2

Disordered electronic systems

The properties of disordered solids are very different from their crystalline counterparts. This chapter focuses on some of the fundamental electronic properties of disordered solids. It begins with an outline of the basic phenomena and types of relevant disordered systems. Then the model Hamiltonians used to study disordered systems are introduced. The basic concepts of the two relatively well-understood limiting cases of strong localization and weak localization are described, followed by an analysis of the general localization problem with emphasis on the metal-insulator transition. The chapter ends with an introduction to percolation theory, which plays an important role in the electric properties of disordered systems.

2.1 Disordered solids

Conventional research in solid state physics focused on ordered crystals in which the electronic states are largely influenced by the symmetries of the crystal lattice and thus take the form of periodic Bloch functions, which extend throughout the entire material. In this picture, the material is either insulating or metallic depending on whether the Fermi level, E_F, lies between bands or within a conduction band. Conductivity in the metal is hindered by the presence of static impurities and by phonons that determine the resistance and its T dependence. In the insulator, conductivity at finite temperature is achieved by thermal activation to the conduction band and thus

$$\sigma \propto \exp\left\{-\frac{E_G}{2kT}\right\} \tag{2.1}$$

where E_G is the gap between bands.

With increasing interest in disordered systems, initiated by the seminal works of Anderson and Mott, it became clear that the behavior of electrons in noncrystalline solids, in which crystal symmetry is absent, can differ greatly from the behavior of

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

electrons in crystals. The basic difference lies in the fact that in a crystal an electron cannot tell the difference between one primitive cell and another so it spreads all over the crystal. In the disordered system, each site looks different from each other site due to a different environment so an electron might prefer one location over others. This basic difference makes the disordered solid much more difficult to treat theoretically. One dramatic consequence of strong disorder is that the electronic wavefunctions are localized in space and are characterized by a localization length, ξ. At zero temperature, the electrons are immobile, and thus the material is an insulator even though E_{F} is within a finite density of states. Conductivity is achieved at finite temperature by phonon assisted tunneling (hopping) between localized states. Thus in these materials, the presence of phonons increases conductivity, opposite to the situation in systems with extended states. Another qualitative difference is that in the disorder-localized systems the conductivity increases with frequency while the opposite is true in extended states systems.

2.1.1 Energy scales

In disordered systems, four energy scales play an important role. The first is the disorder energy, W, which is the size of fluctuations of the local random potential due to lack of translational symmetry. The second is the quantum transfer energy, t (sometimes called hopping energy), between localized states. This is the energy to elastically transfer a charge between two localized states and is determined by the overlaps of the wavefunctions. For example, in a hydrogen molecule, this represents the energy difference between the odd and even wavefunctions. The localization of the wavefunctions depends on the ratio W/t. The limit $W/t \gg 1$ corresponds to the strongly localized regime, whereas $W/t \ll 1$ is the weak localization or dirty metal limit.

The third energy scale is the electronic interaction, $E_{\mathrm{e-e}} \sim e^2/(\kappa r)$, where κ is the dielectric constant and r is a typical distance between carriers. In general, the more localized are the states the smaller is the screening and the larger the importance of interactions. In the weak localization regime, interactions can be usually treated perturbatively, whereas in the strongly localized regime they are similar in magnitude to the disorder energy, and one has to treat both on the same footing.

Two limits are considered throughout this book. The first, $E_{\mathrm{e-e}} \ll W$ will be called the noninteracting case, and the second $E_{\mathrm{e-e}} \sim W$ will be termed the interacting case. In the opposite limiting case $E_{\mathrm{e-e}} \gg W$, the disorder plays a very small role. This case will not be treated in the book.

The fourth energy scale is the thermal energy, $E_{\mathrm{th}} = kT$. This book deals with low temperatures (i.e., the case where $E_{\mathrm{th}} \ll W$).

2.1.2 Types of disordered solids

There are a number of disordered systems that may exhibit hopping conductivity and/or glassy effects and are discussed extensively in this book. Some of their main properties are reviewed below.

- **Amorphous and polycrystalline solids** Anderson insulators composed of non-crystalline metals or semi-metals which become insulating due to the disorder. Polycrystalline samples are obtained from amorphous samples by heating above the crystallization temperature. Disorder in these systems can be varied in a number of ways depending on the specific material; for example, the substrate temperature of thin film evaporation, thermal annealing, carrier concentration manipulation, and film thickness in ultrathin films. In compounds (e.g., indium oxide, zinc oxide), the disorder may be due to lack of chemical stoichiometry; hence, it may be controlled by varying the stoichiometry. The disorder depends on the range of atomic correlations. Long-range order is always absent, but short-range atomic order may exist. An experimental way to determine the degree of disorder is to measure the $R(T)$ curve and to extract, in the insulating phase, the localization length via the hopping conductivity expressions discussed later. These systems undergo a metal–insulator transition as a function of disorder. The position of the transition can be quantified via the Ioffe–Regel parameter, $k_{\mathrm{F}}l$, where k_{F} is the Fermi wavenumber, and l is the mean free path determined by the disorder. The sample is metallic if it obeys the Ioffe–Regel criterion $k_{\mathrm{F}}l > 1$; otherwise, it is an insulator. Finite-sized 2D films undergo a similar transition, where a different criterion for critical disorder is via the sheet conductance, G. For $G > e^2/h$ the film shows metallic-like behavior while in the opposite limit, $G < e^2/h$, the film exhibits insulating behavior. Disordered metals are usually characterized by an occupation of many electrons within a localization volume and a significant overlap between localized wavefunctions. Hence, the transfer energy, t, compares with the disorder energy, W, and must be taken into account.
- **Granular metals** Systems of metallic islands embedded in an insulating matrix. In this case, localization is due to geometrical confinement in the grains and small transfer energy between them. The conductivity is determined by hopping between grains, whereas the conductivity within a grain is taken to be infinite. The transfer energy is determined by the distance between grains, which depends on the percentage of the material occupied by the metal. When the metallic part percolates or the transfer energy is large enough, the system acts as a metal; when the grains do not percolate and the transfer energy is small, it behaves as an insulator. In these systems, each grain contains many electronic states that may be correlated and thus it is important to consider intragrain correlation effects. An important energy in these systems is the charging energy to add a single charged

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

particle to the grain (Coulomb blockade), which may play an important role in the hopping rates. Usually, most of the grains are ionized, and the determination of the charge of each grain is a complicated problem involving also the Coulomb gap. This becomes even more complicated when one considers background charges present in the substrate or in the oxide between grains. These contribute a random potential that affects the charging of the grains.

- **Doped semiconductor** In these systems, conductivity at low temperatures is governed by hopping among impurity states (impurity conduction). The wavefunction of an isolated impurity can, in many cases, be approximated by an hydrogen-like wavefunction with a very large radius. The overlap between wavefunctions, and so the transfer energy varies exponentially with distance. The interaction energy and the disorder energy are usually similar as both arise from Coulomb interactions between impurity sites and they vary as the inverse of the distance. The density of impurities or concentration is a key parameter in this problem. Increasing the impurity concentration increases the transfer energy exponentially and the disorder only algebraically. Thus, there is a critical density above which states are extended and below which states are localized. In these materials the on-site interaction energy is usually much larger than the other energies, and it is assumed that a state can be occupied at most by a single carrier and the occupation on each site can be either 0 or 1. Another relevant parameter in these materials is the compensation, the ratio between minority impurities and majority impurities. It corresponds to the percentage of unoccupied sites of the majority type. The importance of interactions is maximal at half compensation. At small compensation, the carrier concentration equals the minority impurity concentration, which corresponds to the unoccupied sites on the majority impurities. At large compensation the carriers are the occupied sites on the majority impurities. The minority impurities are all ionized and responsible for much of the disorder.
- **Two-dimensional electron gases (2DEG)** These are not disordered systems per se. On the contrary, the main technical effort is invested in maximizing sample mobility, μ. Nevertheless, these systems can undergo a metal-insulator transition because of the ability to vary their carrier concentration and therefore are mentioned in this book in several contexts. A 2DEG is a gas of electrons free to move in two dimensions, but tightly confined in the third. This is achieved by electronic band engineering. The two most common systems are Si MOSFET and GaAs/AlGaAs heterostructures. The former consists of a metal/oxide/silicon trilayer where the metal is a gate electrode, the oxide an insulator and the silicon the disordered system under study. Applying a negative voltage between the metal and the semiconductor bends the electronic bands and confines electrons to a very thin layer (a few nm) adjacent to the oxide. In the latter the different

band structure of GaAs and AlGaAs confines electrons to the interface between the two layers. In these systems, applying a gate voltage to the 2DEG (utilizing metallic electrodes) can be used to finely control the carrier concentration, n. For low n, this allows us to tune the system through the metal–insulator transition.

2.2 Hamiltonians for disordered systems

To avoid undue complications not relevant to the disorder, one generally adopts the tight-binding approximation. The Hamiltonian representing the system can be written in second quantized form as

$$\mathcal{H} = \sum_{n,m} h_{n,m} a_n^+ a_m + \frac{1}{2} \sum_{klmn} V_{klmn} a_k^+ a_m^+ a_n a_l \,. \tag{2.2}$$

where a_m^+, a_m are the creation and annihilation operators of an electron in the state m, and $h_{n,m}$ and V_{klmn} are matrix elements. The first term represents the one-particle energies, and the second the interaction energies. In the tight-binding approximation, one chooses a basis of local functions with one orbital per site described by the site index i and the spin σ.

The most important one-particle contribution is the diagonal part

$$\sum_{i,\sigma} \varepsilon_i a_{i,\sigma}^+ a_{i,\sigma} \tag{2.3}$$

where ε_i is the (generally random) energy on site i. This contribution can be rewritten in terms of the number operator on site i, $n_{i,\sigma} = a_{i,\sigma}^+ a_{i,\sigma}$. If the site wavefunctions are not extremely localized, one has to consider the (generally random) intersite quantum tunneling energy

$$\sum_{i \neq j} \sum_{j,\sigma} t_{i,j} a_{j,\sigma}^+ a_{i,\sigma} \tag{2.4}$$

which is responsible for the delocalization of the wavefunction. In models with sites at random positions, it is customary to consider an exponential dependence of $t_{i,j}$ on distance

$$t_{i,j} = \langle i | \mathcal{H} | j \rangle = I_0 \exp\left\{ -\frac{r_{i,j}}{\xi_0} \right\} \tag{2.5}$$

where ξ_0 is the decay length of the wavefunction of an isolated impurity, and I_0 is a characteristic energy of the order of the Coulomb energy at a distance ξ_0. In models on regular lattices, $t_{i,j}$ is generally approximated by a constant between nearest neighbors and zero otherwise.

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

The most important interaction term is when all four operators are on the same site and corresponds to the interaction of two electrons on the same site (and so with opposite spins)

$$\sum_i U_{i,i} n_{i\uparrow} n_{i\downarrow} \tag{2.6}$$

$U_{i,i}$ is often called the Hubbard energy. The Hubbard model only retains this term and the one-particle contributions of Equations (2.3) and (2.4). Other important interaction contributions are the diagonal elements $m = n$, $k = l$, that is, the direct interaction between two particles on different sites

$$\frac{1}{2} \sum_{i \neq j,\sigma} \sum_{j,\sigma} U_{i,j} n_{i\sigma} n_{j\sigma} \tag{2.7}$$

The corresponding matrix elements are

$$U_{i,j} = \int \mathrm{d}^3 r \int \mathrm{d}^3 r \; |\phi(\mathbf{r} - \mathbf{r}_i)|^2 \frac{e^2}{\kappa |\mathbf{r} - \mathbf{r}\,'|} |\phi(\mathbf{r}\,' - \mathbf{r}_j)|^2 . \tag{2.8}$$

For very localized wavefunctions, this expression can be approximated by $U_{i,j} = e^2/(\kappa r_{i,j})$, with $r_{i,j} = |\mathbf{r}_i - \mathbf{r}_j|$ being the (generally random) intersite distance.

The Hamiltonian in (2.2) also includes exchange interactions involving only two wavefunctions, but coupled in the opposite order than in the direct interactions. They may be very important in certain cases (e.g., in magnetism) but are usually neglected in the electronic glass literature because they are a factor $|t_{i,j}|^2$ smaller than the corresponding direct terms. The most relevant terms of the Hamiltonian are then

$$\mathcal{H} = \sum_{i,\sigma} \varepsilon_i n_{i,\sigma} + \sum_{i \neq j} \sum_{j,\sigma} t_{i,j} a^+_{j,\sigma} a_{i,\sigma} + \sum_i U_{i,i} n_{i\uparrow} n_{i\downarrow} + \frac{1}{2} \sum_{i \neq j,\sigma} \sum_{j,\sigma} U_{i,j} n_{i\sigma} n_{j\sigma} \tag{2.9}$$

The disorder can be positional due to randomness in the location of sites (so-called off-diagonal disorder), which will affect the second and fourth terms in (2.9), or energetic due to a randomness in the site energies (the so-called diagonal-disorder), which will affect the first term in (2.9), or both types can occur together.

When considering long-range interactions, it is interesting to conserve charge neutrality, which can be done automatically by redefining the occupation number as $n_{i\sigma} = a^+_{i,\sigma} a_{i,\sigma} - K$, K being the ratio n/N of the number of electrons (more precisely, majority carriers) to number of sites in the system. Usually one then considers a positive charge $|en/N|$ to reside on every site. Some alternative models are briefly discussed in Pollak (1971).

Dealing with the full Hamiltonian of (2.9) is very difficult, and, depending on application, various simplifications may be appropriate. In many cases the intrasite

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

interaction is much greater than other energies which justifies the elimination of the second term and the prohibition of double occupation of a site. If one also neglects the exchange energy, already not included in (2.9), the spin then becomes unimportant, and for most purposes one can just consider a system of spinless electrons. The above near-ubiquitous approximations simplify the Hamiltonian to

$$\mathcal{H} = \sum_i \varepsilon_i n_i + \sum_{i \neq j} \sum_j t_{i,j} a_i^+ a_j + \frac{1}{2} \sum_{i \neq j} \sum_j U_{i,j} n_i n_j \tag{2.10}$$

In this simplified version, the first term is the random energy, the second term is the intersite tunneling energy, and the last term is the intersite Coulomb interaction energy. For special purposes, $\mathcal{H}$ is further simplified. Keeping the first and third term has been often used as a model for the electron glass whereas using the first two terms was used by Anderson in his classic paper on (single-particle) localization (Anderson, 1958). A further simplification in his model was to assume a structurally ordered system (a simple cubic lattice) and nearest neighbor tunneling only (thus rendering $t_{i,j}$ fixed) and to relegate the disorder to the site energies. Such a model may seem oversimplified, but it does capture the important physical features leading to localization while it still turns out to be a very nontrivial system to solve quantitatively.

Throughout the book macroscopic homogeneity shall be assumed. This means that one can divide the system into a statistically large enough number of equal sized parts such that each part is large enough and statistically identical to the other parts.

2.3 Strong disorder

2.3.1 Strong localization

When the disorder is very strong the electronic wavefunctions are localized. This book focuses on this regime. Localized wavefunctions have large magnitude around a central region and decay exponentially outside this region. Their wavefunctions can be written in the form

$$\psi\ (\mathbf{r}) = \sum_i c_i \phi_i (\mathbf{r}) \exp\left\{ -\frac{|\mathbf{r}\ - \mathbf{r}_i|}{\xi} \right\} \tag{2.11}$$

where i refers to the sites in the solid ($\mathbf{r}_i$ being their locations), c_i are randomly fluctuating amplitudes, $\phi_i(\mathbf{r})$ is the wavefunction at site i, and $\mathbf{r}$ is the "center" of wavefunction . ξ is termed the localization length. It is a measure of the typical size of a state and is a basic lengthscale in the strongly localized regime. It should be noted that the exponential form of the wavefunction in (2.11) is an idealization

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

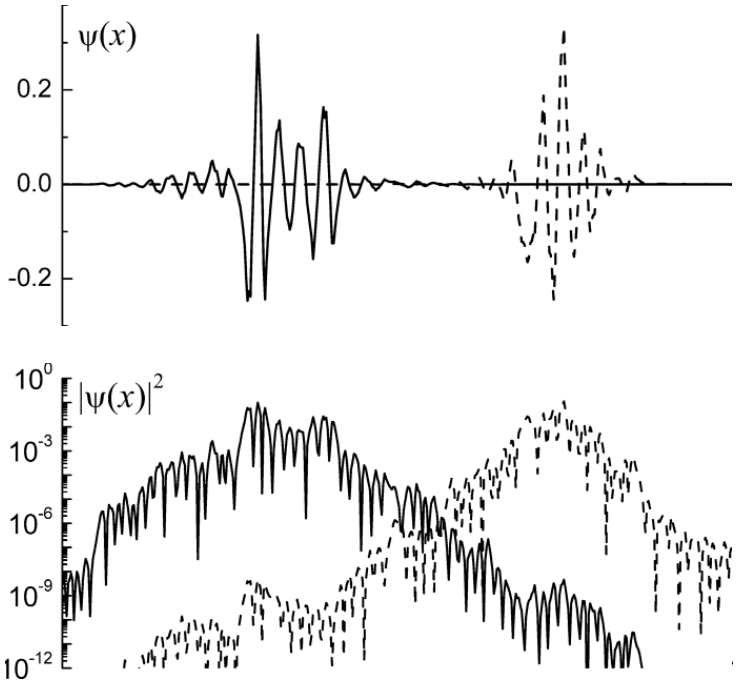

Figure 2.1. Two localized wavefunctions with very similar energy but different spatial location. Bottom part: modulus square of the same two wavefunctions on a logarithmic scale.

since there is not a well-defined center, except for strongly localized wavefunctions (Markos, 2006). More importantly, theoretical calculations, including many that will appear later in the book, assume a single ξ that is taken to be uniform for all sites. This is a very crude approximation that is probably unrealistic and may be one of the reasons for discrepancies between theory and experiment (Somoza and Ortuño, 2005). In the upper part of Figure 2.1 two numerically obtained wavefunctions $\psi\ (x)$ with very similar energy are shown and, in the lower part, their moduli square, $\ln|\psi\ (x)|^2$, are represented on a logarithmic scale. These logarithms decay approximately as straight lines, whose slope is proportional to the localization length.

If the Fermi level lies in a region of localized states, the dc conductivity at $T = 0$ is zero. However, for finite samples, the conductivity is finite and depends on the localization length

$$G \propto \exp\left\{-\frac{2L}{\xi}\right\} \tag{2.12}$$

where L is the size of the system. This dependence of conductivity on length is, in principle, the experimental and numerical way to determine the localization length (Kramer and MacKinnon, 1993).

In the strongly localized regime, each wavefunction can be associated with a site, j, where its amplitude is maximal. One can treat the transfer energy $t_{j,k}$ as a

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

perturbation, and the zero order state is just the site state $|j\rangle$. To first order the state, $|j\rangle$ is

$$|j^{(1)}\rangle = |j\rangle + \sum_k \frac{t_{j,k}}{E_k - E_j}|k\rangle \tag{2.13}$$

where E_j is the energy of site j. To first order, the energies are just the unperturbed energies. It is easy to see that the perturbation expansion converges rapidly if the disorder is strong so higher-order terms beyond (2.13) become unimportant. States that are nearby in space are in general very different in energy and their contribution to (2.13) is small because of the large energy denominator. States that have very similar energies are in general very far apart, and their overlap is exponentially small (Lee and Ramakrishnan, 1985).

If the transfer energy decays exponentially, $t_{j,k} \propto \exp(-r_{j,k}/\xi_0)$, comparing Equation (2.11) with (2.13) one notices that the localization length is, to first order in the perturbation, equal to the single site decay length ξ_0.

2.3.2 Density of states – the coulomb gap

Interactions are usually very important in the localized regime because the low mobility of charges results in a drastic reduction of screening. Then the two most relevant energies are the disorder and the long-range Coulomb interaction. The competition between them leads to a depletion of the single-particle density of states (DOS) near the Fermi energy known as the Coulomb gap. This gap was predicted by Pollak (1970) and its shape was obtained by Efros and Shklovskii (ES) (1975). The single-particle DOS is defined as the distribution of the energy E_i required to add (or remove) an electron to the system in site i, holding the rest of electrons fixed. The energy E_i is defined as the random energy of site i plus its Coulomb interaction energy with all other sites, depending on their occupation in a ground state. The DOS in the Coulomb gap is (see Chapter 4)

$$N(E) \propto |E|^{d-1} \tag{2.14}$$

E is the energy measured with respect to the Fermi level, and d is the dimensionality of the system. In 1D systems, the gap is logarithmic, not constant as could be extracted from (2.14).

In localized interacting systems, the density of excitations cannot be obtained as a convolution of the DOS. If an electron is transferred from site i to site j, the energy of this one-electron hop is

$$\Delta E_{j,i} = E_j - E_i - \frac{e^2}{\kappa r_{i,j}} \tag{2.15}$$

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

Notice that in the definition of the site energies full sites refer to an n-electron system while vacant sites refer to an $n+1$ electron system (i.e., the energy of an electron on a vacant empty site). This is very appropriate for the definition of a one-particle DOS but may become confusing when dealing with excitations that do not change the number of electrons in the system.

The last term in (2.15) is the interaction of an electron with the hole it leaves behind and is responsible for the Coulomb gap. It also causes an increase of short low-energy excitations, with respect to the noninteracting case. This is easy to understand because such an excitation from one site to another is conditioned by a single occupation of the pair of sites. This condition is facilitated by the electron–electron repulsion. A more rigorous treatment of the Coulomb gap will be provided in Chapter 4 where it will also be made clear that this treatment, based on the single-particle DOS, may be modified when one considers correlated effects.

2.3.3 Hopping conduction

At finite temperature, the conduction in the localized phase is by hopping between states. The energy difference between the states is provided by phonons, which interact with electrons by changing their environment. The probability of a hop with an energy increase $\Delta E_{i,j}$ is thus proportional to the probability of finding a phonon of this energy $\propto \exp(-\Delta E_{i,j}/kT)$. The reverse hop involves the emission of a phonon. The probability of a hop also depends on the overlap between the states involved, which according to (2.11) depends exponentially on the distance between them, $r_{i,j}$. The transition rate between two states is of the form

$$\Gamma_{i,j} \propto \exp\left\{-\frac{\Delta E_{i,j}}{kT} - \frac{2r_{i,j}}{\xi}\right\} \tag{2.16}$$

The first exponential factor represents the number of phonons of energy $\Delta E_{i,j}$, the second factor represents the transition matrix element squared in accordance with the golden rule.

Two important conclusions can be extracted from (2.16):

- The rates are exponentially distributed and so mean field approaches are usually not adequate. Instead, percolation approaches are much more appropriate.
- Energy difference and space separation play similar roles and a trade-off between them is to be expected.

At relatively high T that approaches the range of the disorder energy, W, only the spatial factor in (2.16) has to be minimized and hopping is through nearest neighbors. The conduction path is independent of T, which results in activated behavior with the activation energy corresponding to the highest energy jump in

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

the path. However, this cannot extend over a very large range of temperature since the Arrhenius dependence becomes a poor approximation for the Fermi distribution.

In the usually larger range where the energy in (2.16) is significantly smaller than W, there is a trade-off between energy and distance. The conduction path depends on T and so does the typical hopping distance. The latter increases with decreasing T and the activation energy, decreases with decreasing T. This conduction mechanism was discovered by Mott who named it variable range hopping (VRH). In this regime, the conductivity is of the form

$$\sigma \propto \exp\left\{-\left(\frac{T_0}{T}\right)^s\right\} \tag{2.17}$$

where the exponent s depends on dimensionality $s = 1/(d+1)$. Coulomb interactions change the relation between the typical energy of a jump and its distance, modifying the exponent in (2.17), which becomes $s = 1/2$ independent of dimensionality. The value of this exponent was obtained by Efros and Shklovskii (1975) by extending Mott's argument for VRH to the DOS of the Coulomb gap of (2.14). Figure 2.2 shows the resistivity as a function of $T^{-1/2}$ for seven ion-implanted Si:P,B samples (Zhang et al., 1993). The straight lines are fits to (2.17) with $s = 1/2$, corresponding to VRH with interactions. To the extent that the fit is adequate to determine well the value $s = 1/2$, this can be taken as evidence for interaction.

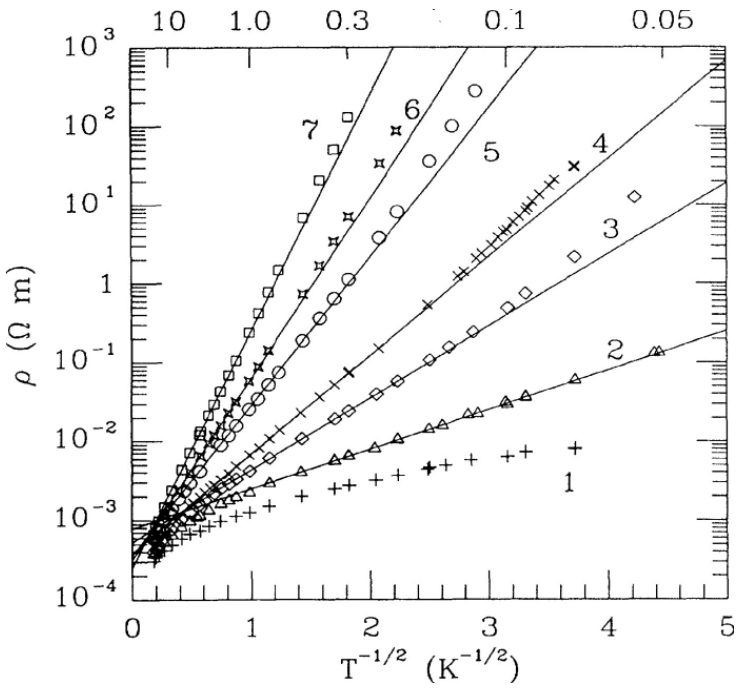

Figure 2.2. Resistivity vs. $T^{-1/2}$ for ion-implanted Si:P,B samples. The straight lines are fits to equation (2.17) with $s = 1/2$, after Zhang et al. (1993). Copyright by the American Physical Society.

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

2.4 Weak disorder

2.4.1 Weak localization

Even for mild degrees of disorder, electronic localization may play a role in the transport properties, but in this regime the influence of disorder amounts to quantum correction to the classical Drude conductivity

$$\sigma_{\mathrm{D}} = \frac{ne^2\tau}{m} \tag{2.18}$$

where n is the carrier concentration, m is the electron, mass and τ is the mean free time. For weakly disordered systems, it is useful to consider two characteristic length scales. The first is the elastic scattering length, L_{el}, which is the average distance between impurities. The second is the inelastic length, L_{in}, which is the distance traversed by a charge between two inelastic scattering events, such as scattering by another electron or by a phonon. The latter is often identified with the phase breaking length, L_ϕ, which is the length over which phase memory is maintained by the particle (though some slight differences may occur between the two length scales). These length scales are associated with the elastic mean free time, τ_{el}, the inelastic time, τ_{in}, and the phase breaking time, τ_ϕ, through the relation $L = \sqrt{D\tau}$ where D is the diffusion constant. In the Drude formula of (2.18),

$$\frac{1}{\tau} = \frac{1}{\tau_{\mathrm{el}}} + \frac{1}{\tau_{\mathrm{in}}}\,. \tag{2.19}$$

If $L_{\mathrm{el}} \ll L_\phi \ll \xi, L$, where L is the sample size, the electric conductivity can be treated by a diffusion approach in which an electron diffuses between static scatterers without losing phase memory for distances smaller than L_ϕ. The probability to traverse from point a to point b should include all possible trajectories and thus is given by

$$P_{ab} = \left| \sum A \right|^2 = \sum |A|^2 + \sum_{\neq\beta} \sum_{\beta} A\ A^*_\beta \tag{2.20}$$

where each A is the amplitude of a possible trajectory. The first term on the right-hand side of (2.20) is the classical probability of propagation and the second term is the quantum contribution due to interference between different electronic trajectories. This is a sum over random phases from different sections of the sample and thus, for most cases, this term averages out to zero. However, there is a special set of trajectories (i.e., closed loops in which the electron returns to its original position), where the interference term adds a finite contribution. This is because every such trajectory is accompanied by its time-reversed trajectory, traveling in an opposite direction and providing the same phase shift. Each such pair of closed

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

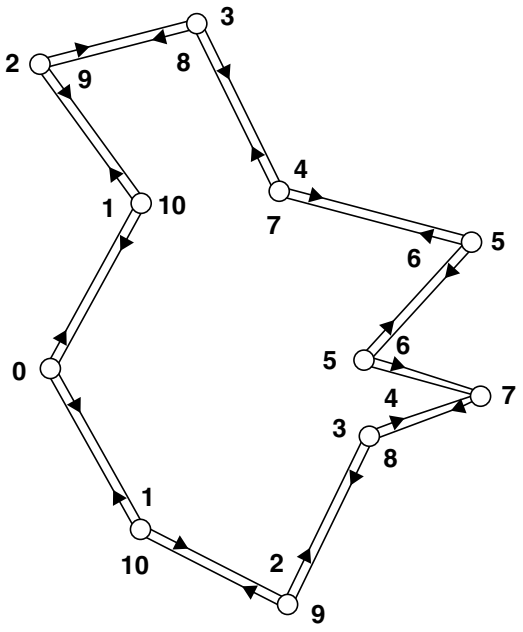

Figure 2.3. An illustration of two trajectories moving in opposite directions, contributing identical phase shifts. After Bergmann (1984). With permission by Elsevier.

loops interferes constructively thus increasing the probability of an electron to remain in its origin. Hence, in the case of quantum diffusion the probability to return to the origin is twice as great as in classical diffusion since the amplitudes add coherently. This phenomenon, which hinders the motion of electrons and adds a negative contribution to the conductivity, is termed "weak localization." A set of two time-reversed trajectories is illustrated in Figure 2.3.

The dimension of a sample governed by these physics is determined by L_ϕ. For example, if the thickness, z, is smaller than L_ϕ, the system can be considered two dimensional as far as weak localization is concerned. Most research on weakly localized systems has been performed on thin films which fulfill the above condition, hence this section will focus on the 2D case. The correction to the conductivity in 2D is given by (Abrahams et al., 1979)

$$\frac{\sigma}{\sigma_{\mathrm{D}}} = 1 - \frac{3}{2k_{\mathrm{F}}^2 z L_{\mathrm{el}}} \ln \frac{L_\phi}{L_{\mathrm{el}}} \tag{2.21}$$

where σ_{D} is the Drude conductivity. Usually $L_\phi \propto T$ where = 1-2; hence,

$$\Delta\sigma = \sigma - \sigma_{\mathrm{D}} \propto \ln \frac{T}{T} \tag{2.22}$$

The above treatment applies to simple disordered systems. Strong spin-orbit scattering adds an opposite contribution to the conductivity named weak antilocalization (Bergmann, 1984). This effect will be ignored here.

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

As disorder is increased and ξ becomes smaller than L_ϕ the system crosses over to the strong localization limit in which conductivity is governed by hopping.

2.4.2 Magnetoresistance

Weak localization is suppressed when a magnetic field, H, is applied, due to the Aharonov-Bohm effect. A magnetic flux that penetrates the loop trajectories induces a different phase shift for the two time-reversed paths. In other words, the magnetic field breaks time reversal symmetry and destroys the constructive interference, thus delocalizing the electronic wave functions. Such a process leads to negative magnetoresistance, which is proportional to H^2 at low fields and saturates at high H. The magnetic field scale of saturation is determined by the magnetic length $L_H = \sqrt{\Phi_0/H}$ where $\Phi_0 = \ /e$ is the flux quantum. For large fields in which $L_H \ll L_\phi$, the delocalization process no longer increases with increasing H.

The weak localization magnetoresistance in 2D is given by (Rammer, 1998)

$$\frac{\Delta\sigma_{\mathrm{WL}}}{\sigma(H=0)} = -\frac{3}{2k_F^2 z L_{\mathrm{el}}}\left[\psi\left(\frac{1}{2}+\frac{3}{4}\frac{L_H^2}{L_{\mathrm{el}}^2}\right)-\psi\left(\frac{1}{2}+\frac{3}{4}\frac{L_H^2}{L_\phi^2}\right)\right] \qquad (2.23)$$

where ψ is the digamma function. Figure 2.4 shows a set of magnetoresistance curves performed on Mn thin films with fits to (2.23).

This "orbital" effect depends on the dimension of the sample. In 2D where $z < L_\phi$, the magnetoresistance will be anisotropic. Application of a field perpendicular to the film will yield a magnetoresistance of the form of (2.23). However, in the case of a parallel field, L_ϕ will be substituted by z, and the field scale for magnetoresistance will be much wider.

2.4.3 Mesoscopic fluctuations

A special case is when all lateral dimensions are smaller than L_ϕ. In this regime, called the mesoscopic regime, the right-hand term in (2.20) will yield a certain value for each sample, and this will determine its conductivity. An ensemble of samples having the same thermodynamic properties (i.e., size, geometry, mean free path) will differ from each other by their conductivity since σ is affected by the off-diagonal phase term, which is determined by the microscopic configuration of the scatterers. The amplitude of these sample-to-sample conductance fluctuations was found to be universal in the sense that they did not depend on geometry, lateral extent (as long as it was smaller than L_ϕ) or resistance and was always of the order of e^2/h (Altshuler, 1985). Hence, these fluctuations were termed "universal conductance fluctuations." The conductivity is so sensitive to the microscopic configuration of impurities that

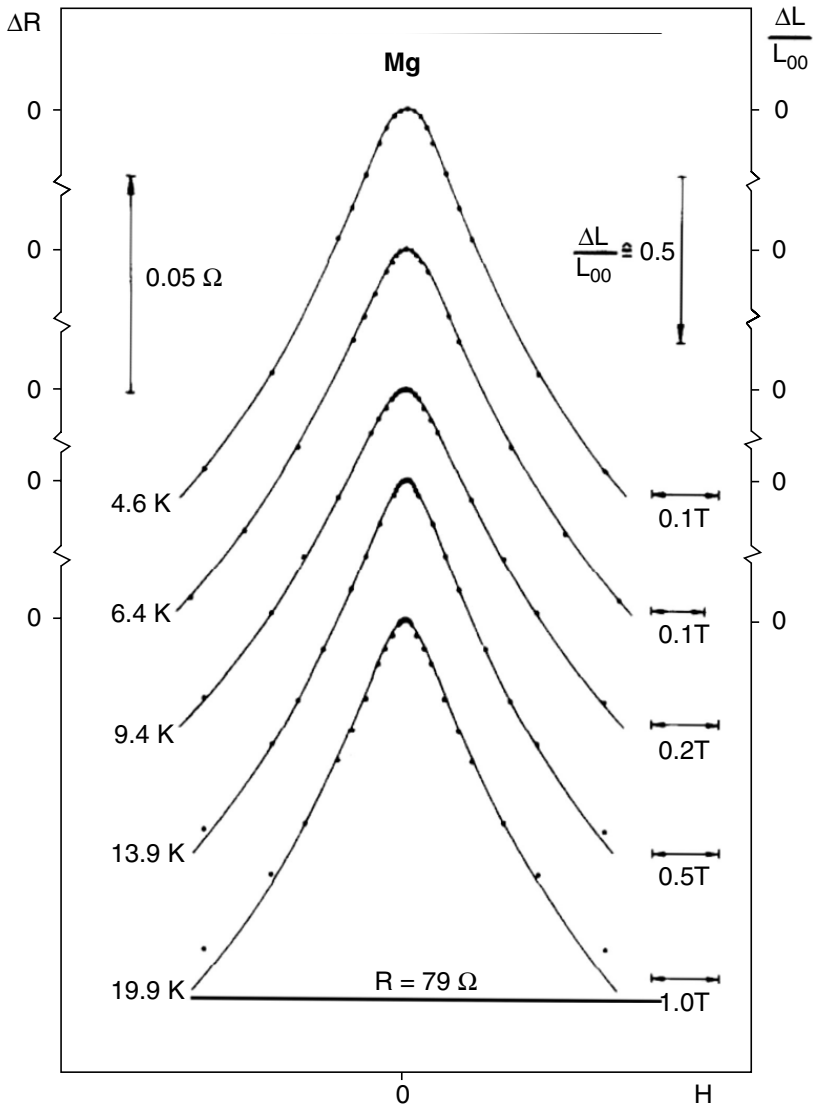

Figure 2.4. The magnetoresistance of a thin Mn film for different temperatures as a function of the applied field. The units of the field are given on the right side of the curves. The points represent the experimental results. The full curves are fits of the theory (Bergmann, 1984). With permission by Elsevier.

shifting so much as a single scatterer by more than a Fermi wavelength will cause a substantial change of order e^2/h in the conductance.

A similar effect can be observed on a single sample by sweeping a bias that changes the interference pattern, such as magnetic field or gate voltage. Magnetic field modifies the interference via the Aharonov-Bohm effect in loops of trajectories. Gate voltage shifts the Fermi level causing the electron to experience a different environment. In both cases, the conductivity exhibits universal conductance fluctuations of amplitude $\sim e^2/h$. Since each microscopic arrangement leads to a different interference pattern, $\sigma(H)$ or $\sigma(V_g)$ is different for each sample, thus

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

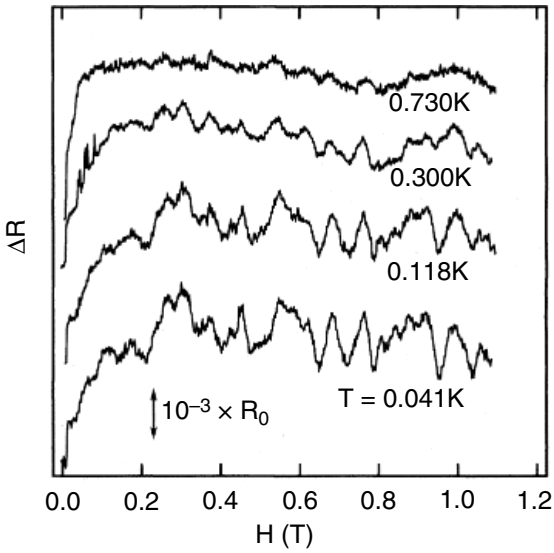

Figure 2.5. Temperature dependence of the magnetoresistance from 0 to 1.2 T of a Au ring. The inside diameter of the ring was 280 nm, and the width of the lines forming the ring was roughly 45 nm. The zero-field resistance of the ring was 7.7 Ω. After Umbach et al. (1984). Copyright by the American Physical Society.

the curve can be viewed as a "fingerprint" of that specific sample. An example for magnetoresistance mesoscopic fluctuations is shown in Figure 2.5.

As the sample size is increased so that $L \gg L_\phi$, the fluctuation amplitude decreases like $(L_\phi/L)^{1/2}$. This represents the fact that the magnitude of the quantum interference contributions decay as the squareroot of the number of independent phase coherent units. Similarly, the temperature dependence of L_ϕ determines the ability to observe fluctuations as T is increased. Since $L_\phi \propto T$, the amplitude of the universal conductance fluctuations decreases with temperature like a power law.

If disorder is increased so that $L_\phi \gg \xi$, quantum interference effects become less prominent. However, small samples in the strong localization regime also exhibit mesoscopic conductance fluctuations. These arise from a different physical origin and will be treated in Chapters 5 and 6.

2.4.4 Density of states – zero bias anomaly

The effect of electron–electron interactions in the weakly disordered regime was considered by Altshuler and Aronov (1985). They showed that the interactions cause a depletion of the DOS around the Fermi level, a phenomenon that became known as the zero bias anomaly (ZBA). In two dimensions (z smaller than the thermal length, $L_T = \sqrt{D/kT}$), the correction to the DOS is given by

$$N(E,T) - N_0 = -\frac{N_0}{8\pi^2 D} \ln \frac{\max(kT,E)}{D^2 k \tau_{\mathrm{el}}} \ln \frac{\tau_{\mathrm{el}} \max(kT,E)}{} \qquad (2.24)$$

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

where E is the energy, N_0 is the 2D normal (non interacting) DOS measured at high energies, D is the diffusion constant, and $\tau_{el} = L_{el}^2/D$ is the elastic mean free time. The ZBA amplitude increases with increasing disorder.

A useful parameter used to define the degree of disorder in dirty metals is the dimensionless conductance

$$g = \frac{G}{G_0} = N_0 D \tag{2.25}$$

where G is the conductance, and G_0 is the quantum of conductance e^2/h. The Altshuler and Aronov theory applies only to mildly disordered metals, $g \gg 1$. As the disorder is increased and g decreases below 1, the sample crosses over to the strong localization regime in which the Coulomb gap, Equation (2.14), determines the interacting DOS. Both the ZBA and the Coulomb gap stem from electron–electron interactions and are due to somewhat similar physics. However, the Coulomb gap is associated with a Hartree term of interactions, which represents the contribution of the classical Coulomb interaction, while the ZBA is a consequence of the exchange term, which corresponds to much subtler quantum effects. In special geometries the two effects can be observed and separated (Bitton et al., 2011).

The logarithmic dependence of the density of states on energy of Equation (2.24) was found in many tunneling experiments into 2D-disordered metal films (for a more detailed explanation on tunneling experiments see Subsection 4.2.2). An early example is shown in Figure 2.6 where the tunneling DOS of a GaAs film is shown for different neutron irradiations. The irradiation introduces impurities in the film, thus enabling to study the effect of disorder on the ZBA amplitude.

Electron–electron interactions affect the conductivity as well. In 2D, they contribute a correction to conductivity, which is proportional to $\log(T)$. Hence, it has the same functional dependence as (2.22) of weak localization. Indeed, both quantum coherence and electron–electron interactions contribute similarly to the 2D $\sigma(T)$ curve and in many cases it is difficult to tell them apart.

2.5 Anderson localization and metal–insulator transitions

After having discussed the two limiting cases of strongly localized states and of extended states with weak disorder, it is interesting to address the case of intermediate disorder. This is a very complicated theoretical problem even in the absence of interactions, but the results most relevant to the issues discussed in this book are fairly well understood by now.

To obtain some physical insight into Anderson localization, it is useful to have a closer look at the perturbation expansion of (2.13) considering only two sites with energy difference $\Delta E_{j,k}$. For strong disorder, most of the links obey the condition

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

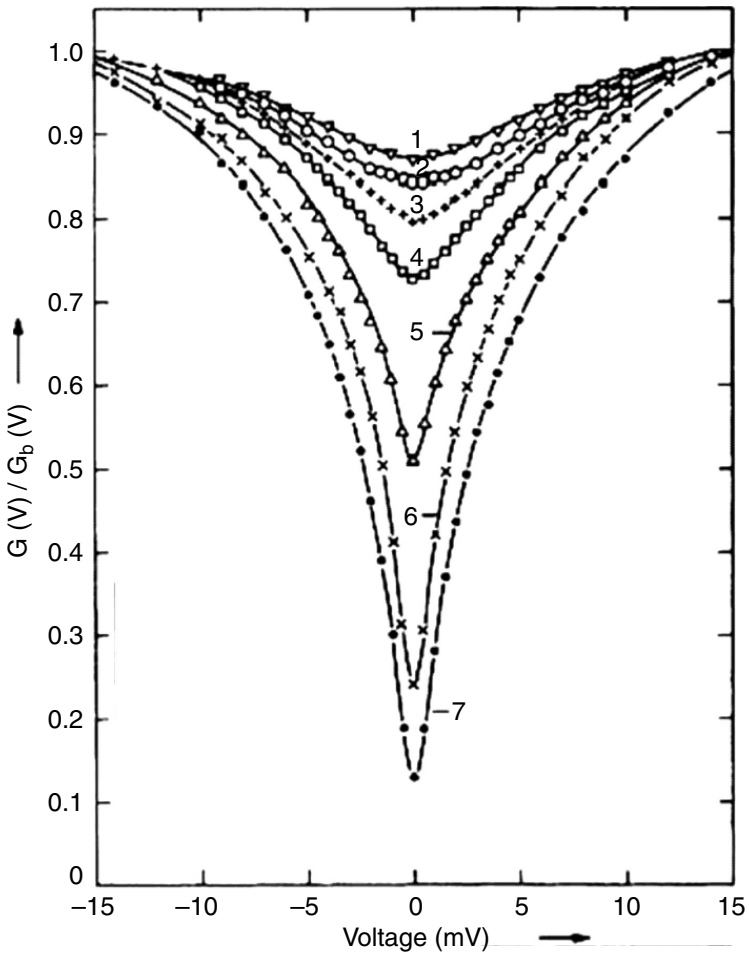

Figure 2.6. Tunneling conductance as a function of voltage for a Pb-GaAs tunnel junction at different neutron irradiation times, increasing disorder successively from 1 to 7 (Mora et al., 1971). Copyright by the American Physical Society.

$\Delta E_{j,k} \gg t_{j,k}$, and the states of the electron are approximately

$$|j\rangle + \frac{t_{j,k}}{\Delta E_{j,k}}|k\rangle, \qquad -\frac{t_{j,k}}{\Delta E_{j,k}}|j\rangle + |k\rangle \tag{2.26}$$

In this case the electron is strongly localized (with a probability $1-(t_{j,k}/\Delta E_{j,k})^2$) on one site or the other. However in some rare links $\Delta E_{j,k} \ll t_{j,k}$, the states for the electron become

$$\frac{|j\rangle + |k\rangle}{\sqrt{2}}, \qquad \frac{|j\rangle - |k\rangle}{\sqrt{2}} \tag{2.27}$$

Here the electron is evenly delocalized on the two sites. The transition between the two cases is reasonably sharp at around $t_{j,k} = \Delta E_{j,k}$.

The links for which $t_{j,k}/\Delta E_{j,k} > 1$ are called resonating bonds, and their number increases with decreasing W/t. Intuitively one may think that if one can pairwise propagate through resonating bonds over a macroscopic distance, then this would

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

constitute a delocalized state. Thinking this way would make delocalization a percolation problem and miss out on the quantum aspects of it. In particular, the resonance condition is not just a property of i and j alone but depends on hybridization with other sites. The interference effects between closed trajectories traveling in opposite directions also decreases delocalization. Thus, the quantum aspect suppresses delocalization. This is most important in 2D where the pairwise approach would allow for delocalization, which is in fact forbidden.

2.5.1 Perturbation expansion

For a better understanding, it is useful to follow Anderson's original ideas on localization based on a perturbation expansion (Anderson, 1958), of which (2.13) is the first-order term. For a noninteracting system,

$$|\widetilde{l}\rangle = |l\rangle + \sum_j \frac{\langle j|V|l\rangle}{E_l - E_j}|j\rangle + \sum_{j,k} \frac{\langle k|V|j\rangle\langle j|V|l\rangle}{(E_j - E_k)(E_l - E_j)}|k\rangle + \cdots$$
$$+ \sum_{j,k,\ldots,n,m} \frac{\langle m|V|n\rangle \cdots \langle k|V|j\rangle\langle j|V|l\rangle}{(E_j - E_k)(E_i - E_j)\cdots(E_i - E_m)}|m\rangle + \cdots \qquad (2.28)$$

where $|\widetilde{l}\rangle$ is an eigenfunction of the Hamiltonian and $|l\rangle$ is the local function on site l. The general term above signifies an admixture of $|m\rangle$ to $|l\rangle$. Divergence of (2.28) conveys delocalization – the very high order terms dominate and the electron must propagate through macroscopic distances from i to m. Thouless (1974) pointed out that percolation is a reasonable approximation to the problem. Figure 2.7 helps to visualize the process. The first term in the perturbation expansion extends the wave function on l to j, j , …, the second term to k, k , …, the third to q, q , …, and so on. If the successive order terms diminish, the wave function $|\widetilde{l}\rangle$ drops off and stays localized; if they increase, the expansion diverges and the wave function becomes delocalized. The numerators (i.e., the tunneling energies) clearly increase with concentration, and the denominators clearly increase with disorder energy. Thus, for low concentrations and large disorder, one expects the states to be localized, while for large concentrations and small disorder, one expects the states to be delocalized.

This simple picture ignores formation of loops namely that the path can intersect itself. This would enable an indefinite propagation in a closed loop with a divergent expansion, (2.28), but not necessarily implying delocalization. Furthermore, the interference between paths that traverse in opposite directions in a loop causes localization and it is the underlying physics of weak localization described in the previous section.

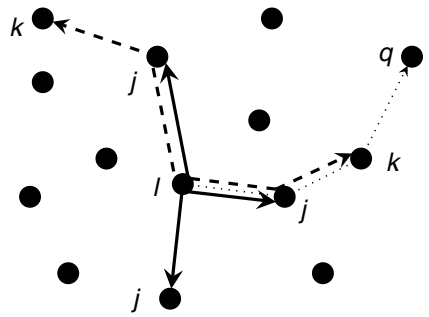

Figure 2.7. Visualization of the processes leading to (2.28). The solid lines represent terms in the first sum which admix functions on close by sites. The dashed lines are for terms through there to further sites, and so on.

Alternatively, the transition can be approached from the extended phase. The elastic scattering length L_{el} decreases with increasing disorder, but it cannot be smaller than the distance between sites r_0. It is thus conceivable that a drastic change of behavior will take place when approaching this limiting value of $L_{\mathrm{el}} \approx r_0$.

2.5.2 Scaling theory

It is clear that the transition must be sensitive to the dimensionality of the system. In a disordered one-dimensional system, it has been shown early on by Mott and Twose (1961) that all states must be localized even for small disorder. In higher dimensions, breaks can be avoided by changing the direction of propagation. It turns out that in 2D the wave functions also must be localized for any disorder, while in 3D a transition can occur. A clear picture of the situation was obtained with the scaling theory of localization (Abrahams et al., 1979). The scaling theory is based on an argument by Thouless (1974), which relates the conductance to a diffusive time across a system of linear size L and an average level spacing $\Delta E(L)$ of (one-electron) levels. One can consider the macroscopic system made up of subsystems of linear dimensions L arranged next to each other. Since we consider a macroscopically homogeneous system, the statistical quantities are the same in all (equally sized) subparts of the system. In order for the electron to propagate from one subsystem to an adjacent subsystem, the appropriate levels have to match, within a certain tolerance. The tolerance (level broadening) is given by the uncertainty principle, namely the magnitude $\delta E(L) = \ /\tau(L)$ where $\tau(L)$ is the diffusion time through the system of size L. The typical level separation is $\Delta E = W/L^d$. Roughly speaking, if $\delta E > \Delta E$, the electron can enter the adjacent site; otherwise, it cannot. Assuming that the motion is diffusive, $\tau(L) = L^2/D_L$, where D_L is the diffusion constant, so $\delta E/\Delta E \propto D_L L^{d-2}$ (i.e., is proportional to the conductance) and is the relevant magnitude to analyze the behavior of the electron.

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

Abrahams et al. (1979) (commonly known as "the gang of four") formulated a scaling theory for the conductance $g(L)$ of a cube of d dimensions ($d = 1, 2, 3$) of linear dimension L at $T = 0$. They considered scaling the dimension of the cube by a factor b, such that the linear dimension is bL, and argued that for any dimensionality the conductance g must be a function of b and of $g(L)$ (i.e., $g(bL) = f[b, g(L)]$). It is convenient to make b infinitesimally larger than 1, $b = 1 + \epsilon$, $\epsilon \ll 1$. Then, with $\epsilon = \mathrm{d}L/L$,

$$g(bL) = g(L + \epsilon L) = g(L + \mathrm{d}L) = f[1 + \epsilon, g(L)] \tag{2.29}$$

and

$$\mathrm{d}L \frac{\mathrm{d}g}{\mathrm{d}L} = g(L + \mathrm{d}L) - g(L) = f[1 + \epsilon, g(L)] - f[1, g(L)] = \frac{\partial f}{\partial b}\epsilon \tag{2.30}$$

or

$$\frac{1}{b}\frac{\partial f}{\partial b}\bigg|_1 = \frac{\mathrm{d}\ln g(L)}{\mathrm{d}\ln L} \equiv \beta[g(L)] \tag{2.31}$$

The main assumption of the scaling theory is that, for any dimension, β remains a function of g only. The dependences of β on L, W, and the Fermi level, all enter implicitly through g. In other words, g is the only relevant scaling variable of the problem.

In the macroscopic limit, conduction at $T \to 0$ requires the existence of delocalized states. Thus, $\lim \beta(L \to \infty) > 0$ allows for delocalization, while $\lim \beta(L \to \infty) < 0$ implies localization. Whether the systems is characterized by localization or delocalization depends on dimensionality, as will be shown. In the limit of small g (large disorder), the conduction is in a strongly localized regime and thus goes as $g_0 \exp(-2L/\xi)$, where ξ is the localization length. On the other hand, in the limit of very large g, corresponding to a weakly disordered system, Ohm's law states that g is proportional to L^{d-2}. These dependences of $g(L)$ in the limits of large and small g give

$$\beta \to \begin{cases} d - 2 & \text{for } g \to \infty \\ c_d + \ln g & \text{for } g \to 0 \end{cases} \tag{2.32}$$

where c_d is a constant that depends on dimensionality. The value of β for intermediate values of g is expected to be a smooth monotonous function of g (and so of $\ln g$). This leads to the interpolation showed in Figure 2.8 for systems of dimensions one, two, and three, from bottom to top, respectively. Support for the previous assumption also came from perturbation theory in the weak disorder regime (Anderson et al., 1979; Gorkov et al., 1979) and a solid justification was provided by the so-called nonlinear σ model (Wegner, 1979; Efetov, 1997).

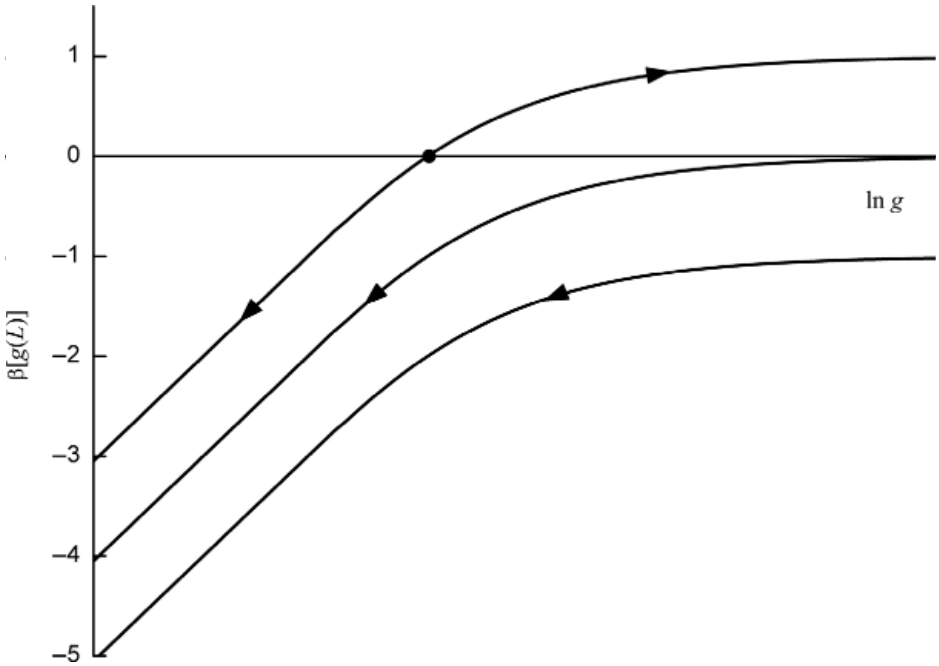

Figure 2.8. β as a function of $\ln g$ for 1D, 2D, and 3D systems, from bottom to top, respectively. The arrows indicate the flow of the conductance with system size.

According to this picture in one dimension, all states are localized, as expected. In two dimensions, all states are localized as well for any finite g (i.e., nonzero disorder) but mildly so at large g. This is so because β is always negative, and increasing the system size causes the conductivity to diminish. The localization length may be larger than L for a finite system in which case the system will behave like a metal in an experiment.

In three dimensions, there is a transition from delocalized to localized states (i.e., a metal-insulator transition) where the curve crosses the horizontal axis. If for a given disorder one starts with $\beta < 0$, increasing the size will make the system more and more insulating. On the other hand, if one starts with $\beta > 0$, the system will become a better conductor as the size increases.

The scaling theory should be modified in the presence of strong spin-orbit coupling. When spin-orbit coupling is important, the scaling curve for 2D systems changes drastically. At very large $\ln g$, β tends to zero from positive values, instead of negative values as in the standard case. At very small values of $\ln g$, the curve must be negative; hence, it must cross the horizontal axis and the system must undergo a metal–insulator transition.

2.5.2.1 Interaction effects

The scaling theory was developed for a non-interacting system. A number of theoretical works in the early 1980s predicted that high enough electron interactions

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

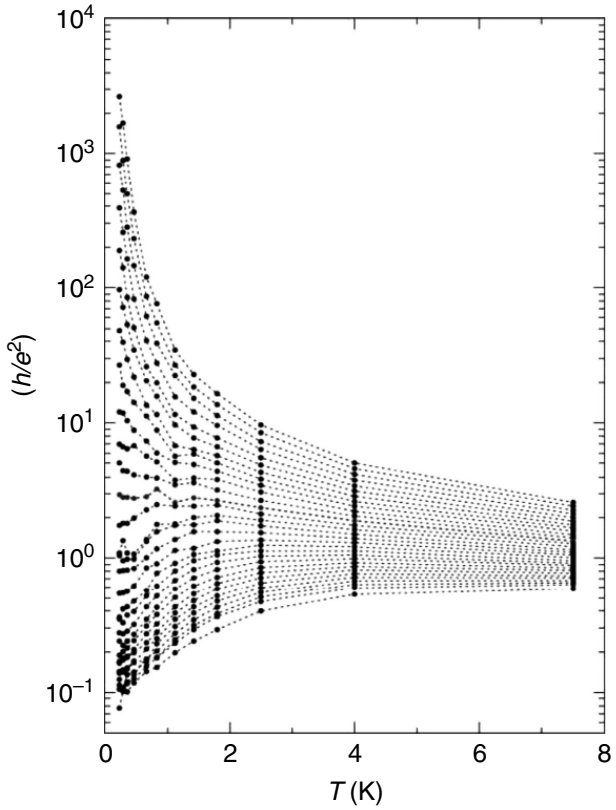

Figure 2.9. Temperature dependence of the resistivity in a dilute lightly disordered Si MOSFET for 30 different electron densities ranging from 7.12×10^{10} cm^2 to 1.37×10^{11} cm^2. From Kravchenko et al. (1995). Copyright by the American Physical Society.

would lead to the scaling of a 2D film to a metallic state at low temperatures. Experimental indications for a metal-insulator transition as a function of carrier density in a two-dimensional film were observed only a decade later, first in Si MOSFET films and later in other materials (for a recent review see Kravchenko and Sarachik 2010). The strength of interaction in 2D films was controlled by applying a back gate voltage, which varied the 2D carrier concentration, n_{2D}. This interaction can be characterized by the dimensionless Wigner-Seitz radius:

$$r_s = \frac{1}{(\pi n_{2D})^{1/2}\xi} \tag{2.33}$$

Figure 2.9 shows the resistivity as a function of T for different n_{2D} of a high mobility ($\mu \sim 4 \times 10^4$ cm^2/Vs) Si MOSFET film. It is seen that for low carrier concentration ($r_s < 10$) the film is insulating; however, for high concentrations, when $r_s > 10$, the resistance drops significantly with decreasing temperature, apparently exhibiting 2D metallic behavior.

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

The reason for the decade's delay of experiment versus prediction was attributed to the advance in semiconductor technology, which enabled the fabrication of very high mobility samples. The high mobility is required in order to be able to measure samples with low enough n so that the interaction energy is large compared to E_F. Notice that the interaction grows in importance with increasing concentration, but the Fermi level increases with concentration faster. This is also why the Wigner transition, described below, occurs at low concentrations.

The "separatrix" resistivity curve, which separates between insulating and metallic curves, was found to be characterized by a sheet resistance $R = h/e^2$ at high temperatures, which approached $3h/e^2$ at low T for all materials independent on the critical n_{2D}.

Various explanations have been suggested for the experimental results and the option of a metallic state in a two-dimensional system is still under debate. Punnoose and Finkel'stein (2001) provided an explanation for the observed metal-insulator transition in 2D by extending the scaling theory to an interacting system. They obtained two scaling expressions for the resistivity, $\rho = (e^2/h)R$, and for the electron–electron scattering amplitude, γ_2

$$-\frac{\mathrm{d}\ln(\rho)}{\mathrm{d}\ln(T\tau_{el})} = \rho^2\left\{n_\nu + 1 - (4n_\nu^2 - 1)\left[\frac{1+\gamma_2}{\gamma_2}\ln(1+\gamma_2) - 1\right]\right\} \tag{2.34}$$

and

$$-\frac{\mathrm{d}\gamma_2}{\mathrm{d}\ln(T\tau_{el})} = \rho\frac{(1+\gamma_2)^2}{2} \tag{2.35}$$

where n_ν is the number of degenerate minima in the electronic spectrum. Equations (2.34) and (2.35) predict specific temperature-dependence curves for the resistance and the interaction amplitude. Figure 2.10 shows the comparison between these theoretical predictions and the experimental results of Anissimova et al. (2007). The agreement is impressive, especially since there are no fitting parameters. It should be noted, however, that the Punnoose and Finkel'stein theory assumes diffusive conductivity, while many of the experimental results were obtained in the ballistic regime where the theory is not applicable.

2.5.3 The anderson metal–insulator transition

The disorder induced transition is called the Anderson transition and arises from a competition between disorder energy and kinetic energy (i.e., between the first and last terms of the Hamiltonian (2.10)). Notice that both parts are one-electron operators so the Anderson localization is a one-particle phenomenon. The most important consequence of a transition between systems that are localized, and systems that are delocalized is the existence of a metal-insulator transition. In the limit of very

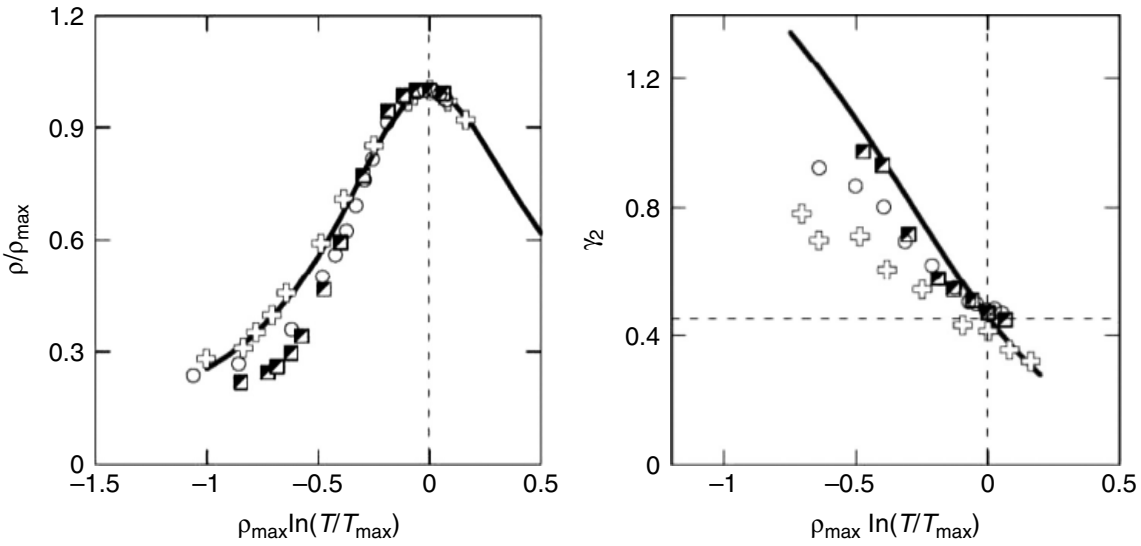

Figure 2.10. Comparison between theory (lines) and experiment (symbols). (a): $\rho/\rho_{\max}$ as a function of $\rho_{\max}\ln(T/T_{\max})$. (b): γ_2 as a function of $\rho_{\max}\ln(T/T_{\max})$. Vertical dashed lines correspond to $T = T_{\max}$, the temperature at which $\rho(T)$ reaches maximum. Electron densities are 9.87 (squares), 9.58 (circles), and 9.14×10^{10} cm^2 (crosses). From Anissimova et al. (2007). Reprinted by permission from Macmillan Publishers Ltd.

low temperatures, the localized system cannot carry current, while the delocalized system will carry current.

The transition is characterized by a set of critical exponents, and the most important for the book is the one describing the behavior of the localization length. In the strongly localized regime, the localization length is equal to the decay length of an isolated state and increases as the disorder decreases, diverging at the transition as a power law

$$\xi \propto (W - W_c)^{-\nu} \tag{2.36}$$

where W_c is the critical disorder for the transition (for a given t) and ν is called the correlation length exponent. The most precise numerical simulation up to date found $\nu = 1.57 \pm 0.02$ (Slevin and Ohtsuki, 1999).

At the transition, the wavefunctions are multifractal, due to the strong fluctuations at criticality. A multifractal wavefunction is characterized by an infinite set of critical exponents describing the scaling of its moments

$$\int \mathrm{d}^3\mathbf{r}|\psi(\mathbf{r})|^{2q} \sim L^{D_q(q-1)} \tag{2.37}$$

where D_q are fractal dimensions. For a localized state $D_q = 0$ and for an extended state $D_q = d$, while at criticality D_q is a nontrivial function of q (Evers and Mirlin, 2008). The most important fractal dimension is D_2, which is the normal fractal

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

dimension d_f that will appear in other contexts. Numerical simulations found that d_f is equal to 1.3 in 3D systems (Mildenberger et al., 2002). Large localized states have also a fractal structure down to distances of the order of the localization length.

The conductivity tends to zero as the transition is approached from the extended phase as a power law

$$\sigma \propto (W_c - W)^s \tag{2.38}$$

The conductivity exponent s is related to the localization length exponent ν through the scaling relation $s = \nu(d-2)$ (Markos, 2006).

In 3D the conductivity tends to zero at the transition as the inverse of the correlation length that diverges as the localization length. This is contrary to the concept of minimum metallic conductivity, proposed by Mott, based on extensive experimental support at early times. Mott argued that the minimum value of the elastic mean free path is $k_F L_{el} \approx 1$, k_F being the Fermi wavenumber, and the corresponding conductivity is the minimum value that a system can have

$$\sigma_{min} \approx 0.025 \frac{e^2}{a} \tag{2.39}$$

In hopping experiments, one often sees that the preexponential factor of (2.17) is close to this value. In 2D a similar argument leads to a minimum metallic conductivity equal to $e^2/$.

The dielectric function also diverges at the transition

$$\kappa \propto (W - W_c)^{-\zeta} \tag{2.40}$$

A rough estimate of ζ from tunneling experiments is $\zeta \approx \nu$ (Lee et al., 1999). The increase of the dielectric constant as the transition is approached means that interaction effects are diminished.

2.5.4 The mott metal – insulator transition

A Mott transition as originally conceived (Mott, 1958) occurs in an ordered (i.e., crystalline) system; however, it can play an important role in disordered systems as well. The Mott metal–insulator transition relates to monovalent systems and arises from a competition between intrasite Coulomb interaction and kinetic energy. It is sometimes called the Mott-Hubbard transition because it can be derived from a Hamiltonian due to Hubbard (1963), which incorporates the intrasite repulsion energy and the kinetic energy, basically the second and the last term of (2.9).

For insight, it is again useful to first consider a two-site system, now with two electrons (e.g., a hydrogen molecule). Coulson and Fisher (1949) discussed in some detail this system, concurrent with Mott's paper (Mott, 1949). The basic idea is the

following: in the absence of electron–electron interaction, the two-electron states correspond to the various possible occupations of the one-electron states of (2.27). The assumption here is that there is no energy disorder. The lowest energy state is then the state with both electrons in $|j\rangle + |k\rangle$:

$$[(|j_1\rangle + |k_1\rangle)(|j_2\rangle + |k_2\rangle) + (|j_2\rangle + |k_2\rangle)(|j_1\rangle + |k_1\rangle)]/\sqrt{2} \tag{2.41}$$

and by symmetry must of course be a singlet. The subscripts 1 and 2 label the two electrons. In comparison with the energy of two electrons on isolated sites the energy is lowered by $2t_{j,k}$ proportional to the overlap $\exp(-r_{j,k}/\xi)$. There are four excited states with one electron in $|j\rangle + |k\rangle$ and one in $|j\rangle - |k\rangle$ (a singlet and a triplet), and a singlet state with both electrons in $|j\rangle - |k\rangle$. Clearly any of the above six states is delocalized in the sense that each electron has a probability $1/2$ to be on either site.

The dominant electron–electron interaction is the intrasite interaction $\approx e^2/(2\kappa\xi)$. Equation (2.41) is a fair description of the ground state if $e^2/(2\kappa\xi)$ is small compared to $t_{j,k}$. On the other hand, if $e^2/(2\kappa\xi)$ is large compared to $t_{j,k}$ (i.e., when the spacing $r_{j,k}$ is large), the electrons prefer to stay on each site separately, that is, in the (singlet) state

$$[|j_1\rangle|k_2\rangle + |j_2\rangle|k_1\rangle]/\sqrt{2} \tag{2.42}$$

or in one of the triplet states

$$[|j_1\rangle|k_2\rangle - |j_2\rangle|k_1\rangle]/\sqrt{2} \tag{2.43}$$

Exchange energy usually makes the singlet state have the lower energy. On this microscopic two-site crystal, the singlet ground state can be considered as antiferromagnetic. The remaining two singlet states are $|j_1\rangle|j_2\rangle$ and $|k_1\rangle|k_2\rangle$ states with both electrons on the same sites (i.e., one negative and one positive ion). In passing, it should be commented that, in analogy with a negative hydrogen ion, one should expect the negative ion to have an appreciably larger localization length, which violates the assumption of the tight-binding approximation. This fact is often suppressed but has been emphasized in some contexts (Kamimura, 1979; Kurobe and Kamimura, 1982).

Expanding now to a macroscopic crystal with N sites, the high-density case becomes a band of $2N$ (the 2 is for two possible spins) states

$$\sum_j \exp\ \{i\mathbf{k}\cdot\mathbf{r}_j\}\,|j\rangle \tag{2.44}$$

filled with N electrons in the lowest energy levels. It is thus a metal since the band is half occupied. In the low-density (large r) case, the small gain in kinetic

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

energy (i.e., small t) loses with respect to the intrasite Coulomb interaction and the electrons are localized on the sites with $n_j = 1$ for all j (i.e., an antisymmetrized product of $|j, \sigma_j\rangle$ where σ_j stands for the spin on site j).

The N-electron ground state has an antiferromagnetic arrangement of spins, due to exchange, and the low-energy excitations within this system of states are spin rearrangements. There are 2^N possible spin arrangements while the total number of states (i.e., the dimension of the Hilbert space) is $(2N!)/(N!)^2$, which is much larger. The other states involve doubly occupied sites and an equal number of empty sites. This is the so-called upper Hubbard band. For large r, there is a gap between this band and the lower 2^N states so such a system is an insulator, and thus there exists an insulator to metal transition as r increases from small to large values.

In disordered systems, Anderson and Mott localizations usually reinforce each other. In impurity conduction, for example, the disorder may produce an overlap of the two Hubbard bands, but the DOS in the overlapping region is small, and it is likely that the states will be Anderson localized, the system being still an insulator.

2.5.4.1 The Wigner transition

The Wigner transition is probably the first proposal of a metal–insulator transition, dating back to 1938 (Wigner, 1938). It is similar to the Mott transition in the sense that it results from a competition between kinetic energy and Coulomb interactions – this time, interactions among free electrons. Although free electrons are not directly a subject of this book, we mention the Wigner transition in passing since localization into a Wigner crystal is sometimes invoked in the literature on electron glasses. Since it also arises from competition between kinetic and Coulomb energies the Coulomb interactions (and with it localization) wins at *low* electron concentration. The reason is that the kinetic energy increases faster with concentration than the Coulomb energy. It should be emphasized that this contrasts with the case of Anderson localized electrons where the role of (intersite) Coulomb interaction becomes more important as the electron concentration *increases* because the electrons are already localized by disorder.

2.6 Percolation theory

As noted in Section 2.3.3, the fact that the hopping rates, given by (2.16), are exponentially distributed makes mean field approaches not suitable for calculating the resistance of a strongly disordered sample. A more appropriate treatment is based on the percolation theory. This treatment provides profound physical insight and shows that the conductivity is governed by a set of critical resistances and that the current flows in a very nonhomogeneous network. The rigorous derivation of

the percolation approach for hopping conductivity will be provided in Chapter 5. Here we outline the basic foundations of percolation theory.

2.6.1 Percolation – basic concepts

Percolation theory addresses the question of transport through strongly inhomogeneous media. It was originally introduced in connection with hydrology of rocks (Broadbent and Hammersley, 1957) and lately this application of percolation theory witnessed a strong revival (Hunt and Ewing, 2009). A simple example of a percolation problem is a space-filling random arrangement of volumes, some fluid-transmitting, the others fluid blocking. The central question is the fraction of volume filled by the fluid-transmitting sections that separates between blocking and transmitting fluid over a macroscopic distance. Several books dedicated to percolation theory exist (Hunt and Ewing, 2009; Stauffer and Aharony, 1992; Efros, 1986) and review articles (Kirkpatrick, 1973).

The simplest cases of percolation problems are defined on a lattice. The lattice consists of a set of lattice points forming a regular array, and bonds between these points. Usually only the bonds between nearest neighbors are considered. In the so-called bond percolation problem, each bond has a certain probability p of connecting two corresponding lattice points, and a probability $1 - p$ of not connecting them. The problem is to know for what value of p there is a continuous connection over macroscopic distances. As p increases, these bonds combine into clusters of two, three, and more bonds. With increasing p, larger and larger clusters are formed until for large enough p a large cluster that spans the entire system is established. For an infinite system, this happens at a sharply defined critical percolation probability p_c. The value of p_c differs from lattice to lattice, and it depends primarily on the coordination number (number of nearest neighbors), z, and on the dimensionality d. Shante and Kirpatrick (1971) pointed out that this approximate dependence is

$$z p_c \approx \frac{d}{d-1} \tag{2.45}$$

Another lattice model is site percolation, where the connectivity criteria is not a property of a bond, but rather a property of a site. When a lattice site allows passage that occurs with probability p, all the bonds linked to this site allow passage up to the midpoint of the bond. Again there is a critical percolation probability p_c, the smallest p for which an infinite cluster exists. p_c is different for bond and site percolation and depends on the type of lattice.

Percolation can be also formulated for random lattices, a problem directly applicable to hopping conduction. Lattice sites are at random, and usually one considers bonds between all pairs of sites. The condition for passage includes the bond length

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

in such a way that longer bonds are less likely to be connecting bonds. If we consider an infinite number of bonds for each site, p_c obviously vanishes. In place of p one uses P, the average number of connecting bonds per site, which is well defined. A simple random lattice model is r-percolation – one inserts bonds with increasing length, beginning from the shortest, up to the length where there is a macroscopic path. Pike and Seager (1974) determined by simulations that in this model $P_c = 2.4$. In all that follows a disordered lattice is considered and therefore P will be used. Similar derivations can be applied to ordered latices where the use of p is more natural.

A first quantity of interest is the mean number of bonds in a cluster, $\langle s \rangle$, defined as

$$\langle s \rangle = \frac{\sum_s s^2 n_s}{\sum_s s n_s} \tag{2.46}$$

where s is the number of bonds in a cluster and n_s is the number of cluster of size s. The denominator in the definition of $\langle s \rangle$ is the total number of connecting bonds in the sample. Above P_c, it is understood that the contribution from the infinite cluster is not included in Equation (2.46). Near P_c, $\langle s \rangle$ diverges as

$$\langle s \rangle \propto |P - P_c|^{-\gamma} \tag{2.47}$$

where γ is a critical exponent. The quantity $\langle s \rangle$ is analogous to the susceptibility in magnetic phase transitions.

To study the typical linear dimension of clusters, one introduces the correlation function $C(r)$, defined as the probability that a bond at distance r from a bond in a certain cluster also belongs to the same finite cluster. The correlation length, L_P, which plays an important role in hopping conduction, is the average distance between two bonds belonging to the same cluster

$$L_P^2 = \frac{\sum_r r^2 C(r)}{\sum_r C(r)} \tag{2.48}$$

Near P_c, the correlation length is the average radius (defined as the mean distance between two bonds in the cluster) of the largest clusters and diverges as

$$L_P \propto |P - P_c|^{-\nu} \tag{2.49}$$

The correlation length exponent ν is $4/3$ in two dimensions and 0.88 in three. Above P_c, L_P is the size of the largest "holes" in the extended cluster.

The average radius $r(s)$ of a large cluster of s bonds is related to the number of bonds s through

$$r(s) \propto s^{d_f} \tag{2.50}$$

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

where d_f is the fractal dimension, which is another critical exponent. The fractal dimension d_f also characterizes the behavior of the extended percolation cluster at P_c, which for a finite system of linear size L contains an average of L^{d_f} bonds. Clusters near P_c are fractals, while far above P_c, they are extended objects with the dimensionality of the lattice.

The correlation length L_P is the only relevant length dominating the critical behavior. As in thermal phase transitions, the existence of only one relevant length implies single-parameter scaling, which results in scaling relations between the critical exponents. Only two of them are independent, and the rest can be obtained by the scaling relations. For example, the exponents defined above are related by

$$\nu(2d_f - d) = \gamma \tag{2.51}$$

2.6.2 Percolation conductivity

One can associate a conductance 1 with every connecting bond and zero with the nonconnecting bonds and ask what is the total conductivity of a macroscopic sample as a function of P when this is connected to two electrodes on opposite sides. Obviously, the conductivity is zero if there is no extended cluster. One could then naively think that the conductivity is proportional to the number of bonds in the percolation cluster. However, this is not the case. Sequences of bonds that are connected to the percolation cluster by only a single bond cannot contribute to the current. These are called dead ends. The collection of bonds through which current flows is known as the backbone cluster. A bond belongs to the backbone cluster if and only if it is connected to the two electrodes by nonintersecting paths.

The percolation conductivity σ goes to zero at P_c as

$$\sigma \propto (P - P_c)^{\lambda} \tag{2.52}$$

This new exponent λ is not directly related to previously defined exponents through scaling relations. To obtain it, extra assumptions must be made about the topology of the backbone cluster. Monte Carlo simulations found $\lambda = 0.975\nu$ in 2D systems (Normand et al., 1988) and 2.28ν in 3D systems (Gingold and Lobb, 1990).

The real structure of the backbone network is very complicated, including structures at many length scales. A useful simple model was proposed by Skal and Shklovskii (1975) and De-Gennes (1976). For $P > P_c$, one can simulate the situation by a periodic network whose nodes are separated a distance L_P with links connecting them. The number of links along the direction of the current is proportional to the length of the system, while the number of links in each section perpendicular to the current is proportional to the cluster area. The difficult task is the calculation of the typical conductance of a link, which can be fairly winding

and which can include parallel paths in some sections. Pike and Stanley (1981) improved the nodes and links model with the inclusion of blobs, sections of links with parallel paths. They also realized that the number of singly connected bonds in a link, which must be roughly proportional to the conductance of the link G_0, grows as $(P - P_c)^1$. The conductivity of the system is

$$\sigma = \frac{G_0 (L/L_P)^{d-2}}{L^{d-2}} = \frac{G_0}{L_P^{d-2}} \propto (P - P_c)^{1+\nu(d-2)} \tag{2.53}$$

in reasonable agreement with the previously mentioned numerical simulations.

2.6.2.1 Exponential spread of resistances

A problem of interest is a network of resistances with an exponentially wide distribution, that is, of the form

$$R_{i,j} = R_0 e^{\eta} \tag{2.54}$$

where η is some random variable, which for this analysis can be assumed to be uniformly distributed in the interval $-\eta_0 < \eta < \eta_0$. The underlying lattice can be of any dimension and can be regular or disordered. Most of the rationale for solving this problem for hopping conduction with large η_0 comes from the fact that a parallel (series) connection of exponentially widely distributed resistances has a resistance nearly equal to the smallest (largest) resistance. Starting by connecting the smallest resistances of the system and gradually connecting larger and larger resistances, one gets finite clusters that increase in size. At and beyond criticality, $p_c = (\eta_c + \eta_0)/(2\eta_0)$, one gets an infinite cluster that can carry current. The critical value, η_c, needed for percolation determines the conductivity of the system as will be shown bellow. In the finite cluster regime, the statistics of clusters allows one to calculate the frequency dependent conductivity.

According to the previous arguments, at criticality the conductivity is still zero because the distance, L_p, between current carrying paths is infinite. To reduce L_P, one needs to increase the critical resistance above R_c to arrive at an optimal resistance, R_{opt}, which minimizes the ratio L_p/R_c with respect to L_p (Friedman and Pollak, 1981). This results in

$$\frac{d\eta_{opt}}{dL_p} = -\frac{1}{L_p} \quad \Rightarrow \quad L_p \propto \eta_0^{\nu} \tag{2.55}$$

The conductivity is then proportional to

$$\sigma \propto \eta_0^{\nu} e^{-\eta_{opt}} \approx \eta_0^{\nu} e^{-\eta_c} \tag{2.56}$$

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press

Usually the increase in η, with respect to η_c, needed to establish the current carrying network is very small and can be neglected as will be discussed in Chapter 5. The previous argument indicates that the prefactor is proportional to η_0^y.

Le Doussal (1989) showed for hierarchical lattices that the exponent y of the preexponential of the conductivity for an exponentially wide distributed resistances is

$$y = (d-2)\nu \tag{2.57}$$

in agreement with the previous intuitive calculation and with Shklovskii and Efros (1984). He also suggested that this result applies to all lattices for $d \leq 6$. The result $y = 0$ is exact for the square lattice, see Strelniker et al. (2005). Also, Equation (2.57) has been verified numerically for several two and three-dimensional lattices (Tyc and Halperin, 1989). Strelniker et al. (2005) studied the distribution function of the resistance of finites samples and found that it depends only on L/η_0 and is approximately log-normal.

https://doi.org/10.1017/CBO9780511978999.002 Published online by Cambridge University Press