovskii and Efros (1984). The mathematical aspects of the model were discussed in detail by Lifshitz in his review (Lifshitz 1965), where this model was used in the study of the energy spectrum of a metal with impurities.

It is expedient to consider a one-electron model of disorder different from the Anderson model. In this model a random potential $V(\boldsymbol{r})$ consists of *equivalent* randomly arranged wells $v(\boldsymbol{r})$ with energy level E_0 in each of them:

$$V(\boldsymbol{r}) = \sum_{\boldsymbol{R}_i} v(\boldsymbol{r} - \boldsymbol{R}_i). \tag{5.38}$$

Disorder in this model is determined by the randomness of the set of vectors $\boldsymbol{R}_i$.

Note: The Anderson model described by eqn (5.1) and the model of structural disorder described by eqn (5.38) deal with qualitatively different types of disorder. One may state that the Anderson model originates from the physics of semiconductors, where the energies of the impurity centers are always different, whereas the model of structural disorder originates from the physics of metals, where, because of pronounced screening, the energies of all the centers are considered to be equal. It was no accident that the potential (5.38) was used in the diffraction theory of electron transport in liquid metals (see Chapter. 1). All the electrons were assumed there to be delocalized and the

potential (5.38) was simply a source of scattering. This corresponded to a large ratio of the attenuation length a_B to the average distance between the wells, $n^{-1/3}$, i.e. to the inequality $a_B n^{1/3} \gg 1$.

As earlier, we assume that beyond the well, the wave function attenuates as

$$\psi \propto \frac{e^{-r/a_B}}{r}, \quad a_B = \frac{\kappa \hbar^2}{m^* e^2}. \tag{5.39}$$

Let the concentration n be low so that the attenuation length is much less than the average distance between the centers, $a_B n^{1/3} \ll 1$. Nevertheless, the wave levels of the neighbors slightly overlap, therefore, the δ-like density of states is spread into a band. Now, divide all the wells into pairs of nearest neighbors. If the distance between the wells in a pair is r_{12}, then, since the wells are resonant due to the overlap of the wave-function tails, eqn (5.39), the level E_0 splits into two:

$$E = E_0 \pm \varepsilon_{1,2}, \quad \varepsilon_{1,2} = J_0 \frac{\exp(-r_{12}/a_B)}{r_{12}\, n^{1/3}} \tag{5.40}$$

and the collectivized wave functions arise described by eqn (5.5). The distance r_{12} in the expression for level splitting (5.40) is normalized to the average distance $n^{-1/3}$ between the centers in order to ensure that J_0 would have dimensions of energy. The J_0 value as well as the a_B value depends on the characteristics of the well, the permittivity κ of the material, the effective mass of the electron, etc.

As is seen from Fig. 5.12, where the resonance pairs are shown by dotted ellipses, not all the centers may be involved in the formation of resonant pairs. Thus, well 2, which is the nearest neighbor of well 3, may have well 1 as the nearest neighbor, so that $r_{12} < r_{23}$. Of these three wells, the resonant shifts are ε_1 and ε_2, whereas the shift ε_3 is nonresonant and is much less than the former two, because $\varepsilon_3 \propto \exp(-2r_{23}/a_B)$. Figure 5.12 shows two such configurations, consisting of three and four wells with $r_{12} < r_{23} < r_{34}$ in the latter configuration.
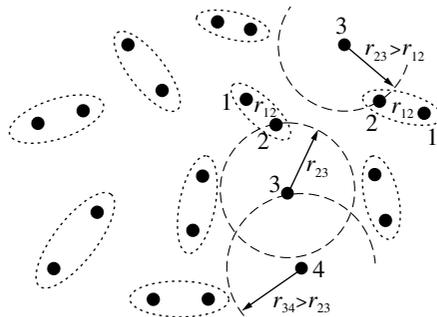
FIG. 5.12.   Random impurity distribution. Pairs of nearest neighbors are shown by dotted ellipses. The centers indicated by the number 3 are adjacent to the pairs, whose wells are indicated by 1 and 2. These centers form triads. Groups with larger number of centers are also possible.

In configurations formed by three and more wells, there always exists at least one resonant pair with the minimum distance between the wells and the maximum shift of the levels. The characteristic width Δ of the resulting function of the density of states is determined by the resonant pairs of wells, with the average distance between the wells being $n^{-1/3}$, so that

$$\Delta \approx J_0 \exp\left(\frac{-n^{-1/3}}{a_\mathrm{B}}\right). \tag{5.41}$$

The tail of the density-of-states function at $|\varepsilon| \gg \Delta$ is formed due to the anomalously close pairs of wells with $r_{12} \ll n^{-1/3}$, whereas the states at low ε, for which $|\varepsilon| \ll \Delta$, are formed due to nonresonant and lone wells.

It is expedient to compare the above transformation of the initially δ-like density of states into the band and the formation of an impurity band in the "donors + acceptors" system (Chapter 3). In Chapter 3, spreading was caused by random electric fields of charged centers. Such a level transformation into a band may be called classical. In the classical broadened impurity band, the metal–insulator transition takes place because of the overlap of the wave functions with a further increase in the concentration of impurity centers. On the contrary, eqn (5.41) is of a quantum nature. The parameter $a_\mathrm{B} n^{1/3}$ controls both the level transformation into a band and the metal–insulator transition.

As is seen from eqn (5.41), the model of structural disorder has no independent parameter equivalent to the bandwidth W in the Anderson model described by eqns (5.1)–(5.3). This fact seems to explain the widespread use of model (5.1)–(5.3) for the description of metal–insulator transitions. In Fig. 5.1, no random classical fields (always present in real systems) are present in explicit form – instead, they seem to be included into the dispersion of well depths. However, it is also possible to represent these field in explicit form, by considering not Fig. 5.1 but another equivalent sketch – Fig. 5.13, which reflects the real situation much better, e.g., the situation in a partly compensated semiconductor, where all the impurity centers are chemically equivalent and long-range electric fields are present.

FIG. 5.13. Alternative representation of the Anderson model: periodically arranged wells have the same depths and are located against the background of a random long-range potential (cf. Fig. 5.1).

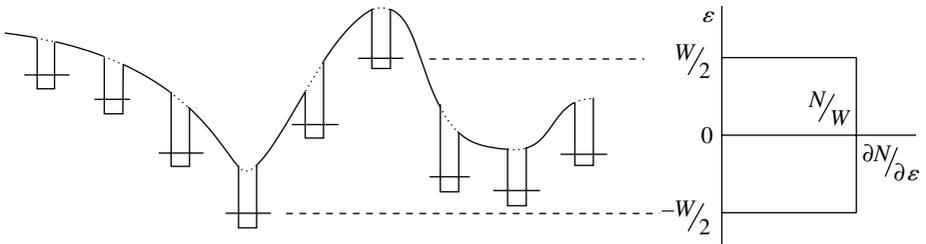The metal–insulator transition in the model of structural disorder has been studied theoretically in less detail than in the Anderson model. However, the computer simulation demonstrated the existence of the transition with respect to the parameter $a_B n^{1/3}$. In addition to the natural control parameters (such as the shape and depth of the well $v(\boldsymbol{r} - \boldsymbol{R}_i)$ and the well concentration n), the model of structural disorder has a latent mechanism of transition control. This control is performed via the correlations on the set of vectors $\boldsymbol{R}_i$. By increasing the correlations, the set of vectors can be transformed from a random into a regular set. This seems to be implemented in some metal alloys and quasicrystals (Chapter 7).

5.6. The Mott transition

Consider a system of donors with electron wave functions (5.39). Compare the Bohr radius a_B with another quantity having the dimensions of length – the screening radius r_e of an electric field used to describe the systems of free carriers. In the Fermi statistics, r_e includes the electron concentration n:

$$r_e = \left(\frac{4me^2 n^{1/3}}{\kappa \hbar^2} \right)^{-1/2} = \tfrac{1}{2}(a_B n^{-1/3})^{1/2}. \tag{5.42}$$

If all the electrons are localized, the system is described by the length a_B; if the electrons are itinerant, it is described by the length r_e. Moreover, there is one more length in both limits – the average distance between the electrons, $n^{-1/3}$. Assume that we gradually increase the donor concentration n. Unless $r_e > a_B$, the screening is inessential, each electron is in the vicinity of its donor, and the material is an insulator. However, if the inequality is changed to $r_e < a_B$, the state of an insulator becomes unstable. If all n electrons abandon their donors, they would not be able to return, because they would not be able to find their donors because of the strong screening. The ionized donors ensure a positive background compensating the negative charge of itinerant electrons. Hence, the equality

$$r_e = a_B, \quad \text{i.e., } \tfrac{1}{2}(a_B n_c^{-1/3})^{1/2} = a_B, \quad a_B n_c^{1/3} = 0.25 \tag{5.43}$$

is the condition for the occurrence of the metal–insulator transition, which occurs when the concentration n attains the critical value n_c. This transition is called the Mott transition.

Note: In this simple speculation, disorder was ignored. The driving force of the Mott transition is the interelectron interaction. It is possible to imagine that ordered donors would form a superlattice; then the concentration n is varied by varying the superlattice period $a = n^{-1/d}$ without introducing any disorder.

To write eqns (5.42) and (5.43), we used eqn (5.39) for the Bohr radius which is valid only for hydrogen-like impurity centers. Another somewhat more general line of reasoning can be proposed. Let the energy of an electron at an impurity center be E_0. The Pauli principle admits the presence of two electrons with
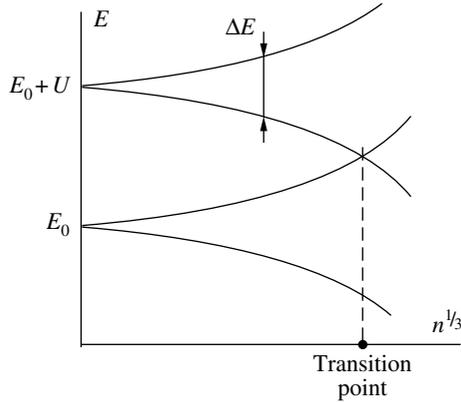
FIG. 5.14.   A Mott transition in the Hubbard scheme.

different spins at this level. However, if the center already has one electron, the other electron may be attached to the center only by it overcoming the electrostatic repulsion. Therefore, the level of the second electron is higher by a value of $U \approx e^2/\kappa a_B$, the so-called Hubbard energy. If the impurity centers have a nonzero concentration, $n \neq 0$, then both levels are spread into minibands, each with nV levels (V is the volume) because of the overlap of the wave functions irrespective of the arrangement of the centers – random or regular. Since the electron concentration is also n, then for nonoverlapping minibands, all the n levels in the lower miniband are filled and all the nV levels in the upper miniband are empty. The material is an insulator with the gap at the Fermi level (Fig. 5.14).

Note: For any level of the lower Hubbard band, only one electron may be localized at a center, but it may have an arbitrarily directed spin. Therefore, the insulator thus obtained is a paramagnetic. Ferro- or antiferromagnetic ordering may arise in this insulator only if additional interactions between the centers take place.

The miniband width ΔE is determined by the transfer integral (5.2),

$$\Delta E \approx 2J \approx 2 \int \psi_i^* \hat{H} \psi_j \ d^3 r.$$

The evaluation of the J integrals is based on eqn (5.39) for asymptotic wave functions. The integral $\int \psi_j^* \hat{H} \psi_j \ d^3 r$ determines an unperturbed energy level and, therefore, is equal to E_0. The replacement of ψ_j^* by ψ_i^* indicates that everywhere, where the integrand has a nonzero value, it is decreased by a factor of $\exp(-r_{ij}/a_B)$. Therefore, we have

$$\Delta E \approx 2J \approx 2 \int \psi_i^* \hat{H} \psi_j \ d^3 r \approx 2E_0 \exp\left(\frac{-r_{ij}}{a_B}\right) \approx 2E_0 \exp\left(\frac{-1}{a_B n^{1/3}}\right). \quad (5.44)$$

An increase in the concentration results in broadening of the bands and their overlap. At the critical concentration n_c the bands merge together when

$$U \approx \Delta E \approx 2E_0 \exp\left(\frac{-1}{a_B n^{1/3}}\right). \tag{5.45}$$

Since $U \simeq E_0$, both quantities are of the order of the atomic energy $e^2/\kappa a_B$, and the numerical value of the above expression can hardly be considerably different from the value 0.25 obtained in eqn (5.43).

Thus, we have considered two types of metal–insulator transitions. Now, compare these types with one another and also with two types of quantum corrections to the conductivity of a dirty metal at low temperatures. The Anderson transitions take place because of disorder and may be well described within the one-electron approximation. The same is true for weak localization. On the contrary, the driving force of the Mott transitions is the Coulomb interactions controlled by screening, whereas the degree of disorder is considerably less important. In the Aronov–Altshuler effect, disorder is necessary solely to provide the diffusion motion of electrons, whereas the quantum correction to conductivity is caused by electron–electron interactions.

For an experimentalist, it is very important to know what type of transition, Mott or Anderson, he deals with. However, in practice, the choice made is not quite convincing. Any change of the electron concentration n is the result of nonstoichiometry or the presence of impurities. Therefore, a change in n is also accompanied by a change in the disorder. On the other hand, an increase in the degree of disorder influences the screening. No matter how strange it may look, the quantitative characteristics of the Mott and Anderson transitions are also similar. Equation (5.3) of the Anderson transition may be written in the form

$$a_B n^{1/3} = -\left(\ln\frac{c_a W}{E_0}\right)^{-1}, \qquad c_a = \left(\frac{J}{W}\right)_{\text{crit}}. \tag{5.46}$$

and eqn (5.45) of the Mott transition in a similar form

$$a_B n^{1/3} = -\left(\ln\frac{c_m U}{E_0}\right)^{-1}, \tag{5.47}$$

where the constant c_m is introduced.

Recall. The formulas for the two types of quantum corrections to conductivity are also similar.

The quantities W, U, and E_0 in the logarithm arguments are of the order of the atomic energy $e^2/\kappa a_B$, whereas the values of c_a and c_m are of the order of unity. As a result, the estimates of the product $a_B n_c^{1/3}$ are practically the same for the transitions of both types (Fig. 5.15). Therefore, the experimental data on the transitions summarized in Fig. 5.16 are not surprising. The proportionality $a_B \propto n_c^{-1/3}$ is confirmed in the range of n_c variation within six orders of magnitude, and the proportionality coefficient determined by the least squares method equals 0.26 (cf. eqn 5.43).

MOTT TRANSITION

Width ΔE is proportional to

$J = E_0 \exp(-1/a_\mathrm{B} n^{1/3})$

$\boxed{\Delta E = U}$

$a_\mathrm{B} n^{1/3} = -(\ln c_m U/E_0)^{-1}$

$\boxed{a_\mathrm{B} n_\mathrm{c}^{1/3} = 0.25}$ *Mott relation*

$a_\mathrm{B} n^{1/3} = -(\ln c_a W/E_0)^{-1}$

Transfer integral
$J = E_0 \exp(-1/a_\mathrm{B} n^{1/3})$
enters parameter W/J

$\boxed{W/J = \mathrm{const}}$

ANDERSON TRANSITION
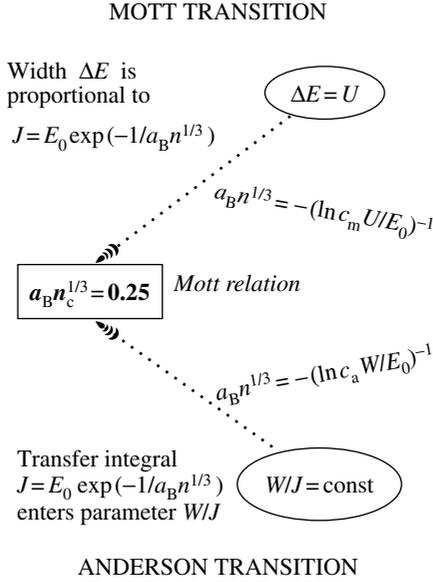
FIG. 5.15. Common and distinguishing features of the expressions for Mott and Anderson transitions.
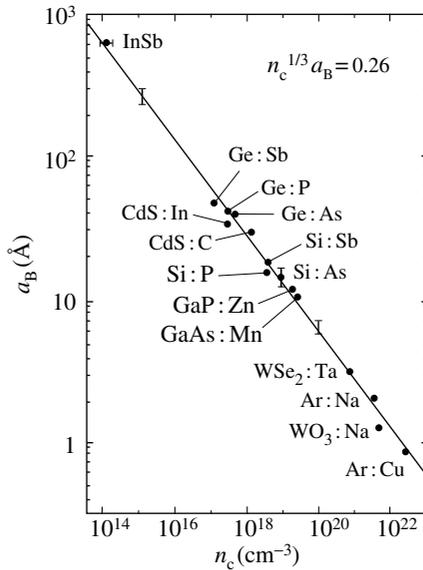

FIG. 5.16. Correlation between the critical carrier concentration and the effective Bohr radius at the points of metal–insulator transitions in 15 materials. The a_B and n_c values of all the materials are determined in independent experiments (Edwards and Sienko 1978).
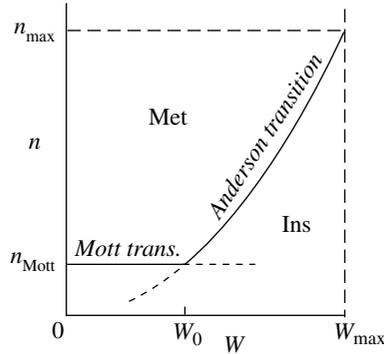
FIG. 5.17.   Concentration vs disorder diagram. Disorder is measured in W units, a_B is assumed to be a constant. The inclined line describes Anderson transitions; the horizontal one describes Mott transitions.

Note: Since the criterion $a_B n^{1/3} = $ const. is valid for both types of transitions, the concentration n may serve as a control parameter not only for the Mott transitions (which is quite natural), but also for the Anderson transitions induced by disorder. In this case, n appears not as the electron concentration but as the concentration of the centers of their localization. The distance between these centers determines the value of the transfer integral.
Note also that the condition for localization of the carriers in the model of structural disorder is the smallness of the same parameter $a_B n^{1/3}$.

Despite the variation of the concentration and the degree of disorder coexist in real transitions, one has to remember the important difference between the Mott and Anderson transitions. This is illustrated by the diagram in Fig. 5.17. The quantity W along the horizontal axis is used as a measure of disorder. The concentration is plotted along the vertical axis. The Bohr radius a_B is assumed to be a constant. The insulator states are on the right of the line of the Anderson transitions described by eqn (5.46); the metal states are on the left of this line. In these coordinates, the line of the Mott transitions described by eqn (5.47) is the horizontal line located the higher, the lower the a_B value. The lines intersect at the point $W_0 = (c_m/c_a)U$. If the disorder is strong, $W > W_0$, the metal–insulator transitions are controlled by the disorder and take place along the line given by eqn (5.46). At concentrations $n < n_{\text{Mott}}$, no metallic state exists irrespective of the level of disorder. The value of n_{Mott} follows from eqn (5.47).

Disorder created by atomic displacements has a certain upper limit, which is attained in the absence of any correlations between the atomic positions. In Fig. 5.17, this limit is indicated as W_{max} on the right-hand edge of the range of possible disorder strength. By the curve of transitions, the disorder W_{max} determines the concentration n_{max}. Then, one may ask what are the electron concentrations n in real metals and alloys. If $n > n_{\text{max}}$, no structural disorder may lead to an Anderson transition.

Note. The quantity W can be considered as a quantitative measure of disorder only conditionally. Therefore, the diagram in Fig. 5.17 is only an illustration. Nevertheless, the question is formulated and the answer may be obtained only by performing the respective experiments. We have already discussed such experiments in Chapter 1 and we shall come back to this problem in Chapter 7.

5.7. The minimum metal conductivity?

The suggested criterion for the transition – the value of the parameter $a_B n_c^{1/3}$ – postulates the existence of a metal–insulator transition, but gives no information on the nature of this transition. The first serious discussions of this problem were based on the analysis of the nature of metal conductivity. The expression for conductivity

$$\sigma = \frac{e^2}{\hbar}\frac{nl}{k_F}, \quad n = (3\pi^2)^{-1}k_F^3, \tag{5.48}$$

proceeds from assuming an electron system to be a gas of charged particles with free path l. Like in a conventional gas, at a certain stage, some constraints appear because of the finite sizes of gas particles. The effective electron size is the de Broglie wavelength $1/k_F$. This is the natural limit from below of the quantity l. For the mean free path $l \sim 1/k_F$, the expression for σ has the value

$$\sigma_{\text{Mott}} \approx (3\pi^2)^{-2/3}(e^2/\hbar)n^{1/3} \approx 0.1(4 \cdot 10^3\Omega)^{-1}n^{1/3}. \tag{5.49}$$

The above value limits from below the conductivity possible within the gas model.

Proceeding from this constraint, Mott assumed that a metal–insulator transition could not be a continuous phase transition. Being applied to an Anderson transition, this assumption is formulated as follows: with an increase in the disorder in the system with a fixed concentration n, the conductivity first drops to a minimum, eqn (5.49), and then should go to zero in a jumping manner. The Mott transition shows an avalanche-like increase in the number of delocalized carriers in the initial scenario: a delocalized electron makes its contribution to screening and, thus, promotes delocalization of the remaining electrons.

Moreover, the main statement forming the basis of the concept of a Mott transition reads that the concentration n cannot be infinitesimal. Substituting n_c from eqns (5.43) or (5.45) into eqn (5.49), one may express σ_{Mott} in terms of a_B. However, a_B cannot have a macroscopically high value, because it is the characteristic size of the wave function of an individual electron. It seems that the value $a_B \approx 600\,\text{Å}$ in InSb (Fig. 5.16) is close to the maximum attainable value, and hence should follow the existence of the absolute minimum value of metal conductivity.

It is natural that the concept of minimum metal conductivity and the nature of a metal–insulator transition were subjected to thorough verification in numerous experiments, which did not confirm the concept and proved that the transition is continuous.
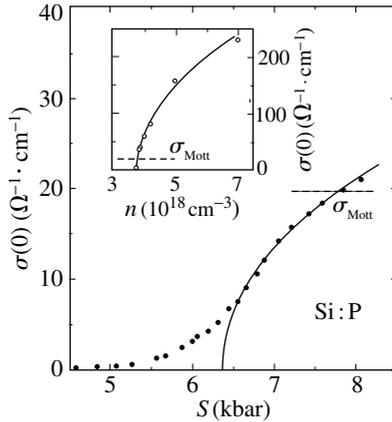
FIG. 5.18. Fine tuning of the metal–insulator transition in Si : P with the aid of pressure S. Dashed lines in the plot and the inset indicate the minimum conductivity σ_{Mott} calculated using the experimental electron concentration (Paalanen et al. 1982).

One of these experiments is illustrated in Fig. 5.18. Since the conductivity of Si : P samples depends on the concentration n (in the inset), it becomes possible to determine the critical concentration n_{c}. However, the insufficiently accurately determined concentrations in each sample do not allow one to determine the law of the conductivity variation in the vicinity of the critical concentration. However, applying a certain pressure to a sample with a concentration slightly lower than n_{c}, it is possible to pass through the interval of σ values from zero to the assumed σ_{Mott} to confirm that there is no jump in conductivity.

Note: Each point on the $\sigma(S)$ plot is obtained by extrapolating the temperature dependence $\sigma(T)$ measured at the given pressure to $T = 0$. A similar plot for any finite temperature would be insufficiently convincing, because an insulator at a finite temperature has a finite conductivity.

References

Anderson, P.W., Thouless, D.J., Abrahams, E., and Fisher, D.S. (1980). Phys. Rev. B **22**, 3519.

Dunlap, D.H., Wu, H-L., and Phillips, P.W. (1990). Phys. Rev. Lett. **65**, 88.

Edwards, P.P. and Sienko, M.J. (1978). Phys. Rev. B **17**, 2575.

Fowler, A.B., Harstein, A., and Webb, R.A. (1982). Phys. Rev. Lett. **48**, 196.

Imry, Y. (1997). *Introduction to Mesoscopic Physics*. Oxford University Press.

Izrailev, F.M. and Krokhin, A.A. (1999). Phys. Rev. Lett. **82**, 4062.

Kramer B. and MacKinnon A. (1993). Rep. Prog. Phys. **56**, 1469.

Kuhl, U., Izrailev, F.M. Krokhin, A.A., and Stöckmann, H.-J. (2000). Appl. Phys. Lett. **77**, 633.

Lee, P.A. and Ramakrishnan, T.V. (1985). Rev. Mod. Phys. **57**, 287.

Lifshitz, I.M. (1965). Sov. Phys. – Uspechi **7**, 549.

Mott, N.V. (1990). *Metal–Insulator Transitions*. Taylor & Francis.

Paalanen, M.A., Rosenbaum, T.F., Thomas, G.A., and Bhatt, R.N. (1982). Phys. Rev. Lett. **48**, 1284.

Shklovskii, B.I. and Efros, A.L. (1984). *Electronic Properties of Doped Semiconductors*. Springer.

Stöckmann, H.-J. (1999). *Quantum Chaos*. Cambridge University Press.

Stöckmann, H.-J., Barth, M., Dörr, Kuhl, U., and Schanze, H. (2001). Physica E **9**, 571.

van Wees, B.J., Kouwenhoven, L.P., et al. (1988). Phys. Rev. B **38**, 3625.

6

SCALING HYPOTHESIS

It would be useful to become acquainted with the problems which are discussed in the previous chapter and in this one set forth in a somewhat different style; see, for instance, the book by Mott (1990) and the review articles of Lee and Ramakrishnan (1985) and Kramer and MacKinnon (1993).

Below, we discuss the phenomenological theory based on the scaling hypothesis. It was created to eliminate the inconsistencies between the simplest ideas on the minimum metallic conductivity and experiments.

6.1. Foundations and formulation of the scaling hypothesis

A clear and accessible discussion of the basic ideas considered in this section exists in the short pioneering article by Abrahams et al. (1979).

A metal–insulator transition is a somewhat unusual phase transition. The basic property, which allows one to distinguish between different states – conductivity – manifest itself only under nonequilibrium conditions. Therefore, one of the reasons to consider this phenomenon as a continuous phase transition is its successful description by the standard mathematical apparatus of the theory of phase transitions. First of all, let us introduce a control parameter x which determines the distance to the transition so that $x - x_\mathrm{c}$ has different signs on opposite sides of the transition. We have learnt in Chapter 5 that this may be the electron concentration n, the strength of the disorder, the magnetic field B, the pressure, etc. However, this cannot be the temperature because the transition takes place at $T = 0$. Then, one has to select from the functions that describe the transport the one which sets the system state, as is usually done in thermodynamics.

Let the samples have the shape of a hypercube with volume L^d in space of dimensionality d. The sample conductance Y has dimensions $[\Omega^{-1}]$ independent of d. We also consider the dimensionless conductance y given by the formula $(e^2/\hbar)y = Y$ in which the combination of the atomic constants $e^2/\hbar$ has dimensions $[\Omega^{-1}]$ and is equal to $2.43 \cdot 10^{-4}\,\Omega^{-1}$. The quantity $e^2/\hbar$ is often called a conductance quantum. Apart from the conductance, the conductivity σ also exists. It has dimensions $[\Omega^{-1}\,\mathrm{cm}^{2-d}]$ dependent on the space dimensionality d. The dimensional Y and dimensionless y conductance are related to the

conductivity by the relationships

$$Y = \sigma L^{d-2}, \quad y = (e^2/\hbar)^{-1}\sigma L^{d-2}. \tag{6.1}$$

At first glance, it seems natural to select the conductivity σ as the basic function of a state. However, it remains the "specific conductivity", i.e., a characteristic of the material, only while the sample size L exceeds a certain internal material parameter with the dimensions of length.

For instance, in the classical physics of metals, the relation between the sample size L and the mean free path l gives rise to the DC size effect: the equation $Y = \sigma L$ relating conductance and conductivity at $d = 3$ is valid only if $L > l$.

One of the major quantities in the theory of phase transitions is the correlation length ξ, which is defined on both sides of the transition. At the transition point, the control parameter x approaches the critical value, $x \to x_c$, and the correlation length ξ diverges, $\xi \to \infty$. The theory of conventional continuous phase transitions is based on the assumption that with the approach to the transition, ξ becomes a characteristic size of fluctuations. A metal–insulator transition is observed at the absolute zero of temperature at which the quantum fluctuations take place instead of thermal fluctuations. In the close vicinity of a transition, the regions with different ground states of the electron system may flicker when the system is close to the transition. They are of typical size ξ and last for a time δt determined by the uncertainty relation $\delta E \, \delta t \sim \hbar$.

Since the correlation length ξ diverges in the vicinity of x_c, the condition $L > \xi$ is necessarily violated, the function σ loses its sense, and the relation between conductance and conductivity is no longer reduced to the factor L^{d-2}. This dictates the choice of the basic function – it is conductance y and not conductivity.

The selection of conductance as the basic function is also favored by important arguments based on the analysis of the wave functions. Following Thouless, we consider the transformation of a discrete electron spectrum at the merging of 2^d hypercubes of volume L^d into one hypercube of volume $(2L)^d$. The wave functions in a large hypercube ψ_{2L} may be considered as linear combinations of the wave functions ψ_L of small hypercubes. In first-order perturbation theory, the coefficients c_i before the corrections to the initial wave functions ψ_L have the form $J/\Delta_i E$, where J is the transfer integral and $\Delta_i E$ is the difference between the energies of the unperturbed states (see, e.g., eqn 5.4). The quantity $\Delta_i E$ is of the order of the level spacing defined by confinement, $\Delta_i E \sim (g_d L^d)^{-1}$, where g_d is the density of states in the space of the corresponding dimensionality. Therefore, we have

$$c_i \simeq \frac{J}{\Delta_i E} \simeq \frac{J}{(g_d L^d)^{-1}} = J(g_d L^d). \tag{6.2}$$

If the transfer integral J is small, the coefficients c_i are also small, and the wave functions are localized in the initial hypercubes L^d. If the integral J has a high value, the wave functions become spread over the whole volume of the

large hypercube $(2L)^d$; whence the assumption that the behavior of the wave functions during duplication of the hypercube size L may be described with the aid of only one parameter J, eqn (6.2).

One can readily see that the transport properties of the volume L^d vary depending on J in a similar way. For high J values, the conductance also has a high value, because the charge may pass from one cube face to the opposite one by the mechanism of metallic conductivity. At low J values, the transport is supported by hopping and tunneling; the lower the J value, the lower the probability of hopping between the centers or tunneling between the faces. Therefore, we may consider conductance as a physically measurable quantity from which the wave functions of electrons in the ground state may be judged. It is conductance and not conductivity, because conductance measured in $e^2/\hbar$ units is dimensionless irrespective of the space dimensionality d and because it retains its sense at an infinitesimal distance from the transition.

Note: If we wish to estimate the density of states in terms of the energy bandwidth W and electron concentration n_d, $g_d \simeq n_d/W$ (cf. eqn 5.1), then, in fact, the parameter given by eqn (6.2) is transformed into the Anderson parameter given by eqn (5.3). This confirms once again the relation existing between the nature of the wave functions and conductance.

The sample conductance y was used as the basic physical function of a state for the first time in the fundamental study by Abrahams et al. (1979) where one-parametric scaling theory of a metal–insulator transition was introduced. Following Abrahams et al. (1979), we depict in Fig. 6.1 the logarithmic derivative of the conductance with respect to sample size L,

$$\beta = \frac{d\ln y}{d\ln L} = \frac{L}{y}\frac{dy}{dL},$$
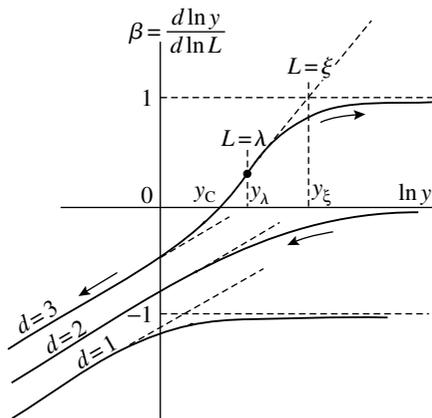

FIG. 6.1. Universal $\beta(\ln y)$ functions for different dimensionality (Abrahams et al. 1979).

as a function of the argument $\ln y$,

$$\frac{d \ln y}{d \ln L} = f(\ln y). \tag{6.3}$$

To clarify the sense of the scaling variable β,

note: If y is the power function of L, i.e., $y \propto L^{\phi(d)}$, then β is the exponent, $\beta = \phi(d)$, and the plot in Fig. 6.1 would become a horizontal line. Equation (6.1) presents just this type of relation.

Any particular sample in the space with dimensionality d is mapped onto the curve $\beta_d(y)$ as a point. Arrows show how the point displaces along the curve with increase of L. According to the scaling hypothesis, the functions $\beta_d(\ln y)$ are universal for any dimensionality d, and the states are determined by only one parameter y. The asymptotic behavior of the universal functions $\beta_d(y)$ at very high and very low values of y may be obtained proceeding from general considerations. At high y values, one may use the macroscopic eqn (6.1) of transport theory, whence

$$\beta_d \to d - 2 \quad \text{as } y \to \infty. \tag{6.4}$$

In the limit of low y values, where all the electrons are localized and their wave functions fade away within distances a_{B}, the final conductance at $T = 0$ is due to the exponential tails of the wave functions at the opposite faces of the hypercube and therefore

$$y = y_0 e^{-L/a_{\mathrm{B}}}, \quad \ln y = \ln y_0 - L/a_{\mathrm{B}}, \quad \beta = \ln y/y_0. \tag{6.5}$$

Connecting the asymptotic lines (6.4) and (6.5) by smooth lines without extrema, we obtain the curves shown in Fig. 6.1. The left-hand parts of all the curves lie in the lower half-plane $\beta < 0$. The disposition of the right-hand part which approaches the asymptotic line $\beta = d - 2$ depends on d. This line is in the upper half-plane $\beta > 0$ if $d > 2$ and it is in the lower half-plane $\beta < 0$ if $d < 2$. The difference is of utmost importance because the line $\beta_d(y)$ with $d > 2$ intersects the abscissa. The existence of the intersection point indicates that the system undergoes a metal–insulator transition considered in detail below. This allows us to state that the two-dimensional case with the asymptotic line $\beta = 0$ is, in fact, the boundary case (see Section 5.1).

The boundary properties of two-dimensional systems stimulated the creation of the so-called "$2 + \varepsilon$ models" in the theory of metal–insulator transitions. In these models, d is considered to be a continuous variable. The expansion in the vicinity of the value $d = 2$ is performed with respect to this continuous variable.

The $\beta_d(y)$ curves are assumed to be the final product of the scaling hypothesis, and we need a prescription for using it. Consider a small cube (or a square film, or a wire) of size L made from the material we are interested in and measure its conductance y at the absolute zero of temperature. This measurement gives us a point on one of the curves in Fig. 6.1. If this point lies in the lower half-plane $\beta < 0$, the material is an insulator. With an increase in L, the conductance y

of the cube becomes exponentially small. If the point obtained lies in the upper half-plane $\beta > 0$, the material is a metal whose conductance y increases with dimension L.

Although the above *gedanken* measurement is unrealistic, the scaling hypothesis has played an important role and allowed one to create a consistent picture from separate facts. However, the hypothesis itself gives rise to the following questions.

  – What are the physical grounds underlying this hypothesis?
  – Are there any theoretical calculations that may confirm this hypothesis?
  – What are the physical consequences of the scaling hypothesis and can they be verified experimentally?

All that was stated above about conductivity, conductance, and asymptotic behavior only partly answer the first question. However, we limit ourselves to this partial answer and refer all those interested in this problem to the original publication (Abrahams et al. 1979) and review article (Lee and Ramakrishnan 1985). However, one circumstance should be emphasized here. *At all the stages of our consideration, e.g., when writing eqn (6.5), we considered the wave function of one electron and ignored the states of all the other electrons and spin interactions. Therefore, strictly speaking, the curves in Fig. 6.1 are related to a system of noninteracting spinless electrons.* As a rule, it is not known *a priori* to what extent one may consider an electron system as a noninteracting one. The general strategy reduces to the following. Comparing the predictions of the scaling hypothesis with the measured data for various electronic systems, one tries to establish the applicability of the hypothesis, prove its universality, and clarify the role of interactions.

Now, consider the second question. The possibilities to compare the scaling hypothesis with direct theoretical calculations are rather limited, but nevertheless exist. The transition from itinerant carriers to completely localized ones, described by the curves $\beta_d(y)$ in Fig. 6.1, begins with weak localization. Therefore, we project onto these curves all the known theoretical calculations of the quantum corrections to conductivity discussed in Chapter 2. At $T = 0$, the inelastic length is $L_\varphi = \infty$. Assuming that an electron which has reached the surface forgets its phase in scattering, we replace L_φ by L in all the formulas of the theory of weak localization. The corrections to $\beta_d(y)$ thus obtained for all three dimensions are listed in Table 6.1, where σ_d is the Boltzmann conductivity in d-dimensional space.

The calculated functions β indicated in the penultimate row of the table are rewritten in the unified form, $\beta_d = (d - 2) - K_d/Y$ in the last row. These functions show the expected asymptotic behavior as $Y \to \infty$ and deviate from it downward at finite Y values, in accordance with the scaling hypothesis (Fig. 6.1). Thus, everywhere where the calculation within the perturbation theory was performed, the result obtained confirms the validity of the scaling hypothesis.

Note: The latter statement is not as indisputable as it may look at first glance. This statement is based on the signs of quantum corrections to the conductivity (2.11) and,

TABLE 6.1.

	$d = 1$	$d = 2$	$d = 3$
σ	$\sigma_1 + K_1 l - K_1 L$	$\sigma_2 - K_2 \ln L/l$	$\sigma_3 - K_3/l + K_3/L$
y	$(\hbar/e^2)(\sigma_1/L - K_1)$	$(\hbar/e^2)(\sigma_2 - K_2 \ln L/l)$	$(\hbar/e^2)(\sigma_3 L + K_3)$
$\beta = \dfrac{d\ln y}{d\ln L} = \dfrac{L}{y}\dfrac{dy}{dL}$	$\dfrac{L}{y}\left(-\dfrac{\hbar}{e^2}\dfrac{\sigma_1}{L^2}\right) = \dfrac{-\sigma_1}{\sigma_1 - K_1 L}$	$\dfrac{1}{y}\dfrac{dy}{d\ln L}$	$\dfrac{\hbar}{e^2}\sigma_3\dfrac{L}{y} = \dfrac{1}{1 + K_3/\sigma_3 L}$
β_d	$-1 - K_1/Y$	$0 - K_2/Y$	$1 - K_3/Y$

The constants K_1, K_2, and K_3 have the dimensions $[\Omega^{-1}]$

in fact, signifies the following. Weak localization is the first step to strong localization proper. However, the quantum correction is rather intricate and, in strong spin–orbit interaction, weak localization changes to antilocalization. Then, how do we make the scaling hypothesis consistent with the spin–orbit interactions? This problem is considered at the end of this chapter after revealing the circumstances under which this consistency may be important.

Now, proceed to the third question – the consequences that follow from the scaling hypothesis.

6.2. Three-dimensional (3D) systems

Consider a small trial cube with edge λ made of some material. If the point that represents the conductance y of this cube on the scaling curve $d = 3$ lies in the lower half-plane $\beta < 0$, then an increase of the cube edge to $L > \lambda$ shifts this point to the left. Increasing L, one may make the conductance infinitely low (*Note:* only at $T = 0$!). Then, the material is an insulator. If the point representing the conductance obeys the inequality $\beta(y_\lambda) > 0$, then an increase in L shifts this point to the right, so that the cube conductance may be made infinitely high. Then, the material is a metal. The intersection of the curve $\beta_3(y)$ with the axis $\beta = 0$ at the point y_c signifies that the system undergoes a metal–insulator transition.

Now, approximate the 3D scaling curve $\beta_3(y)$ in the upper half-plane $\beta > 0$ by a broken line which consists of a segment originating at the point $(\ln y_c, 0)$ with the slope angle $s = (d\beta/d\ln y)_{y_c}$ (it is usually assumed that s is of the order of unity) and the horizontal half-line $\beta = 1$. The conductance value at the salient point is denoted by y_ξ, and the corresponding value of L, by ξ (Fig. 6.1). Because of the specific form of the function β, the segment is described by a differential equation

$$\frac{d\ln y}{d\ln L} = s \ln \frac{y}{y_c}. \tag{6.6}$$

Its general solution has the form

$$\ln \frac{y}{y_c} = \left(\frac{L}{\lambda}\right)^s \ln \frac{y_\lambda}{y_c}. \tag{6.7}$$

Here λ plays the role of the initial conditions set at one point of the segment. At $y = y_\xi$ (salient point), we have

$$\ln\left(\frac{y_\xi}{y_c}\right) = \frac{1}{s}. \qquad (6.8)$$

It follows from eqn (6.7) that then the cube size is equal to

$$\xi = \lambda\left(s\ln\frac{y_\lambda}{y_c}\right)^{-1/s}. \qquad (6.9)$$

With a further increase in the dimension $L > \xi$, the point under consideration moves along the horizontal part of the curve (where conductivity exists) and one can use eqn (6.1) for y. Since ξ is the minimum length at which the notion of conductivity makes physical sense, it is called the correlation length. Finally, substituting eqn (6.8) for y_ξ and $L = \xi$ into eqn (6.1), we obtain

$$\sigma = \left(\frac{e^2}{\hbar}\right)\frac{y_\xi}{\xi} = \left(\frac{e^2}{\hbar}\right)\frac{\ln y_c + 1/s}{\xi} \sim \left(\frac{e^2}{\hbar}\right)\frac{1}{\xi}. \qquad (6.10)$$

The size λ of a trial cube cannot be infinitesimal. A reasonable minimum size is equal to the electron wavelength $\lambda \approx 1/k_F$. If the conductance y_λ of the trial cube lies in the range $y_c < y_\lambda < y_\xi$, then the correlation length is $\xi > l_{\min} \approx k_F^{-1}$ and $\sigma < \sigma_{\text{Mott}}$.

If y_λ is close to y_c, then $(\ln y_\lambda/y_c) \approx (y_\lambda/y_c) - 1$. Assuming that $s = 1$, we obtain from eqn (6.9) the following estimate

$$\xi \simeq \lambda\frac{y_c}{y_\lambda - y_c}. \qquad (6.11)$$

At $y_\lambda \to y_c$, we have $\xi \to \infty$. Then, eqn (6.10) indicates that the conductivity tends to zero, $\sigma \to 0$. Thus, *it follows from the scaling hypothesis that the value of the three-dimensional conductivity of a rather large sample may be infinitesimal, and, thus, the metal–insulator transition is continuous.*

Note: We needed a small trial cube only at the intermediate stage of our consideration. The conclusions about conductivity σ and the nature of the transition are made for macroscopic samples.

The curve $\beta(y)$ is valuable because of its universality, which is important, first and foremost, in the vicinity of a transition. The universality includes the assumption that the curve $\beta_d(y)$ for different systems has the same slope s at the intersection with the $\beta = 0$ axis. For noninteracting electrons, the $2 + \varepsilon$ model yields the slope $s = \varepsilon$. If one assumes that the expansion $d = 2 + \varepsilon$ may be used up to a value $\varepsilon = 1$, then for three-dimensional systems, we have $s = 1$. The s value characterizes the so-called universality class, which depends on the fact whether the interactions change the symmetry properties of the Hamiltonian that describes the system or not. In the last section of this chapter, we consider how the universality may be broken by spin–orbit interactions (see below, Fig. 6.12).

6.2.1. *Critical vicinity of the transition*

The experimental verification of the above statements requires the extrapolation of conductance measured at finite temperatures to $T = 0$. Therefore, first of all, one has to clarify the situation in the vicinity of the transition at $T \neq 0$.

Consider the phase plane (x, T) (Fig. 6.2). Let high x values correspond to metal states, and low ones to insulator states. The phase transition is depicted by an isolated point $(x = x_{\mathrm{c}}, T = 0)$ on the x-axis. Let our sample be always sufficiently large, $L > l, \xi$, so that we may always consider the conductivity σ to exist. Along the abscissa ($T = 0$), the conductivity $\sigma = 0$ at $x \leqslant x_{\mathrm{c}}$ and gradually increases on the right of x_{c}. There is one more peculiar point, x_1, on the right of x_{c} at which

$$\sigma(x_1, T = 0) = \sigma_{\mathrm{Mott}} = \frac{e^2}{\hbar} k_{\mathrm{F}} = \frac{e^2}{\hbar} l^{-1}. \tag{6.12}$$

According to eqn (6.10), in the range from x_{c} to x_1, the conductivity σ of large samples of size $L > l, \xi$ is expressed in terms of the correlation length ξ, whereas for $x > x_1$, the Drude formula is valid:

$$\sigma = \left(\frac{e^2}{\hbar}\right) \times \begin{cases} 0 & x \leqslant x_{\mathrm{c}}, \\ 1/\xi & x_{\mathrm{c}} < x \leqslant x_1, \\ (k_{\mathrm{F}}l)^2/l & x \geqslant x_1. \end{cases} \tag{6.13}$$
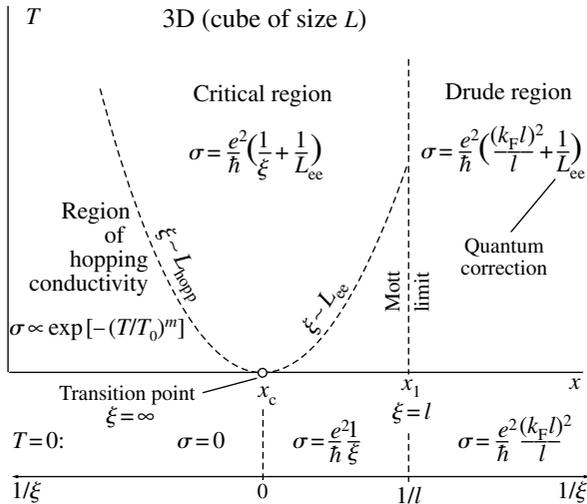
FIG. 6.2. Neighborhood of a metal–insulator transition in a 3D system on the phase plane (x, T), where x is the control parameter. Below the x-axis, the scale of the reciprocal length is shown, which has zero value at the transition point and positive values of the variable $1/\xi$ on both sides of the origin. In the $(1/\xi, T)$ coordinates, both dashed curves on the diagram are cubic parabolas. Between the $(1/\xi)$- and x-axes, the conductivity values $\sigma(T = 0)$ are indicated.

The latter two expressions for conductivity coincide at $x = x_1$, where $\xi = l$ and $k_\mathrm{F} l = 1$.

Let us assume that the transition at x_c is an Anderson–Mott transition. When describing the conductivity at $T = 0$, we used the scaling hypothesis for noninteracting electrons and eqn (6.13). However, when analyzing the vicinity of the transition at finite temperatures, we shall proceed from the quantum corrections due to the interelectron interaction (Section 2.4). Consider the phase plane (x, T) in Fig. 6.2. On the left, deep in the region of an insulator with the Coulomb gap, hopping conductivity takes place, whereas on the right, the Boltzmann expression σ_3 for the classical three-dimensional conductivity with the quantum correction is valid:

$$\sigma = \sigma_3 + \frac{e^2}{\hbar} L_\mathrm{ee}^{-1}, \quad L_\mathrm{ee} = \sqrt{\frac{\hbar D}{T}}. \tag{6.14}$$

The intermediate region in the vicinity of the point x_c is called critical. The interpolation formula for this region has the form

$$\sigma = \frac{e^2}{\hbar}\left(\frac{1}{\xi} + \frac{1}{L_\mathrm{ee}}\right), \quad \sigma > 0. \tag{6.15}$$

Equation (6.15) fits eqn (6.14) along the line $x = x_1$ and yields the appropriate values of conductivity for the segment $x_\mathrm{c} < x \leqslant x_1$ at $T = 0$.

Now, move from right to left along the line $T = \mathrm{const}$. At the far right, the quantum correction is relatively small and electron diffusion proceeds because of impurity scattering, described by the first term of eqn (6.14). Therefore, the diffusion coefficient D that enters L_ee is independent of temperature. However, as soon as we enter the critical region, σ_3 is transformed into $(e^2/\hbar)\xi^{-1}$ and starts rapidly decreasing. Under these conditions, D starts to vary: diffusion seems to proceed at the same fluctuations of the electromagnetic field that determine L_ee and, therefore, becomes temperature-dependent. Then, we may write combined equations for the functions $\sigma(T)$ and $D(T)$, the second equation being the Einstein relation

$$\begin{cases} \sigma = \dfrac{e^2}{\hbar}\left(\dfrac{1}{\xi} + \sqrt{\dfrac{T}{\hbar D}}\right), \\[2mm] \sigma = e^2 g_\mathrm{F} D. \end{cases} \tag{6.16}$$

Here g_F is the density of states at the Fermi level.

Now, exclude D from these two equations and solve the remaining equation with respect to $\sigma(T)$ using the assumption that we are sufficiently close to the transition and that $1/\xi \ll 1/L_\mathrm{ee}$. We obtain the temperature dependence of the conductivity in the right part of the critical region, $x > x_\mathrm{c}$,

$$\sigma(T) = \frac{e^2}{\hbar}\left(\frac{1}{\xi} + (T g_\mathrm{F})^{1/3}\right) \equiv \alpha + \beta T^{1/3}. \tag{6.17}$$

At the transition point, $\xi = \infty$ and $\alpha = 0$.

Note: The power 1/3 arises because the phase-breaking time τ_φ was taken to be equal to the dephasing time $\tau_{\text{ee}} \propto T^{-1}$. For different mechanisms of phase breaking with $\tau_\varphi \propto T^{-\nu}$, the temperature in eqn (6.17) would have power equal to $\nu/3$.

The critical region should also exist on the left of the transition point. Naturally, the power value in $T^{1/3}$ in the temperature dependence should remain intact. However, the constant α in eqn (6.17) becomes negative. To estimate it and establish the boundaries of the critical region, we plot along the abscissa in Fig. 6.2 the value of $1/\xi > 0$, assuming that $1/\xi \propto |x - x_{\text{c}}|$. The right-hand boundary of the critical region is determined by the constraint used in the derivation of eqn (6.17): $\xi = L_{\text{ee}}$. Since at sufficiently low temperatures, only the hopping conductivity exists on the side of the insulator, the position of the left boundary of the critical region may be determined from the comparison of the correlation ξ and hopping $\bar{r}$ lengths, $\xi = \bar{r}$. Using eqn (2.30) for L_{ee} and eqn (4.18) for $\bar{r}$ with ξ substituted for a_{B}, one can readily show that, in the $(1/\xi, T)$ axes, both curves limiting the critical region are cubic parabolas

$$T \propto (g_{\text{F}} \xi^3)^{-1}. \tag{6.18}$$

At the left boundary of the critical region, eqn (6.18), the critical conductivity σ_{crit} transforms into the much smaller hopping conductivity σ_{hopp}. Neglecting the later, we may assume that at this boundary $\sigma_{\text{crit}} \approx 0$ and

$$\alpha \approx -\beta T^{1/3} \approx \left(\frac{e^2}{\hbar}\right)\left(\frac{1}{\xi}\right). \tag{6.19}$$

Combining eqns (6.17) and (6.19), we get for the critical region

$$\sigma_{\text{crit}} = \alpha + \beta T^{1/3}, \quad \alpha = \left(\frac{e^2}{\hbar}\right) \times \begin{cases} 1/\xi & x > x_{\text{c}} \\ -1/\xi & x < x_{\text{c}}, \end{cases} \quad \beta = \left(\frac{e^2}{\hbar}\right)(g_{\text{F}})^{1/3}. \tag{6.20}$$

Note: One may pass from the state of a metal to the state of an insulator (both at $T = 0$) avoiding the phase-transition point x_{c} (dashed line in Fig. 6.3). However, in this case, one has to intersect two cross-over lines, in the vicinity of which the dominating
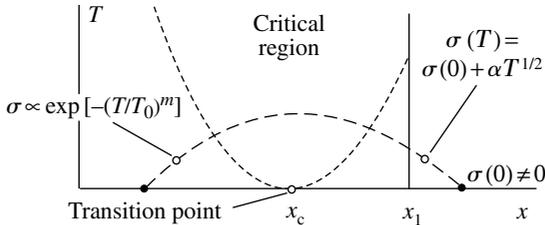

FIG. 6.3.   The path from a metal state to an insulator state through the critical region leaving aside the transition point.
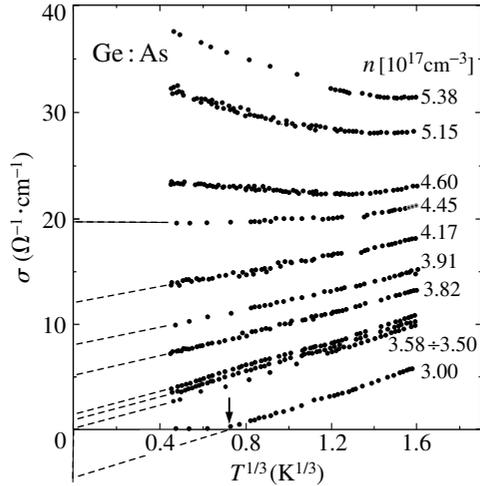
FIG. 6.4. Temperature dependence of conductivity in the vicinity of the metal–
insulator transition for a series of differently doped Ge : As samples (Shlimak
et al. 1996).

processes that determine conductivity alternate. At the concluding part of this path,
the conductivity decreases exponentially with lowering of the temperature.

To observe experimentally a metal–insulator transition, one has to appropri-
ately select the control parameter, measure the temperature dependence of the
conductivity at different values of the selected parameter, construct this depend-
ence in the straightening axes $(T^{1/3}, \sigma)$ and, being certain that the dependence
has the form of straight lines, extrapolate it to $T = 0$ to obtain $\sigma(x, 0)$. The transi-
tion should manifest itself in the values of the function $\sigma(x, 0)$ in the vicinity
of $x = x_c$.

The most natural control parameter is the carrier concentration n determined
by the number of impurities in a sample. Each experimental curve in Fig. 6.4
corresponds to the measurements made along vertical lines on the diagrams
shown in Figs 6.2 and 6.3. According to the experimental data extrapolated to
$T = 0$, the critical concentration of As in Ge : As equals $n_c = 3.5 \cdot 10^{17}$ cm^{-3}. At
concentrations $n > n_c$, extrapolation yields positive $\sigma(0)$ values. The arrow at
the lowest curve $\sigma(T)$ indicates the temperature at which the cross-over happens
from the critical region to the region of hopping conductivity in the sample with
the concentration $n = 3.0 \cdot 10^{17}$ cm^{-3}. The hopping conductivity has a much
lower value than the conductivity in the critical region and, on the scale used in
Fig. 6.4, is almost indistinguishable, so that the dependence $\sigma(T)$ seems to rest
against the abscissa. This has been used above in writing eqn (6.19).

As is seen from Fig. 6.4, the series of the $\sigma(T)$ curves for a Ge : As sample
reveals one more remarkable impurity concentration, $n = 4.45 \cdot 10^{17}$ cm^{-3}, at

which the low-temperature derivative $d\sigma/dT$ changes sign. It is natural to believe that it is the value of the control parameter $x_1 = n_1$ at which eqn (6.12) is valid, because at this concentration, the temperature-dependent classical and quantum corrections to conductivity, having opposite signs, compensate one another. The Drude description of conductivity becomes invalid and the critical region is attained (see also eqn 6.13). In fact, knowledge of n_1 signifies knowledge of $\sigma_{\text{Mott}} \equiv \sigma(n_1) \simeq 20\,\Omega^{-1}\,\text{cm}^{-3}$, so that it becomes possible to indicate the concrete values of the variables at the abstract diagram shown in Fig. 6.2. These are the values of the n_1/n_c ratio and the proportionality coefficient between $1/\sigma$ and ξ in the concentration range $n_c < n < n_1$. It also becomes possible to write the conductivity in the reduced units $\sigma/\sigma_{\text{Mott}}$.

The Ge:As sample with concentration $n = 4.6 \cdot 10^{17}\,\text{cm}^{-3}$ is located deep in the metal region (Fig. 6.4). However, this sample may be transformed into an insulator by applying a magnetic field which, in this case, plays the role of a control parameter. It is seen from Fig. 6.5, that $5\,\text{T}$ is the critical value of the field. In stronger fields (7 and $8\,\text{T}$), the experimental points at the lowest temperatures deviate from straight lines. This is caused by the proximity to the region of hopping conductivity – see Fig. 6.2 (in Fig. 6.4, the similar salient point at $n = 3 \cdot 10^{17}\,\text{cm}^{-3}$ looks like a kink).

On many occasions, the accuracy of conductivity measurements and the range of temperature variations are insufficient for the unique determination of the functional dependence $T^{1/3}$. Sometimes, in order to straighten the curves in the course of processing of the experimental data, one uses not $T^{1/3}$ but $T^{1/2}$, thus continuing using the power that determines the quantum correction in the Boltzmann region at $\Delta\sigma \ll \sigma(0)$ up to the transition point. This alternative
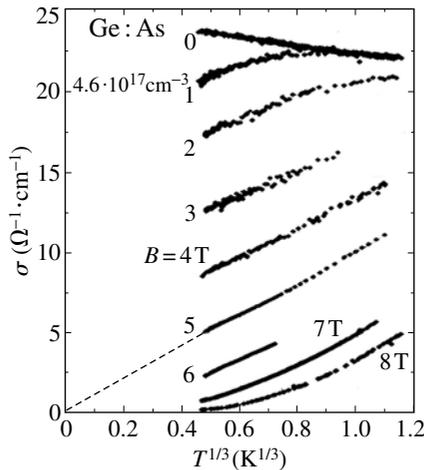

FIG. 6.5. Temperature dependence of conductivity of a Ge:As sample in different magnetic fields (Shlimak et al. 1997).

procedure leaves the value of the control parameter at the transition point almost intact.

Above, we have described the evolution of the transport properties in the vicinity of a metal–insulator transition which corresponds to the description based on the scaling hypothesis. However, this evolution gives no information on the wave functions and the mechanism of conductivity. Now, return to the segment (x_c, x_1) of the abscissa $T = 0$ in Fig. 6.2. Formally, the electrons are delocalized, but since $\sigma < \sigma_{\text{Mott}}$, their wave functions are not conventional Bloch waves. It seems that the wave functions have a structure such that the electrons propagate only in the fractal part of the space with dimensionality $d_f(x) < 3$. Then, at the point x_1, where for certain $d = 3$, the symmetry of the wave functions of the ground state would changed, which is the necessary condition for the occurrence of a phase transition. It is usually assumed that no phase transition takes place at the point x_1. However, the final solution to this problem can be obtained only after the conductivity mechanism in the range (x_c, x_1) becomes clear.

6.3. Two-dimensional (2D) systems

The curve $d = 2$ in Fig. 6.1 lies in the region of negative β values. No matter what material is used for preparing a 2D film, a gradual increase of its size L should lower the conductance value, so that the conductance becomes exponentially low and, in the limit $L \to \infty$ and $T \to 0$, would go to zero. In other words, at a sufficiently low temperature, any sufficiently large film is an insulator. At first glance, this consequence of the scaling hypothesis contradicts the indubitable fact of the existence of metal films. However, this contradiction is not fatal. Since the 2D concentration $n_2 \propto k_F^2$, and the Boltzmann 2D conductivity $\sigma_2 = n_2 e^2 l / \hbar k_F \approx (e^2/\hbar)(k_F l)$, the conductivity of a 2D system with due regard for one of the quantum corrections (2.11) or (2.37) has the form

$$\sigma = \sigma_2 - \frac{e^2}{\hbar} \ln L_\varphi / l \approx \frac{e^2}{\hbar}(k_F l - \ln L_\varphi / l). \qquad (6.21)$$

The above expression is valid only while the correction is relatively small. However, the tendency of the conductivity to decrease with lowering of the temperature remains further anyway. Therefore, eqn (6.21) may be used to estimate parameters at which the state changes considerably. By equating the conductivity (6.21) to zero, a specific L_φ value called the localization length and denoted by the letter ξ is determined:

$$\xi \approx l \exp(k_F l). \qquad (6.22)$$

It is important that the length ξ thus defined depends only on the electron concentration (via k_F) and disorder (via l). Now, assume that the phase breaking takes place as a result of the electron–electron collisions occurring with frequency $\tau_\varphi^{-1} \equiv \tau_e^{-1} = T/(\varepsilon_F \tau)$ under the diffusion conditions (see eqn 2.33). Then, $L_\varphi = (D\varepsilon_F \tau/T)^{1/2}$, and we arrive at the estimate of the temperature T_{cr}

at which $L_\varphi = \xi$,

$$T_{\mathrm{cr}} = \frac{D\varepsilon_{\mathrm{F}}\tau}{\xi^2} \approx \varepsilon_{\mathrm{F}} \exp(-2k_{\mathrm{F}}l). \qquad (6.23)$$

Note: If the frequency τ_φ^{-1} is controlled by another scattering process so that $\tau_\varphi^{-1} \propto T^\nu$, then the exponent in eqn (6.23) equals not 2 but $2/\nu$. This does not considerably influence the further considerations. If the quantum correction to eqn (6.21) is due to the interelectron interference and not to the weak localization, so that $\tau_\varphi^{-1} \equiv \tau_{\mathrm{ee}}^{-1} = T/\hbar$, then we arrive at the result given by eqn (6.23) with the only difference of the factor $\hbar/\tau$ instead of ε_{F} in the pre-exponent.

The cross-over from the logarithmic to exponential decrease in conductivity takes place in the region of the temperatures close to that given by eqn (6.23). However, even at a relatively low "metallic" value of the parameter $k_{\mathrm{F}}l > 5$–10, this temperature becomes unrealistically low, so that, in fact, no localization takes place. Moreover, the cross-over may not come into being because of the finite sample dimensions L. Even after the sample has cooled to the temperature T_{cr}, no cross-over would be observed if $L < \xi$. The cross-over line is shown in Fig. 6.6, where, based on eqn (6.22), the $1/\xi$ value is plotted along the abscissa and indicates a quantitative measure of disorder.

Thus, the length ξ determined by eqn (6.22) is the localization radius at which the given disorder in a two-dimensional system localizes carriers at $T = 0$. In "metallic" films, $l \gg k_{\mathrm{F}}^{-1}$, and the length ξ is exponentially large, so that localization takes place only hypothetically. To observe such localization one needs unrealistically large samples with $L \gg \xi$ and unrealistically low temperatures, eqn (6.23). With an enhancement of disorder and a decrease in the parameter $k_{\mathrm{F}}l$, the limitation for L from below subsides, whereas the temperature given by
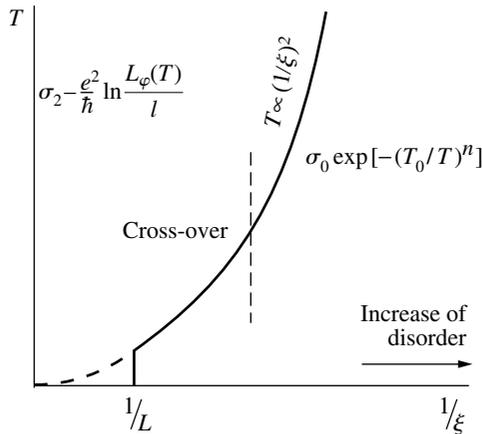

FIG. 6.6. Cross-over from the logarithmic to exponential temperature dependence of conductivity in two-dimensional systems.

eqn (6.23) increases. At $k_{\mathrm{F}}l \gtrsim 1$, both parameters acquire realistic values. It is for this region that the scaling hypothesis predicts for a 2D system with fixed disorder the cross-over from weak to strong localization with lowering of the temperature instead of a metal–insulator transition.

The above considerations are illustrated by experiments on various 2D systems. These systems may be either specially prepared planar interfaces between two media along which a two-dimensional electron well is formed (heterostructures, inversion layers) or, else, thin films.

In the GaAs–Al$_x$Ga$_{1-x}$As heterostructure, the concentration of 2D electrons can be varied with the aid of a gate. In Van Keuls et al. (1997), the whole spectrum of concentrations was obtained on one sample by varying the applied electric field. As is seen from Fig. 6.7, the function $\rho(T)$ obviously has the activation character at the concentration $n \approx 0.6 \cdot 10^{11}\,\mathrm{cm}^{-2}$, whereas at $n \approx 6 \cdot 10^{11}\,\mathrm{cm}^{-2}$ it shows only a weak logarithmic increase in ρ with lowering of T. Although the possibilities for detailed comparison were rather limited in this experiment, the qualitative result obtained was just that which could be expected from the scaling hypothesis.

A more detailed comparison was made for Cu, Ag, and Au films with the thicknesses b ranging from 0.3 to 2.0 nm (Hsu et al. 1995).

Note: The dimensionality of the electron system in a given sample is determined by comparing the sample thickness b with the inelastic length L_φ. Therefore, the films in the experiments (Hsu et al. 1995) were really two-dimensional. At sufficiently low temperatures, much thicker films may also become two-dimensional objects.
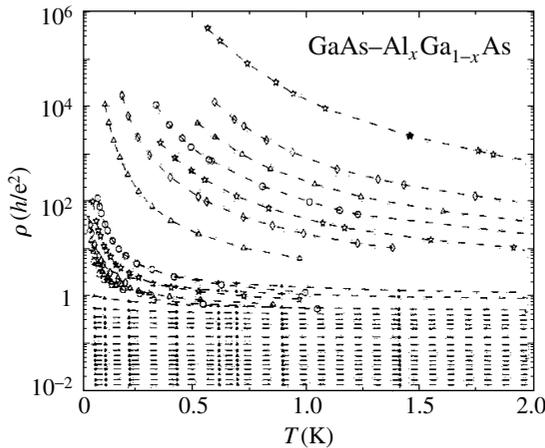

FIG. 6.7. Evolution of the temperature dependence of 2D-gas resistance in a heterojunction with a change of the two-dimensional electron density from $n \approx 0.6 \cdot 10^{11}\,\mathrm{cm}^{-2}$ (the upper curve) to $n \approx 6 \cdot 10^{11}\,\mathrm{cm}^{-2}$ (the lower curve) (Van Keuls et al. 1997).
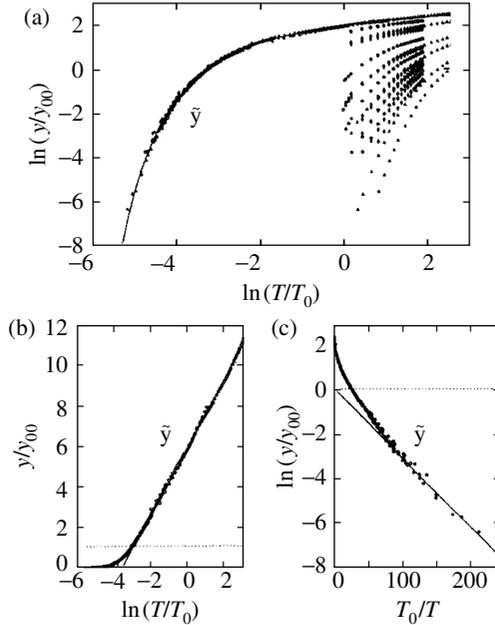
FIG. 6.8.   (a) Temperature dependence of conductance, $\ln y/y_{00}$, for films of different metals with different thicknesses; the parameter $T_0 = 1\,\mathrm{K}$ (right-hand curves). The same data are also shown with T_0 as a free parameter. The horizontal parallel translations bring all the data to one universal curve $\tilde{y}$. (b) The asymptotic behavior of the curve $\tilde{y}$ at high $\ln y/y_{00}$ values indicates the logarithmic dependence typical of weak localization. (c) At low $\ln y/y_{00}$ values, the curve $\tilde{y}$ describes the activation Arrhenius law (Hsu et al. 1995).

The measurements performed on films of the three metals were used to draw a series of curves shown in the right-hand part of Fig. 6.8a. Initially, these curves were drawn by selecting T_0 equal to $1\,\mathrm{K}$ and y_{00} equal to $y_{00} = (2\pi)^{-1}(e^2/\hbar)$. With fixed y/y_{00} ratio, the authors managed to fit all the experimental points to one curve by making T_0 a free parameter, different for different experimental curves. For the given x-scale in Fig. 6.8a, this signifies parallel translation along the horizontal axis for a distance of $-\ln T_0$. Although the measurements on each sample fit only a small segment of the summary curve, this curve encompasses more than three orders of magnitude with respect to T/T_0 and has both weak localization (the curve part $\ln y/y_{00} > 0$, Fig. 6.8b) and activation ($\ln y/y_{00} < -2$, Fig. 6.8c) portions.

In the activation portion (the region of strong localization), the theory of hopping conductivity admits different exponents in the expression for $y \propto \exp[-(T_0/T)^\nu]$. According to Fig. 6.8c, the experimental temperature dependence $y(T)$ has the power $\nu = 1$, which indicates the existence of a rigid gap of width T_0.
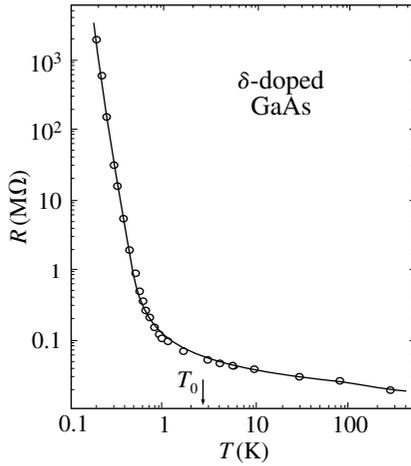
FIG. 6.9.   Temperature dependence of resistance of a strip of width 0.05 μm from δ-doped GaAs (Havin et al. 1998).

Compare: Ultrathin Be films have $\nu = 1/2$ (Fig. 4.7 in Chapter 4) and the Coulomb gap (Fig. 3.10 in Chapter 3 and Fig. B.5 in Appendix B). The materials preserve their individuality even in the form of ultrathin films.

There also exists the observation of the cross-over from weak to strong localization on one sample (Havin et al. 1998). Figure 6.9 shows the resistivity measured on a δ-doped GaAs : Si sample having a quasi-one-dimensional configuration (long strip of width $500\,\text{Å}$) in the range from room temperature to $0.2\,\text{K}$ (more than three orders of magnitude). The 2D-carrier concentration is determined by the initial doping in the δ-layer and cannot be varied during the experiment. Similar to the case of thin metal films, here the "one sample–one concentration" principle is obeyed. The measurements made below $1\,\text{K}$ fit the activation dependence $R \propto \exp(T_0/T)$ with $T_0 \simeq 2.6\,\text{K}$, and above $4\,\text{K}$, the $R(T)$ curve is described by the formulas of the theory of weak localization. It is unclear to what extent the quasi-one-dimensional nature of the sample is important. Possibly, the geometric ratio used was especially favorable for increasing the temperature range of the measurements. In any case, the linear dependence of resistivity on $\ln T$ at temperatures above T_0 is characteristic of two-dimensional samples.

The harmony between the scaling hypothesis and the experimental results was broken by the experimental data on the inversion layers in silicon, where the concentration was controlled by the voltage applied to the gate (Kravchenko et al. 1995). An increase of the 2D concentration of carriers, n_2, in a 2D layer of the samples with extremely high carrier mobility changes the sign of the derivative $\partial\rho/\partial T$ so that at $n_2 \simeq 1.4 \cdot 10^{11}\,\text{cm}^{-2}$, the resistivity decreases several times with
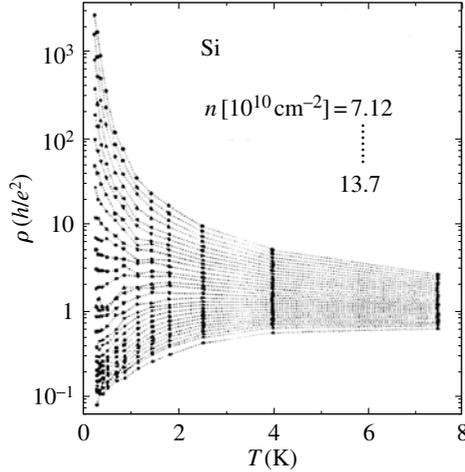
FIG. 6.10. Temperature dependence of resistance of the inversion layer on a Si surface at different carrier densities in the layer (Kravchenko et al. 1995).

lowering of the temperature from $4\,\mathrm{K}$ to several dozens of millikelvins (Fig. 6.10). This behavior is opposite to the behavior of a $\mathrm{GaAs/Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ heterostructure in Fig. 6.7 and is inconsistent with the scaling hypothesis.

The interpretation of the evolution of the derivative $\partial\rho/\partial T$ as a function of the concentration in an inversion silicon layer is the subject of numerous discussions. It is always possible to believe that the experimental temperature was not low enough. However, at present, it seems to be probable that, at a certain intermediate concentration at which the low-temperature derivative $\partial\rho/\partial T \simeq 0$ at the attained temperatures, a metal–insulator transition takes place at $T = 0$. This may signify the existence of interactions in the electron system, which may considerably affect the wave functions of the ground state at least on one side of the transition. One example is well known: the superconducting interaction in the 2D electron system changes the behavior described by the scaling hypothesis and may give rise to a quantum superconductor–insulator transition. A classical illustration of the latter transition is presented in Fig. 6.11 which shows $R(T)$ curves from several amorphous Bi films with different thickness (Haviland et al. 1989).

Although the patterns in Figs 6.10 and 6.11 are qualitatively similar, the behavior of silicon cannot be interpreted as superconductivity. Superconductivity is low sensitive to scattering, whereas, in silicon, the positive derivative $\partial\rho/\partial T$ at high n values is observed only on samples with high mobility. Therefore, these experiments give rise to new questions rather than to answers. What factor is responsible for a positive $\partial\rho/\partial T$ derivative? How would the $\sigma(T) \equiv \rho^{-1}(T)$ function behave with a further lowering of the temperature, i.e., what is the real $\sigma(n, T = 0)$ function, etc.
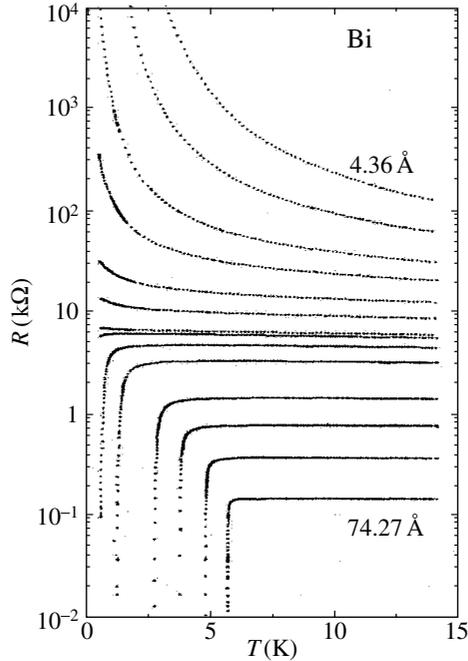
FIG. 6.11. Superconductor–insulator transition in amorphous Bi films of differ-
  ent thickness deposited onto 5 Å-thick amorphous Ge film on a SiO_2 substrate
  (Haviland et al. 1989). Here R is 2D-resistivity.

Note: Up to now, all these questions form a vicious circle. To decide whether a metal–
insulator transition really takes place or not, one has to know the function $\sigma(n, 0)$,
which requires the appropriate extrapolation of $\sigma(T)$ to $\sigma(0)$. This, in turn, requires
a clear understanding of the nature of the temperature dependence $\sigma(T)$, i.e., singling
out of the essential interactions, and the *a priori* knowledge of whether a transition
really takes place or not.

6.4. Scaling and spin–orbit interaction

The existence of the spin–orbit interaction and related antilocalization makes
the question of the universality of the $\beta(\ln y)$ curves in Fig. 6.1 more com-
plicated. The interaction changes the sign of the quantum correction and,
according to Table 6.1 at the beginning of this chapter, also changes the sign
of the derivative in the right-hand part of the $\beta(\ln y)$ curves. This is especially
important in the case $d = 2$, in the range where the curve $\beta_2(\ln y)$ approaches
the line $\beta = 0$. Allowance for the spin–orbit interaction makes the statement
that two-dimensional systems have no metal–insulator transition somewhat
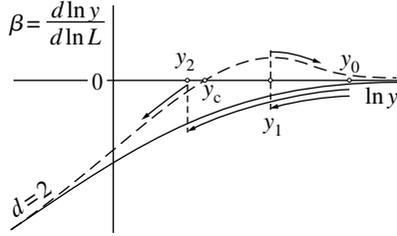dubious.

FIG. 6.12.   Two variants of the $\beta(\ln y)$ curve $(d = 2)$ from Fig. 6.1.

In the two-dimensional case, the quantum correction (2.25) has the form

$$\Delta\sigma_2 \approx -\left(\frac{e^2}{\hbar}\right) \int_{\tau}^{\tau_\varphi} \frac{dt}{t} \left(\frac{3}{2} e^{-t/\tau_{\text{so}}} - \frac{1}{2}\right). \qquad (6.24)$$

Since we consider conductivity at $T = 0$, we seemingly may take $\tau_\varphi = \infty$ as the upper limit of integral (6.24). At the same time, the τ_{so} value remains finite, because the probability of spin flip in elastic scattering is independent of temperature. The factor

$$\frac{3}{2} e^{-t/\tau_{\text{so}}} - \frac{1}{2},$$

in the parentheses in the integrand in (6.24) in the region $t > \tau_{\text{so}}$ is of the order of $-\frac{1}{2}$. This makes the integral diverge at high t values and makes it negative, and the quantum correction to conductivity, positive. The $\beta(\ln y)$ curve acquires a maximum in the upper half-plane $\beta > 0$ and also the critical point y_c. In Fig. 6.12, this curve is shown by the dashed line. The solid line shows the curve constructed without allowance for spin–orbit interaction under the assumption that $\tau_{\text{so}} = \infty$.

However, one should be very careful with directing $T \to 0$, i.e., with substitution $\tau_\varphi \to \infty$ in the upper limit of integral (6.24). This is seen from the following considerations.

Let the classical conductance of a sample be y_0. Calculate the quantum correction to the conductivity by gradually increasing the upper limit τ^* in integral (6.24) beginning with τ. This signifies that we start with the allowance for interference at the smallest diffusion loops and then gradually take into account also the loops with an increasing number of kinks. Until $\tau^* < \tau_{\text{so}}$, the absolute value of the correction to conductivity increases, but remains negative. Therefore, we move along the lower (solid) curve toward lower $\ln y$ values. When τ^* reaches τ_{so}, antilocalization comes into being, and the representative point would jump to the upper (dashed) curve. The problem reduces to the prediction of the point when this jump will take place. The more pronounced the spin–orbit interaction, the lower the τ_{so} value and the earlier the jump takes place in the motion from right to left. The jump at the point $y_1 > y_c$ indicates that with an increase in the film size, the representative point starts moving along the dashed curve to the right, the conductance increases, so that the film remains in the metal state.

The jump at the point $y_2 < y_c$ indicates that the representative point remains in the lower half-plane and continues moving to the left toward the state of an insulator.

To evaluate the critical value of τ_{so}, we note that the critical point is located on the dashed curve approximately at the same y value as the cross-over from the logarithmic to exponential dependence of conductance on temperature on the solid curve defined by eqn (6.23). Using eqn (6.21) and eqn (2.7) for L_φ, we obtain the estimate of τ_{so}^* which brings the point y_1 to the critical value y_c:

$$k_F l = \tfrac{1}{2} \ln \tau_{so}^*/\tau, \quad \tau_{so}^* = \tau e^{2k_F l}. \tag{6.25}$$

In order to move to the right after the jump, i.e., to obtain a two-dimensional metal state, it is necessary that the inequalities $\tau_{so} \ll \tau_{so}^*$ are fulfilled. Since by definition, $\tau_{so} > \tau$, then an admissible range of τ_{so} values is

$$\tau \ll \tau_{so} \ll \tau \exp(2k_F l). \tag{6.26}$$

In strongly disordered films, these inequalities cannot be fulfilled, because $k_F l \sim 1$. In such films, the spin–orbit interaction produces practically no considerable changes.

In films with $k_F l \gg 1$, the problem is academic, because the cross-over temperature T_{cr} defined in eqn (6.23) is unrealistically low. However, in principle, a well-conducting film with a considerable $k_F l$ value may remain metallic at the absolute zero of temperature because of the spin–orbit interaction.

References

Abrahams, E., Anderson, P.W., Licciardello, D.C., and Ramakrishnan, T.W. (1979). Phys. Rev. Lett. **42**, 673.

Haviland, D.B., Liu, Y., and Goldman, A.M. (1989). Phys. Rev. Lett. **62**, 2180.

Havin, Yu., Gershenson, M., and Bogdanov, A. (1998). Phys. Rev. B **58**, 8009.

Hsu S.-Y. and Valles, Jr., J.M. (1995). Phys. Rev. Lett. **74**, 2331.

Kramer B. and MacKinnon A. (1993). Rep. Prog. Phys. **56**, 1469.

Kravchenko, S.V., Mason ,W.E., Bowker, G.E., Furnaux, J.E., Pudalov, V.M., and D'Iorio, M. (1995). Phys. Rev. B **51**, 7038.

Lee, P.A. and Ramakrishnan, T.V. (1985). Rev. Mod. Phys. **57**, 287.

Mott, N.V. (1990). *Metal–Insulator Transitions*. Taylor & Francis.

Shlimak, I., Kaveh, M., Ussyshkin, R., et al. (1996). Phys. Rev. Lett. **77**, 1103.

Shlimak, I., Kaveh, M., Ussyshkin, R., et al. (1997). Phys. Rev. B **55**, 1303.

Van Keuls, F.W., Mathur, H., Jiang, H.W., and Dahm A.J. (1997). Phys. Rev. B **56**, 13263.

CHEMICAL LOCALIZATION

The band theory of metals with its concept of energy-band overlap describes rather than explains the metallic properties of matter. The fundamental reason for the existence of the metallic state is that in an isolated metal atom the valence electrons occupy energy levels close to the upper edge of the potential well, so that in the condensed state any perturbation introduced by neighboring metal atoms leads to delocalization of the valence electrons. From this viewpoint, the grouping of chemical elements into metals and metalloids is caused by the structure of the atoms; metals are in the lower left corner of the Periodic Table, and the boundary between metals and metalloids, which is a diagonal of the Periodic Table, is blurred and extremely conventional.

In metals, valence electrons are itinerant and their concentration n is at least no lower then the concentration of atoms, $n \sim a^{-3}$ where a is the mean interatomic distance. Electron concentration in alloys is of the same order. At the beginning of Chapter 1, materials with such a concentration of itinerant electrons were called *standard metals*. Along with this definition, the question was brought up whether a metal–insulator transition was possible in a material with high electron density (see Section 1.4 and Fig. 1.8 in Chapter 1 and Fig. 5.17 in Chapter 5). It followed from the experimental data presented in Chapter 1 that increase of neither static nor dynamic disorder can lead to a metal–insulator transition. By introducing extreme disorder of any kind into the alloy we only bring it closer to the brink of localization. For a transition to occur, a fraction of the metal atoms must be replaced by metalloid atoms, which drives the concentration n of the itinerant electrons down to below of $n_{\max}$.

Imputing of metalloid atoms can prove to be twice as effective in the sense that the concentration of metal atoms does not always uniquely determine the concentration n of the delocalized or potentially delocalizable electrons. If the metal and metalloid atoms can form stable chemical molecules, metal electrons enter the chemical bonds: from the shallow potential well of a metal atom they go to a much deeper potential well of the molecule and, therefore, remain localized, notwithstanding the surroundings of the molecule. Hence the effective electron concentration n, which affects the position of the material on the metal–insulator phase diagram, decreases even more due to the emergence of chemical bonds.

Bearing in mind the tying-up of a fraction of the valence electrons into chemical bonds, we can formulate the following question: *Is there a way to build deep potential wells using only metal atoms, which would transform a material in*

which there are no metalloid atoms into an insulator in spite of the electron concentration of a standard metal? The experimental data discussed in the present
chapter show that this is possible.

7.1. Intermetallic compounds in two-component melts

*Experimental details may be found in the reviews by Hensel (1979) and van der
Lugt and Geerstma (1987).*

It has been known for a long time that the resistivity of a liquid melt of two
very good metals may change severalfold, even by a factor of 10, depending on
the relative concentration of the two components, and reach its maximum at a
certain rational ratio of the atomic concentrations, such as 1 : 1 or 1 : 3 or 1 : 4.
To get such data one needs to measure the resistivity at a fixed temperature as
a function of the alloy component concentration. As an example, we outline here
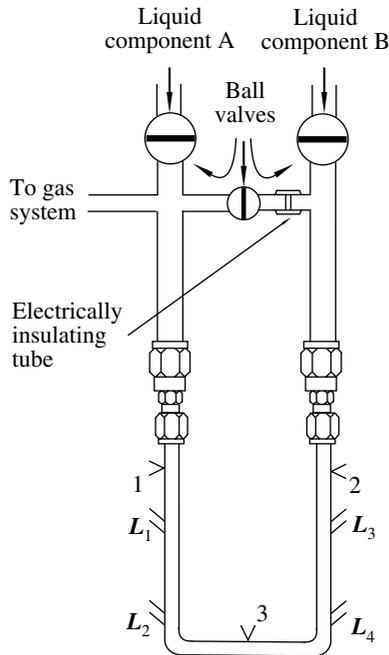the experimental setup which was used for measurements of the Na–Pb system
(Fig. 7.1).

FIG. 7.1. Electrical resistivity cell for measurements of two-component melts
as a function of alloy component concentration (Calaway and Saboungi
1983). 1, 2 and 3 – thermocouples, L_1 and L_2 – cooper leads for resistivity
measurements in the left branch, L_3 and L_4 – the same for the right branch.

The resistivity measurements were performed using a four-probe technique independently in two branches of a U-shaped tube made of low-carbon steel with outer diameter 3 mm and wall thickness 0.4 mm and filled by the melted alloy. Four cooper leads were attached to the outer side of each branch of the tube. The resistance of the empty tube as a function of temperature was determined first; this was used to introduce amendments later. Both sides of the tube were connected to two tubes of larger diameter which served as reservoirs. By pressurizing one reservoir with high-purity helium in order to drive all the alloy to the other side, the mixing of the alloy was achieved. The mixing was repeated until the resistivity measurements on both sides of the tube confirmed that a homogeneous alloy composition has been reached. The final measurements were done after the gas pressure and the melt levels in both branches of the tube were equalized. Then successive addition of a component was made and the whole procedure was repeated.

Figure 7.2 shows the results of measurements of the resistivity of Na–Pb melts at $725\,^\circ$C. Clearly, the concentration ratio $\mathrm{Na:Pb} \sim 4:1$ is marked out. An addition of 20% of lead increases the resistivity compared to that of pure Na by a factor of about 20. Here the peak value of resistivity is of the order of the maximum possible value ρ^* of a standard metal,

$$\rho^* \approx \frac{\hbar}{e^2}\, n^{-1/3} \approx (200\text{–}300)\mu\Omega \cdot \mathrm{cm} \tag{7.1}$$
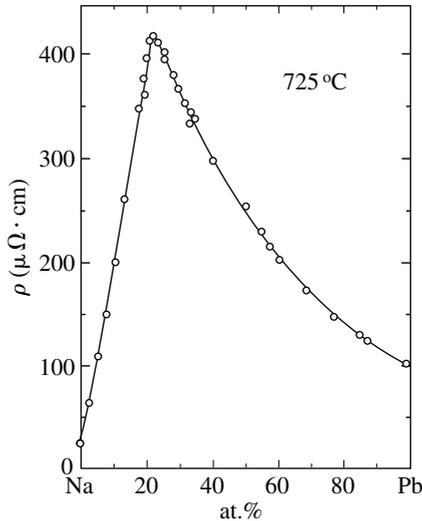

FIG. 7.2.   Resistivity of melts of the Na–Pb system at $725\,^\circ$C. The peak value is reached at a lead concentration $C_{\mathrm{Pb}} = 20\%$, where stable $\mathrm{PbNa_4}$ configurations emerge (Calaway and Saboungi 1983).
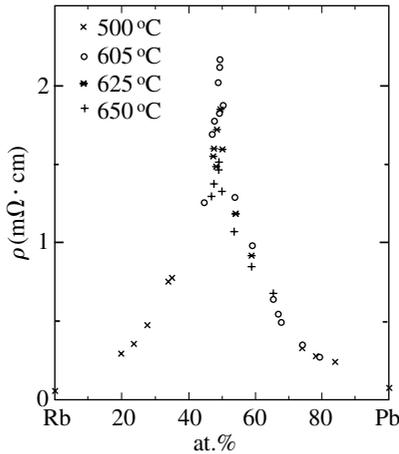
Fig. 7.3. Resistivity of melts of the Rb–Pb system at different temperatures (Calaway and Saboungi 1983). The peak value is reached at lead concentration $C_{\mathrm{Pb}} = 50\%$.

(see eqns 1.5 and 1.6 in Chapter 1). The Li–Pb system behaves in a similar manner.

Replacing Li and Na with a heavier alkali metal, K, Rn, or Cs, changes the resistivity vs. concentration diagram significantly. For example, Fig. 7.3 shows the diagram for the Rb–Pb system. The peak has shifted to another rational ratio of the component concentrations $\mathrm{Rb:Pb} \sim 1:1$, while the peak value of the resistivity increased severalfold. Now this value exceeds the maximum resistivity of a standard metal, eqns (1.4) and (1.6), by a factor of 10. The resistivity peak is rather narrow: a small deviation from the ratio $1:1$ results in a slump of the resistivity. The survey diagrams in Fig. 7.4 show that melts of alkali metals with another tetravalent metal, tin, behave in the same manner. The highest resistivity values are realized in Cs-based melts.

The high values of resistivity mean that near the respective concentration ratios the melt ceases to be a standard metal in the sense that a fraction of carriers in it are bound in some manner and the remaining effective concentration is $n_{\mathrm{eff}} \ll 4 \cdot 10^{22} \mathrm{cm}^{-3}$ (cf. 1.4). Indeed, for the Rb–Pb system, the value $\rho \approx 2200 \, \mu\Omega \cdot \mathrm{cm}$ is 10 times greater than the maximum value for a standard metal, $\rho^* \approx (200\text{--}400) \, \mu\Omega \cdot \mathrm{cm}$. According to eqn (1.5), this implies that the number of free carriers in the melt is no greater than $10^{-2}\text{--}10^{-3}$ of the ordinary number of carriers in a standard metal.

From the rational component-concentration ratios it follows that the increase in resistivity is due to formation of complexes within which most electrons turn out to be traped. The position of the peak in resistivity in Li- and Na-based melts points to the existence of $\mathrm{Na_4Pb}$ and $\mathrm{Li_4Pb}$ compounds in the melts. The five atoms comprising such a compound have eight electrons in their valence
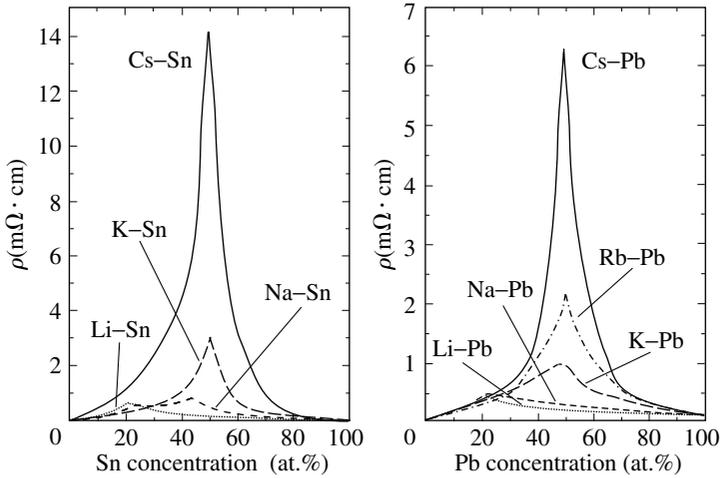
FIG. 7.4. Survey diagrams of the resistivity vs. concentration for A–B melts (A = Pb, Sn, B is an alkali metal) (Xu et al. 1992). The resistivity of alloys with Li and the Pb–Na alloy is a maximum at $C_A = 20\%$; the resistivity of the Sn–Na alloy has two maxima at $C_{Sn} \approx 25\%$ and 45%; the rest have a resistivity maximum at $C_A = 50\%$.
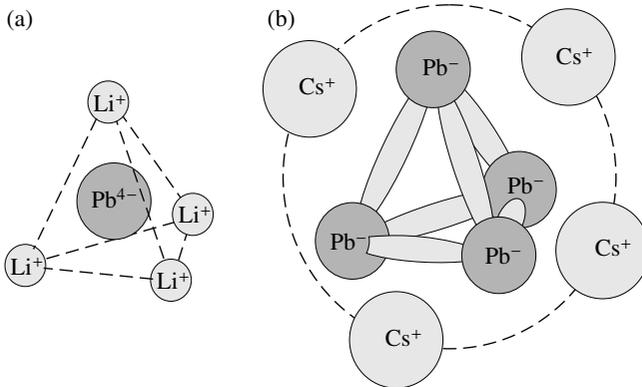
FIG. 7.5. (a) Ionic configurations consisting of Pb and Sn atoms and atoms of a light alkali metal Li or Na; (b) the same with a heavy alkali metal K, Rb, or Cs.

shells. Apparently, they form a single stable outer shell of the Pb^{4-} ion, while four alkali ions held together by Coulomb forces surround that ion. The four ions form a barrier thanks to which the eight electrons in the outer shell of Pb are kept within this electrically neutral atomic configuration and do not participate in conduction (Fig. 7.5a).

An increase in the size of the alkali atoms leads to a qualitative change of the forming compounds, with the ability of these compounds to act as electron traps gaining in strength. Such structures are well known and are called Zintl structural units (named after the German chemist who in the 1930s discovered the rule of formation of ionic configurations). If an electron goes from an alkali atom to the lead atom, the Pb^{1-} ion will have five electrons in the outer shell, the same as in the P or As atoms. As is well known, these two elements form, in the gaseous phase, tetrahedral molecules P_4 or As_4. Here there are eight electrons near each atom: five electrons belonging to the atom proper and one electron from the covalent bonds with each of the three neighbors in the tetrahedron. Pb^{1-} ions also form such tetrahedra, and the total electric charge $-4e$ of such a tetrahedron is balanced by the electric charge of the four alkali-metal ions surrounding it. Sn^{1-} ions form similar tetrahedra $(Sn_4)^{4-}$ surrounded by four alkali ions. It is the structural unit

$$A_4B_4, \qquad A = Pb, Sn, \qquad B = K, Pb, Cs \qquad (7.2)$$
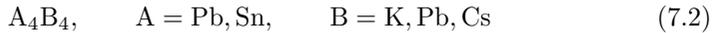
that is the configuration within which 20 valent electrons are locked (Fig. 7.5b).

Binary melts consisting of alkali metals and some other metals behave in a similar manner. The absolute champion when it comes to forming high-resistive melt is the melt of two superb metals, the alkali metal Cs and the noble metal Au in equal concentrations. As Fig. 7.6 shows, the formation of compounds in the melt reduces the conductivity by a factor of 10 000. Here the conductivity is comparable to that of salt melts ($3\Omega^{-1} \cdot cm^{-1}$ for CsAu and $1\,\Omega^{-1} \cdot cm^{-1}$ for the CsCl salt melt).
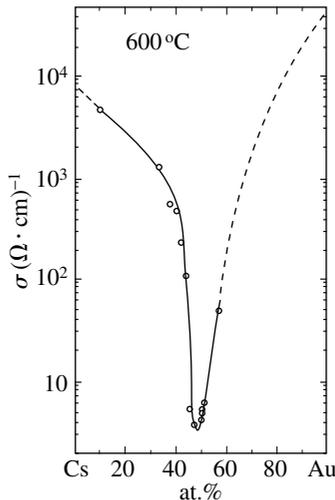

FIG. 7.6.   Conductivity of melts of the Cs–Au system at $600\,°C$ (Hoshino et al. 1975; Hensel 1979).

It is quantum chemistry together with chemical thermodynamics that study where, when, which and how many Zintl structural units appear in a metallic melt and what are the bond energies of these configurations. Naturally, the melts exist at high temperature, so that the curves in Figs 7.3 or 7.6 cannot be accepted as demonstration of the metal–insulator transition we are studying here. For these materials to become objects related to the theory of metal–insulator transitions, they should be quenched into the glass state. Then the low-temperature transport measurements would probably give quantitative characteristics of these traps. However, such experiments apparently are not still done.

In the language of physics, Zintl structural units may be called deep wells of the random potential which trap the electrons in alloys with stoichiometric composition and thus localize them (the potential is random because the positions of Zintl units are assumed to be random in the melt). This legalizes applying the concepts and ideas from the theory of the metal–insulator transition to this phenomenon. Since initially all Zintl traps are identical, the most appropriate concept seems to be the model with structural disorder (see Section 5.5) with the potential

$$V(\boldsymbol{r}) = \sum_{\boldsymbol{R}_i} v(\boldsymbol{r} - \boldsymbol{R}_i), \tag{7.3}$$

where the set of vectors $\boldsymbol{R}_i$ is random and $v(\boldsymbol{r})$ is the potential of a separate Zintl unit. At low concentration n of the units,

$$a_{\mathrm{B}} n^{1/3} \ll 1, \tag{7.4}$$

this model describes an insulator: each electron is localized in its own well $v(\boldsymbol{r} - \boldsymbol{R}_i)$ and its wave function fades outside the well at distance a_{B}.

Compare: The same potential (7.3) was treated in Chapter 1 as one producing only scattering in elemental liquid metals because the inequality opposite to eqn (7.4) was assumed to be valid there and the electrons were *a priori* supposed to be itinerant.

Each configuration A_4B_4 from eqn (7.2) is an almost spherical well for the $4 \cdot 4 + 4 = 20$ valence electrons located inside it on a sequence of energy levels. The decay length a_{B} which enters eqn (7.4) actually refers to the uppermost occupied level. The electrons on deeper levels do not leave the well. This reduces by a factor of 10 the concentration of electrons that can become itinerant and facilitates the metal–insulator transition.

This model allows us to explain qualitatively the temperature dependence of the conductivity of CsAu which is presented in Fig. 7.7. In the liquid state, the conductivity of the melt is controlled by the number of excited carriers; above the melting-point, the conductivity rises along with temperature. The crystallization creates clusters of resonance wells which all have identical and identically located neighbors. The randomness in the location of the wells is partially retained only to the extent to which the alloy is nonstoichiometric and due to the presence of crystal defects and intercrystalline boundaries. The result of the increase in the number of resonant wells is partial delocalization and a tenfold increase in the conductivity of the crystal compared to that of the melt.
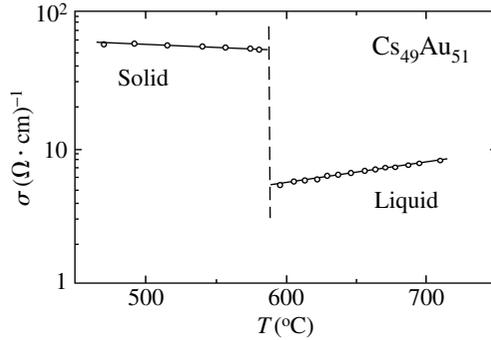
FIG. 7.7.   Temperature dependence of the conductivity of the Cs–Au alloy with
    51% of Au in the liquid and solid states (Schmutzler et al. 1976; Hensel 1979).

However, *note*: as Fig. 7.7 clearly shows, the conductivity of crystalline CsAu is
still about 50 times lower than the minimum conductivity $1/\rho^*$ of a standard metal
(eqn 1.6). It remains unclear to what extent and how the conductivity depends on
deviations from stoichiometry, the number of defects, temperature, and other factors.

7.2. Quasicrystals

*For a comprehensive discussion of different aspects of physics of quasicrystals,
see, for instance, Stadnik (1999).*

Introducing translational symmetry is not the only way to establish long-range
correlations on the set of vectors $\boldsymbol{R}_i$ from the model with structural disorder,
eqn (7.3). Another way is to introduce the quasicrystalline long-range order.
    Translational symmetry, always present in crystals, allows for the existence
of axes of 2-, 3-, 4-, and 6-fold symmetry only. At the same time, it is easy to
imagine that of all possible local configurations of a small number of atoms of the
chemical elements A, B, and C, i.e. $A_nB_mC_p$, the configuration with the lowest
energy has a different symmetry axis, say the axis of 5-fold symmetry. Formation
of a crystal from a material with the composition $A_nB_mC_p$ then becomes a
problem. Sometimes optimal local symmetry is sacrificed, so that a crystal with
a different configuration of the nearest neighbors is formed, with the loss in
local configurations energy balanced by gain caused by translational symmetry.
There is, however, another possibility. Let us arrange optimal configurations of
$n + m + p$ atoms at the sites of a crystal lattice, say, a body-centered cube, as in
Fig. 7.8. Then the loss in energy emerges caused by mismatch and distortions in
the places where these configurations meet, where the short-range order is sure
to be non-optimal. Nevertheless, some substances have such crystal structures.
They are called crystal approximants or crystal prototypes of quasicrystals.
    It occurs, however, that we can do entirely without translational symmetry
by densely packing the space with optimal configurations. That this is possible,
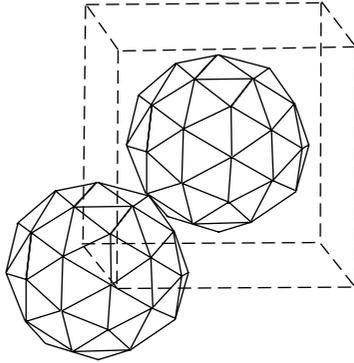
FIG. 7.8. Crystal packing of Mackey icosahedra, closely resembling body-centered cubic packing, in the crystal alloy α(AlMnSi), a quasicrystal approximant (Goldman and Kelton 1993). Each icosahedron consists of more than 50 atoms.

at least theoretically, is demonstrated by the Penrose tiling in the lower part of Fig. 7.9; the plane is covered perfectly (i.e. without gaps and overlaps) by rhombic tiles of two types, with the acute angles equaling $2\pi/5$ and $\pi/5$. Single tiles are depicted in the upper left corner of the same figure. Since the rhombuses adjoin each other at preassigned vertices, the correctly specified functions $F(\boldsymbol{r})$ on the rhombuses remain continuous at the junctions and form a continuous aperiodic function whose separate segments are repeated in the plane an infinite number of times. In the upper right corner of Fig. 7.9 the set of vertices of the rhombuses are depicted to the scale $1:2$. Since there is no translational symmetry, it is rather difficult to notice any correlations in this arrangement. However, there is long-range order in this system: the rhombic tiles are arranged on the plane in a well-distinguishable pattern (although the pattern is not unique).

Quasicrystals are built according to the same principles. Optimal configurations with high-order symmetry axes and with matching configuration-spacers in between, that minimize energy losses at the junctions, form a net which resembles Penrose tiling. The resulting arrangement has no translational symmetry but has long-range order. Many families of such materials are known today. Most of them are metal alloys in the sense that they consist only of metal atoms: Al–Mn, Ga–Mg–Zn, Al–Cu–Fe, Al–Pd–Re, etc. Here the local base configurations may be extremely complex. For instance, in quasicrystals with the Al–Pd–Mn composition one of the local base configurations consists of three sells inserted into each other; altogether there are 51 atoms in this configuration (Fig. 7.10).

Basically, quasicrystals are identified and studied by the X-ray diffraction method. The Fourier transform of the density $\varrho(\boldsymbol{r})$ in a perfect crystal is a sum of an infinite number of narrow peaks (ideally, δ-functions):

$$\varrho(\boldsymbol{r}) = \sum_{\boldsymbol{q}} \varrho_{\boldsymbol{q}} \exp(i\boldsymbol{q}\boldsymbol{r}). \tag{7.5}$$
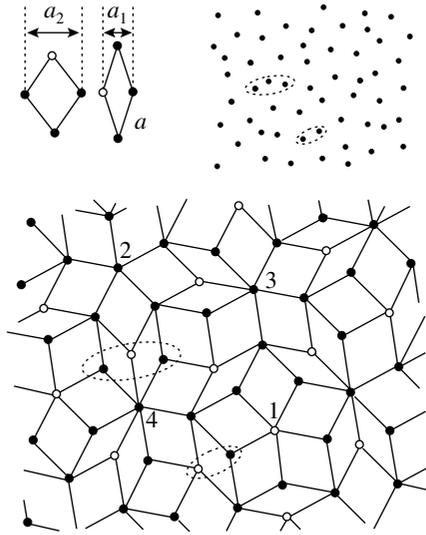
FIG. 7.9. Penrose tiling. Below: tiling a plane without gaps or overlaps by two
   types of rhombic tiles depicted in the upper left corner of the figure, rhombuses
   with equal sides a and acute angles $2\pi/5$ and $\pi/5$, respectively (the vertices
   marked by open circles adjoin each other in the tiling). Upper right corner: the
   set of rhombus vertices of the tiling depicted in the lower part of the figure to
   the scale $1:2$. Although each of the sites 1, 2, 3, and 4 has its nearest neighbors
   only at a distance a, the quality of these neighbors differ considerably (see the
   main text). The dotted ellipses mark a resonant pair of closely located sites
   and a compact triplet of sites.

9 at.   12 at.   30 at.   51 at.
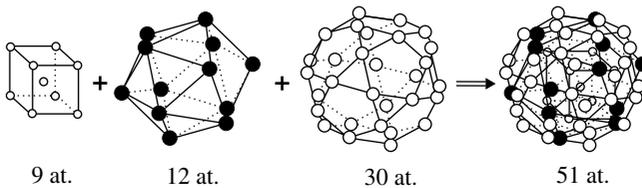
FIG. 7.10. A sequence of atom shells in a Mackey pseudoicosahedron, which
   is the base element of the structure of the Al–Pd–Mn quasicrystal; the total
   number of atoms is 51 (Janot 1996).

The set of vectors $\boldsymbol{q}$ form a lattice in q-space with the same symmetry as the
initial lattice of atoms. To each site of this reciprocal lattice there corresponds
a Bragg reflection in the Laue diffraction pattern. The more perfect the crystal
the shaper the reflections.

   Bragg reflections are not an exceptional property of crystals. For the right
part of eqn (7.5), we can initially select a series of δ-functions $\delta(q)$ in which the

set of vectors $\boldsymbol{q}$ does not possess translational symmetry. By an inverse Fourier transformation we arrive at a function $\varrho(\boldsymbol{r})$ that has no translational symmetry either. The Fourier transforms of quasicrystals are just such series. However, the width of the Bragg reflections is still determined by the imperfectness of the structure, namely, by deviation of the local configurations from the ideal configuration, failure of long-range order because of impurities and vacancies, etc. The sharper the Bragg reflections the closer the quasicrystal is to a perfect one.

The following correlation is typical for metallic single crystals: the higher the quality of the Laue diffraction pattern of a certain substance, the lower the residual resistivity ρ of the crystals. The correlation reflects the wave nature of electrons: the better the conditions for the propagation of an X-ray wave, the smaller the scattering of the Bloch wave. In quasicrystals it is just the opposite: annealing, while increasing the quality of the Laue diffraction pattern, also increases the resistivity. Here the values of resistivity are extremely high. For instance, in quasicrystals of the Al–Cu–Ru composition at 4 K, the resistivity values are as high as $30\,\mathrm{m\Omega \cdot cm}$, which is approximately 100 times higher than the value of ρ^* estimated by eqns (1.5) and (1.6) from the concentration n of the metallic valence electrons.

The properties of an insulator manifest themselves most vividly for the Al–Pd–Re system. Ingots of this alloy can be made by arc melting a mixture of extremely pure Al, Pd, and Re in an atmosphere of pure argon. The resistivity values of this alloy are stably of order 200–$300\,\mathrm{m\Omega \cdot cm}$. After being annealed in vacuum for 24 hours at $980\,^{\circ}\mathrm{C}$ the alloy becomes an icosahedral quasicrystal and the corresponding Bragg pattern appears. After this, the material remains sensitive to low-temperature annealing at $600\,^{\circ}\mathrm{C}$. Such annealing in the course of one to two hours may double or even triple the resistivity at 4 K, with the quality of the Laue diffraction pattern remaining the same or even improving.

The temperature dependence of the resistivity of $Al_{70}Pd_{22.5}Re_{7.5}$ quasicrystals can be described following the theoretical scheme from Chapter 6 used there to describe conduction in the vicinity of metal–insulator transitions in ordinary materials (see eqns 6.14 and 6.20 and Figs. 6.2 and 6.4). Figure 7.11 depicts the conductivity as a function of $T^{1/3}$ (the four lower curves) or $T^{1/2}$ (the three upper curves). To distinguish between the various samples and between the various states of a single sample obtained in the low-temperature annealing process, the value σ_{10} of conductivity at 10 K may be chosen as a parameter (the scales along the horizontal axes in Fig. 7.11 have been selected in such a way that at this temperature the two scales coincide, as they also do at $T = 0$).

Figure 7.11 clearly shows that for all measured functions $\sigma(T)$ a linear extrapolation on the selected scales makes it possible to determine $\sigma(0)$. For the three upper states with $\sigma(0) \gtrsim 6(\Omega \cdot \mathrm{cm})^{-1}$ we can assume that

$$\Delta\sigma \simeq \sigma_{10} - \sigma(0) \lesssim \sigma(0). \tag{7.6}$$

This makes it possible to assume that the temperature-dependent part of the conductivity is a quantum correction, and this is why the function $\sigma(T)$ looks
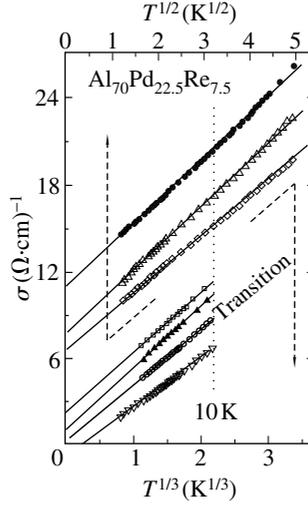
FIG. 7.11.   Temperature dependence of the conductivity of the $Al_{70}Pd_{22.5}Re_{7.5}$ quasicrystal (Wang et al. 1998, 1999, 2000). In the immediate vicinity of the metal–insulator transition, the dependence, when represented by a function of $T^{1/3}$, is a straight line (the four lower states). In the bulk of the metallic region, the dependence becomes a straight line when represented by a function of $T^{1/2}$ (the three upper states). The states can be labelled by the magnitude of the conductivity σ_{10} at 10 K.

like a straight line in the $(T^{1/2}, \sigma)$ plane. For the four lower states with $\sigma_{10} \lesssim$ 12–14 $(\Omega \cdot cm)^{-1}$ we have the opposite of eqn (7.6). This means that these states are in the critical vicinity of the metal–insulator transition. Hence, when depicted in the $(T^{1/3}, \sigma)$ plane, the function $\sigma(T)$ is represented by a straight line:

$$\Delta\sigma \equiv \sigma(T) - \sigma(0) \propto T^{1/3} \qquad (7.7)$$

and we can conclude that the metal–insulator transition occurs at the state with $\sigma(10\,K) \equiv \sigma_{10} \simeq 9\,(\Omega \cdot cm)^{-1}$.

For states with smaller values of σ_{10} low-temperature transport is realized through the hopping conduction mechanism. This is illustrated by Fig. 7.12. The diagram shows that the conductivity of high-resistance $Al_{70}Pd_{22.5}Re_{7.5}$ quasicrystals obeys the Mott law

$$\ln \sigma \sim T^{-1/4}. \qquad (7.8)$$

Such temperature dependence implies that near the Fermi level the density of states of the electronic spectrum has a constant, finite value (cf. eqns 4.17–4.19 and Fig. 4.6).

Thus, low-temperature annealing of $Al_{70}Pd_{22.5}Re_{7.5}$ quasicrystals improves the conditions for the propagation of electromagnetic waves, but impedes electron

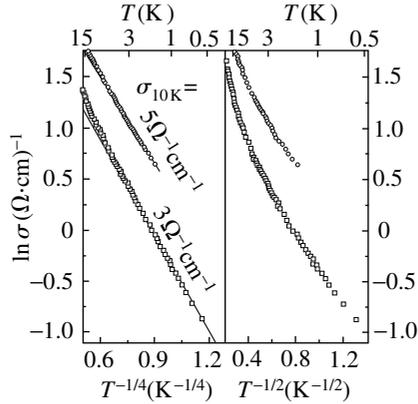FIG. 7.12. Mott law for the conductivity of $Al_{70}Pd_{22.5}Re_{7.5}$ quasicrystals in the insulator region (the conductivities at 10 K are 5 and 3 $(\Omega \cdot cm)^{-1}$) (Wang et al. 1998, 1999, 2000). The dependence becomes a straight line only when $\ln \sigma$ is plotted as a function of $T^{-1/4}$.

movement. The DC transport properties fully correspond to the pattern of metal–insulator transitions as the parameter $a_B n^{1/3}$ decreases (Chapters 5 and 6).

Note: Although the tendency of the resistivity to increase as the Bragg reflections get narrower is a characteristic feature of many families of quasicrystals, so far the metal–insulator transition has been observed only in Al–Pd–Re. The maximum resistivity values at 4 K for the Al–Cu–Fe, Al–Cu–Ru, and Al–Cu–Mn systems are smaller than the value for Al–Pd–Re by factor of 10–100. Even these values are higher than the value of ρ^* which follows from eqn (7.1) by a factor of 10–100.

Let us try to understand how the insulator $Al_{70}Pd_{22.5}Re_{7.5}$ is formed. The $Al_{70}Pd_{22}Mn_8$ structure is the most thoroughly studied one and differs from $Al_{70}Pd_{22.5}Re_{7.5}$ in only one aspect, i.e. Re is replaced by isovalent Mn. The quantitative characteristics of these quasicrystals can be assumed to be the same. The structure of the $Al_{70}Pd_{22}Mn_8$ and $Al_{70}Pd_{22.5}Re_{7.5}$ quasicrystals is based on high-symmetry, close to spherical configurations consisting of 51 atoms (see Fig. 7.10). According to diffraction data, the density of the atoms in these substances is close to $6 \cdot 10^{22} \, cm^{-3}$. Since the atoms of the transition elements "grab" some of the three valence electrons of aluminum, the number of the remaining "potentially metallic" electrons is somewhat smaller than two per atom, i.e. about $10^{23} \, cm^{-3}$. For a substance with such a huge electron concentration to be an insulator, the electrons must reside in deep potential wells, or traps. In intermetallic binary melts, the traps are configurations presented in Fig. 7.5, while in quasicrystals, they are the high-symmetry configuration of Fig. 7.10 which have levels for about 90 of the former valence electrons. Under favorable condition

only one to two electron from the upper levels may leave the trap. Hence the initial electron concentration is reduced by a factor of 100, after which the more or less standard models describing the metal–insulator transition can be employed.

The arrangement of the levels in all atomic configurations that are bricks of the quasicrystal is, in the zeroth approximation, the same. If the configurations were arranged periodically, the levels would transform into bands in accordance with band theory and the electrons from the upper levels would find themselves in the metallic band. However, in a quasicrystal there are many ways in which the neighboring configurations can be arranged in relation to a given configuration. According to the model with structural disorder (Section 5.5), each variant of the surroundings corresponds to a specific shift of the levels in the given configuration. Let us illustrate this using Penrose tiling as an example, for which we return to Fig. 7.9.

The distance between a given site and a neighboring site can be equal to the length a of the rhombus side or to the length of the smaller diagonal of either the narrow rhombus, $a_1 = 0.62a$, or the wide rhombus, $a_2 = 1.18a$. However, the number of variants of the surroundings, which determine the shift of the level of a specific site, is very large. Sites with close neighbors at distance a_1 may form resonant pairs or triplets. For instance, sites 1 and 2 each have five neighbors at a distance a, but all five neighbors of site 1 enter into resonant pairs or compact triplets with pairwise distances $a_1 < a$, while site 2 has no such neighbors; site 3 has six neighbors at a distance a, but three of these neighbors form a compact triplet; site 4 has seven neighbors at a distance a, but six of these neighbors form two compact triplets, etc. As a result, a single level, which initially was the same for all configurations (sites), becomes a band. Whether or not the states in this band are localized depends on the parameter $a_B n^{1/3}$, where the a_B is the decay length of the wave function outside the configuration well.

The very fact that the conductivity of $Al_{70}Pd_{22.5}Re_{7.5}$ is so small is, apparently, caused by the specific combination of the parameters of the configuration well, which makes the decay length a_B smaller than in other quasicrystals. The low-temperature annealing of $Al_{70}Pd_{22.5}Re_{7.5}$ leads to "internal repair" of the configuration wells in the quasicrystal accompanied by a decrease in the leakage of the wave function from the well, i.e. a decrease in the effective decay length a_B.

7.3. Metal–insulator transition in systems with high electron density

Processes that form the electronic spectrum in two-component melts with an alkali metal as one of the components and in quasicrystals have proved to be very similar. In such systems the effective carrier concentration decreases and the screening weakens during material structuring, which makes an ordinary metal–insulator transition possible.

The overall scheme is as follows. Suppose that each configuration contains N valence electrons. The potential produced by the ion cores of the atoms in a

configuration is so strong, i.e., the potential well is so deep, that the electronic spectrum of these N electrons becomes radically transformed and the electrons occupy "positions" on an intrinsic ladder of levels. Only one or two electrons on the upper levels have a chance of leaving the well. As a result, the concentration of "potentially itinerant" electrons becomes of order n/N, where n is the concentration of the "initially metallic" valence electrons. In two-component melts, N is of order 10, while in quasicrystals it is of order 100. Thus, the metal–insulator transition occurs in a system with a reduced carrier concentration.

The same effect can be described differently if we consider configurations as quantum dots in 3D space. The concentration of such dots is of order n/N, with each dot containing N electrons. When one electron leaves a quantum dot, the dot's charge increases by e, and this requires energy of order

$$\varepsilon_{\mathrm{e}} \approx e^2/r, \tag{7.9}$$

where r is the radius of the quantum dot. This quantity is similar to the Hubbard energy, eqn (5.45), in the theory of Mott transitions. At the same time, e^2/r is the Coulomb energy of an isolated metal sphere of radius r carrying charge e or the energy of the capacitor that appears in the theory of the Coulomb blockade in nanostructures (see Section 8.2). On the metal side of the metal–insulator transition, the electric field of a charged dot is screened and the energy ε_{e}, eqn (7.9), is insignificant. On the insulator side, there is no screening by free carriers and the number of charged dots is determined by comparing the energy equation (7.9) with the temperature. When $\varepsilon_{\mathrm{e}} \ll T$, the number ν of charged dots is exponentially small:

$$\nu = (n/N) \exp(-\varepsilon_{\mathrm{e}}/T). \tag{7.10}$$

Since conduction under these conditions is determined by tunneling between charged and uncharged dots, ν acts as the number of carriers. This standard line of reasoning used to describe granular metals (see Section 8.2) determines the activation nature of the conduction.

The importance of replacing n with n/N can be illustrated by the following experimental fact: when an amorphous alloy is transformed by annealing into a quasicrystal and it remains metallic, its resistivity nevertheless often increases severalfold. Localization is also substantially enhanced by the absence of translational symmetry and of universal short-range order in the mutual arrangement of configurations. Irregularities in this mutual arrangement prevent resonant tunneling. Of course, the presence of translational symmetry by itself cannot guarantee metallic conduction. However, near the limiting values of concentrations, $n/N \gtrsim n_{\mathrm{Mott}}$, disorder is essential. Indeed, when the CsAu alloy crystallizes, its resistivity decreases by a factor of 10 (see Fig. 7.7).

Thus, the two classes of condensed media briefly discussed in this chapter provide an affirmative answer to the question about possibility to localize a system of valence electrons in a medium consisting only of metal atoms. Such localization is realized through the formation of molecule-like configurations at least in two cases: in two-component melts with an alkali metal as one of the

components, and in quasicrystals. In addition, irregularity in the relative position of the configurations is needed for the localization.

References

Calaway, W.F. and Saboungi, M.-L. (1983). J. Phys. F: Met. Phys. **13**, 1213.

Goldman, A.I and Kelton, R.F. (1993). Rev. Mod. Phys. **65**, 213.

Hensel, F. (1979). Adv. Phys. **28**, 555.

Hoshino, H., Schmutzler, R.W. and Hensel, F. (1975). Phys. Lett. **51A**, 7.

Janot, C. (1996). Phys. Rev. B **53**, 181.

Meijer, J.A., Vinke, G.J.B., and van der Lugt, W. (1986). J. Phys. F: Met. Phys. **16**, 845.

Stadnik, Z.M. (ed) (1999). *Physical Properties of Quasicrystals*, Springer.

Schmutzler, R.W., Hoshino, H., Fischer, R. and Hensel, F. (1976). Ber. Bunsenges. phys. Chem. **80**, 107.

van der Lugt, W. and Geerstma, W. (1987). Can. Journ. Phys. **65**, 326.

Wang, C.R., Lin, S.T., et al. (1998). J. Phys. Soc. Japan **67**, 2383.

Wang, C.R., Lin, S.T., et al. (1999). J. Phys. Soc. Japan **68**, 3988.

Wang, C.R., Lin, S.T., et al. (2000). J. Phys. Soc. Japan **69**, 3356.

Xu, R., de Longe, T., and van der Lugt, W. (1992). Phys. Rev. B **45**, 12788.

8

GRANULAR METALS

The basic transport properties of granular metals are reviewed by Abeles et al. (1975).

8.1. Morphology and classification

Hereafter, a material consisting of randomly arranged small regions (*grains* or *granules*) with essentially different conductivities (in the limit, a mixture of metal and insulator regions) is referred to as a granular material. A random potential of such a material is usually characterized by a length considerably exceeding the interatomic distances up to macroscopic lengths. Let x be the fraction of d-dimensional space occupied by metal grains. However, the x value is insufficient for the characterization of a material. It is clear that the conductivities of two materials with the same value of x but with spherical and fiber metal inclusions would be different. The material morphology is understood here as the shape of the grains. It depends on numerous factors and may be quite diverse. As an example, Fig. 8.1 shows electron micrographs of indium films deposited onto SiO_2 substrates at room temperature obtained in a scanning electron microscope. Indium does not wet the substrate surface.

Atoms possessing a certain thermal energy reach the substrate and, diffusing along its surface, form small randomly located droplets (Fig. 8.1a). During further deposition, these droplets increase in size and merge together into drops with larger diameters (Fig. 8.1b). Then the metal regions acquire elongated shapes. It seems that with an increase of the areas of the drop contact with the substrate, some regions with substantial cohesion are formed in the drop centers. During further merging of large drops, regions with high cohesion play the role of pinning centers for a moving mass of the material and lower the symmetry of the metal grains thus formed (Fig. 8.1c). And finally, prior to the formation of a continuous metal film, when the relative area of the $1-x$ gaps between the metal regions becomes rather small, these gaps acquire the shapes of relatively thin branching fibers (Fig. 8.1d). This process is also governed by a certain combination of the wetting and cohesion rules for a deposited material and the substrate, but we limit our consideration only to the morphological characteristics of the structure.

Divide the d-dimensional space into elementary volumes a^d and assume that the properties of the material inside these volumes are constant and that the properties of two different volumes are independent. Thus, we have reduced the spatial problem to the problem on a lattice with period a, which allows us to use the simplest models of percolation theory. For the structure shown in Fig. 8.1a,
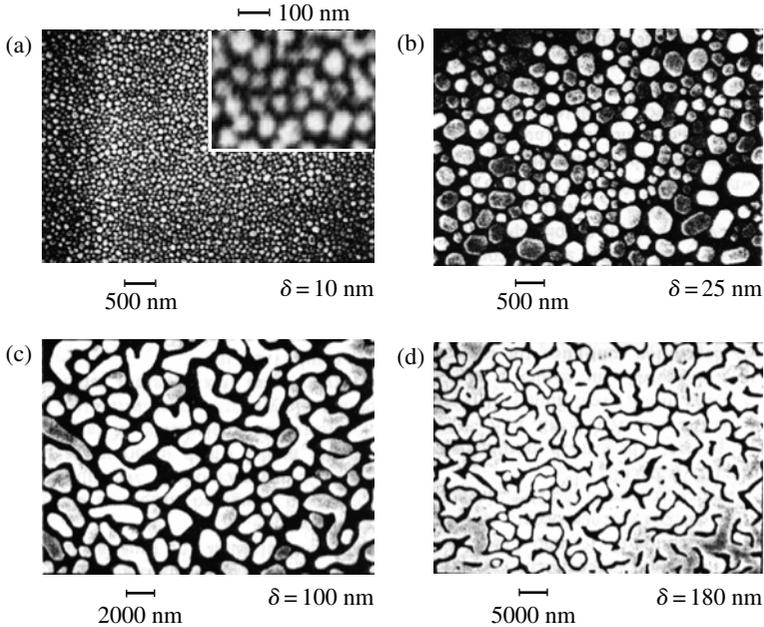
FIG. 8.1.   Micrographs of island indium films obtained in a scanning electron
microscope. Metal regions are light. The scale and the average film thickness δ
are indicated for each micrograph. The inset shows part of the micrograph (a)
at a higher magnification (Yu et al. 1991).

the characteristic size a of metal drops is of the order of $0.05\,\mu$m and for the
structure in Fig. 8.1b, it is of the order of $0.2\,\mu$m. The fact that the variation
of the fraction x of volume of metal also changes the scale is not important.
It is more important that the average transverse size of the metal regions in
Fig. 8.1c is less than the average longitudinal size, $b \approx (2\text{--}3)a$. This signifies that
there exists a correlation between the properties of the b/a neighboring sites on
a square lattice with period of the order of a ($a \approx 1\,\mu$m).

In terms of mathematics, the lowering of the local symmetry of the structure
is described by special correlators, whose introduction considerably complicates
the situation, so that the simplest models of percolation theory (bond and site
problems) become inapplicable. This is one of the possible explanations of the
experimentally established fact that the critical value $x_c = 0.82 \pm 0.02$ for the rel-
ative area of indium on the SiO_2 surface providing percolation is much larger than
the known critical values in similar problems of percolation theory. The second
explanation is the loss of symmetry between the metal and nonmetal regions. For
the structures in Figs 8.1a and 8.1b, the regions between drops have the same
order of magnitude as the drops themselves, whereas in Fig. 8.1d the insulating
regions between drops are much less than the metal regions. Nevertheless, these
small regions between drops still guarantee the insulating properties of the film.

Thus, the critical value x_c is strongly dependent on such physical factors as the accommodation coefficient of the incident atoms to the substrate, surface tension, cohesion forces, etc. Therefore, deposition of other metals under the same conditions is characterized by other x_c values: $x_c = 0.86 \pm 0.02$ for Sn and $x_c = 0.67 \pm 0.02$ for Pb.

Note: Along with metal granules in an insulating matrix, one may also imagine insulator granules in a metal matrix. However, in what follows, when using the term "granule" we always mean "metal granule." Moreover, as has already been indicated, we call a granular material also a material with the structure of the type shown in Fig. 8.1d, which, strictly speaking, has no granules at all.

In the system shown in Fig. 8.1, the role of an insulator separating the metal grains is played by the vacuum. However, this role may also be played by an insulator. If a metal and an insulator making a pair do not dissolve in one another, they may form a mixture of small metal and insulating regions (grains). This mixture, termed a *cermet*, is obtained, e.g., in simultaneous deposition of both components onto an insulating substrate. The scale of the structure thus formed is controlled by the physical–chemical factors acting during deposition. Depending on these factors and also on the deposition time and the film thickness, it is possible to obtain two- or three-dimensional structures. Figure 8.2 shows an electron micrograph of the $Au + Al_2O_3$ cermet in the range of existence of an infinite metal cluster (the "plus" sign is used to distinguish this granular system from the system of Au film deposited onto the Al_2O_3 substrate).

Sometimes, it is also possible to preserve the spherical shape of the granules up to high metal concentrations, $x > x_c$. Figure 8.3a shows an electron micrograph of the structure of a film of granular Al in the matrix of amorphous Ge at the volume Al concentration $x \approx 0.66$ obtained in a transmission electron microscope.
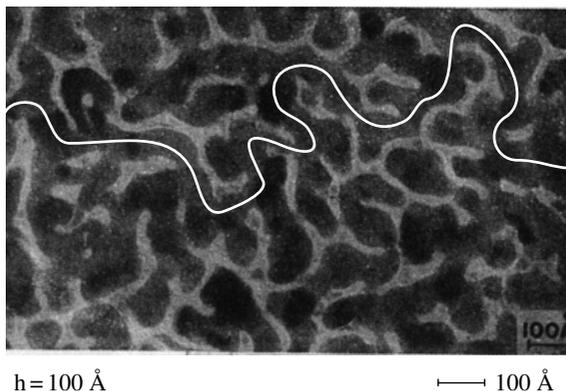

h = 100 Å   ⊢——⊣ 100 Å

FIG. 8.2. Micrograph of a granular $Au + Al_2O_3$ film of thickness h = 100 Å(Abeles 1975). Metal regions are dark. The white line is a continuous pass from left to right.
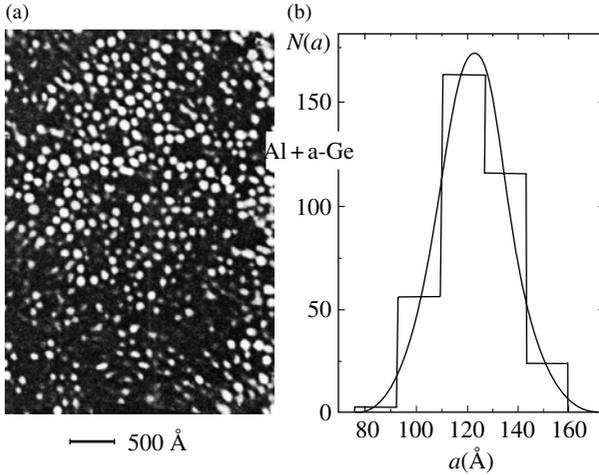
(a)   (b)

FIG. 8.3.   (a) Micrograph of a granular Al + Ge film obtained in a transmission
   electron microscope. Metal regions are light. (b) Histogram of the size distri-
   bution of Al grains and its approximation by a normal distribution (Shapira
   and Deutscher 1983).

It is seen that the Al component of the material consists of spherical granules.
Special measurements allowed one to determine the distribution of granules over
their diameters, which turned out to be rather narrow (Fig. 8.3b).

   All the above systems show a finite conductivity at a certain stage of increas-
ing a relative volume of the metal in the material, in other words, the systems
undergo a metal–insulator transition. This transition is often called a percolation
transition, which implicitly means that this transition is governed by pure geo-
metrical factors, so that the transition is a pure classical macroscopic transition.
Indeed, the percolation laws are invariant with respect to scale, so that it is pos-
sible to imagine percolation, e.g., in a system of metal balls from a ball-bearing
randomly located on a plane and fixed there with solidified paraffin. However, if
the characteristic lengths of the system also include microscopic lengths, some
specific physical factors may also arise and even play a key part. We are interested
just in the latter systems.

   On the other hand, if all the characteristic lengths, such as size of grains, width
of barriers between them, etc., become too small (of the order of interatomic
distances), we come back to a homogeneous disordered material. Theoretical
descriptions of granular and homogeneous disordered materials differ a lot but
their macroscopic physical properties are similar. At least, metallic conductance
and metal–insulator transition exist in both. Hence, it is desirable to set bound-
aries for these different classes of disordered systems. However, the positions of
the boundaries depend on the physical properties we are interested in.

   For an illustration of this statement, let us take an important quantitative
parameter of a granular system – the level spacing $\delta\varepsilon$ caused by the electron

confinement in grains. It can be expressed with the help of the grain volume a^3 and the density of states at the Fermi level, g_F, in a bulky metal,

$$\delta\varepsilon = (g_F a^3)^{-1}. \tag{8.1}$$

To make the necessary estimates, we may assume that $a = 50\,\text{Å}$ and $\delta\varepsilon \approx 10\,\text{K}$.

If a bulky metal is a superconductor with the critical temperature T_c and superconducting gap Δ, the relationship

$$\delta\varepsilon \approx \Delta \simeq T_c. \tag{8.2}$$

determines the minimum size $a_{sc} = (g_F \Delta)^{-1/3}$ of an isolated grain for which the notion of a superconducting state retains its sense. If $a > a_{sc}$, then the grains undergo a superconducting transition at the same temperature as a bulky metal, and the behavior of the material as a whole depends on the interactions between the granules. It is this behavior that is typical of thin Pb films deposited onto mirror-smooth SiO surfaces precooled to helium temperature – see Fig. 8.4a taken from Frydman (2003). With this deposition technique, Pb atoms coagulate into granules which, before merging, amount to a diameter of $200\,\text{Å}$ and a height of 50–$80\,\text{Å}$. All the films whose thickness exceeds some critical value have a superconducting transition at the same temperature $T_c \approx 7\,\text{K}$. The thinner films have no global transition at all but the $\rho(T)$ curves for these films retain some particular feature at the temperature $T \approx T_c$ indicating the existence of the superconducting transition in separate grains.
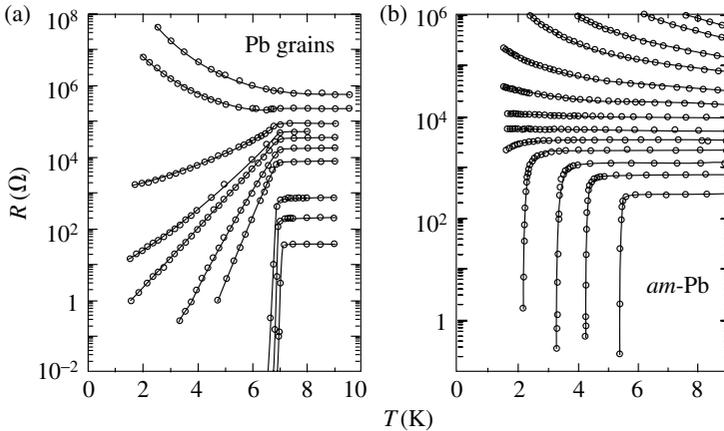
FIG. 8.4. Evolution of the superconducting transition in thin Pb films on SiO substrates with increase of their thickness (from above downward) (Frydman 2003). (a) An example of a granular system. The films are evaporated directly on SiO. The transition temperature is constant. (b) An example of a fine inhomogeneous system. The films are evaporated above a thin layer of amorphous Ge. The transition temperature changes along with the normal resistance.

If the inverse inequality $a < a_{sc}$ is valid, then a separate grain cannot become superconducting on its own. In terms of the superconducting transition, such material is homogeneously disordered and it may have a bulk superconducting transition at some temperature T_c' determined by its average characteristics. The transition temperature may smoothly vary with the change of these characteristics. The Pb films on the SiO substrate evaporated above an additional intermediate 5 Å layer of amorphous Ge demonstrate the correlation between the temperature T_c' and film resistance (Fig. 8.4b).

For a normal metal, the granulation criterion is different and depends on temperature. The relationship

$$\delta\varepsilon \approx T \tag{8.3}$$

determines the minimum grain size at which the notion of a delocalized electron still makes sense. If the range of thermal spread includes only one electron level, then, generally speaking, it should rather be considered as a localized level, and the quantity a should be considered as the wave-function dimension, i.e., the localization length ξ.

The evolution of granular systems may proceed in two possible ways. The first way is associated with the variation in x. Figures 8.1–8.3 illustrate such systems. A metal–insulator transition in these systems seems to be based on percolation. Since the variation in x results in a change of the average concentration of delocalized electrons in the material, it is timely to recollect here the Mott transition. The second possible way of evolution may be described as follows. At a sufficiently high x value, the properties of the barriers between the granules, e.g., the barrier height, would also change. Here, it is also possible to formulate the criterion of a cross-over from a granular system to homogeneously disordered one. This may be made by comparing the level spacing $\delta\varepsilon$ with the transfer integral of the wave functions of electrons from the neighboring granules, which quantitatively describes the efficiency of the insulating barriers. Here, it is timely to recollect the Anderson transition.

In practice, it is rather difficult to distinguish between these types of evolution; however, conditionally, we consider systems of the first type in the next section and systems of the second type in the last section.

8.2. Coulomb blockade and metal–insulator transition

Figure 8.5a shows the resistivity of the cermets in the $Au + Al_2O_3$ system (Fig. 8.2) as a function of the relative metal concentration x measured at two different temperatures.

Note: The range of the resistivity variation along the ordinate exceeds 12 orders of magnitude.

We can clearly distinguish two concentration regions. The region $1 \geqslant x \gtrsim 0.4$ is that of a metal: the resistivity ρ in it is considerably low, only weakly depends
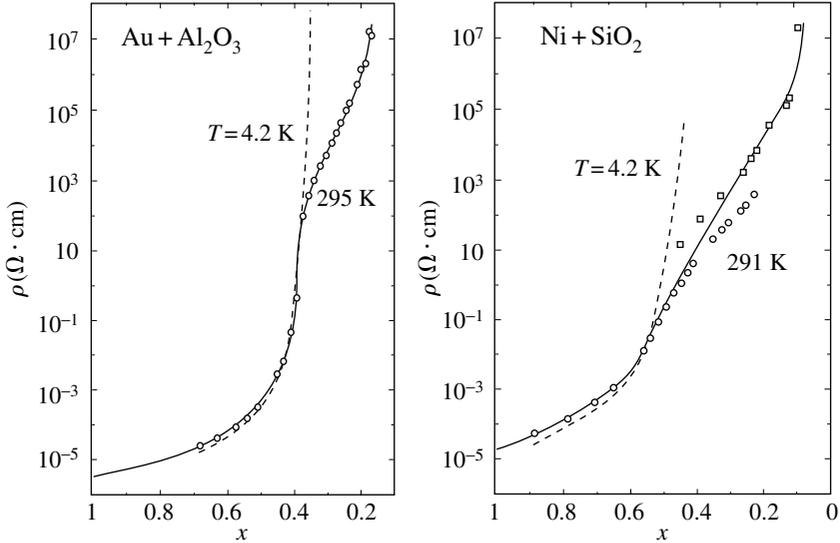
FIG. 8.5.   Resistance of granular $Au + Al_2O_3$ and $Ni + SiO_2$ films as a function of x at two temperatures (Abeles 1975). Data shown by dash lines are obtained at helium temperature, those by solid lines and symbols are obtained at room temperature. Dots and squares on the right plot are taken from two different experiments; the solid line is the result of averaging.

on temperature, and gradually increases with a decrease in x. The boundary between the two regions lies in the vicinity of the value $x = x_c \approx 0.38$. Finally, the insulating region $x \lesssim 0.38$ is characterized by a dramatic increase in resistivity with a decrease in x and a pronounced temperature dependence of resistivity, $\rho(T)$. A similar dependence of $\rho(x)$ in the $Ni + SiO_2$ system is shown in Fig. 8.5b. Qualitatively, this system behaves in the same way as the previous system. In particular, in the vicinity of the critical value x_c, the derivative $\partial\rho/\partial T$ of the function $\rho(x, T)$ changes sign (solid and dash lines cross near x_c). However, the critical value, x_c, for the latter system is different. We have already indicated a similar discrepancy in the x_c values for island films.

The description in terms of the percolation model assumes that at concentrations $x > x_c$, the current lines are inside the metal cluster, and at $x < x_c$, the carriers should tunnel through the insulator, at least at some parts of their path. Then, the temperature dependence $\rho(T)$ in the region where $x < x_c$ should be determined by the insulator properties. However, in fact, this is true only to some extent.

Figure 8.6 shows the resistivity of the materials of the granular $Au + Al_2O_3$ and $Ni + SiO_2$ systems as functions of temperature in the insulating mode, i.e., at $x < x_c$. Since in the temperature range studied, the resistivity changes by several orders of magnitude, it is possible to determine the functional dependence of the
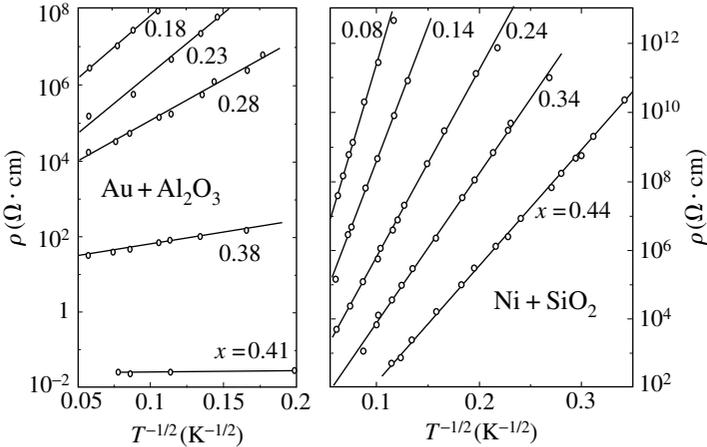
FIG. 8.6.   Resistivity of granular $Au + Al_2O_3$ and $Ni + SiO_2$ films as functions of temperature at low x values, $x < x_c$ (except the curve $x = 0.41 > x_c$ at the left panel) (Abeles 1975).

resistivity $\rho(T)$ as

$$\rho(T) = \rho_0 \exp(T_0/T)^{1/2}. \tag{8.4}$$

This dependence can be reliably distinguished from the dependence of types $\exp(T_0/T)$, or $\exp(T_0/T)^{1/4}$. The corresponding bulky insulators, Al_2O_3, and SiO_2, show no functional dependence determined by eqn (8.4). Moreover, the slope of the straight lines in Fig. 8.6 determined by the value of T_0 from eqn (8.4) depends on x. Therefore, the transport in the materials considered here is controlled not solely by the insulator, but by the entire configuration of the grains in the material. Thus, we arrive at the experimental facts that should be clearly understood and interpreted.

The theoretical model is based on two fundamental assumptions.

1. Tunneling may take place between neighboring granules. In this case, the invariance with respect to the lattice scale characteristic of the percolation problems is lost, and no systems similar to a set of metal balls from the ball-bearings are possible any more. Tunneling from any granule to all the neighboring granules is not necessary any more either. More exactly, we may assume that a set of granules which may exchange carriers via tunneling would form a well-developed infinite cluster. This assumption defines the property which is usually called carrier mobility, $\mu = e\tau/m$, where $\tau = \tau_{tun}$ is the tunneling time for a charge e with effective mass m. Indeed, since tunneling is the main mechanism which constrains the charge motion in space, the charge mobility should be proportional to the

tunneling probability

$$\frac{1}{\tau_{\text{tun}}} \propto \exp(-\beta a), \quad \frac{\beta = 2(2mU)^{1/2}}{\hbar}, \tag{8.5}$$

where a and U are the barrier width and height, respectively.

2. Each charged metal granule induces an electric field in the gap between this granule and other granules and, thus, plays the role of a microcapacitor plate. The capacitance of such a local capacitor is of the order of the product of the granule radius a and the permittivity κ of the surrounding insulator (capacitance of a remote sphere)

$$C \approx \kappa a. \tag{8.6}$$

If the capacitor charge is q, then the field energy in it equals $q^2/2C$. Therefore, for an additional electron $q = e$ to be located at the granule, it must have Coulomb energy $\varepsilon_C \approx e^2/\kappa a$; whence it follows that the charge concentration n is proportional to

$$n \propto \exp(-\varepsilon_C/T) \approx \exp(-e^2/\kappa a T). \tag{8.7}$$

Note: In this case, the material remains electrically neutral, because the numbers of electrons and holes (positively and negatively charged granules) are approximately equal. The energies of both electrons and holes are measured from the Fermi level.

The energy ε_C is far from being low. For a $50\,\text{Å}$ granule at $\kappa \simeq 10$ it is of the order of $300\,\text{K}$. This signifies that, at low temperatures, the number of carriers is exponentially small. It is this fact that limits conductivity, whence the term *Coulomb blockade*. It is usually applied to isolated nanostructures such as pairs of tunneling contacts with low capacitance, with the quantity ε_C describing a certain particular configuration. However, the inequality

$$\varepsilon_C \gg T \tag{8.8}$$

may also determine the properties of the material as a whole.

Formally, eqns (8.5) and (8.7) allow us to single out the most important exponential factors that enter the expression for resistivity, $\rho(T) = \sigma^{-1}(T)$. Since conductivity σ is proportional to the product of concentration n and mobility μ, we obtain

$$\rho \propto (n\mu)^{-1} \propto \exp(e^2/\kappa a T + \beta a). \tag{8.9}$$

Since the exponent in eqn (8.9) consists of two terms and since the length a in one of these terms is in the numerator, whereas, in the other one, it is in the denominator, there exists some magnitude of a

$$a_{\min} = e(\kappa \beta T)^{-1/2}, \tag{8.10}$$

at which the exponent is minimal. The set of grain sizes a in a real material is characterized by a certain dispersion. The existence of the minimum signifies

that the current flows mainly along the chains of grains with $a = a_{\min}$, and the material resistance is described by eqn (8.4) with T_0 equal to

$$T_0 = 2e(\beta/\kappa)^{1/2}. \tag{8.11}$$

Before discussion of the result obtained, we have to make an important stipulation. The length a has different meanings in eqns (8.5) and (8.7). In eqn (8.5), it has the meaning of a gap between grains, whereas in eqn (8.7), it has the meaning of a grain size. Micrographs in Figs 8.1–8.3 show that these two lengths are different. In fact, the above assumption that these lengths are equal may be lifted and replaced by the much more realistic assumption that these lengths are proportional. This signifies that, after scaling, various portions of the grain metal become statistically equivalent. For such a model, the main conclusion remains the same as earlier with the only difference that eqn (8.11) acquires an additional multiplier – a root of the proportionality coefficient. Moreover, the main conclusion remains valid also for any functional relationship between the granule dimension a and the gaps a' between them *at the given x* under the condition that these two quantities are statistically dependent.

Thus, it turned out that at low concentrations of the metal phase, $x < x_c$, the current in the granular material flows nonuniformly and is concentrated mainly in the regions with the optimum average size of the granules. This optimum size depends on the temperature. Therefore, with the change of the temperature, the current distribution over the material should also change.

Compare the tunneling conductivity in a granular system with the hopping conductivity in the presence of a Coulomb gap. Both mechanisms are due to the same type of initial interaction (Coulomb interaction), result in the same functional dependence $\ln \rho \propto T^{-1/2}$, and have similar occasions for changing the main tunneling paths with temperature.

It is no accident that the two problems are similar. Reducing the granule size to zero, we transform the granules into charged or neutral impurity centers. In this limiting transition, one problem should be naturally transformed into the other. This comment evokes some expectations. An insulator with impurity centers has a Coulomb gap and in a metal with large number of impurities, the Coulomb electron–electron interaction results in the inimum density of states at the Fermi level. One has to expect an analogous situation also in a granular material. These predictions are confirmed by a tunneling experiment.

Figure 8.7a demonstrates the tunnel characteristics of the $Al–Al_2O_3–Ni+SiO_2$ structure in which one of the tunnel-junction sides is the film of a granular metal (in our case, the mixture of Ni (metal) and SiO_2 (insulator) not dissolving in one another). The films were 100 Å thick. Since at all the x values, the characteristic sizes of the metal granules were less than 50 Å, then, in terms of the processes forming the electron spectrum, the $Ni+SiO_2$ film should have a three-dimensional structure. The contact plane is a section of this structure. Tunneling may proceed only into the metal granules on the contact plane. The fraction x of these granules in this plane is the same as in the bulk.
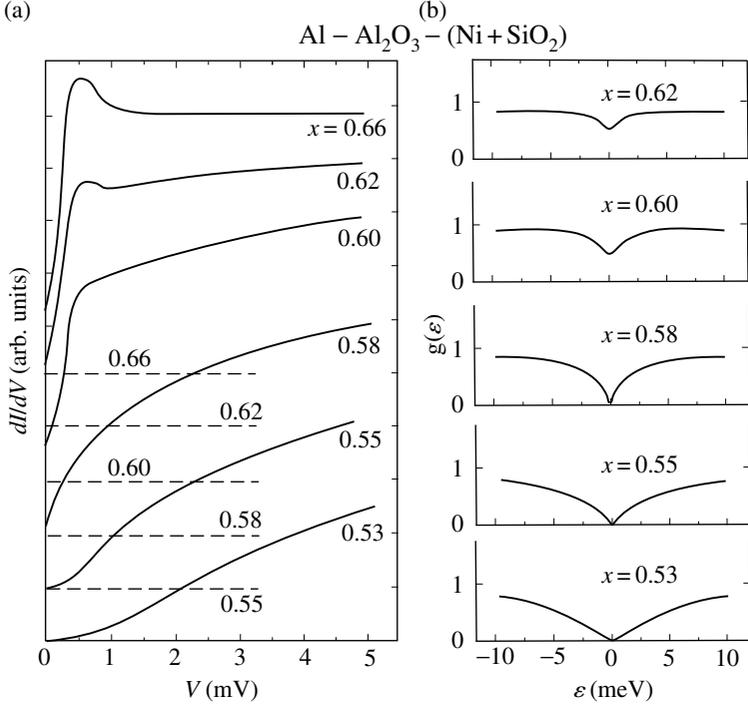
FIG. 8.7.   (a) Tunneling characteristics of the Al–Al$_2$O$_3$–Ni + SiO$_2$ structure
with a granular film as one of the electrodes. For a clearer representation, the
curves for different portion x of the metal surface (and volume) are displaced
upward. The dashed lines indicate the corresponding origin of each curve.
(b) The function $g(\varepsilon)$ for the granular film at different x values extracted from
the curves in panel (a) (Abeles 1975).

At high x values (in fact, already at $x = 0.66$, the upper curve), the presence of
insulating inclusions is not important and Ni + SiO$_2$ behaves as a conventional
metal. The structure of the dJ/dV curve is formed due to superconductivity
of the aluminum counter-electrode (cf. the analogous effect in Fig. B.2 of
Appendix B). The changes in the dJ/dV curves at lower x values are fully con-
trolled by a granular electrode, because there are no changes at the aluminum
electrode. Ignoring the inhomogeneity of the junction, one may use a standard
mathematical procedure consistent with eqn (B.1) in Appendix B for extracting
the density of states,

$$
J(V) \propto \int_{-\infty}^{\infty} g(\varepsilon - eV)g_1(\varepsilon) \left[f\left(\frac{\varepsilon - eV}{T}\right) - f\left(\frac{\varepsilon}{T}\right) \right] d\varepsilon, \quad f(x) = (\exp x + 1)^{-1}.
$$

$$(8.12)$$

Replacing $g_1(\varepsilon)$ with the density of states of the superconducting electrode measured at $x > 0.7$, we may extract the function $g(\varepsilon)$ for the granular electrode from each experimental dJ/dV curve. The result thus obtained is shown in Fig. 8.7b.

Comparing Fig. 8.7 with Figs B.3 and B.4 in Appendix B, we see that the evolution of the function of the density of states with the change in the control parameter in the vicinity of the metal–insulator transition is practically the same for granular and non-granular materials. In both cases, the spectrum has a minimum in the density of states at the Fermi level, which is transformed into a soft gap. According to Fig. 8.7, the critical value of the control parameter in the Ni + SiO$_2$ system equals $x_c \approx 0.56$. The same value is also obtained from the curves in Fig. 8.5b. It is at this x value that the derivative $d\rho/dT$ changes its sign.

Thus, although tunneling occurs into individual granules, the function $g(\varepsilon)$ extracted from the experimental data reflects the state of the material as a whole and even records the metal–insulator transition taking place with the change of x. To interpret this phenomenon, we have to extend eqn (8.12) to the case where one of the electrodes is inhomogeneous. The fact that tunneling occurs only into the metal fraction x of the total contact area is of minor importance: it only reduces the effective area of the contact and increases the total resistivity by a factor of $1/x$. Much more essential is the inhomogeneity of the metal region itself. The part $P < x$ of the electrode area belongs to an infinite metal cluster. If an electron tunnels into this part of the contact, the processes observed are the same as for a homogeneous contact. The corresponding part of the current is described by eqn (8.12) with the density of states $g(\varepsilon) = \text{const.} = g_0$.

The $x - P$ part of the contact consists of individual granules of size a with size distribution function $D(a)$ normalized by the condition

$$\int_0^\infty D(a)a^2 \, da = 1. \tag{8.13}$$

If an electron is to tunnel into such a granule, it should be given an additional energy $\varepsilon_C \approx e^2/\kappa a$. Therefore, the effective density of states g_{eff} that should be substituted into eqn (8.12) instead of g in order to describe the tunneling process is

$$g_{\text{eff}}(\varepsilon) = Pg_0 + (x - P) \int_0^\infty D(a)a^2 g_a(\varepsilon) \, da, \tag{8.14}$$

where the size-dependent density of final states $g_a(\varepsilon)$ is

$$g_a(\varepsilon) = \begin{cases} 0, & |\varepsilon| < \varepsilon_C(a) \\ g_0, & |\varepsilon| > \varepsilon_C(a) \end{cases} \tag{8.15}$$

(hereafter the value $\varepsilon = 0$ corresponds to the Fermi level). This is the function $g_{\text{eff}}(\varepsilon)$ that is extracted from the experimental data with the aid of eqn (8.12) and which is depicted in Fig.8.7b at various x values.

Note: Equation (8.14) could have been transformed further by substituting there eqn (8.15) for $g_a(\varepsilon)$, attracting a model function $D(a)$, etc. However, we consciously avoid these transformations, because eqn (8.6) for capacitance, $C \approx \kappa a$, at the beginning of the chain of these transformations which corresponds to the capacitance of a remote ball is a too-crude approximation. It should be considerably refined to be applied to specific granular structures.

However, even the unamended eqn (8.14) shows that with the disappearance of an infinite cluster (occurrence of a percolation transition and P tending to zero), the density of states at the Fermi level becomes zero.

Thus, neither transport measurements in the vicinity of a metal–insulator transition, nor tunneling experiments allow one to distinguish between a percolation transition in a granular system and, e.g., a Mott transition in a uniformly disordered system. This was to be expected from the most general considerations: the correlation length ξ diverges at the transition. If we are so close to this transition that

$$\xi > a, \tag{8.16}$$

the granulation becomes unimportant. Of course, a should not be too large (as in the system of balls from the ball-bearing), otherwise the transition region would be too small to be implemented in practice.

8.3. Fractal granular metals

Granular materials may be obtained not only by evaporation or sputtering. Other methods of preparation may give rise to structures of other types. As an example, consider the processes of solid-phase amorphization of some metastable alloys.

There are some alloys, in particular, antimony-based ones, Sb–Zn, Sb–Cd, and Sb–Ga alloys, and also the Al–Ge alloy, whose equilibrium phases under atmospheric pressure have low coordination numbers, low densities, and whose Fermi levels are in the forbidden energy band, whereas their high-pressure phases have high coordination numbers, high densities, and the Fermi levels in the region of allowed energies. Cooling a high-pressure chamber with the alloy to liquid-nitrogen temperature, lowering the pressure to normal, and opening the chamber, we obtain a high-pressure phase of the alloy in the metastable state. This metastable metal phase at liquid-nitrogen temperature may be preserved for an infinitely long time. Then, placing the thus obtained sample into a holder with clamping contacts and measuring the temperature dependence of the resistance, $R(T)$, we may reveal a conventional temperature run of the resistance and a superconducting transition at liquid-helium temperature. During sample heating, the metastable state should be transformed into the stable one. Under rather slow heating, the transformation proceeds in two stages. First, at a certain temperature T_1, the thermal energy is sufficient to destroy the metastable lattice, but not sufficient for growth of a stable phase. Therefore, the metastable phase is transformed into the amorphous one with the coordination number, density, and conductivity being the same as in the stable

phase. We call this intermediate phase an insulator and the process, amorphization. With further heating, the sample is crystallized, and an amorphous insulator is transformed into a crystalline insulator. We are not going to discuss this stage and the extremely interesting thermodynamic details of the whole chain of the solid-phase transformations; they may be found in the reviews of the effect.

Since the specific volume of the material considerably increases during its amorphization (in the Sb–Zn and Sb–Ga alloys, the volume increases by about 25%), the amorphous-phase seed in the metastable phase cannot grow isotropically, because of an increase in the local pressure. However, seeds may grow in the shape of flat disks, and their moving growth front is similar to a razor blade cutting the metastable phase. The layer behind the moving edge thickens very slowly, because the surrounding medium is compressed. Nevertheless, the stress increases with time, and, therefore, the layer roughness and its nonuniformly deformed regions act as nuclei of layer branching. Thus, the layers of an insulating amorphous phase propagate due to branching almost without increasing layer thickness, whereas the conducting channels during amorphization become thinner and start winding.

Very slow heating of a sample from liquid-nitrogen temperature accompanied by measuring of its electrical resistance, allow one to fix the moment when resistance starts increasing due to amorphization. The process may be interrupted at any moment by sharp cooling of the sample. Therefore, the transition to the amorphous state may be performed in a stepwise manner. The $R(T)$ curve may be measured at any intermediate state to reveal the occurrence of the superconducting transition.

These measurements yield the following main result. *Gradual transformation may increase the sample resistance by several orders of magnitude practically without a change of the temperature of the superconducting transition T_c.* As an illustration, Fig. 8.8 shows the $R(T)$ curves for the $Sb_{57}Cd_{43}$ and $Sb_{50}Ga_{50}$ alloys. All the curves are normalized to the sample resistance in the given state $R(6\,K)$ at the temperature $T = 6\,K$. The resistance values, $R(6\,K)$, are different for different states, therefore the value $q = \lg(R/R_0)_{T=6\,K}$ at each curve indicates the order of magnitude of the change in the sample resistance R after amorphization in comparison with the sample resistance R_0 in the initial state prior to amorphization. Thus, the value $q = 6.9$ indicates that the resistance at $T = 6\,K$ is by almost seven orders of magnitudes higher than in the initial state. The initial values of resistivity $1/\sigma$ corresponding to $q = 0$ differ from sample to sample but always remain in the range $10 - 100\,\mu\Omega \cdot cm$. This indicates that the mean free path in the initial "homogeneous metal material" is of the order of $l \approx 10k_F^{-1}$.

Note: The superconducting transition in granules in a granular system does not necessarily signify a drop of the total resistance to zero. The main contribution to resistance along the current channels may be endowed either by granules or by space between these granules. If the resistance is controlled by tunneling between granules, then the superconducting transition in granules does not necessarily diminish the total resistance – it may even increase it. This is demonstrated by the curve with $q = 7.8$ in Fig. 8.8b.
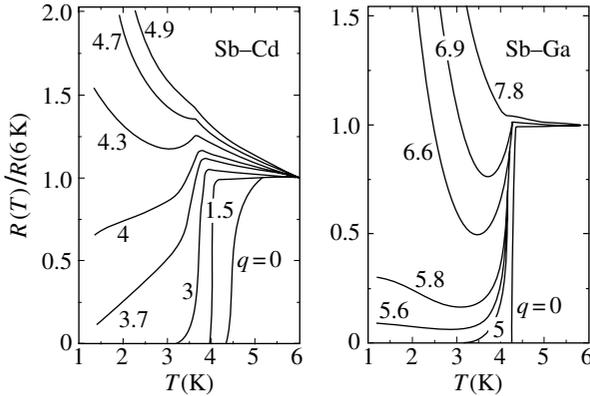
FIG. 8.8. Superconducting transition in the metastable $Sb_{57}Cd_{43}$ and $Sb_{50}Ga_{50}$ alloys at different stages of sample amorphization. All the curves are normalized to the resistance at $T = 6\,K$ at the corresponding amorphization stage (Gantmakher et al. 1993). The parameter q is described in the text.

In accordance with eqn (8.2), the constancy of T_c signifies that granules remain large at all the transformation stages. At the same time, they may become porous and permeated with insulating amorphous phase forming a fractal. The current channel inside such a granule in the normal state is similar to the trajectory of a particle in Brownian motion. We term such materials fractal granular materials. To explain the high q values of these materials, consider Fig. 8.2 where the white curve shows a possible current path from left to right through the labyrinth structure. The length of this channel is about twice as long ($\alpha_L \approx 2$) as the length of the part of the sample shown in Fig. 8.2, whereas its width is about $\alpha_h \approx 0.1$ of the width of this part. Now, assume that the width of the part shown in Fig. 8.2 is the average distance between two through channels (in terms of percolation theory, the correlation length ξ). Then, assuming that Fig. 8.2 shows part of a two-dimensional sample, the net of insulating interlayers in it increases the resistance by a factor $\gamma = \alpha_L/\alpha_h \approx 20$. If we assume that Fig. 8.2 shows a section of a three-dimensional material, we find that the resistance in it increases by a factor of $\gamma = \alpha_L/\alpha_h^2 \approx 200$.

If the granular lead sample (Fig. 8.4) and the Sb-based alloys (Fig. 8.8) underwent no superconducting transition (or if one managed to perform the experiments described above in a strong magnetic field), then these systems would have undergone a metal–insulator transition. However, in actual fact, they underwent a superconductor–insulator transition (we mean here the change in conductivity along a series of curves in Figs 8.4 and 8.8 at $T = 0$). We do not discuss this transition, because superconductivity itself is beyond our consideration. However, we may make an attempt to perceive "through this transition" a metal–insulator transition. With this aim consider the region of a metal–insulator transition for uniformly disordered systems (Fig. 6.2 in Chapter 6).
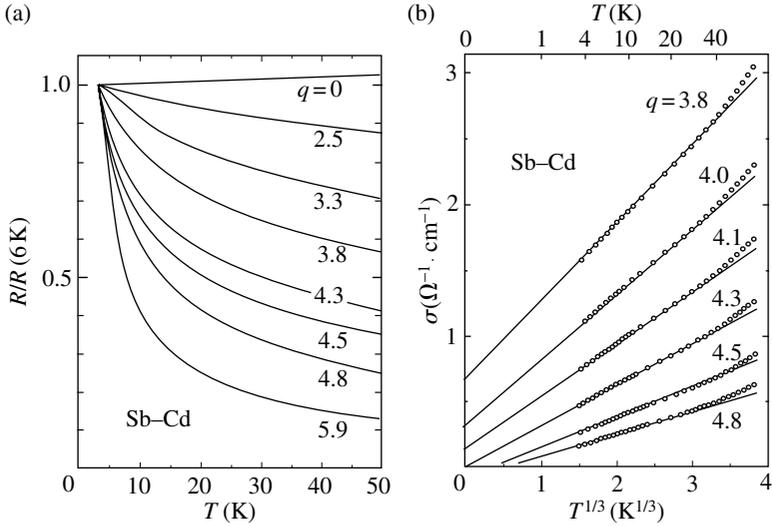
FIG. 8.9. (a) Resistances of metastable $Sb_{57}Cd_{43}$ films in the normal state ($>6\,\mathrm{K}$) as functions of temperature at various stages of sample amorphization. All the curves are normalized to the resistance at $T = 6\,\mathrm{K}$ at the corresponding amorphization stage. The parameter q is described in the text. (b) The same data in ($T^{1/3}$, $\sigma \equiv 1/R$) coordinates (Teplinskii et al. 1992).

The critical region in the vicinity of the metal–insulator transition is limited from the side of high temperatures by the condition

$$\tau_{\mathrm{ee}} > \tau, \tag{8.17}$$

which means that the dephasing time of interfering electrons should exceed the time of their elastic scattering. However, this condition is rather mild because $\tau_{\mathrm{ee}} \approx \hbar/T$ and τ may be as small as $\tau \approx \hbar/\varepsilon_{\mathrm{F}}$, so that eqn (8.17) means only that the temperature should be $T < \varepsilon_{\mathrm{F}}$. Since the temperature of the superconducting transition $T_{\mathrm{c}} \ll \varepsilon_{\mathrm{F}}$, then, generally speaking, in order to distinguish metal from insulator and determine the position of the metal–insulator transition, it should be enough to measure the resistance above T_{c} and extrapolate the temperature dependence $\sigma(T)$ from the region $T > T_{\mathrm{c}}$ to $T = 0$.

Figure 8.9a shows the $R(T)/R(6\,\mathrm{K})$ curves at different stages of amorphization of the Sb–Cd alloy. These curves reflect the changes in resistance at $T > 6\,\mathrm{K}$ and complement the $R(T)/R(6\,\mathrm{K})$ curves for the same alloy in the superconducting region $T < 6\,\mathrm{K}$ (Fig. 8.8a). These data replotted in Fig. 8.9b on the ($T^{1/3}$, $\sigma \equiv 1/R$) axes single out the state with $(R/R_0)_{T=6\,\mathrm{K}} \approx 2 \cdot 10^4$ (i.e. $q = 4.3$) as an in-between state separating the metal and the insulator. This demonstrates that a metal–insulator transition is also possible in fractal granular materials.

At the same time, this transition does not necessarily take place. Figure 8.10 shows an analogous representation of the experimental data in the ($T^{1/2}$, σ) axes
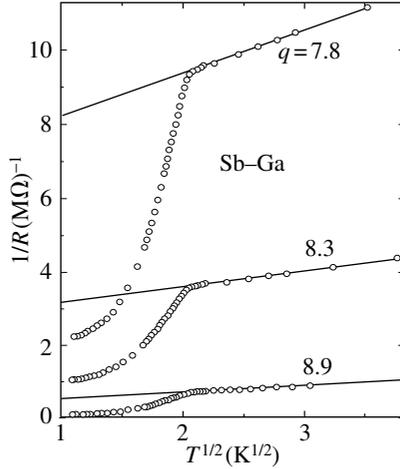
FIG. 8.10.   Extrapolation of conductivities from the region $T > T_c$ to $T = 0$ of
the metastable $Sb_{50}Ga_{50}$ alloy at the completion stages of sample amorphiza-
tion (Gantmakher et al. 1993).

for the most high-resistance states of the Sb–Ga alloy. Even for resistances as
high as $1\,M\Omega$, the extrapolation to $T = 0$ does not reveal a transition to the state
of an insulator. Possibly, the point is that the sizes of fractal granules remain
comparable with the size of the sample, which makes inequality (8.8) invalid.

References

Abeles, B., Ping Sheng, Coutts, M.D., and Arie, Y. (1975). Adv. Phys. **24**, 407.
Gantmakher, V.F., Zverev, V.N., Teplinskii, V.M., Tsydynzhapov, G.E., and
   Barkalov, O.I. (1993). JETP **77**, 513.
Frydman, A. (2003). Physica C **391**, 189.
Yu, X., Duxbury, P.M., Jeffers, G., and Dubson, M.A. (1991). Phys. Rev. B **44**, 13163.
Shapira, Y. and Deutscher, G. (1983). Phys. Rev. B **27**, 4463.
Teplinskii, V.M., Gantmakher, V.F., and Barkalov, O.I. (1992). Sov. Phys. – JETP
   74, 905.

9

INTEGER QUANTUM HALL EFFECT

For a comprehensive discussion of the integer quantum Hall effect, see Prange and Girvin (1990) and Butcher et al. (1993); the latter book also contains a description of the general properties of a two-dimensional electron gas.

The application of a magnetic field B to an ideal system of free electrons gives rise to the finite motion of electrons in the plane normal to this field and quantization of the energy of the transverse motion (Landau quantization). Only the spectrum of the motion along the field remains quasicontinuous. If, in addition, the electrons are confined in a layer with the thickness b of the order of $b \sim 1/k_F$, which is normal to the field, their motion along the field is also restricted. Then, in fact, the electrons are localized and their wave functions are independent of time and have nonzero values only in the vicinity of classical electron orbits.

Under these conditions, the conventional notion of electron localization becomes invalid. In a magnetic field, the electrons in an infinite plane in the absence of any random potential and electric field are localized in cyclotron orbits. Now, a random potential gives rise not to the localization of electrons, as usual, but to their transitions between various cyclotron orbits and, thus, to their delocalization.

The following discussion of the quantum Hall effect, one of the most remarkable discoveries in solid state physics of the second half of the 20th century, is made against this background. The quantum Hall effect (QHE) is an important part of a large field – the physics of two-dimensional systems. Earlier, we only briefly visited this field in Chapter 6 when considering metal–insulator transitions. The present chapter deals only with two-dimensional (2D) systems, but, in accordance with the principles used for selection of the material for this book, we discuss here only the integer quantum Hall effect because it may be considered within the concept of noninteracting electrons. The fractional quantum Hall effect, which would not exist at all under these conditions, is beyond our scope here. Even when discussing the integer quantum Hall effect, we shall focus on those aspects that were among the main topics of the preceding chapters – localization, metal–insulator transitions, scaling, hopping conductivity, percolation, etc.

The first section of this chapter deals with those concepts of the usual metal physics which are necessary for an understanding and description of the QHE and which form the background mentioned above. Then, after setting forth the

basic experimental facts underlying the QHE, we suggest a self-consistent inter-
pretation of this phenomenon. And, finally, the last sections of this chapter are
dedicated to the phase transitions accompanying the QHE.

9.1. Spectrum and dynamics of two-dimensional electrons in strong magnetic fields

*Below, we briefly consider the problems of the classical physics of metals
discussed in detail elsewhere (Abrikosov 1988; Gantmakher and Levinson 1987).*

Consider an infinite homogeneous thin metal film in the plane normal to the
Oz-axis. In a constant magnetic field $\boldsymbol{B} \parallel Oz$, the classical orbit of electrons
having no velocity component along the applied magnetic field is a circle. If one
applies, in addition to the magnetic field $\boldsymbol{B} \parallel Oz$, also an electric field $\boldsymbol{E} \parallel Ox$,
two types of electron motion are superimposed: circular motion with frequency
$\Omega = eB/mc$ and a drift along the direction $Oy \perp \boldsymbol{B}, \boldsymbol{E}$ with drift velocity $\overline{v} =
c(E/B)$ (where c is the velocity of light and m is the electron effective mass).
Since the drift velocity is normal to the electric field, $\overline{\boldsymbol{v}} \perp \boldsymbol{E}$, the linear local
relation of the field $\boldsymbol{E}$ and the current $\boldsymbol{j} = ne\overline{\boldsymbol{v}}$ is described not by a scalar but
by a tensor (here n is the concentration of the itinerant carriers). Therefore,
describing the linear transport properties of a two-dimensional electron gas, one
has to relate the components of the current $\boldsymbol{j}$ and the electric field $\boldsymbol{E}$ either by
the conductivity tensor $\widehat{\sigma}$,

$$\begin{aligned} j_x &= \sigma_{xx} E_x + \sigma_{xy} E_y, \\ j_y &= \sigma_{yx} E_x + \sigma_{yy} E_y, \end{aligned} \qquad \sigma_{xx} = \sigma_{yy}, \quad \sigma_{yx} = -\sigma_{xy} \tag{9.1}$$

or by the resistivity tensor $\widehat{\rho}$,

$$\begin{aligned} E_x &= \rho_{xx} j_x + \rho_{xy} j_y, \\ E_y &= \rho_{yx} j_x + \rho_{yy} j_y, \end{aligned} \qquad \rho_{xx} = \rho_{yy}, \quad \rho_{yx} = -\rho_{xy}. \tag{9.2}$$

Since these tensors are mutually reciprocal, their components are related as

$$\sigma_{xx} = \frac{\rho_{xx}}{\rho_{xx}^2 + \rho_{xy}^2}, \quad \sigma_{xy} = \frac{\rho_{xy}}{\rho_{xx}^2 + \rho_{xy}^2}. \tag{9.3}$$

In the relaxation-time τ-approximation for a free electron gas, the expressions
for σ_{xx} and σ_{xy} may be obtained directly from the equation of motion for a free
electron with due regard for its scattering by introducing a force of effective
friction, $-m\boldsymbol{v}/\tau$, as

$$m\dot{\boldsymbol{v}} = \frac{e}{c}[\boldsymbol{v}\boldsymbol{B}] + e\boldsymbol{E} - \frac{m\boldsymbol{v}}{\tau}. \tag{9.4}$$

Taking into account eqn (9.1) and the fact that the current is $\boldsymbol{j} = ne\overline{\boldsymbol{v}}$, we obtain
from eqn (9.4)

$$\sigma_{xx} = \frac{ne^2\tau}{m} \frac{1}{1 + \Omega^2\tau^2}, \quad \sigma_{xy} = \frac{ne^2\tau}{m} \frac{\Omega\tau}{1 + \Omega^2\tau^2}. \tag{9.5}$$

It follows from the above equations that the longitudinal conductivity of an infinite film containing no impurities ($\tau = \infty$) is $\sigma_{xx} = 0$, whereas the Hall conductivity and resistivity are

$$\sigma_{xy} = \frac{ne\overline{v}}{E} = \frac{nec}{B}, \quad \rho_{xy} = \frac{B}{nec}. \tag{9.6}$$

Usually, a low conductivity indicates a high resistivity and *vice versa*. It follows from eqn (9.3) that the zero σ_{xx} value, $\sigma_{xx} = 0$, at nonzero σ_{xy} value, $\sigma_{xy} \neq 0$, indicates that resistivity also has the zero value, $\rho_{xx} = \sigma_{xx} = 0$. However, it should be taken into account that the conducting region is limited along the Oy-direction. This influences the latter relationship. Moreover, there are always some scattering centers in the bulk. At finite temperatures, their role may be played by phonons. Therefore, the classical longitudinal conductivity has the zero value only approximately, $\sigma_{xx} \approx 0$.

Let a film with scattering centers be a strip elongated in the direction of the Ox-axis and let the boundary conditions set the field component E_x and the current $\boldsymbol{j} = (j_x, 0)$. The drift along the Oy-axis results in the appearance of charges at the strip side edges and the component E_y of the electric field across the strip. Therefore, the electric field in the film has two components, $\boldsymbol{E} = (E_x, E_y)$. The current along the field $\boldsymbol{E}$ is caused by scattering because the center of the circular orbit can be shifted along the field only due to scattering. The ratio of the electric-field components E_x and E_y and the σ_{xx} value are determined by the scattering frequency $1/\tau$.

One more circumstance makes an edge important – in the vicinity of an edge, electrons are scattered more often. Equations (9.1) and (9.2) include the current densities far from the edges. When calculating these densities based on sample sizes and the current measured outside the strip, one has to bear in mind the possible existence of some additional edge currents which do not depend on the frequency $1/\tau$. Let us consider some classical orbits in the film plane which are located in the vicinity of its edge. An electron that once reached the edge and was scattered from it would necessarily be scattered from this edge at the next turn of its trajectory. As a result of this multiple scattering, electrons will move along the edge irrespective of the type (specular or diffuse) of scattering from the edge (Fig. 9.1). Thus, under certain particular conditions, the current along the external electric field in a pure metal will flow mainly along its surface. The possible concentration of the direct current near the surface of a pure metal was first revealed by Azbel in 1963. This phenomenon is called the *static skin effect*. It will be clear from the following that the quantum analogue of the static skin effect is very important for the QHE.

On the terminology: Proceeding from the classical considerations alone, one may state that, far from the edges of an ideal film, the electrons are localized on circular magnetic orbits but that this localization is soft – in an infinitely low electric field, the electrons become delocalized in the direction normal to this field. Moreover, at the film edges,
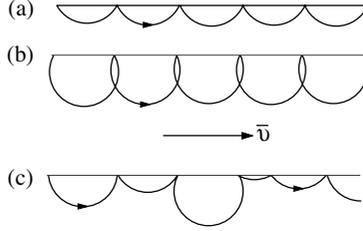
FIG. 9.1. Motion of classical electrons along the film edge in a normal magnetic field (a, b) for specular reflection from the edge and (c) for diffuse reflection. In all cases, the average velocity of electrons, $\bar{v}$, has a nonzero value, is directed along the edge, and, in the first approximation, is independent of the external electric field.

the electrons are delocalized in principle, i.e., they may participate in infinite motion along this edge.

Now, we proceed to the quantum description. An ideal spectrum of spinless electrons in a magnetic field may be represented as a sum of the energies of the transverse motion $\varepsilon_\perp(N)$, which depends on the quantum number N, and the longitudinal motion $\varepsilon_\parallel(k_z)$ as

$$\varepsilon(N, k_z) = \varepsilon_\perp(N) + \varepsilon_\parallel(k_z) = \hbar\Omega\left(N + \frac{1}{2}\right) + \frac{\hbar^2 k_z^2}{2m}, \quad N = 0, 1, 2\ldots, \quad (9.7)$$

where k_z is the wave vector along the magnetic field $\boldsymbol{B}$. The degeneracy γ of the levels at the given N and k_z is

$$\gamma = \frac{S}{2\pi r_B^2}, \quad (9.8)$$

where S is the area of the cross-section of the volume occupied by electrons by the plane normal to field and r_B is the magnetic length or magnetic radius

$$r_B = \left(\frac{\hbar c}{|e|B}\right)^{1/2}. \quad (9.9)$$

If the region occupied by electrons is a slab of thickness b lying in the plane (x, y) and if the field $\boldsymbol{B}$ is applied along the z-axis, then the k_z value in eqn (9.7) is also quantized, $k_z = (2\pi/b)N_z(N_z = 1, 2, \ldots)$ and $\varepsilon(N, k_z) \to \varepsilon(N, N_z)$. The quasicontinuous one-dimensional spectrum $\varepsilon_\parallel(k_z)$ in each magnetic subband with the magnetic quantum number N is transformed into a set of discrete nonequidistant levels

$$\varepsilon_\parallel(k_z) \to \varepsilon_\parallel(N_z) = \frac{\hbar^2}{2m}\left(\frac{2\pi N_z}{b}\right)^2, \quad \Delta\varepsilon_\parallel(N_z) = \varepsilon_\parallel(N_z + 1) - \varepsilon_\parallel(N_z)$$

$$= 2\pi^2(2N_z + 1)\frac{\hbar^2}{mb^2}. \quad (9.10)$$

The spectrum becomes completely discrete, and each level is characterized by two quantum numbers, N and N_z. The degeneracy γ is the same for all levels;

according to eqn (9.8), it depends only on the area S filled with a two-dimensional electron gas and on the magnetic field B. The density of states in such an ideal two-dimensional gas normalized to unit area is a sum of δ-functions

$$\sum_{N,N_z} \frac{1}{2\pi r_{\mathrm{B}}^2}\, \delta(\varepsilon - \varepsilon(N, N_z)). \tag{9.11}$$

The coefficient $n_{\mathrm{L}} = (2\pi r_{\mathrm{B}}^2)^{-1}$ that enters in all the terms of the sum (9.11) is the electron density for any completely filled Landau level. It may be expressed in terms of a quantum of the magnetic flux Φ_0 as

$$n_{\mathrm{L}} = \frac{\gamma}{S} = (2\pi r_{\mathrm{B}}^2)^{-1} = \frac{B}{\Phi_0}, \quad \Phi_0 = \frac{2\pi\hbar c}{e}. \tag{9.12}$$

Equation (9.12) signifies that, for a completely filled Landau level, one magnetic-flux quantum falls to the share of each electron.

Now, let n_3 be the electron concentration in three-dimensional space. Therefore, the electron concentration per unit area of the film is $n_2 = n_3 b$. Then it follows from eqn (9.8) for γ that the number of the filled levels at the temperature $T = 0$ is determined by the filling factor ν,

$$\nu = \frac{2\pi\hbar c n_2}{eB} = \frac{n_2}{n_{\mathrm{L}}}. \tag{9.13}$$

The integer part of the filling factor ν is the number of completely filled levels, whereas the fractional part indicates the existence also of a partly filled level and shows to what extent it is filled. With an increase of the field intensity, the degeneracy γ also increases, whereas the number of filled levels decreases. The next level starts emptying when the fractional part of ν becomes zero in the fields

$$B_i = \frac{2\pi\hbar c}{e} n_2 \frac{1}{i} = \Phi_0 n_2 \frac{1}{i}, \qquad (i = 1, 2, 3, \ldots). \tag{9.14}$$

We are interested in very thin films in strong fields in the case where only several discrete levels are filled. It is convenient to consider a sketch which may be called "a translation from quantum into quasiclassical language". Figure 9.2a shows the section of a Fermi sphere $\varepsilon(k_z, k_\perp) = \varepsilon_{\mathrm{F0}}$ in k-space for the three-dimensional case. With no applied magnetic field, the Fermi radius $\varepsilon_{\mathrm{F0}}$ depends only on the concentration n_3 and sets the energy scale for this diagram. An applied magnetic field quantizes the transverse momentum $k_\perp$. The integers N indicate the allowed values of this momentum and the vertical lines are the sections of the so-called Landau cylinders used in the analysis of the de Haas–Van Alphen effect in the physics of metals. The finite thickness b of the slab quantizes the longitudinal momentum k_z; its allowed values are indicated by numbers N_z. Filled dots inside the circumference of the radius $\varepsilon_{\mathrm{F0}}$ are filled levels with the quantum numbers (N, N_z). There are nine such levels in Fig. 9.2a plus one more level $(N, N_z) = (3, 1)$ which seems to be only partly filled (the gray dot). Since the Fermi level depends on the applied magnetic field and oscillates in the vicinity of $\varepsilon_{\mathrm{F0}}$, i.e., $\varepsilon_{\mathrm{F}}(B) \neq \varepsilon_{\mathrm{F0}}$, it is impossible to determine from the
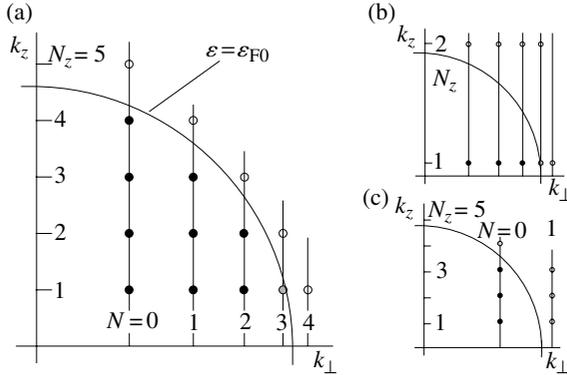
FIG. 9.2. The allowed states in the $(k_\perp, k_z)$ plane with their energies proportional to the squared distance from the origin. Filled circles indicate completely filled levels; partly filled circles indicate partly filled levels, and open circles indicate empty levels. (a) Levels with different magnetic numbers N and size magnetic numbers N_z are filled; (b) only the levels with the minimum size quantum numbers, $N_z = 1$, are filled; and (c) only the levels with the minimum magnetic quantum number, $N_z = 0$, are filled.

sketch to what extent the intermediate level is filled; this may be determined from eqn (9.13).

In stronger magnetic fields, the transition to the ultraquantum limit takes place when only one of the discrete quantum levels is filled, $(0, 1)$. As is seen from Fig. 9.2, this transition may occur in different ways depending on the b value. In ultrathin films and at heterojunctions, where $\Delta\varepsilon_\parallel(1) \geq \varepsilon_{F0}$, only the levels $(N, 1)$ on the horizontal line are filled (Fig. 9.2b). It is these ultrathin objects that are conventionally used in QHE experiments. However, another case is possible where for several levels we have $\varepsilon_\parallel(N_z) < \varepsilon_{F0}$, e.g., if $\varepsilon_\parallel(N_z) < \varepsilon_{F0}$ at $N_z \leq 4$ (Fig. 9.2a, c). Then, with an increase in the field, only some levels of type $(0, N_z)$ on the vertical line remain filled. Since the degeneracy is independent of the quantum numbers N and N_z, eqns (9.8), (9.13), and (9.14) are applicable to both cases. All the following conclusions made in this chapter are also applicable to both cases.

The existence of a spin doubles the number of discrete levels. The sets shown in Fig. 9.2 are implemented for both spin projections but they are shifted by the value of the Zeeman splitting. An additional increase in the number of levels and a more complicated spectrum may be caused by the multivalley electron spectrum of a crystal.

The discrete levels thus obtained correspond to the classical circular orbits of radius $r_N \approx \sqrt{N} r_B$ with fixed centers. (If $N_z \neq 1$, the rotation along the orbit is accompanied by the periodic motion across the film between its surfaces but still remains finite.) A constant uniform electric field $\boldsymbol{E} \parallel Oy$ applied to the film transforms the discrete levels into bands, $\varepsilon(N, N_z) \rightarrow \varepsilon(N, N_z) + Ey$.
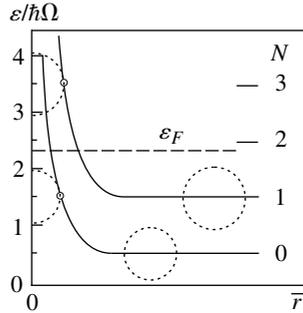
FIG. 9.3.   Energies of 2D electrons at discrete Landau levels as functions of the
distance $\bar{r}$ between the wall of a rectangular potential well and the most remote
point of the classical orbit. Dashed lines indicate the quasiclassical images of
the wave functions – the complete circumferences far from the edge and the
circumference segments in the vicinity of the edge. Open dots indicate the
states with $\bar{r} = r_N$ whose image is an exact semicircle.

Correspondingly, as follows from eqns (9.1)–(9.6), the classical motion in the
direction perpendicular to the electric field $\boldsymbol{E}$ becomes infinite.

The introduction of a finite area S into eqn (9.8) for the degeneracy γ takes
into account the limited size of the conducting two-dimensional region. For sim-
plicity, we assume that the region is a rectangular well, i.e., has vertical walls.
The degeneracy γ is provided by translation invariance, i.e., the equivalence of
orbits having different coordinates of the centers. If the distance from the orbit
center to the edge is less than the classical orbit radius r, the wall imposes a
limitation on the classical electron motion and on its quantum wave function
and the energy of the electron level increases (Fig. 9.3). The abscissa in Fig. 9.3
indicates the distance $\bar{r}$ from the edge to the most remote point of the classical
orbit. It is also assumed that $N_z = 1$ (which corresponds to the case shown in
Fig. 9.2b). In particular, the energy of the orbit at the level N with the center
located at the wall (the classical trajectory is a semicircle) equals the energy of
the orbit far from the wall at the level $N' = 2N + 1$,

$$\varepsilon(N, \bar{r} = r_N) = \varepsilon(N', \bar{r} \gg r_{\mathrm{B}}).$$

As a result of an increase in the energies of the levels in the vicinity of the wall,
all the Landau levels intersect the Fermi level. Thus, all of them may contribute
to the conductivity via the edge current. Again, the transformation of discrete
levels into bands at the film edges immediately results in the appearance of
extended orbits (Fig. 9.1).

The above ideal quantum picture will *a priori* be violated by a random
potential $U(\boldsymbol{r})$ even in the purest real system. Since a magnetic field sets the char-
acteristic length-scale given by eqn (9.9), two limiting types of random potential
should be considered – a potential with large-scale fluctuations with character-
istic lengths $\zeta \gg r_{\mathrm{B}}$ and a short-range potential with $\zeta \ll r_{\mathrm{B}}$. Both potentials

remove the degeneracy of the (N, N_z) levels and give rise to their broadening. Then, Landau levels are transformed into Landau minibands and the problem reduces to the determination of the wave functions in these minibands.

Consider in more detail the long-range potential $U(\boldsymbol{r})$. The effect of such a potential on the electron may be described by the force ∇U lying in the 2Dplane and constant within the size of a classical orbit. Now replace $\boldsymbol{E}$ in eqn (9.4) by ∇U and assume that $\tau = \infty$. Under the action of the force an electron drifts along the direction normal to ∇U, i.e., it moves all the time along the equipotential lines of the function $U(\boldsymbol{r})$. Thus, each electron, for all time, is in its own constant potential U. It is from this potential that one should measure the energy of the levels (N, N_z). As a result, "inhomogeneous broadening" of the initial δ-like degenerate (N, N_z) levels takes place.

In the vicinity of the local extrema, the equipotential lines of the function $U(\boldsymbol{r})$ are closed contours around these extrema. The direction of the electron motion along these contours depends on whether they are located around the local maxima U_+ or the local minima U_-. The characteristic sizes of the contours, $\widetilde{\boldsymbol{R}}$, are the distances within which the wave functions of the corresponding electrons are localized.

On the terminology. We shall call the wave function of an electron in an ideal 2D layer in a strong normal magnetic field a softly localized wave function because, as was shown above, an infinitesimal electric field delocalizes an electron in the direction normal to this field. On the contrary, quasiclassical states on the equipotential lines of finite size $\widetilde{R}$ are truly localized ones because an applied electric field may only deform an equipotential line and the weaker the applied field, the smaller the deformation. Thus, large-scale fluctuations of a potential transform soft localization into true localization at least for some of the electrons in the Landau miniband.

A random 2D potential has a percolation threshold, on approaching to which, infinite clusters with the energies higher and lower than the threshold energy both vanish. The equipotential lines at the threshold are transformed into an infinite net with an infinite number of self-crossings (see Appendix A). At saddle points, the quasiclassical electron velocity equals zero and, therefore, formally, one may consider an electron to be softly localized at the segments of the equipotential line between its self-crossing points. Taking into account that degeneracy at the self-crossing points is removed by an applied electric field and that saddle points may be overcome due to tunneling, the electrons at the percolation level are often called delocalized electrons. Thus, each Landau miniband has one percolation level with delocalized electrons. If the random potential is statistically symmetric, this percolation level is located in the center of the miniband.

Note: generally speaking, one level is not equivalent to one state, because the level may be degenerate. In the context of the integer QHE, this problem is far from being simple. Later we shall return to this problem (see eqn 9.28 and also the discussion in connection with Fig. 9.21).

The width $\delta\varepsilon$ of the energy range with the delocalized wave functions depends on some fine processes such as, e.g., tunneling between two closely located

quasiclassical trajectories in the vicinity of a saddle point (magnetic breakdown). In fact, $\delta\varepsilon$ is the energy uncertainty of any of the delocalized states.

The true localized states may also arise in a short-range potential. Consider a short-range potential $\zeta \ll r_{\mathrm{B}}$ formed due to isolated impurities. The energy of an electron moving in the direct vicinity of an impurity varies, which results in broadening of the Landau level. However, it is not the only effect produced by an impurity. The classification of the levels formed in the vicinity of an impurity may be made on the basis of the model of a hydrogen-like impurity in three-dimensional space. Without a magnetic field, an impurity forms a discrete set of localized states below the continuum of the extended states. Similar to a hydrogen atom, this discrete set is described by a set of quantum numbers. The application of a strong magnetic field results in splitting the continuum of extended states into a set of one-dimensional magnetic Landau subbands. The localized states turn out to be located below each of these subbands because, being detached from an impurity, an electron may be found in any of the Landau subbands. A decrease in the film thickness and transition to a two-dimensional system transforms the Landau subbands into discrete levels at which electrons are softly localized, but the initial true localized states are also preserved. Thus, impurities not only broaden initially discrete (N, N_z) levels but also create discrete (and, therefore, localized) levels in their vicinity.

9.2. Experimental observation of integer quantum Hall effect

The most widespread circuit for transport measurements is a Hall bar (Fig. 9.4a), in which the current J_{12} flows along the x-axis through side contacts 1 and 2, whereas the potential difference is measured between the remaining contacts. The component ρ_{xx} of the resistivity tensor $\widehat{\rho}$ is determined from the potential difference at the contacts located along one edge and the component ρ_{xy} is determined using the contacts located opposite to one another. The formula for ρ_{xx} indicated in Fig. 9.4a is valid under the assumption that the current is uniform over the whole bar. Since the conditions for observation of the QHE do not necessarily guarantee a uniform current, it is better to assume that in the experiments on the QHE, a Hall bar measures the values of the resistance $R_{xx} = V_{34}/J_{12}$.

A circular structure depicted in Fig. 9.4b (the so-called Corbino disk) allows one to measure the component σ_{xx} of the tensor $\widehat{\sigma}$.

Usually, a two-dimensional electron gas is formed due to bending of the bands either at the surface of a semiconductor or at the planar boundaries of specially selected semiconductors. A device of the first type is called a MOS (abbreviation for metal-oxide-semiconductor) structure; a device of the second type, a hetero-structure. In the majority of cases, the upper insulating layer of the measuring structure is additionally coated with a metal film which, together with a 2D gas, forms a plane capacitor. The film is called a gate. Supplying the voltage V_{g} to the gate, one may vary the concentration n of the two-dimensional gas. Therefore, it is possible to vary the number of filled levels in eqn (9.13) not only by varying the
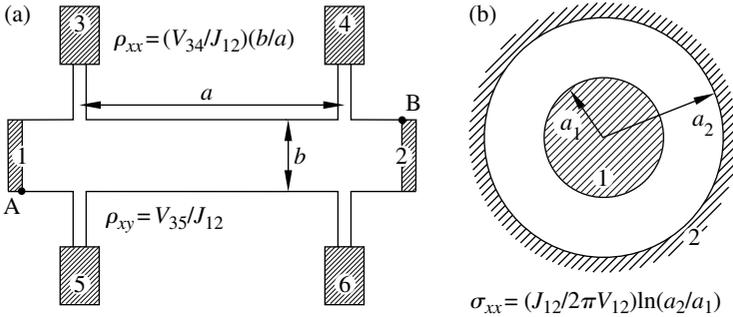
FIG. 9.4.   (a) Hall bar used for the experimental determination of the component ρ_{xx} of the resistivity tensor by measuring the potential difference between contacts 3 and 4 or 5 and 6 and for the experimental determination of the component ρ_{xy} by measuring the potential difference between contacts 3 and 5 or 4 and 6 at the current J_{12} flowing through side contacts 1 and 2; the specific role of the points A and B is considered in the text. (b) Corbino disk with two coaxial circular electrodes of radii a_1 and a_2 used to determine the component σ_{xx} of the conductivity tensor by measuring the potential difference V_{12} and current J_{12} between the coaxial electrodes.

magnetic field B but also by applying a potential difference V_{g} between the gate and the 2D gas. If the gate lies at a distance s from the plane with 2D electrons, then the change of the voltage at the gate, ΔV_{g}, and the electron concentration, Δn, are related by the linear dependence

$$\Delta V_{\mathrm{g}} = 4\pi\,\Delta n\,es. \tag{9.15}$$

Figure 9.5 shows the experimental data (v. Klitzing et al. 1980), which, in fact, were the first observation of the integer QHE and started its experimental study. The curves reveal three main features of the phenomenon: (i) plateaus on the $\rho_{xy}(V_{\mathrm{g}})$ curves in the vicinity of the V_{g} values at which the magnetic field B satisfies eqn (9.14) for fields B_i; (ii) the quantized ρ_{xy} values on these plateaus,

$$\rho_{xy} = \frac{1}{i}\left(\frac{2\pi\hbar}{e^2}\right), \quad (i = 1, 2, 3, \ldots); \tag{9.16}$$

and (iii) the tendency of $R_{xx}(V_{\mathrm{g}})$ to go to zero in the vicinity of the same V_{g} values. All these features are well seen in Fig. 9.6 which shows on an enlarged scale the close vicinity of one of these plateaus (in Figs 9.5 and 9.6, V_{g} values are different because the corresponding experiments were made on different samples).

Figure 9.5 allows one to formulate the first questions that would be answered experimentally prior to any serious discussion of the nature of the integer QHE. Here are the questions and the answers to these questions.

(i) *The accuracy* of the fulfillment of eqn (9.16). The very first study (v. Klitzing et al. 1980) showed that the measured ρ_{xy} values on the plateau
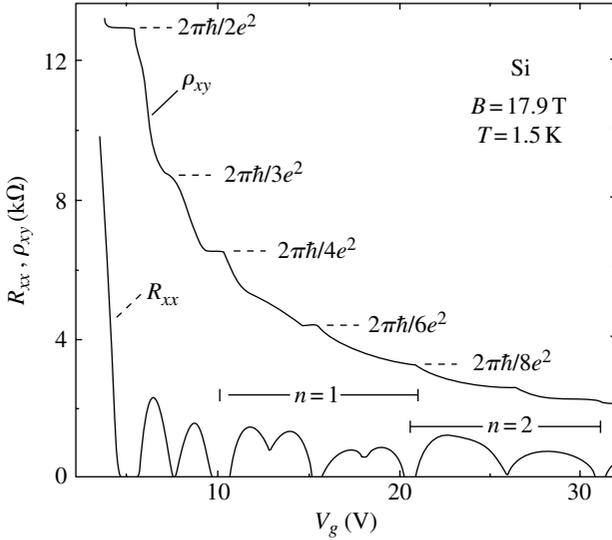
FIG. 9.5. Magnetoresistance R_{xx} and Hall resistivity ρ_{xy} as functions of gate voltage V_g in the MOS structure on the top of a Si surface (v. Klitzing et al. 1980). Magnetic field 17.9 T, temperature $T = 1.5$ K. The device dimensions in the notation of Fig. 9.4a are: $a = 130\mu m$, $b = 50\mu m$, and current $j_{12} = 1\mu A$.
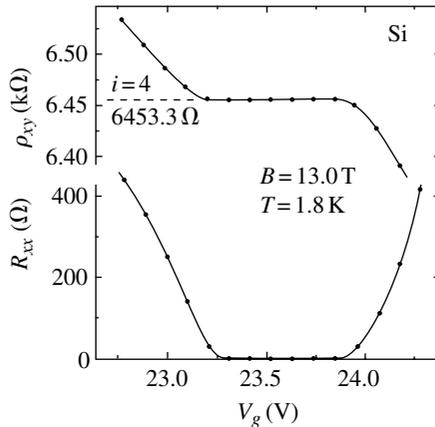

FIG. 9.6. Magnetoresistance R_{xx} and Hall resistivity ρ_{xy} as functions of gate voltage V_g in the MOS structure on a Si surface (v. Klitzing et al. 1980). Magnetic field 13 T, temperature $T = 1.8$ K. The device dimensions $a = 130\mu m$ and $b = 50\mu m$ are the same as in Fig. 9.5, but the distance s to the gate is different. The accuracy of the measurement of Hall resistivity is $0.1\,\Omega$.

coincide with the values given by eqn (9.16) with an accuracy higher than 10^{-4}. Today, reproducibility of the $e^2/2\pi\hbar$ value is about 10^{-8}. This showed that the effect could successfully be used in metrology. However, it was necessary to answer experimentally the question whether the measured ρ_{xy} values also depend on the carrier sign, multivalley nature of the spectrum, and some other factors defined by the crystal structure (it turned out that they do not depend on these factors). Metrology applications also dictated the necessity of studying the possible effect of impurities, temperature, geometric size of a sample, measuring current, and electron concentration (these factors also turned out to be inessential over a rather wide range), the variation of ρ_{xy} along the plateau, etc. The results of such studies may be found in the review articles, e.g., Krasnopolin et al. (1987); Mohr and Taylor (2000). For the following, it is only important that the quantized values given by eqn (9.16) are well distinguished against the background of smoothly varying Hall resistivity $\rho_{xy}(B)$ or $\rho_{xy}(V_g)$.

(ii) *Relative extension of plateaus*, i.e., their width in comparison with the distance between these plateaus either along the field B or the gate voltage V_g. This characteristic is not as stable as the quantized values of Hall resistivity. The plateau width also depends on the material of the two-dimensional layer of carriers, impurities, and temperature. Figure 9.7 shows the QHE curves obtained on a GaAs–Al$_x$Ga$_{1-x}$As heterostructure. The plateaus occupy a larger part of the $\rho_{xy}(B)$ curve so that the curve looks like a staircase with horizontal steps.
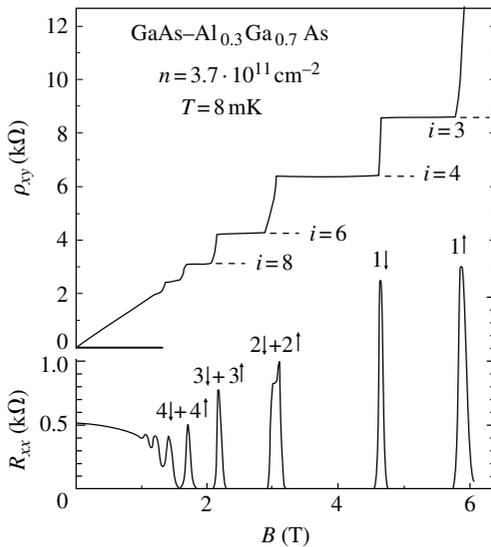

FIG. 9.7. Magnetoresistance R_{xx} and Hall resistivity ρ_{xy} as functions of magnetic field B in a GaAs–Al$_x$Ga$_{1-x}$As heterostructure (Ebert et al. 1982). Electron density in the 2D layer $3.7 \cdot 10^{11}$ cm^{-2}, mobility $\mu = 4.1 \cdot 10^4$ cm^2/V·s, temperature $T = 8$ mK.

The influence of the temperature on the plateau width is illustrated by Fig. 9.8. The $R_{xx}(B)$ function also has a specific structure consisting of rather narrow peaks with the nonzero values being observed only in the intervals between the plateaus. We shall revisit the problem of plateau width several times.

(iii) *Temperature dependence* of longitudinal conductivity in the vicinity of plateaus. In accordance with eqn (9.3), in the vicinity of plateaus, where $\rho_{xx} \approx 0$, one has also to expect that $\sigma_{xx} \approx 0$. However, this formula is valid only for a
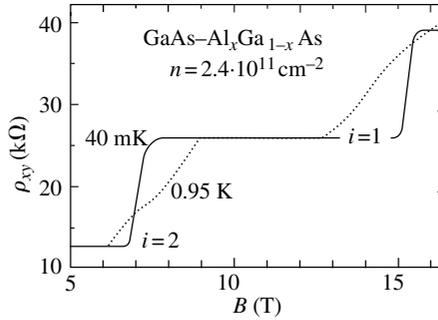

FIG. 9.8. Illustrating narrowing of the Hall plateau $\nu = 1$ in a GaAs–Al_xGa_{1-x}As heterostructure with an increase in temperature (Koch et al. 1991). Electron density in the 2D layer is $2.4 \cdot 10^{11}\,cm^{-2}$, mobility $\mu = 9 \cdot 10^4\,cm^2/V{\cdot}s$.
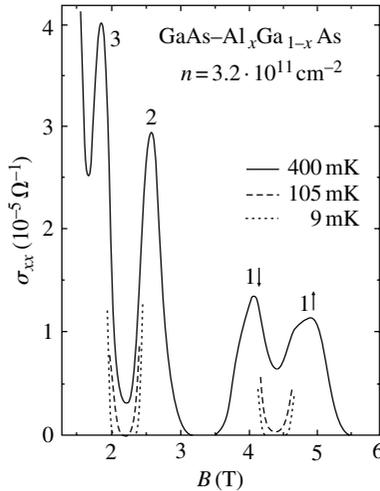

FIG. 9.9. Longitudinal conductivity in a GaAs–Al_xGa_{1-x}As heterostructure at different temperatures; measurements were made on a Corbino disk (Ebert et al. 1983). Electron density in the 2D layer $3.2 \cdot 10^{11}\,cm^{-2}$, mobility $\mu = 10.5 \cdot 10^4\,cm^2/V{\cdot}s$. At the peaks, the uppermost of the fully filled Landau minibands are indicated.

spatially uniform current distribution and, therefore, the simultaneous smallness of the ρ_{xx} and σ_{xx} values, $\rho_{xx} \approx 0$ and $\sigma_{xx} \approx 0$, requires experimental confirmation. Such confirmation came from the experiment on Corbino disks with the direct measurements of the σ_{xx} values. As is seen from the curves in Fig. 9.9 measured on a Corbino disk, the $\sigma_{xx}(B)$ curve consists of a number of peaks. With lowering of the temperature, the conductivity between these peaks (in the field intervals with the Hall plateaus) tends to zero. The law of σ_{xx} decreasing with temperature will be discussed in Section 9.5.

9.3. Mechanism of plateau formation

First, consider the experimental observations of the QHE based on the material of Section 9.1 of this chapter. Let us change the electron concentration of an ideal 2D system by varying the gate voltage V_g. Considering eqns (9.6) and (9.15), we may assume that the Hall conductivity would depend linearly on V_g. Indeed, if the electron spectrum of a 2D gas was continuous, we would have

$$\sigma_{xy} = \frac{nec}{B} = \frac{c(V_\mathrm{g} - V_\mathrm{g0})}{2\pi s B}, \tag{9.17}$$

with voltage V_g0 compensating the contact potentials in the circuit. However, in strong magnetic fields, one of the capacitor plates has a discrete energy spectrum. The energy spectrum of the system consisting of *an ideal 2D gas plus a gate* at three different gate voltages V_g is illustrated by Fig. 9.10.

Under equilibrium conditions with $V_\mathrm{g} - V_\mathrm{g0} = 0$, the Fermi levels at the gate and in the 2D gas lie at the same height and are located opposite to one another.
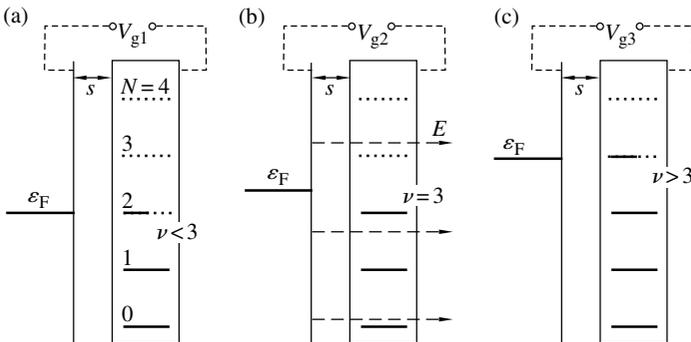

FIG. 9.10.   Fermi level in an ideal 2D gas at various gate voltages V_g. Solid lines indicate the filled states, dashed lines indicate the empty states. (a), (c) filling factors ν are fractional and the Fermi level coincides with the Landau Level; (b) filling factor ν is an integer and the density of states at the Fermi level equals zero.

The position of the Fermi level with respect to the Landau levels is set by the initial concentration n_0. Changes in V_g vary the concentration n. Hence, the V_g changes are described by a displacement of the level ε_F in the gate relative to the 2D gas. In the diagram with the voltage $V_g = V_{g1}$, the partly filled Landau level $N = 2$ coincides with the Fermi level. Upon the attainment of a certain voltage, the level $N = 2$ becomes fully filled so that a further increase in the carrier concentration in the two-dimensional gas becomes possible only if the electrons were localized at the Landau level $N = 3$. This is possible only if the Fermi level in the gate is located in front of this level, i.e., if it is displaced in the upward direction by a value of $\hbar\Omega$ (the right diagram). In turn, this requires a more intense electric field between the gate and the 2D gas. Therefore, in the range of voltages V_g,

$$\Delta_{\mathrm{id}} = \hbar\Omega/e \qquad (9.18)$$

the Fermi level is located in the gap between the third and fourth Landau levels and moves toward the latter (voltage $V_{g2} > V_{g1}$ in the diagram). The two-dimensional gas at V_{g2} is an "insulator" because it cannot absorb additional electrons and screen the external electric field. The additional force lines of the electric field, providing the potential V_g is in the range Δ_{id}, pass through the 2D layer. The electron concentration in the gate increases but it remains constant in the 2D gas with the integer filling number ν. The compensating charge arises in the so-called reservoir outside the 2D layer. The role of such a reservoir may be played by the impurities in the bulk of the substrate of the 2D layer, the opposite surface of this substrate, the contacts, etc. If the Fermi level at the voltage $V_{g3} > V_{g2}$ reaches the next (fourth) Landau level, the charges from the reservoir would return back to the 2D layer. The electron concentration in the layer would first change in a jumping manner and then would smoothly vary in accordance with eqn (9.15).

At a constant charge concentration in the 2D layer, the Hall voltage remains constant. Therefore, in an "ideal material," plateaus with the widths given by eqn (9.18) would be formed in the vicinity of the $V_g^{(i)}$ values determined by eqn (9.14). The plateau heights $\rho_{xy}^{(i)}$ are determined by i completely filled Landau levels and, therefore, would have the exact quantized values given by eqn (9.16).

Once more, these considerations are pertinent to the *gedanken* experiment with an *ideal* 2D layer and hence cannot be accepted as an explanation of the QHE phenomenon. The exact quantized values described by eqn (9.16) were obtained in experiments performed on *real* samples characterized by finite disorder and localized states.

In order to evaluate the relative width of the plateaus, η_{id}, compare the value of Δ_{id} with the change in the voltage Δ_ν necessary for complete filling of one of the Landau levels, i.e., for the change of ν by one and the concentration by $\Delta n = (2\pi r_{\mathrm{B}}^2)^{-1}$. Since the magnetic field enters both expressions, it will be cancelled. Then, assuming that the permittivity of an insulating layer of the

capacitor equals unity, we obtain

$$\eta_{\mathrm{id}} = \frac{\Delta_{\mathrm{id}}}{\Delta_{\nu}} = \frac{\hbar^2/me^2}{s} \, . \tag{9.19}$$

The numerator of eqn (9.19) is the Bohr radius a_{B}. For a hydrogen atom, it equals 0.5 Å. The small effective mass of electrons in the 2D gas may increase the a_{B} value by a factor ranging from 10 to 20, but the spacing s in the capacitor is always much larger. Therefore, the dependence $\rho_{xy}(V_{\mathrm{g}})$ for an ideal 2D gas would have narrow plateaus spaced by considerable distances.

A random potential in the 2D layer broadens the Landau levels into minibands. In accordance with the above consideration, the miniband includes at least one extended state located in the vicinity of the center of a broadened distribution and some localized states at the periphery of this distribution. Suppose the Landau minibands do not overlap so that they are spaced by the intervals of forbidden energies. With an increase in V_{g}, the Fermi level, first, moves over the localized states of the lower Landau miniband and then, after passing the region with zero density of states, over the localized states of the upper Landau miniband (Fig. 9.11). Filling of the localized states would not change the value of the Hall voltage so that the relative width of the plateau, η, becomes equal to

$$\eta = \eta_{\mathrm{id}} + \eta_{\mathrm{loc}} \gg \eta_{\mathrm{id}}, \tag{9.20}$$

where η_{loc} is the fraction of the localized states at the Landau miniband.

Note that the rates of the Fermi level shifts over the energy spectrum of the 2D gas, $d\varepsilon/dV_{\mathrm{g}}$, in the regions shown in Figs. 9.11b and 9.11c are different.
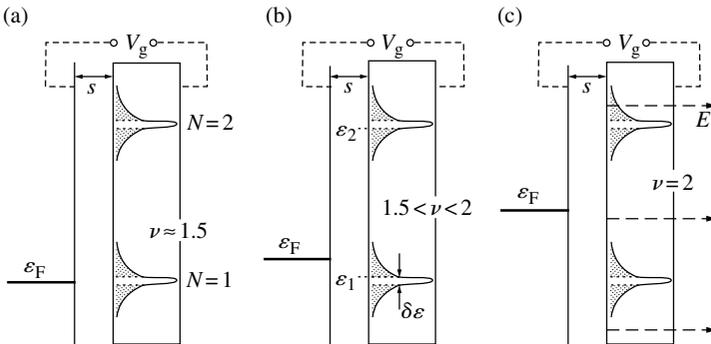

FIG. 9.11.   The same as in Fig. 9.10 for a real 2D gas with a random potential which has transformed the Landau levels into nonoverlapping Landau minibands. The localized states at the band edges are hatched. The lower layer $N = 0$ is not shown. (a) States at the Fermi level are extended; (b) states at the Fermi level are localized; (c) density of states at the Fermi level equals zero when the filling factor ν is an integer.
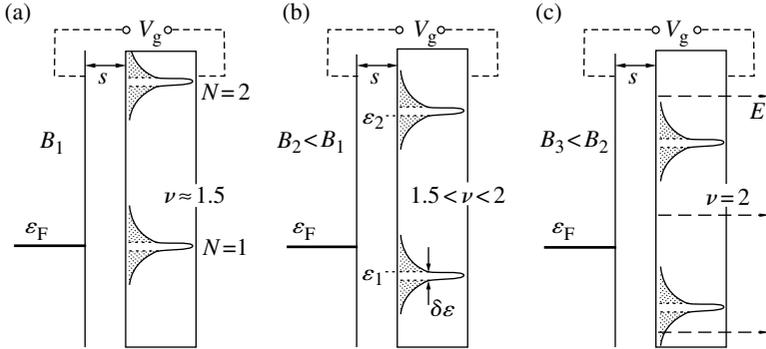
FIG. 9.12.   The same as in Fig. 9.11 but at a constant V_g value and a gradually
decreasing magnetic field B.

All that stated above about plateau formation on the $\rho_{xy}(V_g)$ curves in 2D
devices with a gate is also valid for the $\rho_{xy}(B)$ curves. An increase in V_g is equi-
valent to a decrease in the field B. Because of the lowering of the degeneracy γ,
the number of electrons at lower Landau levels decreases, and the filling factor
of the 2D system increases as well as the number of electrons at the upper level,
$\varepsilon(N) = \varepsilon_F$. Figure 9.12 illustrates the equivalence of the variation of V_g and B.
Since the gate charge does not alter with the variation of the magnetic field, the
Fermi level ε_F should be considered as fixed. If there is no gate at all, the level
ε_F is fixed by the environment of the 2D gas with which the gas is in thermo-
dynamic equilibrium. With a decrease in the field, the degeneracy γ of the levels
also decreases (eqns 9.8 and 9.9) and, to be able to fit in the same number of
electrons, the system of levels has to be displaced downward. Unless the localized
part of the Landau miniband is located in front of the Fermi level, the localized
states are filled up and the Hall resistivity ρ_{xy} remains constant (Fig. 9.12b).
When the density of states at the Fermi level attains the zero value (Fig. 9.12c),
the 2D system becomes insulating and transparent for an electric field, etc.
    Thus, in managing to explain the relative broadening of the plateau by the
existence of localized states, we immediately proceed to the main difficulty
encountered in the interpretation of the QHE. Localized states do not contri-
bute to the Hall current j_y. Moreover, the existence of these states indicates
a decrease in the number of delocalized states participating in the Hall effect.
Therefore, the main problem here is the cause of the *precise* values of the Hall
resistivity on the plateau. Assuming for definiteness that a random potential
$U(\boldsymbol{r})$ includes only large-scale fluctuations, consider this problem in terms of
the concepts formulated by Iordansky (1982), Kazarinov and Luryi (1982), and
Luryi and Kazarinov (1983).
    Figure 9.13a shows a percolation net of equipotential lines at the level $U = U_0$.
Now, apply an electric field $\boldsymbol{E} \parallel Ox$ to a 2D layer. The field removes the
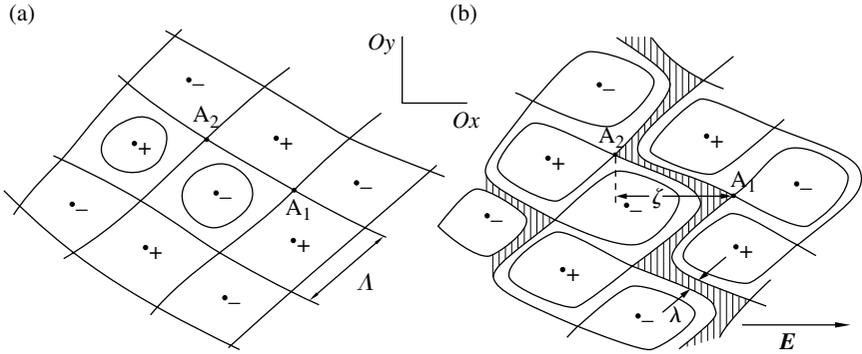degeneracy at the saddle points. The energies at the points A_1 and A_2, which

FIG. 9.13. (a) Net of equipotential lines at the percolation level $U = U_0$ in a 2D electron gas with a large-scale random potential $U(\boldsymbol{r})$; (b) the same in an electric field (Iordansky 1982). The $+$ and $-$ signs indicate the local extrema of the function $U(\boldsymbol{r})$. The equipotential lines, $U = \text{const.}$, in the hatched regions go to infinity.

earlier were located on one potential line, now differ by $eE\zeta$, where ζ is the projection of the distance between these points onto the direction of the field $\boldsymbol{E}$. As is seen from Fig. 9.13b, the net is transformed into a system of zigzag strips with the width λ being dependent on the local value of the random-potential gradient $|\nabla U|_{U_0}$ as

$$\lambda(r) = \frac{eE\zeta}{|\nabla U|_{U_0}(r)} \tag{9.21}$$

where r is the current coordinate along the strip. The equipotential lines inside these strips go to infinity along the Oy direction, and between the strips, these lines form closed contours. Since $\lambda(r) \to 0$ as $E \to 0$, the strip width equals the larger of the λ and r_{B} values.

Now return to Fig. 9.4 and consider a seemingly formal detail. The expression for the Hall voltage,

$$V_{35} = J_{12}\rho_{xy} \tag{9.22}$$

does not include the geometric sizes a and b of the device and the voltage V_{35} depends only on the total current J_{12} flowing through the device, the field B, and the concentration n, which enter ρ_{xy}. This signifies that the regular geometric shape of the Hall bar used is not important when measuring ρ_{xy}. It is only necessary that the 2D gas should "connect" the contacts used in the measurement. *The contributions that come to the transverse conductivity from the percolation net inside a Landau miniband formed by a random potential and from the Landau level under the ideal conditions of an ideal regular potential are the same. A large number of holes in the plane occupied by the ideal 2D gas would not change the ρ_{xy} value, if the boundedness of such a plane is preserved. This is the key point that should be kept in mind for understanding the essence of the QHE.*

Each Landau miniband located below the Fermi level has a narrow layer of delocalized states. In crossed fields, this layer makes the contribution $\Delta\sigma_{xy} = e^2/2\pi\hbar$ to the transverse conductivity. The total value of the transverse conductivity σ_{xy} in the mode of QHE plateaus is obtained by multiplying $\Delta\sigma_{xy}$ by the integer number i of such layers of delocalized states lying below the Fermi level. In the QHE plateaus mode, the longitudinal conductivity is absent, so that

$$\sigma_{xx} = 0, \quad \sigma_{xy} = i\left(\frac{e^2}{2\pi\hbar}\right). \tag{9.23}$$

Let the quantities B and V_g have values such that the upper Landau level is filled only partly (noninteger value of ν in eqn 9.13). For large-scale fluctuations of potential, the area S occupied by the 2D layer is divided into two regions. In one of these regions, the upper Landau level is completely filled with electrons, whereas in the other, the upper Landau level is absolutely empty. As long as the fractional part of the factor ν is small and the region of the 2D layer filled with electrons has only closed equipotential lines of finite size, i.e., only the localized states, these electrons do not affect the transport properties. However, with an increase in the concentration n (or a decrease in the field B), the Fermi level would reach the layer of delocalized states on the percolation net and the contacts of the Hall bar in Fig. 9.4a would be connected by the percolation net. This net covers only a part of the total area and may consist of fine fibers and contain large holes, whose shape is unimportant. Unless the states on this net are filled only partly, the longitudinal conductivity σ_{xx} has a nonzero value and the transverse conductivity σ_{xy}, an intermediate nonquantized value. As soon as all the extended states of the percolating layer become filled up, the transverse conductivity σ_{xy} attains the next plateau.

Equation (9.23) relating σ_{xy} to the number of layers with delocalized carriers in the energy space below the Fermi level explicitly suggests that broadening of each Landau level under the action of a perturbing potential $U(\boldsymbol{r})$ results in the formation of only one layer of delocalized states. Then the question arises whether formation of several layers of extended states with energies $\varepsilon_N^{(1)} < \varepsilon_N^{(2)} < \cdots$ is possible within one Landau miniband. Such layers may be formed but only under specific conditions: (i) if $U(\boldsymbol{r})$ is a periodic potential and (ii) if the magnetic flux Φ through the elementary cell of the potential $U(\boldsymbol{r})$ differs from the quantum of the magnetic flux Φ_0 by a rational factor, i.e., if $\Phi = (p/q)\Phi_0$, where p and q are integers. It follows from the second condition that this potential $U(\boldsymbol{r})$ should itself depend on the magnetic field.

Note that *a priori* we do not consider here a crystal potential against whose background 2D systems are always formed. The crystal potential initially renormalizes the electron spectrum, it may make it multivalley, anisotropic, etc., and may change its effective mass. In the hierarchy of potentials, the crystal potential is the most important. The applied magnetic field acts on the electrons thus renormalized.

Despite all its seeming exotics, the "lattice variant" of the perturbing potential $U(\boldsymbol{r})$ is very important. It is here that the road from the integer to fractional

quantum Hall effect begins. However, the fractional quantum Hall effect is beyond the scope of this book.

9.4. Edge channels

The sense and the role of the edge channels are discussed in detail in the review by Büttiker (1990).

Below, we consider the current distribution in samples in the QHE mode, i.e., at those B or V_g values at which the Hall plateaus are formed with the resistivity given by eqn (9.16) and $R_{xx} \ll \rho_{xy}$. Assume that a random potential is a large-scale one and start our consideration with a Corbino disk (Fig. 9.4b) – the simplest measuring device with the 2D region having no edges except electrodes.

On a Corbino disk, a system of zigzag strips whose width is given by eqn (9.21) arises from the percolation net in the form of nonintersecting closed lines around the internal electrode. The Hall current j_y flows along these lines and avoids the regions with the equipotential lines not encompassing the internal electrode (Fig. 9.14). Under the above assumptions, no current flows between the coaxial electrodes.

The current distribution in a Hall bar (Fig. 9.4a) is more complicated because its contacts are located at the edges of the 2D region in such a way that, moving along the edge, it is possible to pass from one contact to another. As was indicated earlier, the edges of the 2D region have specific properties. Let the boundary be sharp, so that the 2D region would be bounded by vertical walls. It is seen from Fig. 9.3 that in this case, there exists a specific region – a strip with width of the order of r_B along the edges. Inside this strip, the Fermi level is reached by all the Landau levels located below. This strip is a multichannel 1D transport formation with the number of channels equal to the number of Landau levels filled by more than 50% (such systems were described in Section 5.2). It follows from the classical description of the electron motion in this strip that electrons would collide with the edge at each "turn" of their motion and would be displaced at each subsequent turn in the same direction along the channel axis (Fig. 9.1).
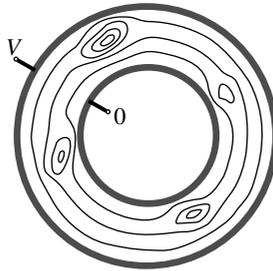

FIG. 9.14. Equipotential lines in a Corbino disk with a large-scale random potential in the QHE mode (Kazarinov and Luryi 1982).

The specific feature of the kinematics of this motion is that these electron collisions with the edge give no rise to backscattering. Therefore, the strip along the edge is not a simple 1D system but an *ideal* multichannel system of the type described by eqn (5.13). In a magnetic field, a finite current would flow over all the channels of this system.

Note that in principle, scattering from one channel to another is possible if $\nu > 2$, although this scattering should change the quantum number N or the direction of the electron spin. However, this scattering would not lead to dissipation. The transition from a channel to the bulk also may take place via large-angle scattering from the impurities located at a distance of the order of r_B from the edge. This scattering process is important: it causes dissipation and washes the current out of the edge channels.

Necessary refinement. The assumption of the existence of a sharp boundary of the 2D region is not confirmed in practice. Usually, there is no vertical potential wall and the width of the region in which the lower Landau levels reach the Fermi level considerably exceeds r_B. As a result, instead of one multichannel strip with i channels, we have i closely located one-channel strips parallel to the edge of the 2D region. It is this picture that is usually invoked in the discussion of real experiments. This problem is considered in detail in the original paper of Chklovskii et al. (1992). This refinement does not matter when the Hall current distribution inside the 2D region is discussed.

Now, let us remove contacts 3–6 from the Hall bar shown in Fig. 9.4a, so that the bar becomes a rectangular two-contact sample shown in Fig. 9.15. As earlier, we assume that $\rho_{xx} = 0$. This signifies that although the current flows along the A′B and B′A edges, no dissipation would take place. Therefore, the potential between the contacts along the edge is constant. By definition, the A′A and B′B edges of the contacts adjacent to the metal are equipotential lines. Therefore, the rectangular contour of the 2D region of a two-contact sample has two equipotential segments, AB and BA, with the potential difference between them equal to V_{12}. At the points A and B, the potential has jumps, and all the equipotential lines with the potentials $0 < V < V_{12}$ would also be concentrated in the vicinity of the points A and B. The configurations of these lines in the
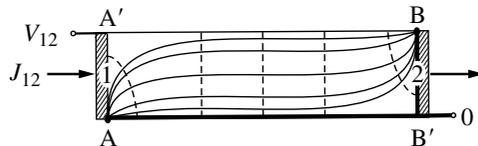

FIG. 9.15.   Field distribution in an ideal rectangular two-contact sample in the QHE mode at $T = 0$. The potential along the solid lines connecting the A and B points is constant. Dashed lines indicate the potential gradient along which the electric field is directed. The density of the equipotential lines (the value of the electric field) is discussed in the text.

absence of a random potential are shown in Fig. 9.15. Such a configuration of
the potential field seems incredible. However, we should like to make a small

lyrical digression concerning general physics. Letting a current J flow along a wire with
resistivity ϱ which is fixed on a wall and forms there an intricate pattern, we formulate
a rather complicated problem for Nature – to find the solution to the Laplace equation
$\triangle\phi = 0$ with the boundary condition that an electric field $E = -\nabla\phi$ on the wire
surface would always be directed along this wire and be equal to $E = J\varrho$. Nature solves
this problem "at the velocity of light" by placing additional static charges on the wire
surface. The field configuration in Fig. 9.15 is no more complicated. The needed charges
are placed along the edges of the rectangle.

In terms of the external circuit, the current J_{12} not shown in Fig. 9.15 plays
the role of the current j_x in the notation of eqn (9.2). However, inside the bar,
it represents in fact the current j_y, because it flows along the equipotential lines
of the electric field. According to eqn (9.16), the ratio V_{12}/J_{12} is equal to

$$\frac{V_{12}}{J_{12}} = \frac{1}{i}\left(\frac{2\pi\hbar}{e^2}\right), \quad i = 1, 2, 3, \ldots \tag{9.24}$$

Let us divide the current J_{12} into i equal parts, $J_{12} = iJ_i$, in accordance with the
number of the Landau minibands below the Fermi level and rewrite eqn (9.24)
in the form

$$J_{12} = \sum J_i = i\left(\frac{e^2}{2\pi\hbar}\right)V_{12}. \tag{9.25}$$

Equation (9.25) shows that J_i may be considered as currents flowing in parallel
over i one-dimensional ideal channels (5.13) between contacts 1 and 2. This
interpretation, first proposed by Büttiker (1990), reduces all the currents in the
QHE mode to edge currents. Each Landau miniband below the Fermi level with
one extended state and the rise to the Fermi level along the sample perimeter is
equivalent to one of these ideal channels.

Pay attention to the similarity of the model of edge channels and the static skin effect,
a classical phenomenon of the current concentration near the surface in pure metals at
low temperatures (Fig. 9.1).

Formally, another interpretation of the flow of the current J_{12} through the
sample is possible. The Hall current contains equal contributions from all i layers
of delocalized states inside Landau minibands. According to Figs 9.10–9.12 all
these i layers are below the Fermi level. This means that the Hall current in
the 2D electron system can flow along the states below the Fermi level. This is
possible because the Hall current is nondissipative.

Consider one of the i filled minibands lying below the Fermi level and represent
the energy ε_N of an electron with the corresponding quantum number N as a
function of the orbit-center position. As a result, we arrive at a surface $\varepsilon(\boldsymbol{r})$ which
reminds us of a U-shaped wash-tub with the bottom corrugated by a random
potential (Fig. 9.16). In this model, the current flows along the equipotential lines
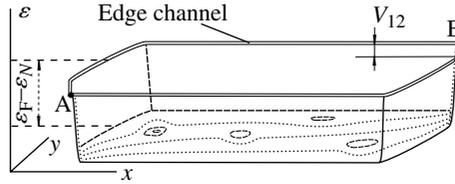(chains of dots) of the wash-tub bottom. Naturally, the final result is the same

FIG. 9.16.   Potential energy of electrons from one of the filled Landau levels as a
  function of their position in the 2D region. Dotted lines are the equipotential
  lines along which the nondissipative currents flow.

as in the model of edge channels: the nondissipative current $J_i = (e^2/2\pi\hbar)V_{12}$
between the points A and B.

   We have discussed two models which give different predictions about the Hall-
current distribution in a 2D plane. According to Thouless (1993), an intermediate
variant takes place in practice. The Hall current indeed flows along the edges of
the 2D region but it decays logarithmically at distances r from the edge much
larger than the magnetic length r_B, as $\log(r/r_B)$. The decay is due to screening
by the 2D electron gas of the field of static electric charges disposed along the
perimeter of the 2D region for satisfaction of the boundary conditions of the
Laplas equation. Because of the screening, the equipotential lines in Fig. 9.15
are arranged densely along the sample edges and rarely in the middle. As the
screening by the 2D electron gas is poor, the line density falls off slowly with
distance r from the edge.

   In essence, the current indeed flows in the bulk but the current density is
proportional to $|\boldsymbol{E}| \propto |\nabla\boldsymbol{U}|$ and therefore falls off with distance from the edge.
The deviations from the model of edge channels are due to the electron scattering
from the edge channels to the bulk. It follows from the aforesaid that both models
can be used to describe the distribution of the Hall currents. The edge channel
model by Bütteker is usually used for a high-mobility 2D gas in wide Hall bars
while the model of currents in the bulk below the Fermi level is applied to a 2D
gas with strong scattering and to narrow bars.

   Now, return contacts 3–6 to the Hall bar, i.e., come back from the configur-
ation in Fig. 9.15 to the configuration in Fig. 9.4. If no current flows through
contacts 3–6, the situation remains practically the same, so that $V_1 = V_3 = V_4$
and $V_2 = V_5 = V_6$ and

$$V_{35} = V_{46} = V_{12} = \frac{J_{12}}{i}\left(\frac{2\pi\hbar}{e^2}\right), \quad V_{34} = V_{56} = 0.$$

   It is worth obtaining direct proof that currents below the Fermi level indeed
exist. The special experiments on a Corbino disk (Dolgopolov et al. 1990,
1992) proved that charge transfer really takes place below the Fermi level.
Instead of applying a potential difference to circular electrodes 1 and 2 of a
Corbino disk (Fig. 9.4b) one may slowly vary a strong magnetic field B normal
to the disk plane. Now the roles of the circular and radial directions are

exchanged: a circular electric field $E_{\text{circ}} \propto \partial B / \partial t$ arises in the disk and the Hall current becomes radial. Since there is no edge connecting the circular electrodes, the Hall current may flow only "through the bulk" and is forced to flow below the Fermi level. This current may be measured by an electrometer switched between contacts 1 and 2.

Note that the mechanism of the appearance of the Hall current $j \propto \sigma_{xy} E_{\text{circ}}$ is such that electrons receive no energy from the electric field E_{circ}. Here comes a new problem: how to measure the current flowing in the external circuit. A conventional amperemeter connected to circular contacts would unavoidably dissipate some energy because of its internal resistance. On the other hand, the charge at the electrodes may be measured by an electrometer only if the charge is not dissipated because of the internal conductivity σ_{xx}. This signifies that a measuring circuit with an electrometer may only be used for measurements on QHE plateaus, where $\sigma_{xx} = 0$.

The results of such experiments are illustrated by Fig. 9.17. It is seen that the charge Q in the electrometer starts increasing as soon as the change in the field brings the sample to the QHE plateau where $\sigma_{xx} = 0$. With the appearance of the finite conductivity σ_{xx} at the opposite end of the plateau, the charge is dissipated. The charge sign depends on the direction of the field change. In the plot shown in Fig. 9.17, with an increase in the field $Q > 0$ and with a decrease in this field $Q < 0$. The rate of charge increase, i.e., the slope of the line, $\Delta Q / \Delta B$, allows one to determine the σ_{xy} conductivity. However, the accuracy of this determination is insufficient for studying the constancy in σ_{xy} along the QHE plateau, but is quite sufficient to distinguish the integers i in eqn (9.23).
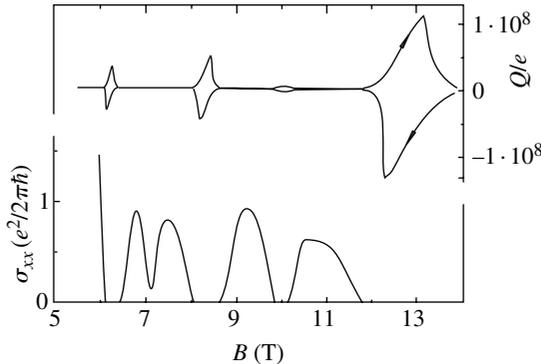

FIG. 9.17. Longitudinal conductivity σ_{xx} as a function of magnetic field (below) and transfer of charge Q between the electrodes of a Corbino disk during field variation in the QHE mode (above) (the directions of changes of the field are indicated by arrows). A Corbino disk with sizes $2a_1 = 225\mu$m and $2a_2 = 675\mu$m was prepared based on a MOS structure on the top of a Si surface. Temperature $T = 25\,$mK (Dolgopolov et al. 1990, 1992).

9.5. Density of states of 2D electron gas in a magnetic field

The density of states schematically shown in Figs 9.11 and 9.12 is of fundamental importance and, therefore, has been repeatedly studied experimentally. First, consider the experimental measurements of the density of localized states.

The $\rho(V_g)$ and $\rho(B)$ curves form plateaus at such V_g or B values at which the Fermi level is located in the region of forbidden or localized states at distances ε_1 and ε_2 from the nearest levels of the extended states. In this case, the dissipative conductivity arises only because of thermal carrier excitation to these levels playing the part of a mobility edge. Hence, one may expect that the changes of the longitudinal resistivity and longitudinal conductivity would obey the activation law

$$\sigma_{xx} \propto \exp[-(\varepsilon^*/T)], \quad \rho_{xx} \propto \exp[-(\varepsilon^*/T)]. \tag{9.26}$$

Figure 9.18 shows the experimental data obtained for a GaAs–Al$_x$Ga$_{1-x}$As heterostructure-based Hall bar. It is seen that the resistivity ρ_{xx} really varies according to the activation law (9.26). The ε^* value determined by the slope of the straight lines in Fig. 9.18a is essentially dependent on the applied magnetic field. The physical processes occurring in this case are well explained based on Figs 9.11 and 9.12. The longitudinal resistivity is inversely proportional to the carrier concentration at the level $N = 2$ or the hole concentration at the level
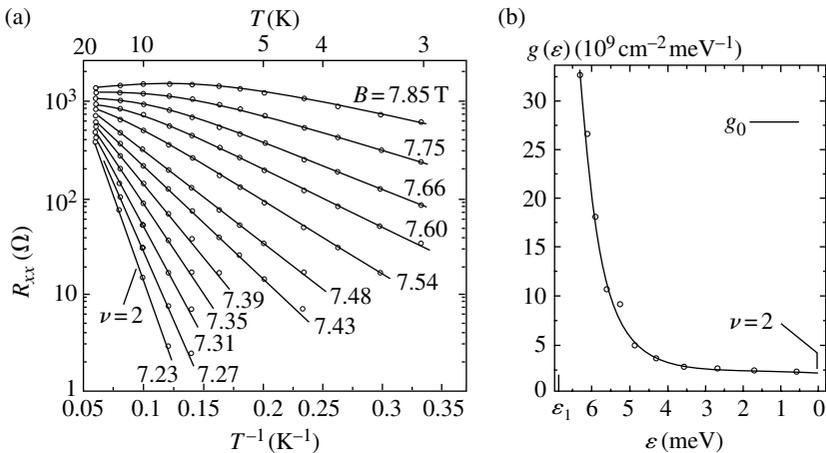

FIG. 9.18. (a) Temperature dependence of the longitudinal resistance of the GaAs–Al$_x$Ga$_{1-x}$As heterostructure in the temperature range 20–3 K at different values of the magnetic field in the vicinity of the filling factor $\nu = 2$. Mobility $5.5 \cdot 10^5$ cm^2/V·s (Stahl et al. 1985). (b) Density of localized states in the gap between the Landau levels determined by Stahl et al. (1985) from experiment. ε_1 is the middle of the Landau miniband $N = 1$ and g_0 is the density of states without a magnetic field.

$N = 1$. Therefore, ε^* which enters eqn (9.26) is equal to

$$\varepsilon^* = \min\left(|\varepsilon_F - \varepsilon_1|, |\varepsilon_F - \varepsilon_2|\right), \quad \varepsilon_2 - \varepsilon_1 \approx \hbar\Omega. \tag{9.27}$$

The latter relationship is accurate within the ratio of the width of the layer of extended states to the cyclotron frequency, $\delta\varepsilon/\hbar\Omega$.

The slope of the straight line $\ln\rho_{xx}$ vs $1/T$ is maximum if the Fermi level is located exactly in the middle between the two mobility edges, which corresponds to an integer value of the filling factor ν. In Fig. 9.18a, the maximum slope is attained at $\nu = 2$. The change of the slope $\Delta\varepsilon^*$ with the variation of B or V_g determines the shift of the Fermi level, $\Delta\varepsilon_F = \Delta\varepsilon^*$. The number of electrons, Δn, at the localized levels in the energy interval $\Delta\varepsilon_F$ is determined either from ΔV_g according to eqn (9.15) or from ΔB according to eqn (9.13). Thus, we immediately obtain the density of states in the energy gap between the Landau levels, $g(\varepsilon) = \Delta n/\Delta\varepsilon$. As is seen from Fig. 9.18b, this density has an unexpectedly high value – about 10% of the density of states g_0 in zero magnetic field.

However, note that the measured density of states also depends on the electrons located in the reservoir. This is especially important if the reservoir levels are located so close to a 2D gas that they are in thermodynamic equilibrium with it.

The existence of a finite density of states in the gap between the Landau levels indicates that considerable lowering of the temperature in the QHE mode should change the conductivity over the delocalized states given by eqn (9.26) to the variable-range hopping conductivity over the states in the direct vicinity of the Fermi level (see Chapter 4). Figures 9.19 and 9.20 (with the range of
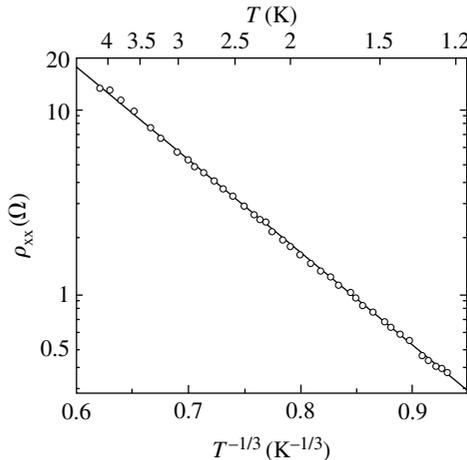

FIG. 9.19. Temperature dependence of the longitudinal resistivity of the GaAs–Al$_x$Ga$_{1-x}$As heterostructure in the QHE mode in the temperature range 5–1 K. Electron concentration $2.4 \cdot 10^{11}$ cm^{-2}, mobility $0.8 \cdot 10^5$ cm^2/V·s (Tsui et al. 1982).
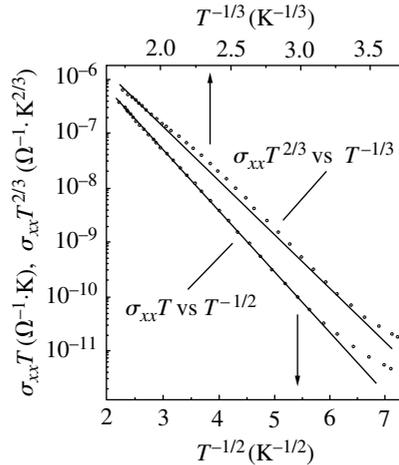
FIG. 9.20.   Temperature dependence of the longitudinal conductivity of the GaAs–Al$_x$Ga$_{1-x}$As heterostructure in the Corbino geometry in the QHE mode at $B = 4.4$ T in the temperature range 0.2–0.02 K. Electron concentration $3.2 \cdot 10^{11}$ cm^{-2}, mobility $0.38 \cdot 10^5$ cm^2/V·s (Ebert et al. 1983). The original $\sigma_{xx}(B)$ curves for this sample at different temperatures were shown in Fig. 9.9.

temperatures lower than the range of temperatures in Fig. 9.18) show that the change of the type of longitudinal conductivity really takes place. The exponent $T^{-1/3}$ in the activation law indicates a constant density of states in the vicinity of the Fermi level (see eqn 4.23), whereas the exponent $T^{-1/2}$ usually indicates the existence of a Coulomb gap (see eqn 4.29). Two different exponents obtained in two different experiments may indicate the change of the hopping conductivity mode during the downward shift along the temperature scale. Indeed, the interval used in Fig. 9.19 is located in the region of lower temperatures than in Fig. 9.20 (cf. Figs 4.3 and 4.4 in Chapter 4). On the other hand, it may also be possible that this fact simply reflects different properties of the two samples.

The problem of the density of states in the vicinity of the levels with extended states is more interesting and, to some extent, is also more important. Figure 9.13a shows the percolation net for a large-scale random potential $U(\boldsymbol{r})$. Formally, a random potential eliminates degeneracy and, thus, makes the level corresponding to the percolation net nondegenerate. However, we have already indicated various factors favorable for delocalization of electrons with energy close to the percolation threshold. These factors are: magnetic breakdown, random short-period fluctuations, finite sample size, finite temperatures, etc. To evaluate the relative number of extended states in one Landau miniband, we assume that the characteristic cell size of the percolation net is Λ. This size is determined by the nature of the potential. Thus, if the random potential $U(\boldsymbol{r})$ is formed due to impurities located in the depth of an insulator at a certain

distance from the 2D layer, then $\Lambda \propto \mathcal{N}^{-1/2}$, where $\mathcal{N}$ is the 2D concentration of these impurities. The total length of the "fibers" of this net per unit area of the 2D layer is of the order of $2/\Lambda$. Assuming that the fiber width is not less than $2r_{\mathrm{B}}$, we may state that the fibers of the infinite net would occupy at least a fraction $4r_{\mathrm{B}}/\Lambda$ of the total area of the 2D layer. It follows from eqn (9.8) that the fraction of the delocalized states in the Landau minibands, η_{ext}, is also of the same order of magnitude,

$$\eta_{\mathrm{ext}} \sim r_{\mathrm{B}}/\Lambda. \tag{9.28}$$

In Figs 9.11 and 9.12, the unhatched area in the center of the density-of-states curve which describes a Landau miniband equals η_{ext} whereas the total area under the curve is unity.

9.6. Chains of phase transitions

Examining the $R_{xx}(B)$ and ρ_{xy} curves in Figs 9.7 and 9.8 as well as the σ_{xy} curves in Fig. 9.9, we immediately come to the conclusion that the states at the QHE plateau are separate phase states with specific transport properties

$$\sigma_{xx} \to 0, \quad \sigma_{xy} = i(e^2/2\pi\hbar), \quad i = 0, 1, 2, 3, \ldots \tag{9.29}$$

The phases with $i > 0$ in eqn (9.29) cannot be considered as insulators because of the properties $\rho_{xx} = 0$ and $\sigma_{xy} = i(e^2/2\pi\hbar) \neq 0$. At the same time, the property $\sigma_{xx} = 0$ does not allow one to consider these states as metallic ones. Therefore, they are called *quantum Hall liquids* with different quantum Hall numbers (Kivelson et al. 1992). Which of the quantum Hall numbers is implemented in practice depends on the filling factor given by eqn (9.13). In order to understand how the smooth variation of ν results in the transition from the quantum number i to the quantum number $i \pm 1$, consider Fig. 9.11 or Fig. 9.12. In these figures, the parameter governing the filling factor ν is the voltage V_{g} or the field B. Let the evolution of the filling factor ν be such that the Fermi level approaches the unhatched region. If the integral density of the extended states η_{ext} is comparatively low, then, with the change of ν, the Fermi level rapidly passes the corresponding region and the change of the governing parameters results in a phase transition from one quantum-Hall-liquid state to another. This transition is accompanied by a jump in the transverse conductivity by the value $|\Delta\sigma_{xy}| = e^2/2\pi\hbar$. In this case, the metallic properties are possessed only by the boundary state, whereas on both sides of the transition, $\sigma_{xx} = \rho_{xx} = 0$. If the density of states in the vicinity of this level is high, then the layer of extended states pins the Fermi level. Then the phase transition is split into two transitions; a metallic state with the Fermi level inside the layer of extended states is formed between two states of the Hall liquid with the i values differing by unity. Which of the two variants is realized in practice depends on a number of factors determining the concrete realization of a random potential.

All this may be illustrated by the experimental curves. In the case of an intermediate metallic phase, the slope of the step of ρ_{xy} and the width of the peak
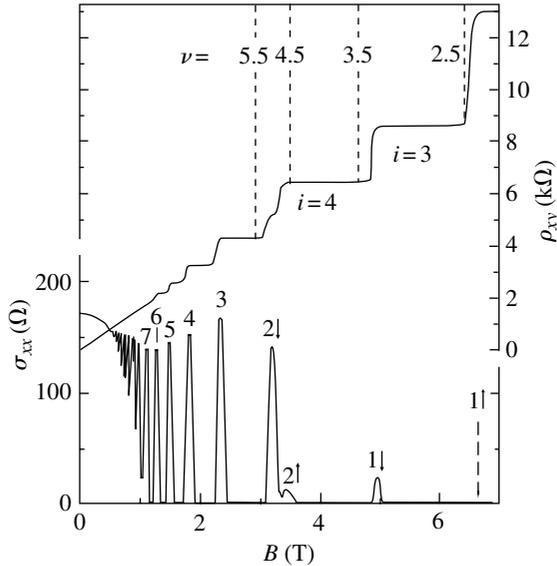
FIG. 9.21. Magnetoresistance ρ_{xx} and Hall resistivity ρ_{xy} of the GaAs–
Al$_x$Ga$_{1-x}$As heterostructure at temperature 50 mK. Electron concentration
$4 \cdot 10^{11}$ cm^{-2}, mobility $0.86 \cdot 10^5$ cm^2/V·s (Paalanen et al. 1982). Peak 1↓ of
the ρ_{xx} curve is comparatively low and peak 1↑ is not formed at all. The ver-
tical dashed lines indicate a half-integer filling factor ν at which the simplest
model would undergo phase transitions with the jumps in the Hall resistivity
ρ_{xy} (the vertical dash lines are plotted after the paper by Cheng in the book
edited by Prange and Girvin (1990) where these experimental data are also
reproduced).

R_{xx} should tend to finite values with lowering of the temperature. The curves in
Fig. 9.7 illustrate such behavior: with lowering of the temperature, the R_{xx} peaks
increase (Ebert et al. 1983). The ρ_{xy} curves in Fig. 9.21 (Paalanen et al. 1982)
behave somewhat differently: the peaks between the plateaus with low i values
decrease and, at low temperatures, completely disappear. This was observed des-
pite the fact that the sample parameters in both experiments were rather close.

If a slope $\partial \rho_{xy}/\partial \nu$ of the step rise between two plateaus tends to a finite
value with lowering of the temperature, it may be used to estimate the fraction
of delocalized states in the Landau miniband, eqn (9.28). In the sample whose
curves are shown in Fig. 9.7, the fraction for the lower minibands amounts to 3%.

The region corresponding to the quantum number $i = 0$, should possess some
specific properties: according to eqn (9.23), all the components of the $\widehat{\sigma}$ tensor
are zeroes: $\sigma_{xx} = \sigma_{xy} = 0$, and diagonal components of the resistivity tensor $\widehat{\rho}$
are infinite: $\rho_{xx} = \infty$. Therefore, this region is named a *Hall insulator*.

The energy diagrams in Fig. 9.11 and Fig. 9.12 with nonoverlapping symmet-
rical Landau minibands lead to the assumption that the transitions should take

place at half-integer filling factors

$$\nu = \nu_c^{(0)} = (i+1)/2, \qquad i = 0, 1, 2, \ldots . \tag{9.30}$$

In Fig. 9.21 illustrating the experiments at a fixed n concentration, eqn (9.30) determines the fields B_i in which one may expect phase transitions and the corresponding steps in the ρ_{xy} curve. As is seen from Fig. 9.21, relationship (9.30) is fulfilled only approximately.

The discrepancy between the ν_c and $\nu_c^{(0)}$ values may be caused by several factors.

First, it may be the asymmetry in the density and the arrangement of the positively and negatively charged sources of a random potential, which, in turn, may give rise to the displacement of the layer of extended states from the center of the Landau miniband. Second, the shift of the ν_c values may be caused by the level broadening exceeding the spacings between the levels. If the tail of the localized states of the miniband $\mathcal{A}$ extends outside the maximum of the density of states of the miniband $\mathcal{B}$ and *vice versa*, then the number of states in the energy layer between two maxima corresponds to the change of the filling factor $\Delta\nu < 1$. This would result in the convergence of the two transitions from the initial positions $(i-1)/2$ and $(i+1)/2$ on the ν scale. This seems to explain the approaching of the transitions $2\uparrow$ and $2\downarrow$ in Fig. 9.21.

In weak magnetic fields, one more cause may give rise to the shift of the phase boundaries toward higher ν values – the so-called *level floating-up* or levitation. Level floating-up was predicted by Khmelnitskii (1984, 1992) based on the assertion that the description of the QHE with delocalized states inside each Landau miniband in the limit $B \to 0$ should be consistent with the scaling hypothesis which states that all the 2D electron states in zero magnetic field should be localized (see Chapter 6). With a decrease in the magnetic field, an ever increasing number of Landau levels descend below the Fermi level, $N \approx \varepsilon_F/\hbar\Omega$. This should also be accompanied by the accumulation of the same number of extended states below the Fermi level. On one hand, the extended states cannot gradually be transformed into localized ones; on the other, no such states would exist in the zero field.

To solve this seeming paradox, let us analyze the transport properties of a 2D electron gas in the region $\Omega\tau < 1$.

On the terminology. The fields $\Omega\tau < 1$ are usually called weak fields. However, in this chapter, we consider the fields $\hbar\Omega(N_0 + 1/2) \gtrsim \varepsilon_F$ where $N_0 = 4$–5 (Fig. 9.2), i.e., the fields in which a magnetic length r_B given by eqn (9.9) only slightly exceeds the average distance $n^{-1/2}$ between 2D electrons, $B \sim (c\hbar/e)n$. Therefore, in this case, the inequality $\Omega\tau < 1$ signifies, in fact, a strong field at very strong disorder, $\hbar/\tau \sim \varepsilon_F$.

Consider the classical equation (9.5) for the components of the conductivity tensor in a magnetic field and the relationship

$$\sigma_{xy} = (\Omega\tau)\sigma_{xx} \tag{9.31}$$

that follows from there.

Equation (9.31) should work locally on a small scale of the order of the magnetic length r_B given by eqn (9.9). On the scales of the order of a sample size, the formulas for the QHE are valid and yield for the transverse conductivity the values (9.23) at the plateaus and the values

$$\sigma_{xy} = \left(i + \frac{1}{2}\right)\left(\frac{e^2}{2\pi\hbar}\right).$$ (9.32)

in between, where the $(i+1)$st delocalized state attains the Fermi level. In order to sew these two scales together, substitute the quantum expression (9.32) for σ_{xy} into eqn (9.31) and retain the classical expression (9.5) for σ_{xx}. Now, solving the equation thus obtained for the electron concentration n that enters the classical expression for σ_{xx}, we obtain

$$n = \left(i + \frac{1}{2}\right)\frac{1}{2\pi r_B^2}\frac{1 + (\Omega\tau)^2}{(\Omega\tau)^2}.$$ (9.33)

Since the Fermi energy in a 2D gas depends linearly on the concentration n, we obtain from eqn (9.33) an expression for the energy $E_\perp$ of the N-th extended state (Fig. 9.22) in the form

$$E_\perp(N) = \left(N + \frac{1}{2}\right)\hbar\Omega\frac{1 + (\Omega\tau)^2}{(\Omega\tau)^2} = \varepsilon_\perp(N)\left(1 + \frac{1}{(\Omega\tau)^2}\right).$$ (9.34)

For convenience, we used in eqn (9.34) the same notation as in eqn (9.7) at the beginning of this chapter (i.e., $\varepsilon_\perp$ and N instead of i).

In the derivation of eqns (9.33) and (9.34), we used a somewhat dubious approach comparing the conductivities determined on different scales. Now, we make another rather arbitrary generalizing assumption that eqn (9.34) is also valid at $\Omega\tau < 1$. Since the fraction in eqn (9.34) at high $\Omega\tau$ values is of the
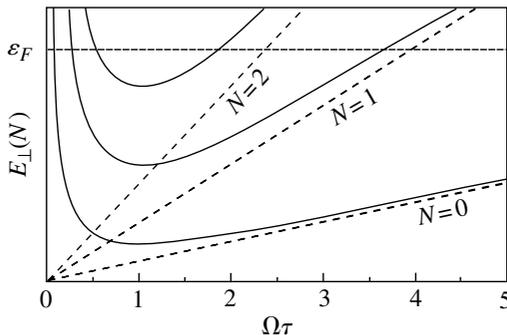

FIG. 9.22.   Level floating-up. Solid curves represent the energies of the extended states as functions of $\Omega\tau$ in accordance with eqn (9.34); dashed straight lines starting from the origin show the should be positions of the half-filled Landau levels in accordance with eqn (9.30).

order of unity and rapidly increases at low $\Omega\tau$, the energy $E_\perp(N)$ increases with a decrease in $\Omega\tau$. This phenomenon is called floating-up of extended levels. With a decrease in the magnetic field, the energy of the extended levels increases and the extended levels, one after another, start crossing the Fermi level.

Note that an increase of $E_\perp(N)$ given by eqn (9.34) in weak fields, i.e., floating-up, takes place in the region $\Omega\tau < 1$, where the Landau classification of the levels is somewhat inadequate. Therefore, eqn (9.34) cannot be considered as the dependence of the energy of a particular extended level on B. However, the existence of the extended levels and their energy do not depend on the basis used for their description. Therefore, one may hope that the scaling procedure used in the transition from eqn (9.31) to eqn (9.33) adequately determines the behavior and the sign of the quantum correction caused by scattering. Thus, the problem reduces to the experimental verification of such behavior.
Note also that eqn (9.34) is written for the energies of extended states and not for the centers of gravity of the Landau minibands.

The speculations underlying the conclusion about level floating-up are not quite rigorous. In fact, this is rather an extension of the scaling hypothesis (Chapter 6) to two-dimensional systems in magnetic fields. For a more rigorous consideration, one has to use much more sophisticated mathematical tools (see, e.g., Pruisken 1990). Therefore, as earlier in Chapter 6, we concentrate our attention on the experimental consequences and verification of the above conclusions.

The experiments confirmed, at least partly, the concept of level floating-up. Floating-up is best illustrated by the measurements of the positions of the maxima of the dissipative component of resistivity ρ_{xx} by varying the field B or the concentration $n \propto V_g$. The maxima are observed at the same B and n values at which the extended states are located at the Fermi level. In the simplest model, these are the straight lines in the (B, n) plane described by eqn (9.30), so that the concentration should decrease linearly with the field. However, it is seen from Fig. 9.23 (Glozman et al. 1995) that at low i values, this is not always so. Indeed, with a decrease in the field, the maximum of the function $\sigma_{xx}(n)$, first, really moves toward lower concentrations, but then shifts back. One may also establish the limits of a possible field decrease: in the fields $\Omega\tau < 1$, the maximum is broadened and the measurement accuracy decreases. Nevertheless, an increase in the energy of the lowest extended state may be reliably observed in weak fields (Fig. 9.24b).

Equation (9.34) and Fig. 9.22 show that the extended delocalized states should preserve their individuality up to the crossing of the Fermi level. Up to now, this theoretical prediction has not been either proved or disproved. It follows from numerous studies that the energies of the extended states cannot be lower than a certain level and that the energy of the lowest of these states increases with a decrease in the field. However, it is still uncertain how these states behave with respect to one another.
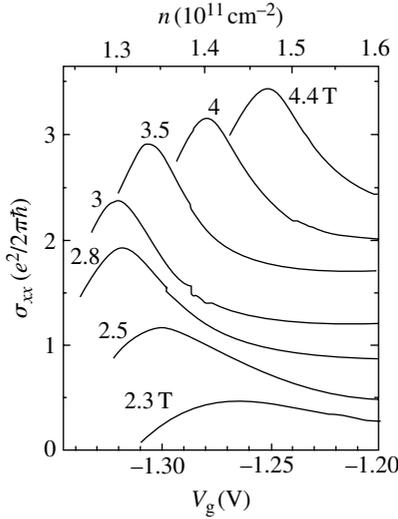
FIG. 9.23.   The maximum of the function σ_{xx} corresponding to the lowest deloc-
alized state in the low-mobility GaAs–Al$_x$Ga$_{1-x}$As heterostructure at different
electron densities (Glozman et al. 1995). The left scale corresponds to the curve
in the field $B = 3\,\text{T}$; for a clearer representation, all the other curves are shif-
ted in the upward (curves in stronger fields) or downward (curves in weaker
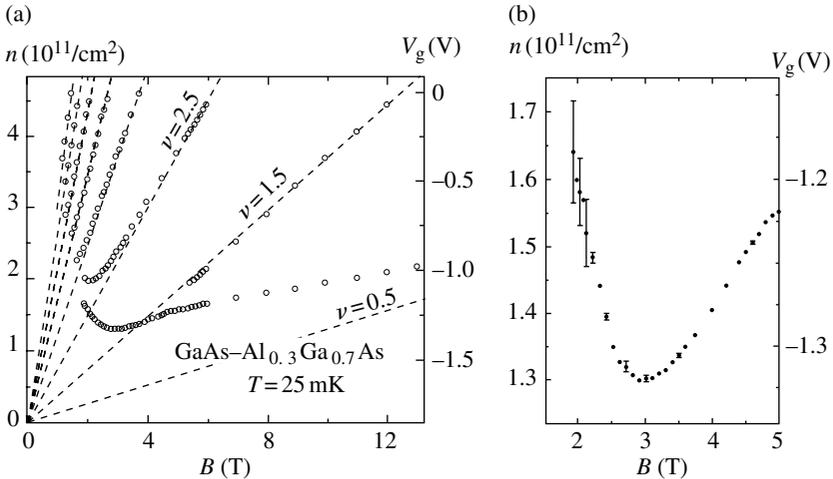fields) directions. Temperature $T = 25\,\text{mK}$.

FIG. 9.24.   (a) Positions of several lower delocalized states in the low-mobility
GaAs–Al$_x$Ga$_{1-x}$As heterostructure in the (B, n) plane (Glozman et al. 1995).
Dashed lines are drawn in accordance with eqn (9.30). (b) The same for the
lowest state on an enlarged scale (Glozman et al. 1995).

9.7. Two-parametric scaling

For different approaches to the concepts underlying two-parametric scaling we recommend the review by Huckestein (1995) and the short article written by the same author (Huckestein 2000).

The scaling hypothesis in the application to a disorder-controlled metal–insulator transition was discussed in Chapter 6. Now, the problem reduces to the generalization of the approach used in Chapter 6 and its application to 2D systems in strong magnetic fields. In fact, we have already started this process by formulating and verifying the assumption that levels may float up. Now, we take the next step and introduce another scaling diagram instead of the diagram with $\beta(y)$ functions used in Fig. 6.1.

When selecting a function that represents the state of a 2D system, there is no need to choose between conductivity and conductance like in the case of a metal–insulator transition, because, in this case, they coincide. However, because of an applied magnetic field, the conductivity has two components, σ_{xx} and σ_{xy}, and the logarithmic derivatives β_{ij} of both with respect to the system length L can be presented as functions of the conductivity (cf. eqn 6.3 in Chapter 6):

$$\frac{d\ln\sigma_{xx}}{d\ln L} = f_1(\sigma_{xx}, \sigma_{xy}), \quad \frac{d\ln\sigma_{xy}}{d\ln L} = f_2(\sigma_{xx}, \sigma_{xy}), \tag{9.35}$$

(it is assumed that the temperature $T = 0$, as it was when eqn (6.3) in Chapter 6 was written). In a metal–insulator transition, the system behavior was determined by only one parameter – the conductance y, whereas in the case under consideration, we have two parameters, σ_{xx} and σ_{xy}; whence the term two-parametric scaling.

Eliminating the variable L from both expressions of eqn (9.35), we arrive at the relationship between σ_{xx} and σ_{xy}, which may be represented in graphical form as curves in the $(\sigma_{xy}, \sigma_{xx})$ plane (Khmelnitskii 1982). This plane is called the phase plane and the curves, the phase trajectories.

On the terminology. A similar representation is widely used in mechanics, where the system motion is described as the trajectory of a mapping point in the $(q, \dot{q})$ plane, where q is the generalized coordinate. This explains the terms "phase plane" and "phase trajectory," where the word "phase" signifies the motion characteristic and not the state of a substance. Since the problem under consideration is from solid state physics and focuses upon the states of substances, then, in order to avoid any misunderstanding, the terms "flow diagram" and "flow trajectories" are often applied to the set of curves describing two-parametric scaling. These terms are also borrowed from mechanics. As will be seen hereinafter, the flow diagram also includes the lines which determine the positions of interphase boundaries in the $(\sigma_{xy}, \sigma_{xx})$ plane and, thus, also represents the phase diagram determining possible transitions between the states of a 2D electron system.

The phase trajectory of a mechanical system starting from the point $(q_0, \dot{q}_0)$ set by the initial conditions determines the system evolution in time, which figures as an implicit parameter. In the QHE flow diagram, the role of time

is played by the system size L, and the initial condition is understood as the values of components of the conductivity tensor, eqn (9.5), which are valid in the classical limit within the length L of the order of a magnetic length r_B given by eqn (9.9). The motion along the flow trajectory signifies a gradual transition to larger lengths L. This should result, first, in the allowance for quantum corrections. A further increase in L should lead to more pronounced corrections and the initiation of the localization processes, which, finally bring the system to the QHE state.

Based on the above reasoning, we may make a draft of the diagram with scaling flow trajectories (Fig. 9.25a). This would be a variant of the scaling hypothesis suitable for 2D systems in a strong magnetic field. As $L \to \infty$, the QHE is observed and conductivity takes the values given by eqn (9.23), so that *all the trajectories* should converge into the points

$$\left(\frac{e^2}{2\pi\hbar} i, 0 \right), \quad i = 0, 1, 2, \ldots \tag{9.36}$$

located on the σ_{xy}-axis. In Fig. 9.25, these are the points A_i, the stationary singular points of the flow diagram. All the trajectories in the vicinity of such a point are directed toward this point. Each of the points A_i corresponds to its
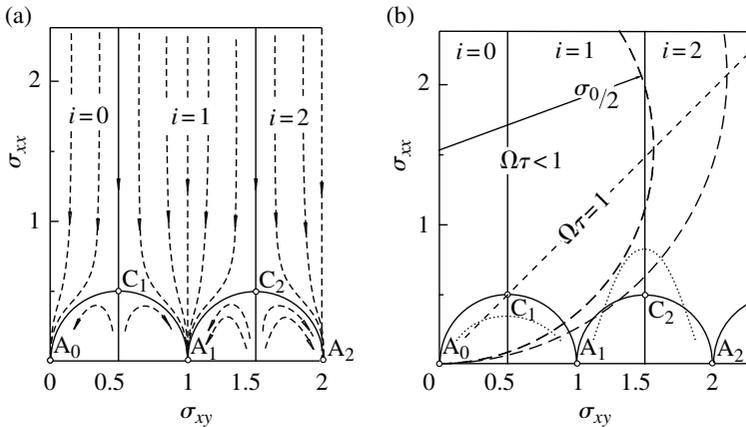

FIG. 9.25. (a) Scaling phase plane, also called a flow diagram. The coordinates are the components σ_{xy} and σ_{xx} of the conductivity tensor in dimensionless $e^2/2\pi\hbar$ units. Solid lines are separatrices, dash lines are the flow trajectories, A_i are stationary singular points, C_i are nonstationary singular points. (b) Positions of the initial points of the flow trajectories determined by n, τ and B values relative to the separatrices in the flow diagram. The dashed line inclined at 45° to the axes separates the regions with $\Omega\tau > 1$ (below) and $\Omega\tau < 1$ (above). The two dashed semicircles, eqn (9.37), are the geometric loci of the points satisfying eqn (9.5) at fixed σ_0 and varying $\Omega\tau$ values. About the dotted curves see the text.

own phase of the quantum Hall liquid with the quantum number $i > 0$ (here the word "phase" has the sense of the state of a substance) or to a Hall insulator at $i = 0$. The phase plane is divided into the regions inside which all the trajectories converge into one of the points A_i. The lines separating these regions are called separatrices.

Let a certain point $(\sigma_{xy}^0, \sigma_{xx}^0)$ be an initial point. The quantum corrections to the longitudinal conductivity were discussed in detail in Chapter 2. Since we discuss here strong magnetic fields, the weak localization may be considered to be destroyed, so that only the interelectron interactions are taken into account. These interactions lead to the logarithmic correction (2.37) to σ_{xx} and in the same approximation they give $\Delta\sigma_{xy} = 0$. Therefore, the initial segments of the flow trajectories are represented by vertical straight lines. Hence, the separatrices are also represented by vertical straight lines.

In each $i < \sigma_{xy}/(e^2/2\pi\hbar) < i + 1$ interval between the stationary points with coordinates given by eqn (9.36), there exists a nonstationary singular point C at a certain $\sigma_{xx} = \sigma_{xx}^* > 0$ in the vicinity of which the flow of trajectories starts branching. Since the quantum phases with different numbers i have similar properties, it is also necessary to require the periodicity of the diagram along the σ_{xy}-axis with period $e^2/2\pi\hbar$. Then the system of separatrices acquires the form of a net represented by solid lines in Fig. 9.25. The vertical separatrices in the flow diagram at the same time divide the regions with different quantum states and, therefore, represent the interphase boundaries of the phase diagram so that the flow diagram is simultaneously a diagram of phase states. The extreme left interphase boundary passing through the point C_1 separates the state of an insulator $i = 0$ from the state of the quantum Hall liquid $i = 1$. According to this diagram, *the insulator $i = 0$ has no boundaries with other states of the Hall liquid, those with $i > 1$*. This signifies that only the $0 \rightarrow 1$ transitions are possible, whereas the $0 \rightarrow 2$, $0 \rightarrow 3$, etc., transitions are forbidden. This is one of the main predictions based on the scaling hypothesis which has to be checked experimentally.

Practically any point in the $(\sigma_{xy}, \sigma_{xx})$ plane may serve as the initial $(\sigma_{xy}^0, \sigma_{xx}^0)$ point of the flow trajectory, i.e., may be regarded as the classical conductivity of the 2D system on small scales. The only exceptions are the points in a narrow strip along the abscissa, where the phase trajectories that describe the fractional quantum Hall effect are located. (This region is beyond our interest; we refer the readers who are interested in this region to the fundamental study by Kivelson et al. 1992).

The arrangement of the initial points becomes clear from Fig. 9.25b. The coordinates of these points depend on three parameters: concentration n, elastic-scattering time τ, and magnetic field B. The product of the first two parameters determines the material conductivity in zero field, σ_0, whereas the two latter parameters enter the dimensionless product $\Omega\tau$. If one fixes both n and τ, thus putting $\sigma_0 = ne^2\tau/m = \text{const.}$, then according to eqn (9.5), with the variation of the magnetic field in the range $0 \leqslant B \leqslant \infty$, the mapping point describes, in

the plane, a semicircle given by the equation

$$(\sigma_{xy}^0)^2 + (\sigma_{xx}^0 - \sigma_0/2)^2 = \sigma_0^2/4. \tag{9.37}$$

The upper part of the semicircle given by eqn (9.37) corresponds to fields $\Omega\tau < 1$, the lower part, to fields $\Omega\tau > 1$. The upper part may be considered as the geometric locus of the initial points of the phase trajectories. It is not so simple about the lower part. When calculating the classical conductivity σ_{xx}^0 with the help of the kinetic equation in the $\Omega\tau \gg 1$ limit, it is important to take into account the quantization of the electron spectrum ignored in relationships (9.4) and (9.5). As the density of states at the Fermi level goes to zero at the integer filling factors $i = 1, 2, 3\ldots$, such quantization makes the conductivity σ_{xx}^0 equal to zero as well, whereas at the half-integer values, $\nu = (i+1)/2$, the " classical conductivity of electrons with the quantized spectrum" attains a maximum. Therefore, one cannot use the lower segment of the semicircle for determining the initial points of the trajectories. The curves that may be used instead of the lower segment of the semicircle are shown by dots in Fig. 9.25b.

Thus, the draft of the diagram is completed and we may proceed to its experimental verification.

Note that since the n, τ and B values determine the initial point of the flow trajectory, their variation results not in motion along the trajectory but in transitions between different trajectories. Therefore, the dependence of the resistivity-tensor components on B or V_{g} (the most widespread type of measurements) is inefficient for studying the diagram structure. For example, it is clearly seen from Fig. 9.21 that the phase transition denoted as 1↓ is shifted with respect to the half-integer filling factor $\nu = 3.5$. However, one may readily see that even this scarce information will be lost in projecting the experimental points from Fig. 9.21 onto the phase diagram.

To move along the trajectory, one has to change the length L of the system still remaining at the temperature $T = 0$. This procedure cannot be performed in practice. However in a real experiment (at a finite temperature), the quantum coherence is lost within the diffusion length, L_φ, determined by eqn (2.7) in Chapter 2. If $L_\varphi < L$, the scaling relationships are determined by the diffusion length. This allows one to move along the phase trajectories by varying the temperature. This requires that the diffusion length should considerably exceed the elastic mean free path l, so that at any temperature, the following inequalities would be fulfilled:

$$l \ll L_\varphi(T) < L. \tag{9.38}$$

Recall: When discussing the effect of the weak localization on the scaling curves in Section 6.1 (see there Table 6.1), we used the opposite substitution: assuming that $T = 0$, we replaced L_φ by L in the formulas of weak localization.

Now, proceed with the experiment on heavily silicon-doped GaAs films (Murzin et al. 2002) for which the inequality $\Omega\tau < 1$ is valid in all the attainable magnetic fields. The 25–40 nm thick GaAs films with the Si (donor impurity) concentration about $1.5\cdot10^{17}\,\mathrm{cm}^{-3}$ were grown by molecular beam epitaxy. This
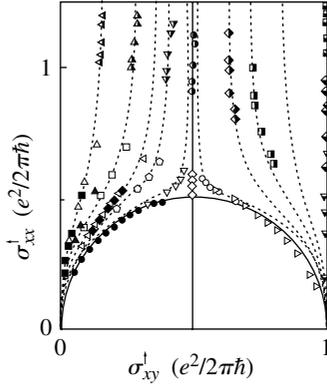
Fɪɢ. 9.26. Scaling flow diagram of heavily doped GaAs films (Murzin et al. 2002). For symbols see the text (the series of equivalent symbols are obtained by temperature variation in the range from 4.2 K to 40 mK). Dashed flow trajectories and the semicircular separatrix $(\sigma_{xy} - 1/2)^2 + \sigma_{xx}^2 = 1/4$ are drawn based on scaling theory (Huckestein 1995, 2000; Pruisken 1990).

method of sample preparation ensures a rather uniform impurity distribution and electron concentration along the 2D layer. From the ρ_{xx} and ρ_{xy} values measured at a certain fixed temperature, the σ_{xx} and σ_{xy} values were calculated. Then, each of these values was divided into two under the assumption that the electronic system consists of two noninteracting subsystems with differently oriented spins, which give the same contributions to the total conductivity. Then the points obtained were plotted onto the flow plane. The series of equivalent symbols in Fig. 9.26 represent the measurements made on the same film in a fixed field at different temperatures. Different symbols indicate the points measured in different fields (from 0.9 to 6.0 T), and, finally, differently filled symbols (filled, filled from below, filled on the right, and empty) correspond to measurements made on four different films. The experiments confirm the common structure and the form of flow trajectories in the vicinity of the interphase boundary $0 \leftrightarrow 1$, including the region $\Omega\tau < 1$.

In a certain sense, it is more important to check the *basic* predictions of the two-parametric scaling hypothesis in the region $\Omega\tau < 1$. Unfortunately, the achievements here are rather modest. Let us move from right to left along this region of the diagram, e.g., along the upper segment of the semicircle $\sigma_0 = $ const. by decreasing the magnetic field (Fig. 9.25b). If the phase boundaries $\sigma_{xy} = (i + 1)/2$, $i = \ldots 2, 1, 0$ are crossed because of a decrease in the field, the Hall conductivity should decrease by $e^2/2\pi\hbar$. Let us specify this prediction on a hypothetical sample for which $\Omega\tau \approx 1$ at the filling factor $\nu \approx 4$. On decreasing from an infinite field $B = \infty$, the resistivity $\rho_{xy}(B)$ would decrease stepwise from high values to $\rho_{xy} \approx 6.5\,\mathrm{k\Omega}$ as in the case of a sample of higher quality in Fig. 9.21. However, a further decrease in the magnetic field would lead in the

hypothetical sample to *an increase* of ρ_{xy} *in discrete steps* instead of a stepwise decrease of ρ_{xy} in the sample with a weak random potential in Fig. 9.21. Up to now, nobody has managed to experimentally observe such a *reverse stepped structure* of the $\rho_{xy}(B)$ function in weak fields. In terms of the level floating-up, this signifies that nobody has managed to observe how the individual extended levels float up one after another and cross the Fermi level (cf. Fig. 9.22 at the end of the previous section).

The experiments performed on pure samples for which the pure limit $\Omega\tau > 1$ is attained in strong fields did not clarify the two-parametric scaling. Some experimental results agree with the scaling diagram, whereas some other experimental results contradict it. There are claims that the transitions exist from the state with $i=0$ directly to the states with $i=2$, $i=3$, etc., which contradict the diagram of two-parametric scaling. These statements are made under the assumption that it is possible to establish the fact of a phase transition and identify the phases based on the sign of the derivatives $\partial\rho_{xx}/\partial T$ at low temperatures. However, this assumption itself should be verified and proved.

The problem of the $0 \to (i>1)$ transitions is closely related to the question of whether only the lowest of the extended levels may float up (which is seen in Figs 9.23 and 9.24) or whether the following extended levels also float up separately. This is well illustrated by the diagram in Fig. 9.27 (Hilke et al. 2002). The open dots are plotted based on the positions of the steps in ρ_{xy} and indicate the transitions between different quantum liquids. Filled dots are plotted based on the analysis of the $\partial\rho_{xx}/\partial T$ derivatives. Examining Figs 9.23 and 9.24, one
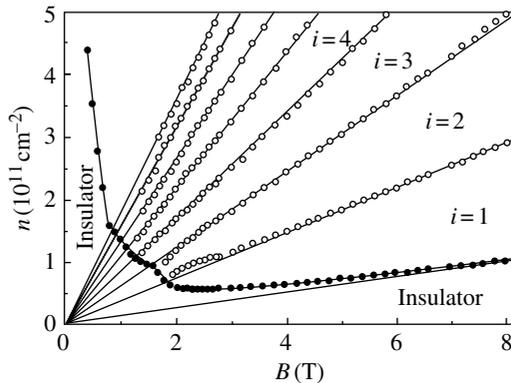

FIG. 9.27. Transitions between different quantum Hall phases of a two-dimensional electron gas in heterostructure Ge–SiGe against the background of the fan of the half-filled Landau levels in accordance with eqn (9.30). Open dots are obtained from the positions of the steps of the ρ_{xy} function; filled dots, from analysis of the intersections of the $\rho_{xx}(B)$ isotherms at different temperatures (Hilke et al. 2002).

can clearly see that, in fact, the filled dots point to floating up of the lowest of the extended levels. If the branches of the empty dots rest on the branch with filled dots so that the levels merge together, then the boundary appears between an insulator and the states with $i > 1$. This makes possible the corresponding $0 \rightarrow (i > 1)$ transitions. If these branches bend up as the levels do in Fig. 9.22, the $0 \rightarrow (i > 1)$ transitions are impossible.

However, since, in any case, the insulator state is adjacent to the ordinate axis, the QHE phase diagram in the limit $B \rightarrow 0$ is consistent with the statement of the scaling hypothesis of the localization of a 2D electron gas in zero magnetic field.

References

Abrikosov, A.A. (1988). *Fundamentals of the Theory of Metals*. North-Holland.

Butcher, B., March, N.H., and Tosi, M.P. (eds) (1993). *Physics of Low-dimensional Structures*. Plenum Press.

Büttiker, M. (1990). Adv. Solid State Phys., **30**, 40.

Chklovskii, D.B., Shklovskii, B.I., and Glazman, L.I. (1992). Phys. Rev. B **46**, 4026.

Dolgopolov, V.T., Zhitenev, N.B., and Shashkin, A.A. (1990). JETP Letters **52**, 196.

Dolgopolov, V.T., Shashkin, A.A., Zhitenev, N.B., Dorozhkin, S.I., and v. Klitzing, K. (1992). Phys. Rev. B **46**, 12560.

Ebert, G., v. Klitzing, K., Probst, C., and Ploog, K. (1982). Solid State Commun. **44**, 95.

Ebert, G., v. Klitzing, K., Probst, C., Schuberth E., Ploog, K. and Weimann, G. (1983). Solid State Commun. **45**, 625.

Gantmakher V.F. and Levinson, Y.B. (1987). *Carrier Scattering in Metals and Semiconductors*. North-Holland.

Glozman, I., Johnson, C.E., and Jiang, H.W. (1995). Phys. Rev. Lett. **74**, 594.

Hilke, M., Shahar, D., Song, S.H., Tsui, D.C., and Xie, Y.H. (2000). Phys. Rev. B **62**, 6940.

Huckestein, B. (1995). Rev. Mod. Phys. **67**, 357.

Huckestein, B. (2000). Phys. Rev. Lett. **84**, 3141.

Iordansky, S.V. (1982). Solid State Commun. **43**, 1.

Kazarinov R.F. and Luryi, S. (1982). Phys. Rev. B **25**, 7626.

Khmelnitskii, D.E. (1982). JETP Lett. **38**, 552.

Khmelnitskii, D.E. (1984). Phys. Lett. **106A**, 182.

Khmelnitskii, D.E. (1992). Helv. Physica Acta **65**, 164.

Kivelson, S., Lee, D.-H., and Zhang, S.-C. (1992). Phys. Rev. B **46**, 2223.

v. Klitzing, K., Dorda, G., and Pepper, M. (1980). Phys. Rev. Lett. **45**, 494.

Koch, S., Haug, R.J. v. Klitzing, K., and Ploog, K. (1991). Phys. Rev. B **43**, 6828.

Krasnopolin, I.Y., Pudalov, V.M., and Semenchinskii, S.G. (1987). Instrum. Exper. Techn. **30**(6), 1275.

Luryi, S. and Kazarinov R.F. (1983). Phys. Rev. B **27**, 1386.

Mohr, P.J. and Taylor, B.N. (2000). Rev. Mod. Phys. **72**, 351.

Murzin, S.S., Weiss, M., Jansen, A.G.M., and Ebert, K. (2002). Phys. Rev. B **66**, 233314.

Paalanen, M.A., Tsui, D.C., and Gossard, A.C. (1982). Phys. Rev. B **25**, 5566.

Prange, R.E. and Girvin, S.M. (eds) (1990). *The Quantum Hall Effect*, 2nd edn. Springer-Verlag.

Pruisken, A.M.M. (1990). In *The Quantum Hall Effect*, 2nd edn (eds Prange, R.E. and Girvin, S.M.). Springer-Verlag.

Stahl, E., Weiss, D., Weimann, G., v. Klitzing, K., and Ploog, K. (1985). J. Phys. C: Solid State Phys. **18**, L783.

Thouless, D.J. (1993). Phys. Rev. Lett. **71**, 1879.

Tsui, D.C., Störmer, H.L., and Gossard, A.C. (1982). Phys. Rev. B **25**, 1405.

Appendices

This page intentionally left blank

A

ELEMENTS OF PERCOLATION THEORY

Elementary mathematical treatments of percolation theory are given in the books by Stauffer and Aharony (1994) and Feder (1988). In applications to the problems of solid state physics, percolation theory is considered by Shklovskii and Efros (1984). Some of the physical applications are also discussed in the review by Isichenko (1992).

Similar to other branches of mathematics, percolation theory is used in various fields of human activity. Below, we briefly consider the main concepts of percolation theory and the consequences that follow from it mainly for the transport problems in inhomogeneous media we are interested in. We consider a medium as a discrete lattice and formulate two simple types of problems. Let us color some randomly chosen lattice sites and use the fraction x of color sites as the main independent parameter. Each site is connected with the nearest neighbours. A connected set of one kind of sites (e.g., all color or all uncolor) is called a cluster. Two color sites belong to the same cluster if it is possible to connect these sites by a continuous chain of color neighbors. Then, the problem of sites may include the determination of an average number of sites in a cluster, the size distribution of clusters, the formation of an infinite cluster, and the determination of the fraction of color sites in it. It is also possible to randomly color the bonds between the neighbor sites and to assume that a cluster contains the sites connected by chains of color bonds. Then again, we have to determine an average number of sites in a cluster, size distribution of such clusters, etc. but now for the problem of bonds.

In some problems, it is natural to rename color sites or bonds into open ones assuming that a flow (e.g. the electrical current) can pass only through open elements. If all the sites (bonds) are closed, $x = 0$, the lattice is a model of an insulator. If all the sites (bonds) are open, $x = 1$, and a current may flow over the open sites and conducting bonds, the lattice is a model of a metal. At a certain critical value $x = x_{\mathrm{c}}$, a percolation transition takes place – a geometrical analogue of a metal–insulator transition. It is most important to consider the percolation theory in the vicinity of this transition, whereas far from such a transition, it is sufficient to consider the approximation of an effective medium.

A.1. Approximation of effective medium

This problem is considered in detail in review article by Kirkpatrick (1973).

Consider an isotropic orthogonal lattice of dimension d with a period a, where all the sites are connected with their nearest neighbors by independent bonds
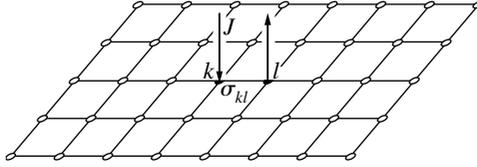
FIG. A.1.   The conductivity σ_{kl} of the kl-bond for a current J flowing through two neighboring sites k and l of a square lattice is shunted by the conductivity of bonds σ_{shunt} in the remaining part of the lattice.

having conductivities σ_{kl}. As an example, Fig. A.1 shows a square lattice with $d = 2$. Knowing the dispersion in σ_{kl} values, one has to calculate an average σ_m value which determines the conductivity of the net

$$\sigma = \sigma_m a^{2-d}. \tag{A.1}$$

Assume first that all the σ_{kl} are equal, $\sigma_{kl} = \sigma_m$. Now attach electrodes to two neighboring sites k and l of the lattice and let the current J flow through the circuit. Let us make a two-step calculation of the current flowing through the bond σ_{kl}. First, remove the electrode l to infinity. Then the current J flowing in through the electrode k is uniformly distributed over the z bonds coming to the site k, so that the current through each bond is J/z. (The parameter z for the square net in Fig. A.1 is $z = 4$; for a three-dimensional cubic lattice, it is $z = 6$.) Now, let us remove the electrode k to infinity, and let the current J flow out through the electrode l returned back from infinity to its initial place. Current J flows to the site l again through z bonds. The current through the bond σ_{kl} is exactly the same as in the first case, J/z, and flows in the same direction. The superposition of these two configurations yields the current distribution we are interested in: the current J through the electrodes at the sites k and l and no current at the infinity. In this case, the current that flows through the kl bond is

$$i_{kl} = \frac{2J}{z}. \tag{A.2}$$

Then, it follows from Ohm's law that

$$\frac{2J}{z\sigma_m} = \frac{J}{\sigma_m + \sigma_{\text{shunt}}}, \tag{A.3}$$

where the denominator in the right-hand side includes the sum of the conductivity of the kl-bond, $\sigma_{kl} = \sigma_m$, and the conductivity σ_{shunt} of the remaining net shunting this bond. The latter can be determined from eqn (A.3),

$$\sigma_{\text{shunt}} = \sigma_m(z/2 - 1). \tag{A.4}$$

Equations (A.2)–(A.4) determine the current distribution between the bond connecting the sites k and l and its environment. The conductivity σ is determined by the quantity σ_m which enters eqn (A.1).

Now, assume that all σ_{kl} are different. Then, it follows from Ohm's law that the following equations are valid for any pair of neighboring sites k and l:

$$\frac{i_{kl}}{\sigma_{kl}} = \frac{J - i_{kl}}{\sigma_{\text{shunt}}}, \qquad i_{kl} = J\frac{\sigma_{kl}}{\sigma_{kl} + \sigma_{\text{shunt}}}. \tag{A.5}$$

These equations should be averaged over all possible pairs of sites. The averaging scheme is called the effective-medium approximation. This approximation is based on two assumptions:

(i) Equation (A.4) for σ_{shunt} is also valid in the system with random σ_{kl}, because shunting is provided by a large number of bonds over which averaging is performed; whence the name of this approximation – the effective-medium approximation; sometimes, it is also called the mean-field approximation. The quantity σ_m defined in eqn (A.4) determines, as earlier, the conductivity in accordance with eqn (A.1).

(ii) Supplying successively the electrodes to a large number of site pairs, we obtain that the average current $\langle i_{kl}\rangle$ is determined by eqn (A.2) derived for a homogeneous medium

$$\langle i_{kl}\rangle = \frac{2J}{z}. \tag{A.6}$$

Averaging eqn (A.5) based on the above two assumptions, we obtain the final expression

$$\frac{2}{z} = \left\langle \frac{\sigma_{kl}}{\sigma_{kl} + \sigma_m(z/2 - 1)} \right\rangle. \tag{A.7}$$

To solve any problem, we have to substitute the distribution function for σ_{kl} into eqn (A.7). If $\sigma_{kl} = 1$ with probability x and $\sigma_{kl} = 0$ with probability $1 - x$, we obtain the solution in the form of a straight line with the slope dependent on the number of nearest neighbors z,

$$\frac{2}{z} = x\frac{1}{1 + \sigma_m(z/2 - 1)}, \qquad \sigma_m = \frac{x - 2/z}{1 - 2/z}. \tag{A.8}$$

Figure A.2 shows the dependence obtained for square ($d = 2$, $z = 4$, dash lines) and cubic ($d = 3$, $z = 6$, solid lines) lattices. The zero value of the function $\sigma_m(x)$ signifies the occurrence of a percolation transition. In the vicinity of this transition, the effective-medium approximation becomes invalid because in this case both assumptions underlying this approximation are unjustified. Therefore, the true critical value x_c or the so-called percolation threshold lies not at the intersection point of $\sigma_m(x)$ and the abscissa, but to the left of it. This becomes clear from the next section.

Note: Figure A.2 also shows the functions $\sigma_m(x)$ for the case where σ_{kl} changes its value (with probability $1 - x$) from unity to $1/2$ and not to zero. It is seen that they are
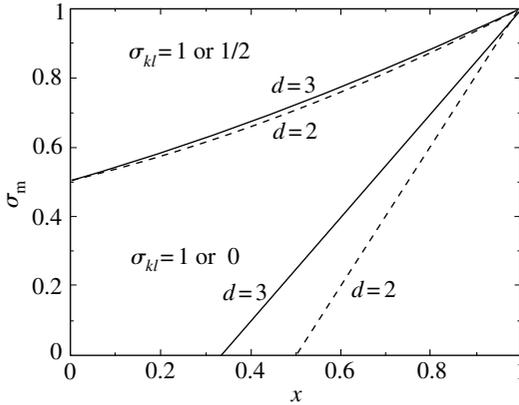
FIG. A.2. Conductivity of square and cubic lattices in the efficient-medium approximation in the case where the $(1 - x)$ bonds are broken (two lower curves) or possess conductivity twice lower than the conductivity of the remaining bonds (two upper curves).

no longer straight lines. At such σ_{kl} values, no percolation transition can take place, and the effective-medium approximation is applicable in the whole range of x variation.

A.2. Percolation thresholds

In the vicinity of a percolation transition, the effective-medium approximation becomes invalid, and the exact results may hardly be obtained by analytical methods. The analytical solutions are obtained for a one-dimensional lattice $(d = 1)$ and an infinite Bethe lattice $(d = \infty)$. The solutions for $d = 2$ were obtained only for some simple lattices. In other cases, the problems are usually solved numerically by computer simulation. For most widespread 2D and 3D lattices, the percolation-threshold values $x_c^{(s)}$ for the site problem and similar $x_c^{(b)}$ values for the bond problem are listed in Table A.1.

The x_c values listed in Table A.1 obey some empirical rules. In the bond problem, the product I_b of the number of bonds z (connecting each site with all the nearest neighbors) by the critical value $x_c^{(b)}$ depends only on the dimension d and is $I_b = z x_c^{(b)} \approx d/(d - 1)$. In the site problem, the corresponding invariant $I_s = f x_c^{(s)}$ has the coefficient f instead of z. This coefficient is equal to the fraction of the volume occupied by contacting spheres of equal radii located at all the lattice sites (for 2D lattices, spheres are reduced to circles, and a volume to an area). As is seen from Table A.1, for 2D lattices, this coefficient is maximal for a triangular lattice; for 3D lattices, it is maximal for a face-centered lattice. This explains the term "close-packed lattice."

The values of the empirical invariants I_s and I_b show that the fundamental parameter in percolation problems is not the lattice symmetry or the number of nearest neighbors but the space dimension. Let color (open) sites be occupied

T<small>ABLE</small> A.1.

	Site problem				Bond problem		
	$x_{\mathrm{c}}^{(\mathrm{s})}$	f	$I_{\mathrm{s}} = f x_{\mathrm{c}}^{(\mathrm{s})}$	Lattice type	$x_{\mathrm{c}}^{(\mathrm{b})}$	z	$I_{\mathrm{b}} = z x_{\mathrm{c}}^{(\mathrm{b})}$
$d = 2$	0.59	0.79	0.47	Square lattice	0.5^*	4	2.0
	0.5^*	0.91	0.46	Triangular lattice	0.35^*	6	2.1
	0.7	0.61	0.43	Honeycomb lattice	0.65^*	3	2.0
$d = 3$	0.31	0.52	0.16	Primitive cubic lattice	0.25	6	1.50
	0.25	0.68	0.17	Body-centered lattice	0.18	8	1.44
	0.20	0.74	0.15	Face-centered lattice	0.12	12	1.44
	0.43	0.34	0.15	Diamond structure	0.39	4	1.56

An asterisk indicates that the marked x_{c} value was calculated analytically.

by conducting spheres with diameters that ensure the contact of the spheres occupying the neighboring sites. Then percolation takes place due to the formation of infinite chains of contacting spheres. The invariant values show that a percolation transition may take place in $d = 2$ space, i.e. in a plane, if about 45% of the plane is conductive and it may also take place in $d = 3$ space, if 16% of the space is conductive.

The site problem may be modified somewhat if for the cluster affiliation it is enough for a site to be among the second, third, etc., nearest neighbors of the site which belongs to the cluster. Thus, the number of neighbors the given site is connected with increases, the conditions for the attachment of the given site to a cluster are facilitated and, as a result, the critical x value is reduced. In Fig. A.3, this is illustrated on the square lattice. The open sites are indicated by crosses. Considering three coordination layers of sites inside the circle of nearest neighbors (Fig. A.3a), we see that four open sites from the total ten form the cluster. If one increases the region of strong interaction to four layers (Fig. A.3b), then seven open sites of the total ten enter one cluster. Similar data for three different three-dimensional lattices are given in Table A.2, which lists the values of the critical concentrations in cluster formation under the assumption that joining is allowed for sites from one, two, or three layers of the nearest neighbors.

The last line of Table A.2 indicates the critical values of the product zx,

$$zx \approx \frac{4\pi}{3} \left(\frac{r}{a}\right)^3 x = \frac{4\pi}{3} r^3 N. \tag{A.9}$$

Here r is the radius of the sphere that envelopes all the layers of sites to which the cite in the origin is connected, a^3 is the unit-cell volume, i.e., the volume per site, and $N = x/a^3$ is the number of color sites per unit volume, i.e., the concentration independent of the lattice parameters. The approximate equation (A.9) is fulfilled the more rigorously, the larger the layer number, ν, and the radius r. Therefore, for sufficiently large r, the $zx_{\mathrm{c}}^{(s\nu)}$ values in the last line of the table should be independent of the chosen lattice. Computer simulation
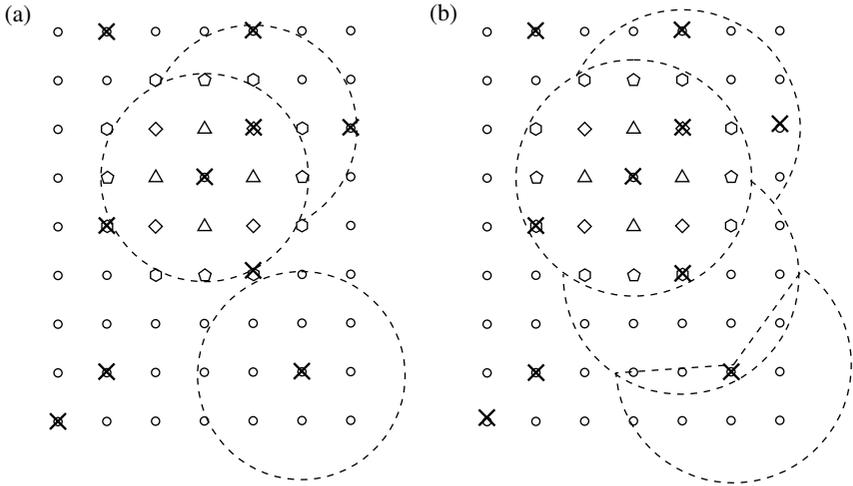
FIG. A.3. Square lattice in which crosses indicate open sites and the circles single out the sites interacting with the site in the center. On the left: circles include three coordination layers of the site in the center (triangles, squares, and pentagons). On the right: the same with four coordination layers (triangles, squares, pentagons, and hexagons).

TABLE A.2.

	Primitive cubic			Body-centered			Face-centered		
Number of layers ν	1	2	3	1	2	3	1	2	3
Number of neighbors z	6	18	26	8	14	26	12	18	42
$x_c^{(s\nu)}$	0.31	0.14	0.10	0.25	0.175	0.095	0.195	0.14	0.06
$zx_c^{(s\nu)}$	1.84	2.45	2.52	1.94	2.45	2.47	1.84	2.45	2.52

confirms this statement and also shows that, as $\nu \to \infty$, $zx_c^{(s\nu)}$ tends to its limit, $B_c^{(3)} = 2.7$.

The universal nature of the $zx_c^{(s\nu)}$ limit for different lattices is not accidental. At large r, one may displace the open sites from their positions by distances $s \ll a \ll r$ without any serious consequences. These displacements signify the transition to a new problem, *percolation in a system of random sites*, equivalent to the site problem in a lattice but at a rather high ν. For percolation at random sites, we have

$$\frac{4\pi}{3} r^3 N_c = B_c^{(3)} = 2.7; \qquad \pi r^2 N_c = B_c^{(2)} = 4.4. \tag{A.10}$$

The expression in the left-hand side of the second expression in eqn (A.10) is written for a 2D lattice with circle area instead of sphere volume; the numerical value $B_2^{(c)} = 4.4$ is obtained by computer simulation.

In eqn (A.10), r is the maximum distance at which sites remain connected (interaction radius). Then, relationships (A.10) determine the critical concentration N_c. However, the problem may be formulated differently. At the given concentration N, eqn (A.10) determines the percolation radius r_c, i.e., the minimum interaction radius that ensures percolation.

A particular class of percolation problems is formed by *continuum problems*. Let us set a random continuous function $U_{min} \leqslant U(\boldsymbol{r}) \leqslant U_{max}$ with the average value $\overline{U(\boldsymbol{r})} = 0$ and statistical properties, which are invariant with respect to the transformation $U \to -U$ in the space of dimension d. We call any bound domain in which the inequality $U(\boldsymbol{r}) \leqslant u$ is fulfilled a color cluster and any domain in which the inverse inequality, $U(\boldsymbol{r}) > u$, is valid, an uncolor cluster. The clusters thus defined are analogous to the clusters in the lattice problems. Now, denote the total volume of all the color clusters normalized to the unit volume by $S_1(u)$ and similarly normalized total volume of uncolor clusters, by $S_2(u)$, so that $S_1 + S_2 = 1$. Now, displace the level u from U_{min} in the upward direction. At sufficiently small u, $u \gtrsim U_{min}$, the number of color clusters is rather small and the clusters themselves are also small, but there is one infinite uncolor cluster. With an increase in u, the average size of color clusters increases. The formation of an infinite color cluster at $u = u_{c1}$ and $S_1(u) = S_{c1}$ indicates the occurrence of a percolation transition. At the opposite end of the interval of possible potential-energy values at $u \lesssim U_{max}$, there exists an infinite color cluster. With a decrease in u in this domain, the number of uncolor clusters and their average value increase unless at certain values, $u = u_{c2}$ and $S_2(u) = S_{c2}$, an infinite uncolor cluster is formed. The statistical symmetry of a random potential $U(\boldsymbol{r})$ yields $S_{c1} = S_{c2}$. The concrete S_{c1} and S_{c2} values depend on the space dimension d. We are interested first of all in the cases with $d = 3$ and $d = 2$ considered in brief below.

In the three-dimensional case, $d = 3$, computer simulation yields the values $S_{c1} = S_{c2} = 0.17$. In the range $u_{c1} < u < u_{c2}$, there are both color and uncolor infinite clusters. The continuum three-dimensional percolation model may be used to describe the localization of a classical electron: if $U(\boldsymbol{r})$ is the potential energy of an electron, then the color clusters $U_{min} \leqslant U(\boldsymbol{r}) \leqslant u$ are domains accessible for an electron with energy $\varepsilon = u$. The percolation transition at $u = u_{c1}$ signifies that an electron with the energy $u > u_{c1}$ moving by the laws of classical mechanics may go to infinity.

For a clearer presentation, when passing to a two-dimensional space, $d = 2$, we invoke geographical terminology and assume that $U(\boldsymbol{r})$ is the height of a certain local point with the coordinate $\boldsymbol{r}$ of the district. The local maxima and minima of the $U(\boldsymbol{r})$ function, U_+ and U_-, are the heights of the hills and the depths of the hollows, respectively. Now assume that the water level in this district changes. As long as the level is low, the color clusters are lakes in the vicinity of the points U_- and the coastlines of these lakes are the equipotential lines of the

potential $U(\boldsymbol{r})$ passing around the points U_-. At a low water level, the territory may be crossed by dry land. With a gradual rise of water, its level attains a local saddle point of the function $U(\boldsymbol{r})$, and the lakes would merge together. Further increase of the water level forces the third lake to merge with the first two, and so on. In the reverse limiting situation – a very high water level – only some small islands (uncolor clusters) are seen around the U_+ vertices above the water surface, and the territory may be crossed only by water. Now, the equipotential lines became island perimeters. Therefore, there exists a certain intermediate water level at which the coastline passes through the whole territory although, at some places, where the water level reaches the height of the saddle point, the width of the connecting water neck is confined into a point.

An infinite cluster is statistically isotropic. If such a cluster exists, then the territory may be crossed along this cluster in any direction. On the other hand, if at a certain water level, the territory may be crossed over dry land, then certainly it cannot be crossed by water along the perpendicular direction. This signifies that for $d = 2$, two infinite clusters cannot coexist. Now, consider also the statistical symmetry with respect to the $U \to -U$ transformation. We obtain

$$S_{c1} = S_{c2} = 0.5. \tag{A.11}$$

If there is no statistical symmetry, the percolation threshold may be displaced, so that eqn (A.11) becomes invalid. However simultaneous existence of two infinite clusters is still forbidden, so that, instead of eqn (A.11), we have only the condition $S_{c1} + S_{c2} = 1$.

It is interesting to compare the S_c values obtained for both dimensions with the invariant I_s in the site problem for regular lattices which has the same physical meaning (see Table A.1). In both cases, the S_c values are somewhat higher. However, for $d = 3$, the difference between these values is at the error level, whereas for $d = 2$ it exceeds the possible error. The lattice gives rise to some correlations and elements of regularity in the structure of a random function. This results in some new features in the structure of color clusters and displaces the point of the formation of an infinite cluster.

A.3. Region close to the percolation transition

To study the region around a percolation transition, we introduce several important functions of concentration x of open sites (for definiteness, we consider here the site problem). Let n_s be a number of s-site clusters per lattice site, so that at $x < x_c$ (i.e., in the absence of an infinite cluster), we have

$$\sum_s s n_s = x, \quad x < x_c. \tag{A.12}$$

If $x > x_c$, a certain part $P(x)$ of open sites is included in an infinite cluster:

$$P(x) = \begin{cases} x - \sum_s s n_s, & x \geqslant x_c \\ 0, & x < x_c. \end{cases} \tag{A.13}$$

The function $P(x)$ is called the strength of the infinite cluster.

Now, define the correlation function $q(r)$ as the probability that a site located at distance r from an open site of a finite cluster is also open and belongs to the same cluster. The obvious properties of the function $q(r)$ are $q(0) = 1$ and $q(a) = x$ (a is the lattice parameter). Since, by definition, the cluster is finite, then $q(r) \to 0$ at $r \to \infty$. The average number $S(x)$ of sites in a finite cluster (to which the initial open site also belongs) is $\sum_r q(r)$, and the summation is performed over all the lattice sites.

The quantity $S(x)$ may also be represented directly in terms of the size distribution of clusters. The probability that an arbitrary site belongs to a cluster with site number s is sn_s, and the probability that this site belongs to a certain finite clusters is $\sum_s sn_s$. Therefore, the ratio $w_s = sn_s / \sum_s sn_s$ is the probability for a cluster, to which an arbitrarily selected open site belongs, to contain s sites. Making such a choice many times, we arrive at the average value

$$S(x) = \sum_r q(r) = \sum_s sw_s = \frac{\sum_s s^2 n_s}{\sum_s sn_s}. \tag{A.14}$$

The function $S(x)$ is called the average cluster size. (Note that we consider here *not a linear size, like a diameter*, but an average number of sites in a finite cluster.)

The linear size, in turn, is defined by the ratio

$$\xi^2 = \frac{\sum_r r^2 q(r)}{\sum_r q(r)} = \frac{\sum_r r^2 q(r)}{S(x)}. \tag{A.15}$$

The function $\xi(x)$ is called the correlation length. In fact, it is the average linear size of an average cluster "typical" for the given x. It is important that the function $\xi(x)$ is defined on both sides of the percolation threshold. At $x > x_c$, the finite clusters are located in the "holes" of an infinite cluster. Therefore, the correlation length in this region is usually interpreted as an average hole size in an infinite cluster. The function $S(x)$ is also defined on both sides of the percolation threshold.

The function $P(x)$ goes to zero at the percolation threshold, where the functions $S(x)$ and $\xi(x)$ go to infinity. The main postulate underlying the theoretical description of a percolation transition reads that with the approach to the threshold all these functions vary as powers of the distance to the threshold,

$$\left. \begin{aligned} P(x) &\propto (x - x_c)^\beta, & x > x_c, \\ S(x) &\propto |x - x_c|^{-\gamma} \\ \xi(x) &\propto |x - x_c|^{-\nu} \end{aligned} \right\} \quad \begin{aligned} &\text{The exponents } \gamma \text{ and } \nu \text{ have the same values} \\ &\text{on both sides of the threshold.} \end{aligned} \tag{A.16}$$

The above functional dependence has been repeatedly checked experimentally. Computer simulations not only confirm the power dependence, but also demonstrate the universal nature of the exponents β, γ and ν dependent only on the space dimension d and not dependent either on the lattice symmetry or the problem type. This is the characteristic property of the theory of phase transitions where analogous exponents are called critical indices. Critical indices

TABLE A.3.

		$d = 2$	$d = 3$
Power of an infinite cluster, P	β	5/36	0.417
Localization length ξ	ν	4/3	0.875
Average number of sites in a cluster, S	γ	43/18	1.795

are also universal and depend on the dimension alone. Thus, it has been reliably established that, in terms of the rigorous mathematical description, a percolation transition is analogous to a second-order phase transition.

In this case,

– the fraction x of open sites (bonds) plays the role of temperature;
– the power $P(x)$ of an infinite cluster is analogous to the order parameter;
– the function $\xi(x)$ is the localization length in the transitions of both types;
– the average cluster size $S(x)$ should be compared with a thermodynamic function, e.g., with susceptibility in a magnetic transition.

The critical indices in the vicinity of the percolation transition in two-dimensional systems, $d = 2$, are obtained analytically, and, at $d = 3$, numerically (Table A.3).

A.4. Example: Electrical conductivity of a strongly inhomogeneous medium

Concluding this appendix, consider the application of percolation theory to the problem of electron transport. We consider a model example, but it clearly reveals the essence of the approach used. Consider a primitive cubic lattice with period a in three-dimensional space ($d = 3$).

Let the resistances of the bonds in the lattice be within an *exponentially large* range of values

$$R = R_0 \exp u, \quad 0 \leqslant u \leqslant u_0, \ \ u_0 \gg 1, \tag{A.17}$$

and let u be a random quantity that may take any value from the allowed interval with probability $F(u)$. We select a certain rather low u' value from this interval and preserve all the bonds with the resistances varying from R_0 to $R_0 \exp u'$, whereas all the bonds with higher resistances are temporarily cut. The fraction of preserved open bonds is

$$x = \int_0^{u'} F(u)\, du, \quad \int_0^{u_0} F(u)\, du = 1. \tag{A.18}$$

Let $x < x_c$. This signifies that the set of resistances used cannot ensure a finite lattice conductivity. To attain the threshold $x = x_c$, this set should be complemented with a certain number of bonds with higher resistances. (We recall

that, in accordance with Table A.1, $x_c = 0.25$ in the bond problem for a primitive cubic lattice). Of course, we add the lowest possible resistances from the remaining ones by gradually increasing u' in eqn (A.18). After reaching the threshold, the corresponding u' value is denoted by u_c. For definiteness, let all the u values from the allowed range be equally probable, so that the distribution function is $F(u) =$ const:

$$F(u) = \begin{cases} 0, & u > u_0 \\ 1/u_0, & u \leqslant u_0. \end{cases} \tag{A.19}$$

Then, the critical value of the parameter u is

$$u_c = x_c u_0. \tag{A.20}$$

The resistances $R_0 \exp u_c$ included in the consideration at the concluding stage, connect large finite clusters into an infinite one. They are switched-on in series with all the other resistances and, at the same time, their resistances are higher than all the other resistances. Therefore, the lattice resistivity is controlled by these resistances and is proportional to $\exp u_c$.

Since a current should flow along the bonds that belong to an infinite cluster, the threshold should be increased up to $u'' = u_c + \Delta u$, so that the strength $P(x)$ of an infinite cluster should take a non-zero value. Then it follows from eqns (A.19) and (A.20) that $x - x_c = \Delta u / u_0$ and the correlation length is $\xi = a(u_0/\Delta u)^\nu$.

To estimate the Δu value and, in particular, to prove that $\Delta u \ll u_0$, consider the distribution of the current flowing along an infinite cluster. The larger part of an infinite cluster in the vicinity of a threshold is occupied by former finite clusters attached to the main part by *one* open bond. This is well seen in Fig. A.4 showing the experimental implementation of an infinite cluster in the vicinity of a percolation threshold obtained by computer simulation of the site problem on a 160×160 lattice for $x = 0.6$. The cluster is assumed to be infinite since it connects the opposite (left and right) edges of the lattice.

Each gray part of the cluster is connected to the main part via one open site. The points of attachment are conditionally indicated by black circles through which no current can flow. To ensure the current flow, the part should be attached to the basic circuit at least at two points. The parts which do not meet these requirements are called dead ends. In addition to the dead ends shown in Fig. A.4, there are also many small dead ends. The part of an infinite cluster remaining after the exclusion of the dead ends is a current-carrying part. The sites included in this part are shown by black dots in Fig. A.4.

The white parts in the square in Fig. A.4 have the same average density of open sites as in the regions covered by an infinite cluster. Numerous finite clusters in these parts are not shown. The obvious asymmetry of the shown infinite cluster is explained by the fact that the correlation length ξ is larger than the square side because of the system proximity to the threshold, $\xi > 160$. The current-carrying part of an infinite cluster for a very large lattice may be represented as a net of current-carrying channels situated the correlation length ξ apart (Fig. A.5). In this current-carrying net, both the unit-cell size, ξ, and the
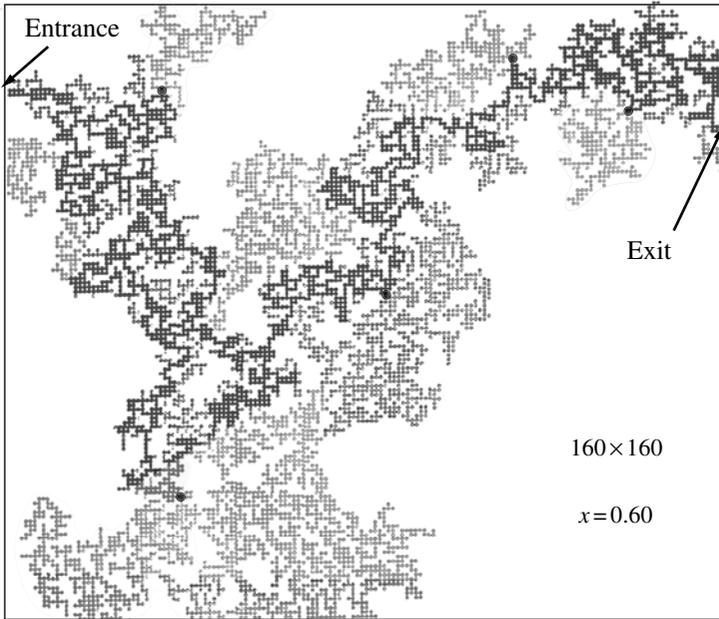
FIG. A.4. Experimental implementation of an infinite cluster in the vicinity of
the percolation threshold obtained by computer simulation of the site problem
on a 160×160 lattice presented in the book by Feder (1988). Only open sites
that belong to an infinite cluster are shown. Dark dots are the sites that belong
to the current-carrying core through which the current may flow from right to
left. Light dots are connected with the current-carrying skeleton via only one
open site; they are included in dead ends. The largest dead ends are shown
by the gray color and the places of their attachment to the current-carrying
skeleton are indicated by black circles.

resistance between two sites, $R_\xi \approx e^{u_c + \Delta u}$, depend on the boundary value u''.
The resistivity of the lattice represented as the resistance of the current-carrying
net is (see, in particular, eqn A.1)

$$\rho = R_\xi \xi = R_0 e^{(u_c + \Delta u)} a \left(\frac{u_0}{\Delta u} \right)^\nu = R_0 a u_0^\nu e^{u_c} \frac{e^{\Delta u}}{(\Delta u)^\nu}. \qquad (A.21)$$

The fraction in the right-hand part of this expression has a minimum at

$$\Delta u = \nu = 0.875 \approx 1 \ll u_0, \qquad (A.22)$$

which signifies that including into the net of the resistances with the u values from
a very narrow interval $u_c < u < u_c + 1$ decreases the lattice resistance because of an
increase in the strength P of an infinite cluster and a decrease in the correlation
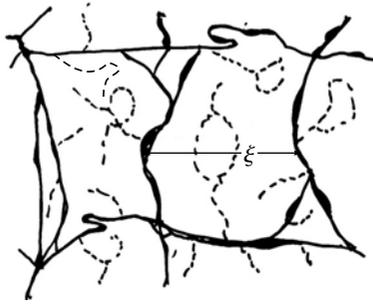
FIG. A.5.   Structure of the core of a current-carrying percolation cluster.

length ξ. The further including of resistances is inefficient because they considerably exceed $R_0 \exp u_c$ and cannot shunt the critical current-carrying net already formed. Within the accuracy of a numerical factor, the lattice resistance is

$$\rho = R_0 a u_0^{\nu} \mathrm{e}^{u_c}. \tag{A.23}$$

References

Feder, J. (1988). *Fractals*. Plenum Press.

Isichenko, M.B. (1992). Rev. Mod. Phys. **64**, 961.

Kirkpatrick, S. (1973). Rev. Mod. Phys. **45**, 574.

Shklovskii, B.I. and Efros, A.L. (1984). *Electronic Properties of Doped Semiconductors*. Springer.

Stauffer, D. and Aharony, A. (1994). *Introduction to Percolation Theory* 2nd edition. Taylor & Francis.

B

TUNNELING CHARACTERISTICS

Two books about tunneling phenomena, Burstein and Lundquist (1969) and Solymar (1972), published many years ago are still very useful. The latest achievements in this field can be found in the review by Aleiner et al. (2002).

Let two conducting materials (M_1 and M_2 in Fig. B.1) be in contact with one another via an insulating layer I so thin that it admits electron tunneling. The whole device is called a *tunnel junction* and the two conducting materials are called electrodes. In equilibrium, the Fermi levels in the electrodes coinside. An electron with energy ε with respect to the Fermi level $\varepsilon_F = 0$, which was in M_1, on the left of the barrier prior to tunneling, after tunneling has the same energy but is located in M_2, on the right of this barrier. The equilibrium is dynamic in the sense that there are electron flows in the opposite directions, but the total flow through the barrier I equals zero.

Compare the tunneling considered in this appendix and the hopping conductivity considered in Chapter 4. In both cases, an electron passes under the barrier through a classically inaccessible region. However, the hopping conductivity proceeds due to quantum transitions between localized states, whereas the tunnel current through a junction arises due to quantum transitions between the delocalized states on different sides of the barrier.

In most instances, the resistance of the electrodes is much lower than the resistance of the junction. Therefore the potential difference V applied to the
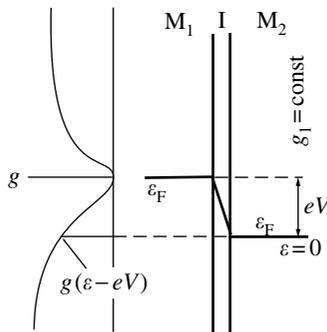
FIG. B.1. Energy scheme of current flow through a tunnel junction M_1–I–M_2.

electrodes is concentrated within the tunneling gap. The tunnel junction becomes out of balance, and the tunneling current J,

$$J(V) \propto \int_{-\infty}^{\infty} g(\varepsilon - eV) g_1(\varepsilon) \left[f\left(\frac{\varepsilon - eV}{T}\right) - f\left(\frac{\varepsilon}{T}\right) \right] d\varepsilon, \qquad \text{(B.1)}$$

starts flowing through the junction (Fig. B.1). The integrand in eqn (B.1) describes the resultant flow of electrons with energy ε from the left to the right caused by the relative shift of the Fermi distribution $f(x) = (\exp x + 1)^{-1}$ on both sides of the barrier by eV. The densities of states g and g_1 enter the integrand as cofactors, because they determine the number of electrons participating in tunneling and the number of states to which tunneling is possible. We omit the proportionality coefficient which should take into account the area of the tunneling contact and the barrier transparency.

Expression (B.1) is considerably simplified if one of the electrodes is a conventional metal with $g_1(\varepsilon) = \text{const.}$ and the temperature is so low that the Fermi distribution may be considered as a step. Then

$$J(V) \propto g_1 \int_0^{eV} g(\varepsilon) \, d\varepsilon. \qquad \text{(B.2)}$$

The addition of an AC voltage $V_\omega \sin \omega t$ to the DC voltage V results in the appearance of an alternating current with frequency ω and amplitude J_ω proportional to the derivative dJ/dV,

$$J(V + V_\omega \sin \omega t) = J(V) + \frac{dJ}{dV} V_\omega \sin \omega t. \qquad \text{(B.3)}$$

The above experimental trick, called the *modulation* method, is rather popular. It follows from eqns (B.2) and (B.3) that

$$J_\omega \propto \frac{dJ}{dV} \propto g(eV). \qquad \text{(B.4)}$$

This allows one to use tunneling measurements for the direct determination of $g(\varepsilon)$.

Expression (B.4) and the procedure for obtaining the function $g(\varepsilon)$ from the experimental data are somewhat complicated by the finite temperature $T \neq 0$,

$$J_\omega \propto \frac{dJ}{dV} \propto - \int_{-\infty}^{\infty} g(\varepsilon) \, d\varepsilon \frac{\partial}{\partial(eV)} f\left(\frac{\varepsilon - eV}{T}\right), \qquad \text{(B.5)}$$

but the study of the density of states in the vicinity of the Fermi level with the aid of the current–voltage characteristics of the tunnel junction is possible at finite temperature too.

Note: Tunneling experiments are much more complicated than the scheme considered above. One has to create a reliable and reproducible 10–15 Å-thick tunneling break between the electrodes. This experimental problem was solved only in the 1960s.
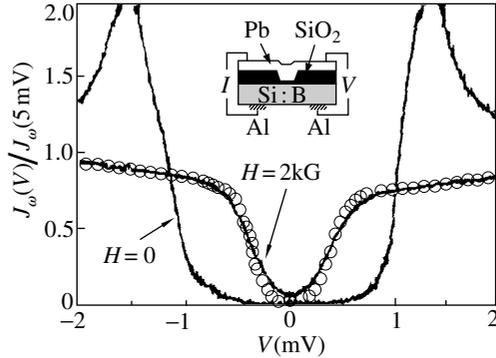
FIG. B.2.   Differential conductivity as a function of bias at the tunnel junction
Pb–SiO$_2$–Si : B. Circles represent the results of the recalculation of the curve
in the field $H = 2$ kG, which eliminates the effect of the thermal broadening of
the distribution function (Massey and Lee 1995). Note the shift of the circles
with respect to the solid line.

Tunneling experiments on superconductors allowed one to directly measure super-
conducting gaps and resulted in the discovery of the Josephson effect. Thus, these
experiments considerably influenced the development of low-temperature physics.

   Being applied to normal metals, the tunneling method allowed one to reveal
the minimum in the density of states at the Fermi level in the spectrum of dirty
metals and the Coulomb gap in the spectrum of localized states and also to
follow the transformation of the former into the latter.
   The revealing of the Coulomb gap in Si : B is illustrated by Fig. B.2. The
Pb–SiO$_2$–Si : B structure with one electrode superconducting and the other a
classical p-type semiconductor was used. The intricate shape of the curve $g(V)$
in zero magnetic field is explained by the use of superconducting lead as the
electrode, whereas Si : B plays the part of a metallic counterelectrode. The corres-
ponding curve depicts the superconducting gap in Pb with the density-of-states
maxima on both sides of the curve. As was to be expected, the curve is sym-
metric with respect to the Fermi level $\varepsilon_F = 0$. In this case, the measurement of
the superconducting gap is, in fact, a calibration experiment. It is an additional
demonstration of the method validity. In a field of 2 kG, the superconductivity
of lead is completely destroyed, and the density of states in lead becomes energy
independent. Now, the lead electrode becomes a counterelectrode, and the min-
imum on the curve $dJ/dV(V)$ is due to the existence of a parabolic Coulomb
gap in Si : B. The 2 kG field is too weak to influence the Coulomb gap. As is seen
from the curve, the gap width is about 1 meV ≈ 10 K (i.e., 5 K on both sides of
the Fermi level).
   The curves in Fig. B.2 were obtained at 1.15 K. To obtain these curves at lower
temperatures is rather difficult, because of an exponential increase in the bulk

resistance of Si : B. A decrease in the potential takes place in a considerable part of the electrode and not within the tunneling gap. However, instead of lowering the temperature in the experiment, it is possible to account for it with the aid of mathematics. A simple transformation of eqn (B.5) allows one to extract the function $g(\varepsilon)$ (empty circles) from the experimental $J_\omega(V)$ curve (solid lines in Fig. B.2). It is seen that as long as the temperature is not too high, it only slightly smoothens the function $g(\varepsilon)$.

The main impurities in Si : B are acceptors and one may directly apply the model of an impurity band at low doping, within which the Coulomb gap was obtained (see eqn 3.22 in Chapter 3). However, the Coulomb gap is also observed in materials of different types, e.g., in binary amorphous films with one metal and one nonmetal components and ultrathin films. We shall illustrate this with several examples.

Figure B.3 shows the tunneling characteristics of the $Ge_{1-x}Au_x$–Al_2O_3–Al structure. Here, Al is a counterelectrode, and $Ge_{1-x}Au_x$ is the material to be studied. The set of characteristics demonstrate the evolution of the spectrum with a decrease in the Au concentration x. The minimum of the curve observed at high x values corresponds to the minimum at the Fermi level in the spectrum. It is due to the interaction of diffusing electrons. This effect, which exists in dirty metals due to disorder, was considered in the last section in Chapter 2. However, the model considered there was obtained within the framework of perturbation theory under the assumption that the corrections to the function $g(\varepsilon)$ were small. The curves in Fig. B.3 not only confirm the existence of the minimum in the spectrum, but also demonstrate the evolution of the minimum with the approach to the metal–insulator transition. At the lowest Au concentration, $x = 0.08$, the $g(\varepsilon)$ function goes to zero at $\varepsilon = 0$, and, in the vicinity of this point, is described by a parabola. This is a Coulomb gap.
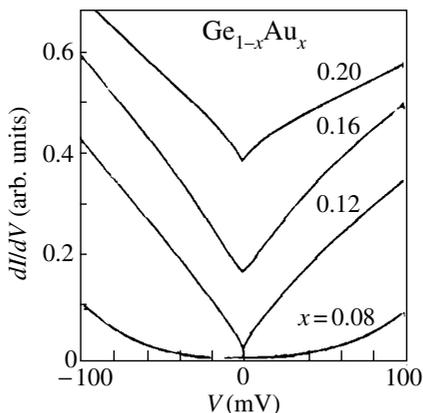
FIG. B.3. Tunneling characteristics of the $Ge_{1-x}Au_x$–Al_2O_3–Al structure (McMillan and Mochel 1981).

Note: The model of the impurity band in a partly compensated semiconductor (within which a Coulomb gap was obtained in Chapter 3) cannot be directly applied to $Ge_{1-x}Au_x$ at low x, because this material does not contain well-defined "donors" and "acceptors," impurity band, random electric field, etc. Therefore, the Coulomb gap revealed in $Ge_{1-x}Au_x$ indicates that a Coulomb gap can arise in different classes of random potentials.

Thus, the series of curves in Fig. B.3 demonstrate three important results – the presence of the minimum at the Fermi level in the spectrum of a dirty metal, which arises due to interelectron interactions (Chapter 2); the existence of a Coulomb gap in the spectrum of a strongly disordered insulator (Chapter 3); and the transformation of the former into the latter, which accompanies the metal–insulator transition driven by the change of the concentration of the metal chemical elements in an alloy (Chapter 5).

A similar spectrum structure and evolution were also observed in amorphous $Si_{1-x}Nb_x$ alloys (Fig. B.4). In the experiments performed on these amorphous alloys, one more interesting feature was observed – the critical vales of concentration determined from the variations of conductivity, x_c, and the tunneling, $x_c^{(t)}$, are somewhat different. The curve with concentration x_c is indicated by an arrow in Fig. B.4. For this state, the function $\sigma(T) \to 0$ with $T \to 0$. However, at this concentration, a finite density of states $g(\varepsilon_F) \neq 0$ is observed at the Fermi level. The density of states $g(\varepsilon_F)$ becomes zero at lower concentration $x = x_c^{(t)}$.

From the definition of a metal given at the beginning of Chapter 5 it follows that the true critical concentration is x_c. The Fermi level of an insulator may
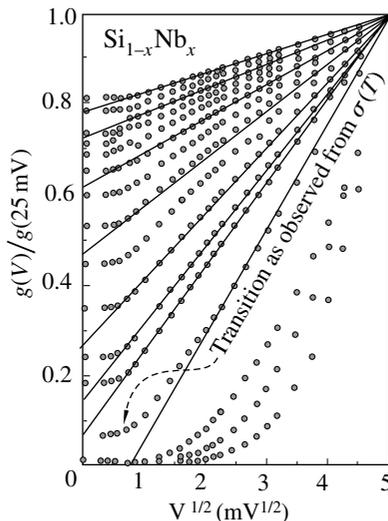
FIG. B.4. Tunneling characteristics of the $Si_{1-x}Nb_x$–Al_2O_3–Al structure (Hertel et al. 1983).

have a finite density of states if these states are localized. Therefore, in fact, the conductivity and tunneling measurements do not contradict each other.

At the same time, the discrepancy between x_c and $x_c^{(t)}$ can be interpreted in a different way. Let us give some comments on measurements of the tunneling current. In the consideration which allowed us to write eqn (B.1) for the tunneling current, it was implicitly assumed that electrons in both electrodes were noninteracting quasiparticles. Hence, the total electron energies on the left and on the right of the barrier were equal to $E = \sum \varepsilon_i$ (the sum of quasiparticles energies) and the tunneling process changed the total energies E on both sides of the barrier by the energy of a tunnelled quasiparticle. The total energy E in the system of interacting particles depends on their number, $E = E(n)$. Then, part of the energy of the electromagnetic field is spent in the change of the total energy of the electron system with the variable number of particles. To avoid any misunderstanding, the density of states g_i in eqn (B.1) (which is measured experimentally) is called the tunneling density of states.

Note: Various types of interelectron interactions act differently on the tunneling density of states. As an example, consider the Coulomb interaction of a tunneling electron with all the other electrons. After tunneling, both electrodes acquire spatial charge inhomogeneities, which have to be resolved. The longer the time it takes to resolve them, the more pronounced the inhomogeneity hinders tunneling, manifesting itself as a decrease in the effective density of states. The resorption time τ_{res} depends, in particular, on the character of electron motion – ballistic or diffusion. In diffusion motion, this time depends on the electron energy measured from the Fermi level (Chapter 2). Therefore, this effect may result not only in the renormalization of the tunneling density of states but also in the change of energy dependence of the function $g(\varepsilon)$.

A similar observation was made in the doped semiconductor Si : B. The detailed measurements illustrated in Fig. B.5 show that here as well $g(\varepsilon_F)$ goes to zero already in the insulator region where $n \approx 0.9 n_c$ (the critical concentration n_c is determined from the dependence of conductivity $\sigma(0)$ on n). The discrepancy in the determination of the critical concentration is about 10%.

However, the minimum of $g(\varepsilon)$ at the Fermi level for the state with $n = n_c$ (where n_c was determined as the critical concentration from the conductivity measurements) turned out in Fig. B.5 to be broader than for the states with other, both lower and higher, electron concentrations. Therefore, the tunneling measurements also single out the "transport value" $n = n_c$ defined above as the true value.

Specific objects of tunneling experiments are ultrathin films. The control parameter in such experiments may be the film thickness, and the quantitative characteristic, the resistivity at a certain fixed temperature. The thicknesses of Be films, whose tunneling characteristics are presented in Fig. B.6, range within 15–20 Å. The indicated values of resistivity were measured at 50 mK. Tunneling structures were created by dosed oxidation of films in air and the subsequent deposition of an Ag layer onto the oxidized film. The thinnest Be film had a spectrum with a Coulomb gap, with the density of states in the vicinity
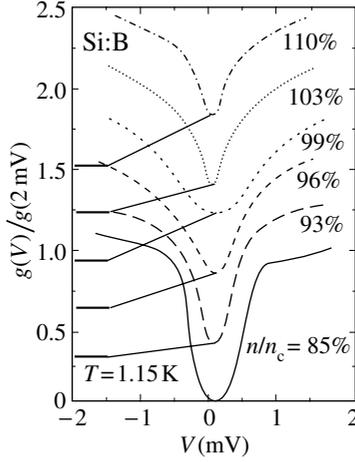
FIG. B.5.   Differential conductivity of the tunnel junction with a Si : B electrode
(Massey and Lee 1996). For a clearer representation, the curves are displaced
along the ordinate axis, but the origin of each curve is indicated on the ordinate
axis. The correction which eliminates the effect of the thermal broadening of
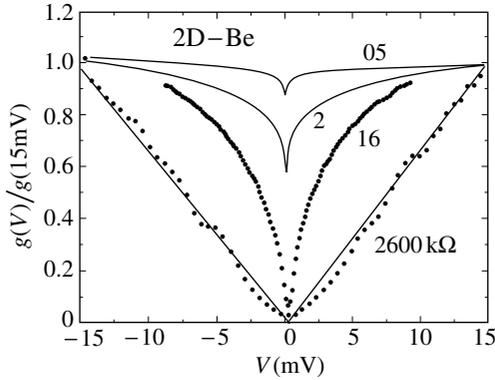the distribution function is included.

FIG. B.6.   Tunneling conductivities of junctions to Be films with different thick-
ness are measured at the temperature $T = 50$ mK. Since the resistivity at
$T = 50$ mK is too high, the measurements for the film with the highest
resistance were performed at $T = 700$ mK (Butko et al. 2000).

of ε_F varying linearly with the energy, in full accordance with the prediction of
eqn (3.23) in Chapter 3. The energy spectra of thicker films showed a narrow
minimum in the vicinity of ε_F. Thus, the ultrathin films, where the thickness
determines the effective disorder and the Anderson transition, also have a min-
imum on the metal side of the transition and the Coulomb gap on the insulator

side of this transition. The specific feature of the curves in Fig. B.6 is that they represent the spectra of certain two-dimensional system.

Figure B.1 and eqns (B.1)–(B.5) describe the simplest scheme of tunneling experiments in which the finite dimensions of electrodes and the area of the tunnel junction are not taken into consideration. All the specific features of the current–voltage characteristics in this scheme are associated with the density of state of a bulky material. Now, let us assume that one of the electrodes is a small metal grain, say, a sphere of radius a. If an electron tunnels in this grain, the latter becomes charged with an electric field around it. The field energy of a single metal sphere with charge e is

$$U = \frac{e^2}{2C}, \qquad C = \kappa a, \tag{B.6}$$

where C is the capacitance of a remote sphere and κ is the dielectric constant of the surrounding insulator. At low temperature, $T \ll U$, the tunneling is possible only if the voltage between the sphere and the bulky electrode exceeds U/e, $V = e/C = e/\kappa a > U/e$. This voltage threshold is called a Coulomb blockade (cf. eqns (8.6)–(8.8) in Chapter 8).

To observe the tunneling effect, both electrodes should be supplied with contacts. To avoid the enlargement of a small electrode by such a contact, two tunnel junctions should be switched in series with a small metal "island" located between these junctions. A Coulomb blockade is a phenomenon widely studied and used in various nanodevices, where the role of the intermediate electrode is usually played by a single grain. Here, another limiting case will be presented – Coulomb blockade in a system with a very large number of grains. Figure B.7 demonstrates the schematic of the experiment and results
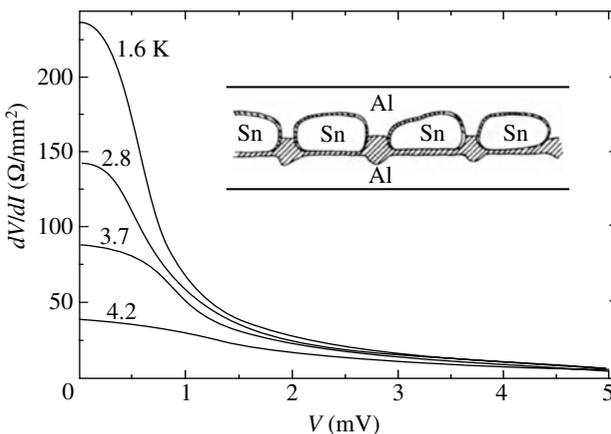
FIG. B.7.   Tunneling characteristics of the Al–Sn grains–Al structure at different temperatures. The structure is shown in the inset (Giaever and Zeller 1968). Hatched area is aluminum oxide (insulator).

obtained. A slightly oxidized surface of an aluminum film was coated with tin, which formed individual islands with the average size being dependent on the amount of deposited tin. Then the sample was subjected to further oxidation. As a result, a thick aluminum oxide layer was formed between the tin islands, whereas the tin particles were coated with a thin tin oxide layer. Finally, an aluminum layer was deposited to create the upper electrode (inset in Fig. B.7).

The curves shown in Fig. B.7 were obtained in a magnetic field sufficiently intense to destroy superconductivity. As all the electrodes are good metals, the current–voltage characteristics of the structure would be horizontal lines. The drop in conductivity observed at low voltages and becoming more pronounced with lowering of the temperature is a consequence of the smallness in the structure of the intermediate electrode – set of tin grains – in other words, the consequence of a Coulomb blockade.

In the experiments illustrated by Fig. B.7, the grains form a two-dimensional layer. A set of grains may also form a three-dimensional conglomerate or a so-called granular metal, which is considered in Chapter 8. In a granular metal, measurements of the effective density of states performed with the use of tunneling are also possible. They are presented in Chapter 8. There, a granular metal is used as one of the two main electrodes, in full accordance with the scheme in Fig. B.1, and the Coulomb interaction enters the problem differently (see Fig. 8.7 and eqns 8.14 and 8.15).

Considering the further development of the tunneling method and its possibilities, one has to recognize that eqn (B.1) is valid only under the assumption that, in tunneling, the wave vector is not preserved. The change of the component normal to the contact surface is quite natural and is caused by the inhomogeneity of the space along the corresponding direction. The tangential component varies because of the rough junction interface. In principle, tunneling which preserves the tangential component of the wave vector (*coherent tunneling*) would be possible if one manages to prepare a junction with an atomically smooth interface.

Coherent tunneling was in fact observed in special experiments (Eisenstein et al. 1991) whose schematic and results are shown in Fig. B.8. Tunneling proceeded through a thin barrier between two two-dimensional quantum wells. The wells consisting of two 140 Å thick GaAs layers with a 70 Å thick AlAs barrier in between were prepared by molecular beam epitaxy. Both sides of this structure were coated with $Al_{0.3}Ga_{0.7}As$ layers. The total thickness of this five-layer sandwich was about $50\,\mu m$. Indium contacts 1 and 2 were burned into both ends of the plate to provide the electrical contact with both wells filled with a two-dimensional electron gas. The metal gates were deposited onto the upper and lower surfaces (see inset in Fig. B.8). The gates a_1 and a_2 were used to divide the two-dimensional gas in both wells into two parts and, thus, to change the parallel connection of the indium contacts by two two-dimensional wells to in-series connection with a tunneling gap in between. The gates t_1 and t_2 allow one to vary the carrier concentration in both lower and upper wells.

Preparing such a structure, one has to solve two complicated experimental problems, which, in turn, requires a rigorous control of the molecular beam
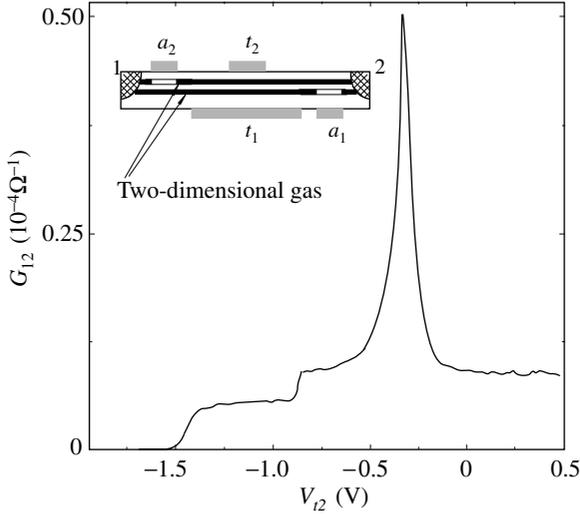
FIG. B.8. Conductivity G_{12} between contacts 1 and 2 proportional to the tunnel current J as a function of the voltage at the gate t_2 with fixed voltage at the gate t_1. The inset shows that the "lower electron gas" is separated from contact 2 by the voltage at the gate a_1, whereas the "upper electron gas" is separated from contact 1 by the voltage at the gate a_2 (Eisenstein et al. 1991). The temperature $T = 1.5$ K, the longitudinal voltage $v = 0.1$ mV.

epitaxy. First, it is necessary to obtain an atomically smooth interface between the GaAs and AlAs layers over quite a large area. Second, the electron densities in both wells should be made approximately equal.

Including contacts 1 and 2 into an external circuit with a negligibly low longitudinal voltage v adjusts the Fermi levels on both sides of the barrier and ensures the fulfillment of the conservation of energy. In a two-dimensional gas, the wave vector is located in the plane and is equal to

$$k_F = (2\pi n)^{1/2}. \tag{B.7}$$

Therefore, coherent tunneling in a nonzero magnetic field is possible only if the electron concentrations in both wells are the same.

The curve in Fig. B.8 was obtained at the longitudinal voltage $v = 0.1$ mV and at a certain fixed voltage at the lower gate V_{t1}. This curve demonstrates the dependence of the tunnel current $J = vG_{1,2}$ on the voltage V_{t2} at the upper gate, i.e., on the electron concentration n_2 in the upper well. The maximum of the tunnel current arises at equal electron concentrations in both wells, $n_1 = n_2$. As was to be expected, with the change of V_{t1}, the maximum is shifted. The relative value of the maximum determines the ratio between coherent and incoherent tunneling.

The two steps on the left-hand side of the $J(V_{t2})$ curve have a natural explanation. The first step is formed when the electron concentration under the gate t_2 in the upper well decreases to such an extent that, under the gate, we have an insulator. Then the area of the tunnel junction decreases by about a factor of 2. The second step is formed at a voltage V_{t2} such that the electric field due to the gate t_2 divides the electron gas in the lower well and, thus, disrupts the electric circuit between contacts 1 and 2.

References

Aleiner, I.L., Brouwer, P.W., and Glazman, L.I. (2002). Phys. Rep. **358**, 309.

Burstein, E. and Lundquist, S. (eds) (1969). *Tunneling Phenomena in Solids*, Plenum.

Butko, V.Yu., DiTusa, J.F., and Adams, P.W. (2000). Phys. Rev. Lett. **84**, 1543.

Eisenstein, J.P., Pfeiffer, L.N., and West, K.W. (1991). Appl. Phys. Lett. **58**, 1497.

Giaever, I. and Zeller, H.R. (1968). Phys. Rev. Lett. **20**, 1504.

Hertel, G., Bishop, B.J., Spencer, E.G., Rowell, J.M., and Dynes, R.C. (1983). Phys. Rev. Lett. **50**, 743.

Massey, J.G. and Lee, M. (1995). Phys. Rev. Lett. **75**, 4266.

Massey, J.G. and Lee, M. (1996). Phys. Rev. Lett. **77**, 3399.

McMillan, W.L. and Mochel, J. (1981). Phys. Rev. Lett. **46**, 556.

Solymar, L. (1972). *Superconductive Tunneling and Applications*, Chapman and Hall.

MATERIALS INDEX

The numbers in the list are numbers of the figures, not pages.

SUBJECT INDEX

Level spacing—*see* Size quantization
Local configuration
  of atoms 127, 130, 132
  of impurities 47
Localization parameter, $a_{\rm B}n^{1/3}$ 93, 94, 97

Mean free path 1, 5, 7, 84, 103
Metals
  2D 119, 122
  amorphous 8, 216
  at $T = 0$, definition of 74, 112
  liquid 2, 124
  standard 1, 7, 123
Microwave billiard 90
Mobility edge 78, 90, 180
Modulation 213

Nonlocality principle 79

Parallel conductivity channels
  in metals with strong disorder 10
  in semiconductors 62
Penrose tiling 131, 132

Random electric field 46
Random potential
  long-range 162, 172
  short-range 162
Radius
  Bohr, $a_{\rm B}$, 46, 59, 92
  localization—*see* Length localization
  magnetic, $r_{\rm B}$ 26, 159
  orbit of electron from N-th Landau
     level, $r_{\rm N}$ 161

  percolation, $r_{\rm c}$ 61, 205
  screening, $r_{\rm e}$ 94
Reservoir—*see* Thermostat
Resonant wells 77, 92

Short-range order 3, 5
Size quantization 78, 143, 159, 161
Soft gap 54, 72, 150, 214 218
Soft localization 158, 163
Static skin-effect 158, 177
Structural disorder 4, 91, 129, 136
Structure factor 4
Shunting resistance 10
Superconductivity 119, 143, 149, 152, 214

Time
  dephasing, $\tau_{\rm ee}$ 35
  magnetic, $\tau_{\rm B}$ 26
  phase-breaking, τ_{φ} 16
  resorption, $\tau_{\rm res}$ 217
  scattering
     elastic, τ 10, 16
     electron-phonon, $\tau_{\rm ph}$ 13, 40
     el-el, $\tau_{\rm e}$ 36
     el-el, ballistic, $\tau_{\rm e}^{\rm ball}$ 37, 41
     el-el, diffusion, $\tau_{\rm e}^{\rm diff}$ 37, 41
  spin-orbit, $\tau_{\rm so}$ 29
  transport, $\tau_{\rm tr}$ 40
  tunneling, $\tau_{\rm tun}$ 146
Thermostat (or reservoir) 31, 78, 170
Transfer integral 76, 93, 95

Universality class 108