Carlo Jacoboni

SPRINGER SERIES IN SOLID-STATE SCIENCES 165

Theory of Electron Transport in Semiconductors

A Pathway from Elementary Physics to Nonequilibrium Green Functions

Springer Series in
SOLID-STATE SCIENCES 165

Springer Series in

SOLID-STATE SCIENCES

The Springer Series in Solid-State Sciences consists of fundamental scientific books prepared by leading researchers in the field. They strive to communicate, in a systematic and comprehensive way, the basic principles as well as new developments in theoretical and experimental solid-state physics.

Carlo Jacoboni

Theory of Electron Transport in Semiconductors

A Pathway from Elementary Physics to Nonequilibrium Green Functions

With 216 Figures

Springer

Prof. Carlo Jacoboni
Università di Modena e Reggio Emilia, Dipartimento di Fisica
Via Campi 213/A, 41125 Modena, Italy
E-mail: carlo.jacoboni@unimore.it

Springer Series in Solid-State Sciences ISSN 0171-1873
ISBN 978-3-642-10585-2 e-ISBN 978-3-642-10586-9
DOI 10.1007/978-3-642-10586-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010933971

Cover design: eStudio Calamar Steinen

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my students

Preface

This book originated out of a desire to provide students with an instrument which might lead them from knowledge of elementary classical and quantum physics to modern theoretical techniques for the analysis of electron transport in semiconductors. The book is basically a textbook for students of physics, material science, and electronics. Rather than a monograph on detailed advanced research in a specific area, it intends to introduce the reader to the fascinating field of electron dynamics in semiconductors, a field that, through its applications to electronics, greatly contributed to the transformation of all our lives in the second half of the twentieth century, and continues to provide surprises and new challenges.

The field is so extensive that it has been necessary to leave aside many subjects, while others could be dealt with only in terms of their basic principles.

The book is divided into five major parts. Part I moves from a survey of the fundamentals of classical and quantum physics to a brief review of basic semiconductor physics. Its purpose is to establish a common platform of language and symbols, and to make the entire treatment, as far as possible, self-contained. Parts II and III, respectively, develop transport theory in bulk semiconductors in semiclassical and quantum frames. Part IV is devoted to semiconductor structures, including devices and mesoscopic coherent systems. Finally, Part V develops the basic theoretical tools of transport theory within the modern nonequilibrium Green-function formulation, starting from an introduction to second-quantization formalism.

Preparing this text has been a very long and at times painful task, especially when it became obvious that it simply could not cope with the overly large ambitions of the original project. I am deeply grateful to my family for understanding and accepting with love my absorption in writing it over a period of several years. In this endeavor, I have been helped by many colleagues. In particular, I thank Antonio Abramo, Andrea Bertoni, Paolo Bordone, Rossella Brunetti, Fabrizio Buscemi, Mauro Ferrario, Fabio Giovanardi, Chihiro Hamaguchi, Paolo Lugli, Giampiero Ottaviani, Enrico

Piccinini, Maria Prudenziati, Lino Reggiani, Susanna Reggiani, Fausto Rossi, Massimo Rudan, Alice Ruini, who have read various parts of the manuscript and/or suggested many improvements, both topical and stylistic.

The main contribution to this text, however, has come from my many students, undergraduate, graduate, and postdocs alike, who since many decades have accompanied my research and teaching activity with intelligence, curiosity, and affection. Without them not only would this book not have been conceived, but my activity itself as represented in it simply would not exist. I cannot name all these people: they have been so numerous that to mention some would inevitably mean being unfair to the others. Some are now well-known scientists around the world, some have taken a way that took them far off, out of sight but certainly not out of mind, and others remained close and continue to share with me our daily work.

This book is dedicated to every one of them.

Modena *C. Jacoboni*
May 2010

Contents

Part V Quantum Transport with Nonequilibrium Green Functions

Symbols and Abbreviations

2DEG	Two-dimensional electron gas
AFM	Atomic-force microscopy
$A(\boldsymbol{k}, \omega)$	Spectral density
$\boldsymbol{A}$	Vector electromagnetic potential
$\mathcal{A}$	Fourier amplitude of the perturbation Hamiltonian
$\langle \mathcal{A} \rangle_{\Psi}$	Expectation value of the observable $\mathcal{A}$ on the state $\lvert \Psi \rangle$
$[\mathcal{A}, \mathcal{B}], [\mathcal{A}, \mathcal{B}]_-$	Commutator $\mathcal{AB} - \mathcal{BA}$
$[\mathcal{A}, \mathcal{B}]_+$	Anticommutator $\mathcal{AB} + \mathcal{BA}$
$[u, v]_{q,p}$	Poisson bracket
a , $\mathsf{a}^\dagger$	Annihilation and creation phonon operators, bosonic annihilation and creation operators
a	Lattice constant
$\boldsymbol{a}$	Acceleration
$\boldsymbol{a}_i$	Unit vectors of direct lattice
$a_{n\boldsymbol{k}}$	Coefficient of the expansion of a wavepacket in the n-th band in Bloch states
BE	Boltzmann equation
BH	Brooks-Herring
BO	Bloch oscillations
$\boldsymbol{B}$	Magnetic induction field
$\boldsymbol{b}_i$	Unit vectors of reciprocal lattice
b	Cutoff of Coulomb potential in CW approach
bcc	Body-centered cubic
b , $\mathsf{b}^\dagger$	Fermionic annihilation and creation operators
BZ	Brillouin zone
BJT	Bipolar junction transistor
C	Elastic constant, autocorrelation function
$C_n(\boldsymbol{R})$	Coefficient of the n-th band in Fourier series
CNT	Carbon nanotube
CW	Conwell-Weisskopf
c	Light velocity in vacuum

$c_{n\boldsymbol{k}}$	Coefficient of the wavefunction in the n-th band in Bloch-state series
$\mathsf{c}\,, \mathsf{c}^\dagger$	Annihilation and creation operators
D	Density of states in frequency space, diffusion coefficient
$(D_t K)$	Deformation potential constant for optical phonon scattering
$(D_t K)_i$	Deformation potential constant for intervalley phonon scattering
$\boldsymbol{D}$	Electric induction field
DFT	Density functional theory
$\boldsymbol{E}$	Electric field
EMC	Ensemble Monte Carlo
$\boldsymbol{E}_e$	Electric field produced by an electron
$\boldsymbol{E}_p$	Electric field associated to a polarization field
E_H	Hall field
E_{ij}, E_1	Deformation potential constant for acoustic-phonon scattering
$-\mathrm{e}$	Electron charge
e^*	Effective charge on atoms of compound materials
$\boldsymbol{e}$	Polarization direction
e_{ijk}	Piezoelectric constant
F	Helmholtz free energy
$F(\boldsymbol{r}, t)$	Envelope function
$\boldsymbol{F}$	Force
FQHE	Fractional quantum Hall effect
FET	Field-effect transistor
f	Distribution function
f_F	Fermi-Dirac distribution
f_M	Maxwell distribution
f_w	Wigner function
f_1	Non-equilibrium correction to the distribution function
f_{nm}	Transformation matrix from matrix elements to Weyl-Wigner transform
fcc	Face-centered cubic
G	Gibbs free energy, spectral density, contour-ordered Green function, conductance
$G(\epsilon)$	Integrated density of states
$\boldsymbol{G}$	Reciprocal lattice vector
$\mathcal{G}$	Overlap integral
GF	Green function
G^a	Advanced Green function
G^r	Retarded Green function
$G^>$	Green function G greater
$G^<$	Green function G less
G^t	Time-ordered Green function
$G^{\bar{t}}$	Anti-time-ordered Green function
g	Generation rate, Green functions, density of states
$\boldsymbol{H}$	Magnetic field

H	Hamiltonian function
H_{ep}	Electron Hamiltonian in a polarization field
$\mathcal{H}$	Hamiltonian operator
$\mathcal{H}_\circ$	Unperturbed Hamiltonian operator
$\mathcal{H}_c$	Crystal Hamiltonian operator
$\mathcal{H}_e$	Electron Hamiltonian operator
$\mathcal{H}_{ep}$	Electron–phonon interaction Hamiltonian
$\mathcal{H}_p$	Free-phonon Hamiltonian
$\mathcal{H}'$	Interaction Hamiltonian operator
HEMT	High-electron-mobility transistor
h	Planck constant
$\hbar$	Planck reduced constant $\frac{h}{2\pi}$
I	Current
IQHE	Integer quantum Hall effect
$\boldsymbol{j}$	Current density
$\mathcal{J}$	Current operator
$\boldsymbol{j}_D$	Drift current density
K_B	Boltzmann constant
$\boldsymbol{k}$	Wavevector, crystal wavevector
$\boldsymbol{k}^*$	Herring-Vogt transformed crystal wavevector
$\boldsymbol{k}_r$	Relative crystal momentum of two colliding electrons
k_ℓ	Longitudinal component of crystal wavevector
k_t	Transverse component of crystal wavevector
$\boldsymbol{L}$	Angular momentum
L	Lagrangian function, string length, crystal length
$\mathcal{L}$	Liouvillian operator
LA	Longitudinal acoustic
LED	Light-emitting diode
LCAO	Linear combination of atomic orbitals
LO	Longitudinal optical
LSI	Large-scale integration
l	Mean free path
ℓ	Polarization index
M	Mass
M_i	i-th moment of Boltzmann equation
$\boldsymbol{M}$	Magnetization field
MBE	Molecular beam epitaxy
MC	Monte Carlo
MOCVD	Metal-oxide chemical vapor deposition
MWCNT	Multiwalled carbon nanotube
MESFET	Metal-semiconductor field-effect transistor
MOSFET	Metal-oxide-semiconductor field-effect transistor
m	Mass, effective mass
m_a	Acceleration effective mass
m_c	Conductivity effective mass

m_d	Density-of-state effective mass
$m_\circ$	Free electron mass
m_e	Electron effective mass
m_h	Hole effective mass
m_ℓ	Longitudinal effective mass
m_t	Transverse effective mass
m^*	Relative effective mass
$\mathcal{N}$	Normal product
N	Total number of particles, total number of unit cells
N_A	Concentration of ionized acceptors
N_D	Concentration of ionized donors
$\mathcal{N}$	Number-of-particles operator
N_e	Number of electrons
N_h	Number of holes
$N_{\boldsymbol{q}\ell}$	Number of phonons of mode $(\boldsymbol{q}\ell)$
N_{op}	Number of optical phonons
NEGF	Non-equilibrium Green functions
NDC	Negative differential conductivity
NDM	Negative differential mobility
n	Density of particles, number of particles, occupation number
n_e	Electron density
n_h	Hole density
n_I	Ionized-impurity density
$n_{\boldsymbol{q}}$	Phonon occupation number
n , $\mathcal{N}$	Number operators, density operator
OPW	Orthogonalized plane wave
P	Probability
$P(\boldsymbol{k}), P(\epsilon)$	Total scattering rate from state $\boldsymbol{k}$ with energy ϵ
$P(\boldsymbol{k},\boldsymbol{k}')$	Scattering rate from state $\boldsymbol{k}$ to state $\boldsymbol{k}'$
$P^{(d)}(\boldsymbol{k},\boldsymbol{k}')$	Scattering rate for deformation-potential phonons
$P_a^{(d)}(\boldsymbol{k},\boldsymbol{k}')$	Scattering rate for deformation-potential acoustic phonons
$P_{ae}^{(d)}(\boldsymbol{k},\boldsymbol{k}')$	Scattering rate for deformation-potential acoustic phonons in elastic approximation
$P_i(\boldsymbol{k},\boldsymbol{k}')$	Scattering rate for ionized-impurity scattering
$P_i^{(BH)}(\epsilon)$	Integrated scattering rate for ionized-impurity scattering in BH approach
$P_i^{(CW)}(\epsilon)$	Integrated scattering rate for ionized-impurity scattering in CW approach
$P_{ae}^{(d)}(\epsilon)$	Integrated scattering rate for deformation-potential acoustic phonons in elastic approximation
$\boldsymbol{P}$	Total linear momentum, polarization field
$\mathcal{P}$	Polarization operator
p	Canonical momentum, piezoelectric constant, probability
p	Momentum operator

$\boldsymbol{p}$	Linear momentum
$\boldsymbol{p}^*$	Momentum path variable
Q	Generalized force, heat
QBE	Quantum Boltzmann equation
QD	Quantum dot
QHE	Quantum Hall effect
QW	Quantum well
QWR	Quantum wire
q	Lagrangian or Hamiltonian coordinate, particle charge
$q_\circ$	Inverse screening length
$\boldsymbol{q}$	Vector in reciprocal space, phonon wavevector
q	Coordinate observable
$\boldsymbol{R}$	Direct lattice vector
R	Hall constant, reflection coefficient, resistance
RTD	Resonant tunnel diode
r	Recombination rate, random number between 0 and 1
$\boldsymbol{r}$	Position
$\boldsymbol{r}^*$	Position path variable
r	Position operator
r_H	Hall factor
S	Entropy
$S(\omega)$	Surface of constant frequency in reciprocal space
$\mathcal{S}, s$	Scattering matrix
sc	Simple cubic
$\mathcal{T}$	Time-ordering operator
$\mathcal{T}_c$	Contour time-ordering operator
T	Kinetic energy, absolute temperature, tension, transmission coefficient
T_e	Electron temperature
T_n	Noise temperature
T_{ij}	Herring and Vogt transformation matrix
$\boldsymbol{T}$	Torque, crystal translation, symmetry translation, vector of direct lattice
TA	Transverse acoustic
TO	Transverse optical
T_{op}	Equivalent temperature of optical phonons
t	Time
t^*	Time in path-variable formulation
$u_{n\boldsymbol{k}}(\boldsymbol{r})$	Periodic part of Bloch wavefunction of band n
U	Generalized potential
$\mathcal{U}(t, t_\circ)$	Evolution operator
$\mathcal{V}$	Potential energy operator
$\mathcal{V}_\mathrm{eff}$	Effective potential energy operator in pseudo-potential theory
$\mathcal{V}_p$	Pseudo-potential energy
V	Potential energy, electric potential, crystal volume

V_A,V_B,V_v	Potential of atomic species A, of atomic species B, and of virtual-crystal atom
V_b	Built-in potential
V_c	Volume of crystal unit cell
$V_i(\boldsymbol{r})$	Potential of an ionized impurity
$V_{cr}(\boldsymbol{r})$	Periodic potential energy of an electron in a crystal
$V(\boldsymbol{G})$	Fourier coefficient of the expansion of $V_{cr}(\boldsymbol{r})$
V_H	Hall potential
$\mathcal{V}_w$	Scattering potential in the Wigner-function formulation
$\boldsymbol{v}$	Velocity
v_d	Drift velocity
v_g	Group velocity
v_l	Sound velocity for longitudinal modes
v_s	Sound velocity
v_t	Sound velocity for transverse modes
v_φ	Phase velocity
VLSI	Very-large-scale integration
W	Work
$W(\boldsymbol{k},\boldsymbol{k}')$	Transition frequency
w_p, w_n	Widths of space-charge regions
WF	Wigner function
WS	Wannier-Stark
x	Position coordinate
Y	String displacement
$\boldsymbol{y}(\boldsymbol{r})$	Atom displacement field
Z	Partition function, number of charges on impurities
Z_f	Number of equivalent final valleys of an intervalley transition
α	Normal coordinate of lattice vibrations, nonparabolicity parameter
β	$= 1/(K_\mathrm{B}T)$, warm-electron coefficient
Γ	Total scattering rate including self-scattering, imaginary part of self-energy
$\Delta\epsilon_g$	Band gap
δ_{ij}	Kronecker delta
$\delta(x)$	Dirac delta function
ϵ	Energy
ϵ_c	Bottom energy of conduction band
ϵ_e	Electron energy
ϵ_h	Hole energy
ϵ_v	Top energy of valence band
$\epsilon_\circ$	Energy of an electron with wavevector equal to the inverse screening length
ϵ_{so}	Split-off energy of valence bands
ϵ_L, ϵ_Γ, ϵ_Δ	Bottom energies of valleys at L, Γ, and Δ, respectively
$\epsilon_n(\boldsymbol{k})$	n-th energy band

ϵ_F	Fermi level
ε	Dielectric constant or electric permittivity
$\varepsilon(0)$	Dielectric constant at low frequency
$\varepsilon(\infty)$	Dielectric constant at high frequency
$\varepsilon_\circ$	Vacuum permittivity or electric constant
ε_r	Relative dielectric constant
$\zeta(z)$	Orthogonal wavefunction
η	Amplitude of chain oscillations
Λ	Spectral width, function generating a gauge transformation
λ	Wavelength, total scattering rate
μ	Magnetic permeability, electrochemical potential, carrier mobility
$\mu_\circ$	Vacuum permeability or magnetic constant
μ_e	Electron electrochemical potential
μ_d	Drift mobility
μ_H	Hall mobility
μ_h	Hole electrochemical potential
μ'	Differential mobility
ν	Frequency
Ξ_d, Ξ_u	Deformation potential constants in ellipsoidal valleys
ξ	Amplitude of string and chain oscillations
ρ	Density, charge density, density of points in phase space, string linear density, density matrix, resistance in units of $h/2e^2$
Σ	Contour-ordered self energy
Σ^a	Advanced self energy
Σ^r	Retarded self energy
$\Sigma^>$	Greater self energy
$\Sigma^<$	Less self energy
Σ^t	Time-ordered self energy
$\Sigma^{\bar{t}}$	Anti-time-ordered self energy
σ	Electrical conductivity, scattering cross section
τ	Momentum relaxation time
τ_ϵ	Energy relaxation time
$\tilde{\tau}_v$	Characteristic time of velocity relaxation by collisions
Φ	Phonon field operator
$\Phi(p,t)$	Wavefunction in p-representation
$\phi(\boldsymbol{r})$	Scalar electromagnetic potential
$\varphi(\boldsymbol{k})$	Coefficient of the scalar electromagnetic potential in Fourier series
χ_e	Electric susceptibility
χ_m	Magnetic susceptibility
$\|\Psi\rangle$	State vector
$\|\Psi_S(t)\rangle, \|\Psi(t)\rangle$	State vector in Schrödinger picture
$\Psi(\boldsymbol{r},t)$	Wavefunction in $\boldsymbol{r}$-representation, electron wavepacket

$\Psi_p(q)$	Eigenfunction of p in q-representation
$\Psi(\boldsymbol{r}), \Psi^\dagger(\boldsymbol{r})$	Annihilation and creation field operators
$\psi(\boldsymbol{r})$	Time-independent wavefunction
$\psi_{n\boldsymbol{k}}(\boldsymbol{r})$	Bloch wavefunction of band n with crystal wavevector $\boldsymbol{k}$
$\lvert\psi_{n\boldsymbol{k}}\rangle$	Bloch state of band n with crystal wavevector $\boldsymbol{k}$
$\lvert\tilde{\psi}_{\boldsymbol{k}}\rangle$	Pseudo Bloch state with crystal wavevector $\boldsymbol{k}$
$\lvert\Psi_H\rangle$	State vector in Heisenberg picture
Ω	Number of accessible states
ω	Angular frequency $2\pi\nu$
ω_c	Cyclotron frequency
ω_{ac}	Frequency of acoustic modes
ω_{op}	Frequency of optical modes
ω_s	Real part of self-energy

Part I

Basic Concepts in Semiconductor Physics

1

Survey of Classical Physics

In this first chapter and in the following one, the fundamentals of classical and quantum physics will be reviewed. Obviously, the purpose is not to provide an exhaustive (or even partial) treatment of these subjects: the readers are supposed to be already familiar with them. We simply intend to recall the main concepts and to define the symbols that will be used in the rest of the book. Many excellent textbooks have been written on classical and quantum physics. We may refer, for example, to Goldstein [168] and Jackson [202] for the former, and to Messiah [306], Schiff [398] or Greiner [172] for the latter.

1.1 Newton Dynamics

Linear Momentum

The fundamental law of nonrelativistic classical mechanics is *Newton second law of motion* for a particle of mass m subject to a force $\boldsymbol{F}$:

$$\boxed{\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \boldsymbol{F} = m\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}t^2}} \tag{1.1}$$

where $\boldsymbol{p}$ is the linear momentum, or simply the momentum, of the particle:

$$\boldsymbol{p} = m\boldsymbol{v}\ , \qquad \boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}.$$

Here, $\boldsymbol{r}(t)$ and $\boldsymbol{v}(t)$ are the position and velocity of the particle at time t.

If a system is composed of many particles, a total linear momentum is defined as the sum over the particles

$$\boldsymbol{P} = \sum_i \boldsymbol{p}_i,$$

where $\boldsymbol{p}_i$ is the momentum of the i-th particle. In such a case, the force acting on each particle is the sum of the forces external to the system and those due to other particles. According to *Newton third law of motion*, the forces that two particles exert on each other are equal and opposite and lie along the line joining the two particles. As a result,

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \boldsymbol{F}^{(e)},$$

where $\boldsymbol{F}^{(e)}$ is the sum of the external forces acting on all the particles of the system.

Angular Momentum

The angular momentum $\boldsymbol{L}$ of a particle with linear momentum $\boldsymbol{p}$ with respect to point $\boldsymbol{O}$ is defined as

$$\boldsymbol{L} = \boldsymbol{r} \times \boldsymbol{p},$$

where $\boldsymbol{r}$ is the vector from $\boldsymbol{O}$ to the particle position. In the same way, if a force $\boldsymbol{F}$ is applied to a particle in $\boldsymbol{r}$, the momentum $\boldsymbol{T}$ of this force (or torque) with respect to point $\boldsymbol{O}$ is defined as

$$\boldsymbol{T} = \boldsymbol{r} \times \boldsymbol{F}. \tag{1.2}$$

Observing that $\boldsymbol{v} \times \boldsymbol{p} = \boldsymbol{v} \times m\boldsymbol{v} = 0$, from Newton second law it follows immediately that

$$\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t} = \boldsymbol{T}. \tag{1.3}$$

If a system is composed of many particles, a total angular momentum is defined as the sum over the particles

$$\boldsymbol{L} = \sum_i \boldsymbol{L}_i,$$

where $\boldsymbol{L}_i$ is the angular momentum of the i-th particle. The application of Newton second and third laws yields

$$\frac{\mathrm{d}\boldsymbol{L}}{\mathrm{d}t} = \boldsymbol{T}^{(e)},$$

where $\boldsymbol{T}^{(e)}$ is the total momentum of the external forces acting on the system.

1.2 Work and Energy

The kinetic energy of a particle with mass m and velocity $\boldsymbol{v}$ is defined as

$$T = \frac{1}{2}mv^2.$$

This is a scalar quantity and should not be confused with the torque $\boldsymbol{T}$ in (1.2), which is a vector quantity.

If $\boldsymbol{F}(\boldsymbol{r})$ is the force acting on a particle in $\boldsymbol{r}$, and the particle moves from $\boldsymbol{r}_1$ to $\boldsymbol{r}_2$ following a path $\boldsymbol{s}$, the work performed by the force on the particle along $\boldsymbol{s}$ is defined as

$$W = \int_{\boldsymbol{s}} \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{r}.$$

From Newton second law, it follows immediately that the work performed over a particle produces an equal change of its kinetic energy:

$$W = T_2 - T_1.$$

If a particle is moving in a force field such that the work performed along any close trajectory is zero,

$$W = \oint \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{r} = 0,$$

then the force field is said to be conservative, and a potential-energy field $V(\boldsymbol{r})$ can be defined such that

$$\boldsymbol{F} = -\nabla V.$$

In this case, the total energy of the particle

$$\epsilon = T + V \tag{1.4}$$

is constant.

In a many-particle system, the kinetic energy is the sum of the kinetic energies of all the particles,

$$T = \sum_i T_i = \sum_i \frac{1}{2} m_i v_i^2.$$

If both the applied (external) forces that act on the particles and the forces due to particle interactions (internal) are conservative, then the total potential energy of the system is given by:

$$V = \sum_i V_i + \frac{1}{2} \sum_{i \neq j} V_{ij},$$

where V_i is the potential energy of the i-th particle due to the external forces, and V_{ij} is the potential interaction energy of the pair of particles i and j. The factor $1/2$ is inserted since each pair of particles is present twice in the sum. Energy conservation, given by (1.4) for a single particle, still holds for the many-particle system.

1.3 Hamiltonian Formulation of Dynamics

In a system composed of n particles that can move separately, even though interacting with each other, the number of coordinates necessary to describe the configuration of the system is $3n$. These quantities are not enough to indicate how the system will evolve, since the differential equations of motion are of second order with respect to time, as shown in (1.1). Thus, also the velocities of the particles must be assigned. This situation is described by saying that the state of the system is defined by the positions and velocities of all its particles.

If, however, the particle positions are subject to given constraints, as it happens, for example, in rigid bodies where the distances between all the particles are fixed, then the number of *degrees of freedom* of the system is reduced. In such a case, the configuration of the system is described by a certain number of parameters q_i, called *generalized coordinates*. The number of independent generalized coordinates necessary to describe the configuration of the system is the number of its degrees of freedom.

The positions of all particles of the system are functions of the generalized coordinates, so that the state of the system is described by the values of all the q_i and their time derivatives $\dot{q}_i$. The dynamical equations of motion in terms of such variables are known as the Lagrange equations. For a conservative system, the Lagrangian function is defined as the difference between the kinetic and the potential energy of the system:

$$L(q_i, \dot{q}_i) = T(q_i, \dot{q}_i) - V(q_i),$$

and the Lagrange equations of motion are

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0. \tag{1.5}$$

The Lagrange equations can be written also for a nonconservative system if a generalized potential function $U(q_i, \dot{q}_i, t)$ can be defined such that the forces applied to the system are given by

$$Q_i = -\frac{\partial U}{\partial q_i} + \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial U}{\partial \dot{q}_i}\right). \tag{1.6}$$

In this case, the Lagrangian function is defined as

$$L = T - U,$$

and the equations of motion are still the Lagrange equations (1.5).

The *momenta p_i, conjugate to the generalized coordinates q_i*, are defined by means of the Lagrangian function, as

$$p_i = \frac{\partial L(q_i, \dot{q}_i, t)}{\partial \dot{q}_i}. \tag{1.7}$$

The Hamiltonian function of the system is then defined as

$$H(q_i, p_i, t) = \sum_i \dot{q}_i p_i - L(q_i, \dot{q}_i, t). \tag{1.8}$$

As can be seen from the l.h.s. of the above equation, H is defined as function of the generalized coordinates q_i and their conjugate momenta p_i. Thus, in the functions on the r.h.s., $\dot{q}_i$ must be replaced with its expression in terms of the q_i and p_i obtained from (1.7).

It may be important to note that the analytical forms of the Lagrangian and Hamiltonian functions are their crucial properties in the theory, rather than their particular numerical values.

In general, the Hamiltonian of a system coincides with its energy, but this is not always necessarily true (see [168] Sect. 7-3).

At this point, we are in the position to write the Hamilton dynamical equations,

$$\boxed{\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}} \tag{1.9}$$

For purely mechanical systems, they are equivalent to Newton laws, but they can also be derived, along with Lagrange equations, from some variational principles that may be used in more general physical systems [168].

In the Hamiltonian formulation of classical physics, the state of a system with N degrees of freedom is described by the set of $2N$ values (q_i, p_i). These may be considered as the coordinates of a point representative of the state of the system in a $2N$-dimensional space called the *phase-space* of the system.

1.4 Canonical Transformations

The generalized coordinates and their conjugate momenta which describe the state of a physical system are not unique. Given a set of q_i and p_i a transformation may be considered to new variables q'_i and p'_i defined by the functions

$$q'_i = q'_i(q_j, p_j, t), \qquad p'_i = p'_i(q_j, p_j, t). \tag{1.10}$$

Such a transformation is said to be canonical, and it is of interest, if the new variables are canonical, i.e., if a function K exists such that the equations of motion in the new set are the Hamilton equations:

$$\dot{q}'_i = \frac{\partial K}{\partial p'_i}, \qquad \dot{p}'_i = -\frac{\partial K}{\partial q'_i}.$$

The function $K(q'_i, p'_i, t)$ plays the role of the Hamiltonian function in the new set of variables.

It can be shown [168] that the transformation from the values of the Hamilton coordinates at time t to the same variables at time t',

$$q'_i(t) = q'_i(q_j(t), p_j(t), t) = q_i(t'), \qquad p'_i(t) = p'_i(q_j(t), p_j(t), t) = p_i(t')$$

is a canonical transformation. In particular, this is true for the transformation that associates with the q_i and p_i at time t their initial values. Thus, the state of the system can be described by the values of the set of $q_{i\circ}$ and $p_{i\circ}$ at the initial time,

$$q_i'(t) = q_i(t_\circ) = q_{i\circ}, \qquad p_i'(t) = p_i(t_\circ) = p_{i\circ}. \tag{1.11}$$

If the value of a physical quantity A at time t is needed, the inverse transformation must be used:

$$A(q_i(t), p_i(t)) = A(q_i(q_{j\circ}, p_{j\circ}, t), p_i(q_{j\circ}, p_{j\circ}, t)). \tag{1.12}$$

Note that while in the original description the state of the system is described by time-varying canonical coordinates and the physical quantities are given functions of such coordinates, after the canonical transformation in (1.11) the state of the system is defined by a set of constant canonical coordinates, while the physical quantities (including $p_i(t)$ and $q_i(t)$), are given by functions of these coordinates that depend explicitly on time as effect of the dynamics. A similar situation exists in connection with the Schrödinger and Heisenberg pictures of quantum mechanics (see Sect. 2.2).

1.5 Small Oscillations

A system is in a stable equilibrium state when its generalized coordinates have values $q_i^{(e)}$ corresponding to a minimum of its potential energy and the kinetic energy is zero. If the system is slightly displaced from that position and then left alone, it will perform small oscillations about the equilibrium position. A set of generalized coordinates can be found, called *normal coordinates*, such that the dynamics described in terms of these coordinates, correspond to n independent harmonic oscillators, if n is the number of degrees of freedom of the system. In fact, if the potential energy is expanded around the equilibrium configuration $q_i^{(e)}$, it is given, to second order, by

$$V(q_1, \ldots, q_n) \approx V(q_1^{(e)}, \ldots, q_n^{(e)}) + \sum_i \left(\frac{\partial V}{\partial q_i}\right)_{(e)} \theta_i + \frac{1}{2} \sum_{ij} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)_{(e)} \theta_i \theta_j,$$

where

$$\theta_i = q_i - q_i^{(e)}$$

are the deviations of the coordinates from their equilibrium values. The first term in the above equation represents the value of the potential energy at the equilibrium configuration. Since V is defined with an arbitrary zero, this value can be made to vanish. The first derivatives in the second term are zero owing to the condition of minimum potential energy, so that we are left with the quadratic term

$$V = \frac{1}{2} \sum_{ij} v_{ij} \theta_i \theta_j,$$

where $v_{ij} = (\partial^2 V/\partial q_i \partial q_j)_{(e)}$. Similarly, the kinetic energy can be put in the form

$$T = \frac{1}{2}\sum_{ij} t_{ij}\dot{\theta}_i\dot{\theta}_j.$$

Both matrices v_{ij} and t_{ij} are symmetric and it can be shown [168] that with a suitable canonical transformation of the generalized coordinates they can be put simultaneously in a diagonal form. In the new *normal coordinates* η_i, the Lagrangian is given by

$$L = T - V = \frac{1}{2}\sum_i \mu_i\dot{\eta}_i^2 - \frac{1}{2}\sum_i \beta_i\eta_i^2,$$

where μ_i and β_i are the diagonal elements of the matrices v_{ij} and t_{ij} transformed into the normal coordinates. The conjugate momenta, according to (1.7), are given by

$$\pi_i = \frac{\partial L(\eta_i, \dot{\eta}_i, t)}{\partial \dot{\eta}_i} = \mu_i\dot{\eta}_i \ ,$$

and, according to (1.8), the Hamiltonian is then given by

$$H(\eta_i, \pi_i) = \sum_i \dot{\eta}_i\pi_i - L(\eta_i, \dot{\eta}_i) = \frac{1}{2}\sum_i \frac{1}{\mu_i}\pi_i^2 + \frac{1}{2}\sum_i \beta_i\eta_i^2.$$

This Hamiltonian is the sum of separate Hamiltonians for each normal coordinate and its conjugate momentum. This means that each normal coordinate follows its own dynamics, which, moreover, is the dynamics of a harmonic oscillator. In fact, Hamilton equations yield

$$\dot{\eta}_i = \frac{\partial H}{\partial \pi_i} = \frac{\pi_i}{\mu_i},$$

already known, and

$$\dot{\pi}_i = -\frac{\partial H}{\partial \eta_i} = -\beta_i\eta_i.$$

These equations are the dynamical equations of a harmonic oscillator: by substitution of the time derivative of the first into the second one, we obtain

$$\ddot{\eta}_i = -\frac{\beta_i}{\mu_i}\eta_i,$$

with solution

$$\eta_i(t) = A_i\cos(\omega_i t + \phi_i), \qquad \omega_i = \sqrt{\frac{\beta_i}{\mu_i}}.$$

Each normal coordinate evolves as an independent harmonic oscillator.

1.6 Maxwell Equations

The electric field $\boldsymbol{E}$ and the magnetic induction field $\boldsymbol{B}$ are defined through the force (*Lorentz force*) they exert on a test charge q:

$$\boxed{\boldsymbol{F} = q\,[\,\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}\,]} \tag{1.13}$$

This expression must be considered in the limit of a test charge so small that the sources of the electric and magnetic fields are not altered by its presence. Here, as in general in this book, we use the International System of Units (SI), recommended by the *Conférence Générale des Poids et Mesures* since 1960.

Sources of the electromagnetic fields are charges and currents. The dynamics of electric and magnetic fields, or electrodynamics, is described by Maxwell equations. If we assume a charge density $\rho(\boldsymbol{r}, t)$ and a current density $\boldsymbol{j}(\boldsymbol{r}, t)$ in vacuum, i.e., in otherwise empty space, Maxwell equations are

$$\boxed{\begin{array}{c} \nabla \cdot \boldsymbol{B} = 0 \\ \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \\ \varepsilon_\circ \, \nabla \cdot \boldsymbol{E} = \rho \\ \frac{1}{\mu_\circ} \, \nabla \times \boldsymbol{B} - \varepsilon_\circ \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{j} \end{array}} \tag{1.14}$$

where $\varepsilon_\circ$ and $\mu_\circ$ are the electric permittivity and the magnetic permeability of free space, respectively. If these equations are considered inside a material, ρ and $\boldsymbol{j}$ contain also charges and currents induced in the medium by the external applied fields. If a polarization field $\boldsymbol{P}$ is defined as the dipole moment per unit volume inside the medium, and a magnetization field $\boldsymbol{M}$ is defined as the magnetic moment per unit volume inside the medium, a polarization charge is generated as

$$\rho_P = -\nabla \cdot \boldsymbol{P},$$

and a magnetization current is generated as

$$\boldsymbol{j}_M = \nabla \times \boldsymbol{M}.$$

These charges and currents are added to the external ρ and $\boldsymbol{j}$ and the last two Maxwell equations in (1.14) become

$$\nabla \cdot \boldsymbol{D} = \rho, \tag{1.15}$$

$$\nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{j}, \tag{1.16}$$

where $\boldsymbol{D}$ and $\boldsymbol{H}$ are the electric induction field and the magnetic field, respectively:[1]

$$\boldsymbol{D} = \varepsilon_\circ \boldsymbol{E} + \boldsymbol{P}, \qquad \boldsymbol{H} = \frac{1}{\mu_\circ}\boldsymbol{B} - \boldsymbol{M}. \tag{1.17}$$

The polarization $\boldsymbol{P}$, the magnetization $\boldsymbol{M}$, and the current density $\boldsymbol{j}$ are induced by the applied fields. Their dependences upon the applied fields are characteristic of each material and are described by the so-called *constitutive equations*. In the simplest case of linear materials, the following equations hold:

$$\boldsymbol{P} = \chi_e \varepsilon_\circ \boldsymbol{E}, \qquad \boldsymbol{M} = \chi_m \boldsymbol{H}, \qquad \boldsymbol{j} = \sigma \boldsymbol{E}. \tag{1.18}$$

The proportionality coefficients χ_e, χ_m, and σ are called *electric susceptibility*, *magnetic susceptibility*, and *electric conductivity*, respectively. If the above equations (1.18) are used in the definition (1.17) of $\boldsymbol{D}$ and $\boldsymbol{H}$, linear relations result between $\boldsymbol{D}$ and $\boldsymbol{E}$ and between $\boldsymbol{H}$ and $\boldsymbol{B}$:

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} = \varepsilon_r \varepsilon_\circ \boldsymbol{E}, \qquad \boldsymbol{B} = \mu \boldsymbol{H} = \kappa_m \mu_\circ \boldsymbol{H}.$$

Here ε is the permittivity or dielectric constant of the material; ε_r is the relative dielectric constant; μ is the magnetic permeability, and κ_m the relative permeability. In a linear homogeneous medium, Maxwell equations can then be rewritten as

$$\nabla \cdot \boldsymbol{B} = 0, \tag{1.19}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0, \tag{1.20}$$

$$\varepsilon \, \nabla \cdot \boldsymbol{E} = \rho, \tag{1.21}$$

$$\frac{1}{\mu} \nabla \times \boldsymbol{B} - \varepsilon \, \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{j}. \tag{1.22}$$

These equations are very similar to the original "microscopic" Maxwell equations (1.14) with the electric permittivity and the magnetic permeability of free space substituted by equivalent quantities of the material.

1.7 Electromagnetic Potentials and Gauge Transformations

It is often convenient to reduce the four first-order differential Maxwell equations to two second-order equations by the introduction of the *electromagnetic potentials*. Since the divergence of the curl of any vector field is zero, the first

[1] In different systems of units, not only the electromagnetic units change, but also the equations of the present section are formally different (see, e.g., [202]).

Maxwell equation is automatically verified if we define a vector field $\boldsymbol{A}(\boldsymbol{r},t)$, called *vector potential*, such that

$$\boxed{\boldsymbol{B}(\boldsymbol{r},t)=\nabla\times\boldsymbol{A}(\boldsymbol{r},t)} \tag{1.23}$$

With this position, the second homogeneous Maxwell equation in (1.14) becomes

$$\nabla\times\left[\boldsymbol{E}+\frac{\partial\boldsymbol{A}}{\partial t}\right]=0$$

and is again automatically satisfied if we define a scalar field $\phi(\boldsymbol{r},t)$, called *scalar potential*, such that

$$\boldsymbol{E}+\frac{\partial\boldsymbol{A}}{\partial t}=-\nabla\phi(\boldsymbol{r},t),$$

since the curl of the gradient of any scalar field is zero. In terms of the electromagnetic potentials $\boldsymbol{A}$ and ϕ, the electric field is then given by

$$\boxed{\boldsymbol{E}=-\nabla\phi(\boldsymbol{r},t)\;-\;\tfrac{\partial\boldsymbol{A}}{\partial t}} \tag{1.24}$$

The electromagnetic potentials are not uniquely defined. In fact, $\boldsymbol{E}$ and $\boldsymbol{B}$ are left unchanged by the following transformations, called gauge transformations:

$$\boldsymbol{A}\to\boldsymbol{A}'=\boldsymbol{A}+\nabla\Lambda,\qquad\phi\to\phi'=\phi-\frac{\partial\Lambda}{\partial t}, \tag{1.25}$$

where Λ is an arbitrary function of $\boldsymbol{r}$ and t.

The freedom implied by the gauge transformations can be used to prescribe that the potentials satisfy the Lorentz condition

$$\nabla\cdot\boldsymbol{A}+\varepsilon\mu\frac{\partial\phi}{\partial t}=0. \tag{1.26}$$

We can still perform a gauge transformation (1.25) and preserve the Lorentz condition if we request that the function Λ verifies the condition

$$\nabla^2\Lambda-\varepsilon\mu\frac{\partial^2\Lambda}{\partial t^2}=0.$$

The electric and magnetic fields given by the electromagnetic potentials in (1.23) and (1.24) satisfy already the first two homogeneous Maxwell equations. If they are introduced in the last two Maxwell equations, they yield:

$$\boxed{\begin{array}{c}\nabla^2\phi-\varepsilon\mu\frac{\partial^2\phi}{\partial t^2}=-\frac{1}{\varepsilon}\rho\\[2ex]\nabla^2\boldsymbol{A}-\varepsilon\mu\frac{\partial^2\boldsymbol{A}}{\partial t^2}=-\mu\boldsymbol{j}\end{array}} \tag{1.27}$$

where use has been made of the Lorentz condition. These are the wave equations that in free space predict a velocity of electromagnetic waves given by

$$c = \frac{1}{\sqrt{\varepsilon_\circ \mu_\circ}}.$$

1.8 Hamiltonian of a Charged Particle in an Electromagnetic Field

A charged particle in an electromagnetic field is subject to the Lorentz force (1.13). This force depends on the particle velocity, so that, to write a Lagrangian, it is necessary to find a suitable function U such that (1.6) is satisfied. It is easy to verify that such a function is

$$U = q(\phi - \boldsymbol{A} \cdot \boldsymbol{v}).$$

The Lagrangian is then

$$L = T - U = \frac{1}{2} m v^2 - q\phi + q\boldsymbol{A} \cdot \boldsymbol{v}.$$

Following the procedure indicated in Sect. 1.3, we have the canonical momenta

$$p_i = \frac{\partial L}{\partial \dot{q}_i} = m v_i + q A_i, \tag{1.28}$$

and the corresponding Hamiltonian, from (1.8), is

$$\boxed{H = \tfrac{1}{2m} \left(\boldsymbol{p} - q\boldsymbol{A}\right)^2 + q\phi} \tag{1.29}$$

This Hamiltonian will be used to study the dynamics of a charged particle in a crystal subject to an electromagnetic field.

2

Fundamentals of Quantum Mechanics

The first half of the twentieth century has witnessed two major revolutions in our understanding of the physical world. One of them, quantum mechanics, deals with the description of physical systems and their dynamics; the other, relativity, concerns the nature and properties of space-time. Here, we will summarize the basic principles of quantum mechanics, again without any claim of completeness, and will not deal with the theory of relativity, since it will not be used in this book.

Quantum mechanics was developed to explain a number of phenomena incomprehensible in the frame of classical physics: thermal radiation, spectroscopy, atomic physics. After some attempts that lasted more than a quarter of a century, two successful formulations were independently given by Schrödinger and by Heisenberg, immediately proved to be equivalent. Soon after, Dirac provided a framework of the new theory that includes the two previous formulations as particular cases among an infinite number of possible ones. His book [120] is one of the greatest monuments of the scientific literature of all times. Here, we shall often use Dirac formalism since it is simpler and more intuitive.

The success obtained by quantum physics is extraordinary: no phenomenon has yet been found, which contradicts the predictions of quantum mechanics. Any experimental result from subatomic to solid-state physics, chemical physics or biophysics has confirmed the validity of quantum mechanics, within the reached precision.

This outstanding success has been obtained paying a high price: we had to give up the idea of describing reality with a unique mental model, independent of the measurement performed on the physical system under investigation. Furthermore, we had to give up the idea of a complete determinism: while the dynamical evolution of an *unobserved* quantum system is described by a deterministic equation, the results of measurements have an intrinsic element of casuality. This also means that the description of the dynamics of a process of measurement cannot be the same as that of the dynamics of other unobserved evolutions: a measurement is not a natural event of the same type of

any other unobserved phenomenon. This is the most difficult aspect to accept of quantum physics from the epistemological point of view, and many theoretical physicists and philosophers of science are still working hard on it (see, for example, [166, 493]).

The following sections contain a review of the basic elements of quantum mechanics, with the purpose of recalling the main concepts and establishing the symbols that will be used in the body of the text. The interested reader not familiar with quantum mechanics may find excellent textbooks on the subject, such as [306, 398], and many others.

2.1 The First Postulates

One of the major novelties of quantum physics with respect to classical physics is that two possible states of a system can be linearly combined to yield another possible state of the system. This requirement is suggested by the experimental evidence of interference phenomena, where a single particle seems to follow different "trajectories" at the same time. Vectors are mathematical objects that can be combined linearly to give other vectors, so that abstract vectors appear to be the most natural mathematical objects to use to represent the states of a physical system. This justifies the first part of the following first postulate of quantum physics.

Postulate 1: Mathematical Description of Physical Systems

A physical system is associated with a vector space in which vectors represent the states of the system, with the specification that proportional vectors represent the same state, and observables represent dynamical variables.

Let us recall that the vector space of quantum mechanics is defined in the domain of complex numbers and that an observable is by definition a linear Hermitian operator with a complete set of eigenstates. The basic elements of the theory of vector spaces and the Dirac formalism used in this book are given in Appendix A. Since vectors that differ by a multiplicative constant represent the same physical state of the system, it is useful to consider state vectors $|\Psi\rangle$ normalized to unity:

$$\langle\Psi|\Psi\rangle = 1.$$

This condition leaves the arbitrariness of a phase factor $e^{i\theta}$ with θ real. Once this phase factor is chosen at the set up of a theoretical elaboration, it will not affect measurable results if kept consistently.

The second part of the above postulate, i.e., the use of linear operators to represent dynamical variables, is more difficult to justify intuitively. We may say that dynamical variables have to do with changes of the state of the

system,[1] as operators do with vectors, but probably the only fair statement is that the above postulate is justified by the consequences derived by it, together with the other postulates of the theory, in perfect agreement with all known experimental results.

It seems reasonable to assume that in certain states a given physical quantity may have a unique, well-defined, value. However, since the state of a system may be a superposition of states with different values for that physical quantity, we cannot say that in general the measurement of a dynamical variable will lead to a unique, well predictable, result. This leads to the second postulate of quantum mechanics, which introduces the expectation value of such a measurement.

Postulate 2: Of the Mean Value

The measurement of a dynamical variable represented by the observable $\mathcal{A}$ in a state represented by the vector $|\Psi\rangle$ does not lead, in general, to an univocally predictable result. The expectation value of such a measurement is given by

$$\langle \mathcal{A} \rangle_\Psi = \frac{\langle \Psi | \mathcal{A} | \Psi \rangle}{\langle \Psi | \Psi \rangle}.$$

From this postulate, two consequences can be derived, both extremely important.

1. A measurement of a dynamical variable $\mathcal{A}$ will certainly give as result a number a if, and only if, the state $|\Psi\rangle$ on which the measurement is performed is an eigenstate of the observable $\mathcal{A}$ belonging to the eigenvalue a. (See Appendix A; note that the eigenvalues of a Hermitian operator are real.)
2. The possible results of a measurement of a dynamical variable $\mathcal{A}$ are its eigenvalues; the probability that a given eigenvalue will be the result of a measurement of $\mathcal{A}$ on a state $|\Psi\rangle$ is given by the sum, over the degenerate eigenstates belonging to that eigenvalue, of the square moduli of the coefficients of the expansion of $|\Psi\rangle$ over the ortho-normalized eigenstates of $\mathcal{A}$.

The statement of this theorem is much more cumbersome than its content. Let us try to illustrate it in a simpler way. We are going to perform a measurement of the dynamical variable represented by the observable $\mathcal{A}$ on the state of the system represented by the vector $|\Psi\rangle$. Being an observable, $\mathcal{A}$ has a complete set of eigenstates orthogonal to each other, and they can be normalized to unity. Let us call a_i the eigenvalues and $|\varphi_i^{(r)}\rangle$ the eigenvectors, where r specifies the eigenvectors possibly belonging to the same degenerate eigenvalue. In formulas,

$$\mathcal{A}|\varphi_i^{(r)}\rangle = a_i |\varphi_i^{(r)}\rangle, \qquad \langle \varphi_i^{(r)} | \varphi_j^{(s)} \rangle = \delta_{ij}\delta_{rs}.$$

[1] For example, linear momentum has to do with translations of the system, and angular momentum with its rotations.

Since the eigenvectors $|\varphi_i^{(r)}\rangle$ form a complete basis, we may expand the state of interest as

$$|\Psi\rangle = \sum_{i,r} C_{i,r}|\varphi_i^{(r)}\rangle. \tag{2.1}$$

The theorem above states that the result of the measurement is one of the values a_i, and the probability of occurrence of each a_i is

$$P(a_i) = \sum_r |C_{i,r}|^2.$$

If the state $|\Psi\rangle$ is an eigenstate of $\mathcal{A}$, only the corresponding eigenvalue can be the result of the measurement so that the first theorem above is a particular case of the second one. Owing to its importance, however, it deserves a statement in itself.

Postulate 3: Contraction of the State at the Measurement

When we measure a dynamical variable $\mathcal{A}$, the disturbance involved in the act of measurement causes the system to collapse into the projection of the state onto the subspace of the eigenvalue obtained as result of the measurement.

In the case of the example in (2.1), if the result of the measurement is the eigenvalue a_i, the state of the system after the measurement is

$$|\Psi'\rangle = \sum_r C_{i,r}|\varphi_i^{(r)}\rangle.$$

For the new state vector to be normalized to unity, the coefficients $C_{i,r}$ must of course be renormalized.

The collapse of the state at the measurement process is the most debated assumption of quantum mechanics. Until today, however, it has been proved to be consistent with all experimental findings.

2.2 Equations of Motion

2.2.1 Pictures and Representations

We know that unitary transformations leave eigenvalues, linear combinations, and scalar products of vectors unaltered (see Appendix A). Thus, if a set of state vectors and observables are used to describe states and dynamical variables of a physical system, we may equally well use a different set obtained from the first one by means of a unitary transformation. A choice of vectors and observables describing states and dynamical variables is called a *picture*. By means of a unitary transformation, we move from one picture to another.

Two types of pictures are most often used in quantum mechanics. In the *Schrödinger pictures*, the time variation of the system due to the dynamics is assigned to the state vectors, while dynamical variables not depending explicitly upon time are described by constant observables. In the *Heisenberg pictures*, on the contrary, state vectors are assumed to be constant, and the dynamical evolution of the system is assigned to the observables:

$$\begin{aligned} &\text{S. picture}: \ |\Psi_S(t)\rangle \ \text{ for states}, \quad \mathcal{A}_S \quad\ \ \text{for dyn. variables},\\ &\text{H. picture}: \ |\Psi_H\rangle \ \ \text{for states}, \quad \mathcal{A}_H(t) \ \text{for dyn. variables}. \end{aligned}$$

We also know (see Appendix A) that vectors and linear operators in a vector space can be specified by their components and matrices with respect to a given set of basis vectors. Thus, once the picture is chosen, we still have the choice of the basis to represent vectors and observables with numbers. The choice of the basis is called a *representation*.

2.2.2 Evolution Operator and Its Equation of Motion

Let us work, for the time being, in a Schrödinger picture, and let $|\Psi_S(t)\rangle$ be the state vector of the system at time t. We then define the *evolution operator* as the operator $\mathcal{U}(t, t_\circ)$, which yields the state vector at time t when applied to the state vector at time $t_\circ$:

$$\boxed{|\Psi_S(t)\rangle = \mathcal{U}(t, t_\circ)|\Psi_S(t_\circ)\rangle} \tag{2.2}$$

This operator must be linear, to preserve the superpositions of states, and must be unitary, to preserve the normalization of the state vectors.

The equation of motion for the evolution operator is the dynamical postulate of quantum mechanics:

Postulate 4: Equation of Motion

The evolution operator verifies the following differential equation and initial condition:

$$\boxed{\mathrm{i}\hbar \tfrac{\partial}{\partial t}\mathcal{U}(t, t_\circ) = \mathcal{H}_S(t)\ \mathcal{U}(t, t_\circ), \qquad \mathcal{U}(t_\circ, t_\circ) = 1} \tag{2.3}$$

If the Hamiltonian does not depend on time, the solution of (2.3) is

$$\boxed{\mathcal{U}(t, t_\circ) = \mathrm{e}^{-\frac{\mathrm{i}\mathcal{H}}{\hbar}(t-t_\circ)}} \tag{2.4}$$

If the initial state is an eigenstate $|\Psi_\epsilon(t_\circ)\rangle$ of the Hamiltonian belonging to the eigenvalue ϵ, (2.4) yields

$$|\Psi_\epsilon(t)\rangle = \mathrm{e}^{-\frac{\mathrm{i}\epsilon}{\hbar}(t-t_\circ)}|\Psi_\epsilon(t_\circ)\rangle,$$

which gives the well-known relation between energy and frequency:

$$\boxed{\epsilon = \hbar\omega} \tag{2.5}$$

2.2.3 Equation of Motion in Schrödinger and Heisenberg Pictures

If (2.3) is applied to the state vector at time $t_\circ$, the equation of motion for the vector states in a Schrödinger picture is obtained:

$$\boxed{\mathrm{i}\hbar\frac{\partial}{\partial t}|\Psi_S(t)\rangle = \mathcal{H}_S(t)\ |\Psi_S(t)\rangle} \tag{2.6}$$

From the Schrödinger picture, we move to the Heisenberg picture through the unitary transformation $\mathcal{U}^\dagger(t,t_\circ)$. State vectors and observables are given by

$$|\Psi_H\rangle = \mathcal{U}^\dagger(t,t_\circ)|\Psi_S(t)\rangle = |\Psi_S(t_\circ)\rangle, \qquad \mathcal{A}_H(t) = \mathcal{U}^\dagger(t,t_\circ)\ \mathcal{A}_S\ \mathcal{U}(t,t_\circ).$$

In the Heisenberg picture, the equation of motion is an equation for the dynamical variables. It is called the Heisenberg equation:

$$\boxed{\mathrm{i}\hbar\frac{d}{dt}\mathcal{A}_H(t) = [\mathcal{A}_H,\mathcal{H}_H] + \mathrm{i}\hbar\frac{\partial\mathcal{A}_H}{\partial t}} \tag{2.7}$$

The commutator accounts for the time variation due to the dynamics. The last term takes into account the possibility that the dynamical variable $\mathcal{A}$ is defined with an explicit dependence on time, so that also in the Schrödinger picture it would be time dependent. The last term in the above equation is then the transformed of the time derivative of $\mathcal{A}_S$ into the Heisenberg picture.

2.2.4 Interaction Picture

Often, the total Hamiltonian of the system of interest contains two parts:

$$\mathcal{H} = \mathcal{H}_\circ + \mathcal{H}',$$

where $\mathcal{H}'$ is a perturbation applied to an "unperturbed", time independent Hamiltonian $\mathcal{H}_\circ$. In such a case, it may be convenient to use a picture, called *interaction picture*, obtained from the Schrödinger picture by means of the unitary transformation $\mathcal{U}_\circ^\dagger(t,t_\circ)$, where $\mathcal{U}_\circ(t,t_\circ)$ is the evolution operator relative to the unperturbed Hamiltonian $\mathcal{H}_\circ$:

$$|\Psi_I(t)\rangle = \mathcal{U}_\circ^\dagger(t,t_\circ)|\Psi_S(t)\rangle, \qquad \mathcal{A}_I(t) = \mathcal{U}_\circ^\dagger(t,t_\circ)\ \mathcal{A}_S\ \mathcal{U}_\circ(t,t_\circ). \tag{2.8}$$

Both observables and state vectors are time dependent in the interaction picture. They carry the dynamical effect of the unperturbed Hamiltonian and of the perturbation, respectively.

2.3 Heisenberg Uncertainty Relations

Since, according to the second postulate, a measurement of a dynamical variable $\mathcal{A}$ in a well precise state $|\Psi\rangle$ does not yield, in general, a well precise result, the distribution of the possible values has a standard deviation. If two dynamical variables are represented by two observables $\mathcal{A}$ and $\mathcal{B}$ that do not commute, then their standard deviations are connected by an uncertainty relation [306]. More precisely, if

$$[\mathcal{A}, \mathcal{B}] = \mathrm{i}\hbar\, \mathcal{C},$$

then the standard deviations ΔA and ΔB of the possible results of their measurements are subject to the condition that

$$\Delta A\, \Delta B \geq \frac{1}{2}\hbar\, |\langle \mathcal{C} \rangle| \,. \tag{2.9}$$

If, by means of a suitable experimental apparatus, we decrease the uncertainty on one of them, the procedure increases the uncertainty on the other one in such a way that the inequality (2.9) holds still true. This means that we cannot measure the two quantities at the same time with arbitrary precisions. In brief, we say that two noncommuting observables are *incompatible*.

In the common situation in which the two quantities, such as the coordinate q and its conjugate momentum p, have a commutator given simply by $\mathrm{i}\hbar$, we have

$$\boxed{[\mathsf{q}, \mathsf{p}] = \mathrm{i}\hbar, \qquad \Delta q \Delta p \geq \tfrac{1}{2}\hbar} \tag{2.10}$$

The time-energy uncertainty relation cannot be treated on the same ground since the time in this theory is a real number and cannot have nonzero commutators with any observable. Nevertheless, a very similar relation holds, which derives from the same general relation (2.9) applied to Heisenberg equation of motion (2.7). If we define the characteristic time of variation of a dynamical variable in a state $|\Psi\rangle$ as the time τ necessary for its mean value to change of a quantity of the order of its standard deviation, then

$$\tau \Delta\epsilon \geq \frac{1}{2}\hbar \tag{2.11}$$

for any dynamical variable $\mathcal{A}$, where $\Delta\epsilon$ is the standard deviation of the possible results of an energy measurement on the state $|\Psi\rangle$. Thus, we may formulate the time-energy uncertainty relation saying that *the time necessary for any variable of a system to change appreciably is related to the uncertainty on the energy of the system by the relation (2.11)*.

In particular, if the system is in an eigenstate of its Hamiltonian, no dynamical variable can change in any finite amount of time, and for this reason the eigenstates of the Hamiltonian are called *stationary states*.

2.4 How to Deal with a General Quantum-Mechanical Problem in a System with a Constant Hamiltonian

A general quantum-mechanical problem may be formulated, in the Schrödinger picture, as follows: given the initial state $|\Psi_S(t_\circ)\rangle$ of the system, which is its state at a successive time t? In particular, what is the probability that at time t a dynamical variable $\mathcal{A}$ will assume a given value?

For the solution, assuming that the Hamiltonian $\mathcal{H}$ is time independent, we must first solve the eigenvalue equation for $\mathcal{H}$, also called the *time-independent Schrödinger equation*,

$$\boxed{\mathcal{H}\ |\varphi_i\rangle = \epsilon_i |\varphi_i\rangle} \tag{2.12}$$

i.e., we have to find all eigenvalues ϵ_i (for simplicity we assume here that these are not degenerate) and eigenvectors $|\varphi_i\rangle$. Then we expand the initial state in series of such eigenvectors, assumed to be normalized to one:

$$|\Psi_S(t_\circ)\rangle = \sum_i C_i |\varphi_i\rangle, \qquad C_i = \langle\varphi_i|\Psi_S(t_\circ)\rangle.$$

We know that this is possible since, for the first postulate, the Hamiltonian is an observable. Finally, we write the state vector at time t as the same linear combination, multiplying each term by the proper frequency phase factor:

$$|\Psi_S(t)\rangle = \sum_i C_i \mathrm{e}^{-\mathrm{i}\frac{\epsilon_i}{\hbar}(t-t_\circ)} |\varphi_i\rangle. \tag{2.13}$$

We leave to the reader the simple proof that (2.13) is the solution of the Schrödinger equation (2.6) with initial condition $|\Psi_S(t_\circ)\rangle$.

As to the second part of the problem, we first solve the eigenvalue equation for the observable $\mathcal{A}$:

$$\mathcal{A}\ |a_i\rangle = a_i |a_i\rangle. \tag{2.14}$$

For the second theorem in Sect. (2.1), we know that the possible outcomes of the measurement of $\mathcal{A}$ are its eigenvalues. Finally, for the same theorem, the probability of obtaining a given eigenvalue is the squared modulus of the coefficients of the expansion of the vector state at time t in the eigenvectors of $\mathcal{A}$, namely, if the $|a_i\rangle$ are normalized,

$$P(a_i) = |\langle a_i|\Psi_S(t)\rangle|^2.$$

If the eigenvalue a_i is degenerate the sum over the eigenstates belonging to that eigenvalue must be performed.

The present section indicates how to solve, in principle, any well-posed quantum mechanical problem with constant Hamiltonian. However, the calculations necessary to apply the above procedure are almost always out of any possible analytical realization. Approximations and numerical solutions have to be performed in even the simplest practical cases.

2.5 The {q} Representation: Wave Mechanics

Hamiltonian and Observables

Let us consider a point-like particle in space subject to a potential energy $V(\boldsymbol{r})$. Its state is classically described by the position coordinate $\boldsymbol{r}$ and its conjugate momentum $\boldsymbol{p}$. To describe its dynamics in quantum mechanical terms, the rule is to write for the quantum Hamiltonian the same expression as for the classical case, for example

$$\mathcal{H} = \frac{\mathsf{p}^2}{2m} + \mathcal{V}(\mathsf{r}), \tag{2.15}$$

where r and p are now the position and momentum operators, obeying the commutation relations in (2.10), and $\mathcal{V}(\mathsf{r})$ is the potential-energy operator corresponding to the function $V(\boldsymbol{r})$.

For the sake of simplicity, let us consider a one-dimensional case. From the commutator in (2.10), many properties can be obtained. In particular, we have

$$[\mathsf{q}, \mathcal{F}(\mathsf{q},\mathsf{p})] = \mathrm{i}\hbar \frac{\partial \mathcal{F}}{\partial \mathsf{p}}, \qquad [\mathsf{p}, \mathcal{F}(\mathsf{q},\mathsf{p})] = -\mathrm{i}\hbar \frac{\partial \mathcal{F}}{\partial \mathsf{q}}, \tag{2.16}$$

where $\mathcal{F}$ is an operator function of q and p [306]. The position operator has a continuous, nondegenerate spectrum of eigenvalues, given by the entire real axis, so that the orthogonality relation is expressed by the Dirac delta function (see Appendix A):

$$\mathsf{q}|q\rangle = q|q\rangle, \qquad \langle q|q'\rangle = \delta(q - q'). \tag{2.17}$$

The completeness of the basis $|q\rangle$ is expressed by the spectral decomposition of the unity (cf. (A.14) of Appendix A)

$$\int_{-\infty}^{\infty} |q\rangle \, \mathrm{d}q \, \langle q| = 1. \tag{2.18}$$

Wavefunction and Schrödinger Equation

Given the state of the system under consideration in the Schrödinger picture $|\Psi_S(t)\rangle$, in the $\{\mathsf{q}\}$ representation, i.e., in the basis formed by the eigenvectors of q in (2.17), it is represented by the coefficients

$$\Psi(q,t) \equiv \langle q|\Psi_S(t)\rangle. \tag{2.19}$$

The function in (2.19) is called the *wavefunction* of the particle, and the $\{\mathsf{q}\}$ representation of quantum mechanics in Schrödinger picture, which uses the wavefunctions to describe the states of the systems, is the *wave mechanics* originally proposed by Schrödinger.

The effect of the basic operators q and p on the wavefunction is found by applying the operators to the vector states and then evaluating the wavefunction of the new vector. The result is [306] that the effect of the application of the operator q is simply given by the multiplication by the number q, while the application of the operator p yields the derivative with respect to q times $(-\mathrm{i}\hbar)$:

$$\mathsf{q} \;\rightarrow\; q, \qquad\qquad \mathsf{p} \;\rightarrow\; -\mathrm{i}\hbar\frac{\partial}{\partial q}. \tag{2.20}$$

At this point, it is immediate to see that if the Hamiltonian is that given in (2.15), then the Schrödinger equation (2.6) in wave mechanics is

$$\boxed{\mathrm{i}\hbar\tfrac{\partial}{\partial t}\Psi(q,t) = -\tfrac{\hbar^2}{2m}\tfrac{\partial^2}{\partial q^2}\Psi(q,t) + V(q)\Psi(q,t)} \tag{2.21}$$

If an electromagnetic field is present in the system under consideration, the Hamiltonian is given by (1.29) of the previous chapter and is gauge dependent. Thus, also the wafunctions depend on the chosen gauge. If a gauge transformation (1.25) is performed, the wavefunctions are transformed as follows [172]:

$$\Psi(\boldsymbol{r},t) \rightarrow \Psi'(\boldsymbol{r},t) = \mathrm{e}^{\mathrm{i}q\Lambda/\hbar}\Psi(\boldsymbol{r},t). \tag{2.22}$$

It is a simple change of phase so that all physical quantities evaluated by means of the wavefunctions remain unaltered.

Eigenfunctions of the Momentum and the $\{\mathsf{p}\}$ Representation

In the $\{\mathsf{q}\}$ representation, the eigenfunctions $\Psi_p(q)$ of the momentum are easily obtained by application of the rules in (2.20). In fact, the eigenvalue equation

$$-\mathrm{i}\hbar\frac{\partial}{\partial q}\Psi_p(q) = p\,\Psi_p(q)$$

has the immediate solution given by the *plane wave*

$$\Psi_p(q) = \frac{1}{\sqrt{2\pi\hbar}}\mathrm{e}^{\mathrm{i}\frac{p}{\hbar}q}. \tag{2.23}$$

The spectrum of the eigenvalues is again the whole real axis, and the above plane waves are ortho-normalized to the delta function, as shown by (A.27) of Appendix A.

If we now want to move from the $\{\mathsf{q}\}$ representation to the $\{\mathsf{p}\}$ representation, we have to perform a change of basis: by inserting (2.18), we obtain

$$\Phi(p,t) = \langle p|\Psi_S(t)\rangle = \int_{-\infty}^{\infty} \langle p|q\rangle \mathrm{d}q \langle q|\Psi_S(t)\rangle.$$

The last scalar product in the integral is the wavefunction (2.19), and the first one is the complex conjugate of $\langle q|p\rangle$, i.e., of the wavefunction (2.23) of the eigenstate of the momentum. Thus,

$$\Phi(p,t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\frac{p}{\hbar}q} \Psi(q,t)\, \mathrm{d}q.$$

In a similar way, using the completeness of the $|p\rangle$ basis, we obtain

$$\Psi(q,t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\frac{p}{\hbar}q} \Phi(q,t)\, \mathrm{d}p.$$

The last two expressions can be recognized as the Fourier transform and the Fourier integral of the wavefunction, respectively.

2.6 Identical Particles and Pauli Exclusion Principle

In the previous pages, we have considered the wave mechanics of a system with only one degree of freedom. In case of a system formed by a particle in 3-dimensional space, very few changes are needed: instead of the generic coordinate q, we may consider the position vector $\boldsymbol{r}$. The commutation relation in (2.10) holds for each coordinate, and the other commutators are zero.

More often, however, we have to deal with more than one particle. The wavefunction becomes

$$\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, t).$$

If the particles are distinguishable, the above wave function has the following interpretation:

$$|\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, t)|^2 \, \mathrm{d}\boldsymbol{r}_1 \, \mathrm{d}\boldsymbol{r}_2 \, \ldots$$

is the probability of finding particle 1 in $\mathrm{d}\boldsymbol{r}_1$ around $\boldsymbol{r}_1$, particle 2 in $\mathrm{d}\boldsymbol{r}_2$ around $\boldsymbol{r}_2$, and so on, so that the normalization is

$$\int\int \ldots \int |\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, t)|^2 \, \mathrm{d}\boldsymbol{r}_1 \, \mathrm{d}\boldsymbol{r}_2 \, \ldots = 1.$$

If the particles are identical, the very fact that they are indistinguishable has important consequences in quantum mechanics. Since a particle has no definite trajectory, when we find a particle of a many-identical-particle system in a position $\boldsymbol{r}$ we cannot say which particle of the system we have detected. This is not due to our ignorance, but to the fact that during the evolution the identical particles have no identity. In this case if we invert two position arguments in the many-particle weavefunction, the probability density must remain unaltered:

$$|\Psi(\boldsymbol{r}_1, \ldots \boldsymbol{r}_i, \ldots \boldsymbol{r}_j, \ldots, t)|^2 = |\Psi(\boldsymbol{r}_1, \ldots \boldsymbol{r}_j, \ldots \boldsymbol{r}_i, \ldots, t)|^2.$$

More precisely, if the particles have an integer spin (including zero) they are called *bosons*, and the above relation is guaranteed by the symmetry of the wavefunction:

$$\Psi(\boldsymbol{r}_1, \ldots \boldsymbol{r}_i, \ldots \boldsymbol{r}_j, \ldots, t) = \Psi(\boldsymbol{r}_1, \ldots \boldsymbol{r}_j, \ldots \boldsymbol{r}_i, \ldots, t).$$

If, instead, the particles have half-integer spins, as in the case of electrons that have spin 1/2, they are called *fermions*, and the above relation is guaranteed by the antisymmetry of the wavefunction:

$$\Psi(\boldsymbol{r}_1, \ldots \boldsymbol{r}_i, \ldots \boldsymbol{r}_j, \ldots, t) = -\Psi(\boldsymbol{r}_1, \ldots \boldsymbol{r}_j, \ldots \boldsymbol{r}_i, \ldots, t).$$

The above-cited symmetry properties have important consequences in the statistical physics of systems composed of identical particles. Consider for simplicity the case of only two particles and expand the two-particle wavefunction in the linear combination of products of wavefunctions $\varphi_i(\boldsymbol{r})$ of a single-particle orthonormal basis:

$$\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2) = \sum_{ij} C_{ij} \varphi_i(\boldsymbol{r}_1) \varphi_j(\boldsymbol{r}_2).$$

In the case of fermions, the antisymmetry requires that

$$C_{ij} = -C_{ji},$$

and, in particular, $C_{ii} = 0$. This means that two identical fermions cannot occupy the same single-particle state. This is the well-known *Pauli exclusion principle*. Its importance cannot be overestimated: it is necessary to explain the atomic structure, the periodic table of elements, and the chemical bond. Thus, the whole structure of matter, including that of biological systems, depends critically on it.

Several applications of the principles of quantum theory of particular interest for the subject treated in this book are reported in the appendices: potential steps, barriers, and wells are treated in Appendix B, the harmonic oscillator is presented in C, quantization in a magnetic field in D, and perturbation theory in E.

3

Fundamentals of Statistical Physics

3.1 Introduction

When we know completely the initial condition of a physical system, quantum mechanics may predict exactly its future (apart from the partial unpredictability of the results of a measurement process). If the system is only partially known, however, we are not forced to give up completely the idea of making predictions on its future behavior. *Statistical physics* is the branch of physics that deals with systems only partially known. The lack of maximum knowledge of the initial condition of the system is often due to the extremely large number of degrees of freedom, as in the case of gases, but this is not necessarily always true, and statistical physics works equally well for simple systems when, for any reason, our knowledge is less than complete. Following the classical text of Tolman [448] we may say that *the general nature of the statistical mechanical procedure for the treatment of complicated systems consists in abandoning the attempt to follow the precise changes in state that would take place in a particular system, and in studying instead the behavior of a collection or* ensemble of systems *of similar structure to the system of actual interest, distributed over a range of different precise states. From a knowledge of the* average behavior *of the systems in a* representative ensemble, *appropriately chosen so as to correspond to the partial knowledge that we do have as to the initial state of the system of interest, we can then make predictions as to what may be expected on the average for the particular system which concerns us.*

The fundamental theoretical tool of statistical physics is thus the *statistical ensemble*, a "mental collection" of an arbitrary large number of systems, all prepared in the same way as the actual system of interest. Only the *accessible states*, i.e., the states compatible with the partial knowledge we have of the system, will be represented in the ensemble. More precisely, taking at random a system in the ensemble, the probability of finding it in a given state is equal to the probability of finding the actual system in that given state. This statement may be considered the definition of the statistical ensemble. How

to actually work with it and how to use it for predicting the average behavior of the system of interest is the subject of statistical physics. In the following sections, the main principles and results of this theory will be briefly reviewed. Once again, the interested reader is referred to the standard textbooks, such as [370, 448], or [194], for a more complete treatment of the subject and, in particular, for the equivalence of the physical quantities defined here in the frame of statistical physics, such as temperature and entropy, with the same quantities as defined in thermodynamics.

3.2 Liouville Theorem

Let us consider a classical system whose state is defined by the Hamiltonian coordinates (q_i, p_i), and consider an ensemble of such systems. As a mental collection, a statistical ensemble contains a number of systems as large as necessary, without problems. Let $\rho(q_i, p_i, t)$ be the density of points in phase space representative of the states of the systems in the ensemble. Its total time derivative, describing the rate of change of ρ as seen by any given point along its motion, is given by

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \sum_i \left(\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right) + \frac{\partial \rho}{\partial t} \ . \tag{3.1}$$

Using Hamilton equations, the above may be written as

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \sum_i \left(\frac{\partial \rho}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial \rho}{\partial p_i} \frac{\partial H}{\partial q_i} \right) + \frac{\partial \rho}{\partial t} \ . \tag{3.2}$$

The Poisson bracket of two functions u and v of phase space is defined as [168]

$$[u, v]_{q,p} = \sum_i \left(\frac{\partial u}{\partial q_i} \frac{\partial v}{\partial p_i} - \frac{\partial u}{\partial p_i} \frac{\partial v}{\partial q_i} \right) \ .$$

Thus, (3.2) may be written as

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = [\rho, H]_{q,p} + \frac{\partial \rho}{\partial t} \ ,$$

whose similarity with Heisenberg equation (2.7) is evident. The quantum analog of the above equation will be seen in Sect. 16.2.

Let us now consider the representative points contained in a region V of phase space and follow their motion for a time δt . Their orbits cannot cross because given initial conditions determine univocally the system orbit. Thus, after δt all and only the points initially in V will be in a region V' as indicated in Fig. 3.1.

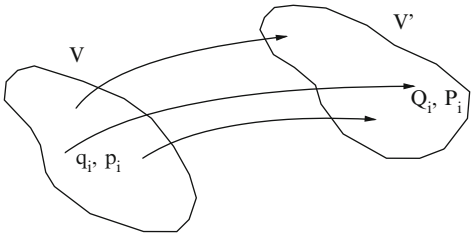


Fig. 3.1. For Liouville theorem - Trajectories of points in phase space

Let us consider the coordinate transformation from the set of initial values (q_i, p_i) to the final values (Q_i, P_i). Using Hamilton equations (1.9), these transformations are written as

$$\begin{cases} Q_i = Q_i(q_i, p_i) = q_i(t+\delta t) = q_i(t) + \dot{q}_i \delta t = q_i(t) + \dfrac{\partial H}{\partial p_i}\delta t \\ \\ P_i = P_i(q_i, p_i) = p_i(t+\delta t) = p_i(t) + \dot{p}_i \delta t = p_i(t) - \dfrac{\partial H}{\partial q_i}\delta t \end{cases} \tag{3.3}$$

The evaluation of the volume of V' requires the Jacobian of the transformation, given by

$$J(Q_i, P_i; q_i, p_i) = \frac{\partial(Q_i, P_i)}{\partial(q_i, p_i)} = \begin{vmatrix} \frac{\partial Q_1}{\partial q_1} & \frac{\partial Q_1}{\partial q_2} & \dots & \frac{\partial Q_1}{\partial p_1} & \dots \\ \frac{\partial Q_2}{\partial q_1} & \frac{\partial Q_2}{\partial q_2} & \dots & \frac{\partial Q_2}{\partial p_1} & \dots \\ \dots & & & & \\ \frac{\partial P_1}{\partial q_1} & \dots & \dots & \frac{\partial P_1}{\partial p_1} & \dots \\ \dots & & & & \end{vmatrix}$$

$$= \begin{vmatrix} 1 + \frac{\partial}{\partial q_1}\frac{\partial H}{\partial p_1}\delta t & \frac{\partial}{\partial q_2}\frac{\partial H}{\partial p_1}\delta t & \dots & \frac{\partial}{\partial p_1}\frac{\partial H}{\partial p_1}\delta t & \dots \\ \frac{\partial}{\partial q_1}\frac{\partial H}{\partial p_2}\delta t & 1 + \frac{\partial}{\partial q_2}\frac{\partial H}{\partial p_2}\delta t & \dots & \frac{\partial}{\partial p_1}\frac{\partial H}{\partial p_2}\delta t & \dots \\ \dots & & & & \\ -\frac{\partial}{\partial q_1}\frac{\partial H}{\partial q_1}\delta t & \dots & \dots & 1 - \frac{\partial}{\partial p_1}\frac{\partial H}{\partial q_1}\delta t & \dots \\ \dots & & & & \end{vmatrix}$$

Let us consider all the products whose sum gives the value of the determinant. The diagonal terms yield the product

$$\left(1 + \frac{\partial}{\partial q_1}\frac{\partial H}{\partial p_1}\delta t\right)\left(1 + \frac{\partial}{\partial q_2}\frac{\partial H}{\partial p_2}\delta t\right) \dots \left(1 - \frac{\partial}{\partial p_1}\frac{\partial H}{\partial q_1}\delta t\right) \dots \tag{3.4}$$

The only term of zero order in δt in the Jacobian comes from this product, when all terms equal to one are considered, and it is equal to unity. The terms of first order coming from the product in (3.4) are opposite in pairs, because of the equality of mixed derivatives. Thus, the terms of first order

from this product give zero contribution. On the contrary, terms of first order could come only from the product in (3.4). In fact if all terms but one are taken in the diagonal, the only possibility is to take also the last one of the diagonal for the rules of construction of a determinant. Thus, if a term is taken off diagonal, at least another term off diagonal must be present in the product. Since, however, the off diagonal terms are all proportional to δt, these products are of second order. The conclusion is that to first order in δt the Jacobian is equal to unity, i.e., the derivative of the Jacobian with respect to time is zero. This implies that the volume of V' is constant, equal to the original value of V. If the arbitrary volume V is constant and the number of representative points contained in it is constant, then the density around each point is constant, and this is the content of *Liouville theorem*. The total time derivative in (3.1) of the density of points in phase space is zero:

$$\boxed{\frac{\partial \rho}{\partial t} + \sum_i \left[\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right] = 0 \qquad \text{or} \qquad \frac{\partial \rho}{\partial t} = [H, \rho]_{q,p}} \tag{3.5}$$

If ρ is a function of (q_i, p_i) only through the Hamiltonian, i.e., through the energy, than the sum in the above equation vanishes:

$$\sum_i \left(\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right) = \sum_i \left(\frac{\partial \rho}{\partial H} \frac{\partial H}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial \rho}{\partial H} \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q_i} \right) = 0.$$

The Liouville theorem then says that

$$\frac{\partial \rho}{\partial t} = 0.$$

In other words, if ρ is function of energy alone, it describes a situation of equilibrium since it is constant in each point of phase space.

3.3 The Fundamental Hypotheses of Statistical Mechanics

When we have a partial knowledge of the system of interest, some points of phase space will represent states compatible with the partial knowledge we have of the system, others will not.

The fundamental hypothesis of classical statistical physics is the *equal a priori probabilities in phase space*. According to this hypothesis *in the statistical ensemble of an isolated system all points in the accessible region of phase space are represented with equal probabilities.* In quantum statistical mechanics, this postulate is complemented with the hypothesis that *all phases of the states in the ensemble are present with equal probabilities.*

The above postulate is coherent with an important result of analytical mechanics: if in a region of phase space the density of points of an ensemble is uniform, it remains uniform after a canonical transformation, so that the postulate does not depend on the choice of the Hamiltonian coordinates.

3.4 Main Definitions and Results of Statistical Mechanics

Entropy

If the states accessible to an isolated system form a discrete set, the statistical definition of entropy is simply

$$\boxed{S \equiv K_{\mathrm{B}} \ln(\Omega(\epsilon))} \tag{3.6}$$

where K_{B} is Boltzmann constant and $\Omega(\epsilon)$ the total number of states accessible to the system with energy ϵ. If, as it happens in classical mechanics, the states form a continuum, Ω must be taken proportional to the accessible volume in phase space. The particular value of the entropy depends on the proportionality constant assumed, but the variations of entropy do not. The uncertainty relations of quantum mechanics (see Sect. 2.3) suggest to take this constant as equal to the inverse of the Plank constant to the power of the number of degrees of freedom, since the values of q_i and p_i cannot be simultaneously specified better than $\delta q_i \delta p_i \sim \hbar$. With this clarification, in the following we will call Ω the "number of states" also in classical statistical mechanics.

Temperature and Thermal Equilibrium

For an isolated system in equilibrium, with energy ϵ, the temperature is defined by the relation

$$\boxed{\beta \equiv \frac{1}{K_{\mathrm{B}} T} \equiv \frac{\partial}{\partial \epsilon} \ln(\Omega(\epsilon)) = \frac{1}{K_{\mathrm{B}}} \frac{\partial S}{\partial \epsilon}} \tag{3.7}$$

This definition requires a continuous variation of $\Omega(\epsilon)$ upon energy. If this is not so, we have to resort to the temperature of the thermal bath which the system has been put in contact with, as described below.

If an infinitesimal amount of energy ΔQ is transferred to a system, the variation of its entropy is

$$\Delta S = K_{\mathrm{B}} \Delta(\ln \Omega) = K_{\mathrm{B}} \frac{\partial}{\partial \epsilon} (\ln \Omega) \Delta Q = \frac{\Delta Q}{T},$$

as in thermodynamic theory.

Let us consider two systems at equilibrium, with energies $\epsilon^{(1)}$ and $\epsilon^{(2)}$, respectively. They have temperatures and entropies

$$\beta^{(1)}, S^{(1)}, \quad \text{and} \quad \beta^{(2)}, S^{(2)}.$$

Now we put them in contact with each other so that they can exchange any amount of energy, maintaining the total energy constant. We also assume that the interaction energy of the two subsystems is negligible, so that the exchange of energy is the only effect of their contact, and the states of the two subsystems are still well defined. What will it be the most probable partition of the total energy between the two subsystems? To answer this question, let us consider that the probability that the energy of the first subsystem is $\epsilon^{(1)}$ is

$$P(\epsilon^{(1)}) = \frac{\Omega(\epsilon^{(1)})}{\Omega_t},$$

where $\Omega(\epsilon^{(1)})$ is the number of accessible states of the total system with an energy $\epsilon^{(1)}$ in subsystem 1 and Ω_t is the total number of accessible states of the total system. The maximum value of this probability is obtained when

$$\frac{\partial}{\partial \epsilon^{(1)}} P(\epsilon^{(1)}) = 0,$$

or

$$\frac{\partial}{\partial \epsilon^{(1)}} \ln P(\epsilon^{(1)}) = 0 = \frac{\partial}{\partial \epsilon^{(1)}} \left\{ \ln \Omega(\epsilon^{(1)}) - \ln \Omega_t \right\} = \frac{\partial}{\partial \epsilon^{(1)}} \ln \Omega(\epsilon^{(1)}), \quad (3.8)$$

since the second term does not depend on $\epsilon^{(1)}$. Since the two subsystems can exchange only energy, we have, with obvious symbols,

$$\Omega(\epsilon^{(1)}) = \Omega^{(1)}(\epsilon^{(1)})\, \Omega^{(2)}(\epsilon^{(2)}).$$

Now we consider that $\epsilon^{(1)} + \epsilon^{(2)} = \epsilon_t$ and therefore $\partial/\partial\epsilon^{(1)} = -\partial/\partial\epsilon^{(2)}$. The condition in (3.8) then yields

$$\frac{\partial}{\partial \epsilon^{(1)}} \ln \Omega^{(1)}(\epsilon^{(1)}) = \ln \frac{\partial}{\partial \epsilon^{(2)}} \Omega^{(2)}(\epsilon^{(2)}).$$

This means that

$$\boxed{T^{(1)} = T^{(2)}, \qquad S^{(1)} + S^{(2)} = \text{maximum}} \quad (3.9)$$

The most probable partition is the one that makes the two temperatures equal and the total entropy maximum.

Chemical Potential and Particle Equilibrium

Let us now consider a situation similar to that of the previous section, with the difference that now the two subsystems can exchange not only energy, but also particles. To analyze this situation, it is necessary to introduce a new statistical quantity. Let $\Omega(n, \epsilon)$ be the number of accessible states of a system,

when it contains n particles and energy ϵ. If n is large, we may assume that $\Omega(n, \epsilon)$ is a continuous function of n and define the *chemical potential* of that system as

$$\boxed{\mu(n, \epsilon) = -\frac{1}{\beta}\frac{\partial}{\partial n}\ln \Omega(n, \epsilon) = -T\frac{\partial S}{\partial n}} \tag{3.10}$$

To obtain some insight into the meaning of this quantity, let us consider that the entropy (3.6) of a system formed by a gas with n particles is a function of its energy, volume, and number of particles:

$$S = S(\epsilon, V, n), \qquad \mathrm{d}S = \frac{\partial S}{\partial \epsilon}\mathrm{d}\epsilon + \frac{\partial S}{\partial V}\mathrm{d}V + \frac{\partial S}{\partial n}\mathrm{d}n.$$

If the number of particles is varied keeping volume and entropy constant, the above equation yields

$$\mathrm{d}\epsilon = \mu\, \mathrm{d}n,$$

where use has been made of (3.7) and (3.10). Thus, the chemical potential is the energy necessary to add a particle to the system, keeping volume and entropy constant.[1]

As in the case of thermal equilibrium, consider now two systems containing particles of the same type. Put them in contact and let them exchange both energy and particles. As before, we assume that the interaction energy between the two subsystems is negligible. The question we ask now is: what are the most probable partitions of the total energy ϵ_t and of the total number of particles n_t between the two subsystems? The probability that the subsystem 1 contains energy ϵ_1 and n_1 particles is given by the number $\Omega(n_1, \epsilon_1)$ of states accessible to the combined system corresponding to energy ϵ_1 and n_1 particles in the first subsystem, divided by the total number Ω_t of states accessible to the total system:

$$P(n_1, \epsilon_1) = \frac{\Omega(n_1, \epsilon_1)}{\Omega_t}\ .$$

The condition of maximum probability is

$$\begin{cases} \frac{\partial}{\partial \epsilon_1}\ln P(n_1, \epsilon_1) = 0 \\ \\ \frac{\partial}{\partial n_1}\ln P(n_1, \epsilon_1) = 0 \end{cases} \quad \text{or} \quad \begin{cases} \frac{\partial}{\partial \epsilon_1}\ln \Omega(n_1, \epsilon_1) = 0 \\ \\ \frac{\partial}{\partial n_1}\ln \Omega(n_1, \epsilon_1) = 0 \end{cases}$$

Since the contact has the only effect of exchanging energy and particles, we may write

$$\Omega(n_1, \epsilon_1) = \Omega^{(1)}(n_1, \epsilon_1)\, \Omega^{(2)}(n_t - n_1, \epsilon_t - \epsilon_1), \tag{3.11}$$

[1] If the system is kept at constant volume and temperature, the chemical potential is the Helmholtz free energy F necessary to add a particle; if the system is kept at constant pressure and temperature, the chemical potential is the Gibbs free energy G necessary to add a particle [370].

where $\Omega^{(1)}$ and $\Omega^{(2)}$ refer to the two subsystems, separately. The condition of maximum probability becomes

$$\begin{cases} \frac{\partial}{\partial \epsilon_1} \ln \Omega^{(1)}(n_1, \epsilon_1) + \frac{\partial}{\partial \epsilon_1} \ln \Omega^{(2)}(n_t - n_1, \epsilon_t - \epsilon_1) = 0 \\ \\ \frac{\partial}{\partial n_1} \ln \Omega^{(1)}(n_1, \epsilon_1) + \frac{\partial}{\partial n_1} \ln \Omega^{(2)}(n_t - n_1, \epsilon_t - \epsilon_1) = 0 \end{cases}$$

or

$$\begin{cases} \frac{\partial}{\partial \epsilon_1} \ln \Omega^{(1)}(n_1, \epsilon_1) = \frac{\partial}{\partial \epsilon_2} \ln \Omega^{(2)}(n_2, \epsilon_2) \\ \\ \frac{\partial}{\partial n_1} \ln \Omega^{(1)}(n_1, \epsilon_1) = \frac{\partial}{\partial n_2} \ln \Omega^{(2)}(n_2, \epsilon_2) \end{cases}$$

i.e.,

$$\boxed{\beta^{(1)} = \beta^{(2)} \quad \text{and} \quad \mu^{(1)} = \mu^{(2)}} \tag{3.12}$$

The most probable partition of energy and number of particles between the two subsystems put in contact is that which makes equal both the two temperatures and the chemical potentials. Keeping in mind the physical meaning of the chemical potential, the latter condition indicates that the two systems have no energetic "interest", at equilibrium, to move one particle from one system to the other.

3.5 Thermal Bath

If one of the two subsystems considered in the previous section is so "large", i.e., has so many degrees of freedom (so large heat capacity), that it can exchange any amount of energy with the other subsystem without appreciably varying its internal energy, then it is called a *heat reservoir* and the "small" subsystem is said to be in contact with a *thermal bath*.

The definition of temperature given above in (3.7) has a well-defined meaning for a system with a well-defined energy, i.e., for an isolated system. If the system is put in contact with a thermal bath, it does not have a well-precise energy nor, therefore, a well-defined temperature. The temperature in this case is a property of the combined system, the one of interest plus thermal bath. Once the system is again isolated, it will have again well-defined energy and temperature. The latter, however, is not necessarily the same of that of the thermal bath, even though this is the most probable, as seen in the previous section. There are possible differences that are, in general, extremely small.

3.6 The Three Fundamental Statistical Ensembles

In Sect. 3.1, the statistical ensembles were presented as large mental sets of systems obtained as replicas of the actual system of interest. All these replicas should be thought of as being prepared with the same operations as those

actually performed to realize the real system. There are several ways that can be followed in this process. In the simplest case, we generate the system(s) with well-defined energy and number of particles. The resulting ensemble is called *microcanonical*. Its use is rather limited, since it corresponds to a physical situation rare and difficult to realize. The second, most common, case prepares the systems putting them in contact with a thermal bath that can exchange with the system only any amount of energy. This produces the *canonical* ensemble. The third ensemble, called *grand canonical* or *macrocanonical*, is obtained by putting the systems in contact with a special reservoir that can exchange not only energy, but also particles. The fluctuations of energy in the systems of a canonical ensemble and the fluctuations of energy and of particle number in the grand-canonical ensemble are often very small so that many statistical results are independent of the ensemble used.

3.6.1 Microcanonical Ensemble

In the microcanonical ensemble, all systems have the same energy $\overline{\epsilon}$, and, if the system of interest is made of particles, all the systems in the ensemble have the same number of particles. In classical physics, since energy is a continuous variable, it is more convenient to assume, without practical consequences, that the systems of the ensemble have energies contained in an interval $\delta\epsilon$ small with respect to the energy variations of interest. The probability that a system of the ensemble is found in a particular state is constant for all states with energy inside $\delta\epsilon$, and zero for the other ones. If we indicate with $\rho(q_i, p_i)$ the density of points representing the states of the ensemble in phase space, the microcanonical ensemble is then defined by

$$\rho(q_i, p_i) = \begin{cases} \text{const.} & \text{if} \quad \overline{\epsilon} \leq \epsilon(q_i, p_i) \leq \overline{\epsilon} + \delta\epsilon \\ 0 & \text{otherwise} \end{cases} .$$

This density is function of energy alone and therefore, as seen at the end of Sect. 3.2, it represents a system in equilibrium.

3.6.2 Canonical Ensemble

The canonical ensemble is used when the system of interest has been put in contact with a thermal bath. Therefore, all the systems of the ensemble have been able to exchange with the heat reservoir any amount of energy.

It is now important to find the state distribution of the systems of the ensemble, that is, the probability that a system of the ensemble, once disconnected from the thermal bath, is found in a given state, as this will be the probability that our actual system can be found in that given state. To find this probability, let $\Omega^{(s)}(\epsilon_s)$ be the number of states accessible to our system when it has energy ϵ_s, $\Omega^{(b)}(\epsilon_b)$ the number of states accessible to the thermal

bath when it has energy ϵ_b. The number of states $\Omega(\epsilon_s, \epsilon_b)$ accessible to the combined system for the partition $\epsilon_s + \epsilon_b = \epsilon_\circ$ of the total energy $\epsilon_\circ$ is such that

$$\ln \Omega(\epsilon_s, \epsilon_b) = \ln \left\{ \Omega^{(s)}(\epsilon_s)\, \Omega^{(b)}(\epsilon_b) \right\} = \ln \Omega^{(s)}(\epsilon_s) + \ln \Omega^{(b)}(\epsilon_b).$$

Now we know that the very definition of thermal bath requires that $\epsilon_\circ = \epsilon_s + \epsilon_b \gg \epsilon_s$, so that we can write

$$\ln \Omega^{(b)}(\epsilon_b) = \ln \Omega^{(b)}(\epsilon_\circ) + \frac{\partial}{\partial \epsilon} \ln \Omega^{(b)}(\epsilon_\circ)(\epsilon_b - \epsilon_\circ) = \ln \Omega^{(b)}(\epsilon_\circ) + \beta(\epsilon_b - \epsilon_\circ).$$

Substituting this expression in the previous one, we have

$$\ln \Omega(\epsilon_s, \epsilon_b) = \ln \Omega^{(s)}(\epsilon_s) + \ln \Omega^{(b)}(\epsilon_\circ) - \beta \epsilon_s,$$

or

$$\Omega(\epsilon_s, \epsilon_b) = \Omega^{(s)}(\epsilon_s) \Omega^{(b)}(\epsilon_\circ) \mathrm{e}^{-\beta \epsilon_s}.$$

Thus, the probability that a system in the ensemble has energy ϵ_s is

$$P(\epsilon_s) = Const\ \Omega^{(s)}(\epsilon_s)\ \mathrm{e}^{-\beta \epsilon_s}. \tag{3.13}$$

Note that here β represents the temperature of the thermal bath.

The probability that the system is found in a specific state s with energy ϵ_s is the probability that the state has energy ϵ_s, given above, divided by the number $\Omega^{(s)}(\epsilon_s)$ of states with that energy:

$$P(s) \propto \mathrm{e}^{-\beta \epsilon_s},$$

or

$$\boxed{P(s) = \tfrac{1}{Z} \mathrm{e}^{-\beta \epsilon_s}, \quad Z = \textstyle\sum_s \mathrm{e}^{-\beta \epsilon_s}} \tag{3.14}$$

This is called the *canonical distribution*, and Z is the *partition function*, such that $\sum_s P(s) = 1$.

3.6.3 Grand Canonical Ensemble

The grand canonical ensemble is specific for systems composed of particles and is considered when the system can exchange, with the heat reservoir, not only energy, but also any number of particles. In the combined system, the total number of particles is constant, as is the energy.

Assume, for simplicity, that there is only one type of particles. We have seen already that the most probable situation after contact is that the system of interest has the same temperature and the same chemical potential as the thermal bath. But now our aim is to find the probability that the system is found in a state s with energy ϵ_s and n_s particles.

In analogy with the canonical case, we assume that the energy ϵ_b of the thermal bath and its number of particles n_b are much larger than the same quantities of the system of interest:

$$\epsilon_t = \epsilon_b + \epsilon_s \gg \epsilon_s, \qquad n_t = n_b + n_s \gg n_s.$$

Thus,

$$\ln \Omega^{(b)}(\epsilon_b, n_b) = \ln \Omega^{(b)}(\epsilon_t, n_t) + \frac{\partial}{\partial \epsilon} \ln \Omega^{(b)}(\epsilon_t, n_t)(\epsilon_b - \epsilon_t) + \frac{\partial}{\partial n} \ln \Omega^{(b)}(\epsilon_t, n_t)(n_b - n_t),$$

or

$$\ln \Omega^{(b)}(\epsilon_b, n_b) = \ln \Omega^{(b)}(\epsilon_t, n_t) + \beta_b(\epsilon_b - \epsilon_t) - \beta_b \mu_b(\epsilon_t, n_t)(n_b - n_t).$$

Substitute this in (3.11):

$$\ln \Omega(\epsilon_s, n_s) = \ln \Omega^{(s)}(\epsilon_s, n_s) + \ln \Omega^{(b)}(\epsilon_t, n_t) - \beta_b \epsilon_s + \beta_b \mu_b n_s$$
$$\Omega(\epsilon_s, n_s) = \Omega^{(s)}(\epsilon_s, n_s)\ Const\ \mathrm{e}^{\beta_b(-\epsilon_s + \mu_b n_s)}.$$

In conclusion,

$$P(\epsilon_s, n_s) \propto \Omega^{(s)}(\epsilon_s, n_s)\, \mathrm{e}^{-\beta(\epsilon_s - \mu n_s)},$$

where the label "b" has been dropped, being clear the meaning of the equation. The probability that the system of interest is found in a state s with n particles and energy ϵ is given by the *grand-canonical distribution*

$$\boxed{P(s) = \tfrac{1}{Z}\mathrm{e}^{-\beta(\epsilon_s - \mu n_s)}, \quad Z = \textstyle\sum_s \mathrm{e}^{-\beta(\epsilon_s - \mu n_s)}} \tag{3.15}$$

If several types of particles are present with $n^{(i)}$ particles and chemical potentials $\mu^{(i)}$, the result is the same with the substitution of μn_s with the sum of the products $\mu^{(i)} n^{(i)}$.

3.7 Equilibrium Particle Distributions in Ideal Gases

In the previous section, the probability distribution of the accessible states of a system was found by means of the statistical ensembles, on the basis of the fundamental hypothesis of statistical physics. At this point, it is possible to apply the results just obtained to the case of ideal gases, meaning gases formed by noninteracting particles, to find out how the particles are distributed within the gas. In classical terms, no other assumptions are necessary. In quantum statistical physics, however, when the particles forming the gas are indistinguishable, they have no identity, and the symmetry of the wavefunction depends on the spin of the particles, as discussed in Sect. 2.6. As we shall see below, this fact produces very significant differences in the statistical properties of the gas.

3.7.1 Classical Gas: Maxwell–Boltzmann Distribution

The system of interest is the entire gas with a large number of particles. In principle, the ensemble is formed by systems all formed by the entire gas and all prepared in the same way as the actual system. In the case of an ideal classical gas, however, we may also consider the system formed by a single particle of the gas. Since it interacts with the other particles only exchanging energy, the entire gas may be considered as an ensemble of systems formed each by a single particle, and apply the results of the canonical ensemble. The probability that a particle occupies a state of energy ϵ is then given by the canonical distribution (3.14) which in this context is called the Maxwell–Boltzmann distribution:

$$\boxed{P(\epsilon) = C\mathrm{e}^{-\beta\epsilon}} \tag{3.16}$$

where C is a normalization constant. Since the particles are independent, this distribution represents also, with the appropriate normalization, the average density of particles in a point of the phase space with energy ϵ.

In the other approach, which considers the entire gas as a system of the ensemble, the state of the system may be characterized by the number of particles occupying a given region of the phase space, characterized by an energy ϵ and we reach the same conclusion, i.e., this number of particles is proportional to the Maxwell–Boltzmann distribution [448].

3.7.2 Bose Distributions

In treating a gas of identical particles in quantum terms, we cannot use the above simplified procedure, as a particle has no identity with respect to the other particles of the gas. We must consider the statistical ensemble of systems formed, each, by the entire gas. Since the particles are not interacting, the state of the system may be characterized by the number n_s of particles in each single-particle state s with energy ϵ_s.[2] For reasons of space we shall not report, here, the derivations of the quantum distributions of the identical particles of a gas. They can be found in many good books of statistical physics, such as [370, 448]. Furthermore, these concepts will be reconsider in Chap. 24, in connection with the equilibrium Green functions of independent-particle systems. We simply report in the following the results of the statistical derivations.

For a gas of bosons, nothing prevents the possibility to find any number of particles in the same quantum state. If we assume that the number N of particles is fixed, the average number of particles in a single-particle state with energy ϵ_i, i.e., the average *occupation number*, is given by the *Bose-Einstein distribution*

[2] This concept will be better clarified in Chap. 23, where the second-quantization formalism will be introduced.

$$\boxed{\overline{n_i} = \frac{1}{e^{\beta(\epsilon_i - \mu)} - 1}} \tag{3.17}$$

where μ is the parameter that takes care of the normalization to the total number of particles $N = \sum_i \overline{n_i}$. μ can be recognized as the chemical potential defined in (3.10).

If, at a given energy, the number of particles in the gas can vary, as in the case of photons or phonons (energy quanta of vibrations of the electromagnetic field or of atoms in a lattice, respectively), the Bose distribution becomes [448]

$$\overline{n_i} = \frac{1}{e^{\beta\epsilon_i} - 1}, \tag{3.18}$$

used by Planck to explain the thermal radiation.

3.7.3 Fermi Distribution

If the particles of the gas are fermions, Pauli exclusion principle (see Sect. 2.6) prevents two particles from occupying the same state. The average occupation number of the i-th state is the *Fermi–Dirac distribution*:

$$\boxed{\overline{n_i} = \frac{1}{e^{\beta(\epsilon_i - \mu)} + 1}} \tag{3.19}$$

The chemical potential μ is again determined by the normalization condition, as for the Bose distribution.

Both Bose and Fermi distributions will often be used in the following of this book since they describe the statistical distributions of phonons and electrons in semiconductors, respectively.

3.7.4 Classical Limit

For energies much higher than the chemical potential ($\epsilon - \mu \gg KT$), both Bose and Fermi distributions tend to the classical Maxwell–Boltzmann distribution since the exponential in the denominator becomes dominant with respect to unity:

$$\overline{n_i} \approx e^{-\beta(\epsilon_i - \mu)} = e^{\beta\mu} e^{-\beta\epsilon_i}.$$

This is coherent with the fact that at these energies the average occupation numbers become much less than 1, and the particle indistinguishability does not play a significant role since the particles do not try to occupy the same state. This last equation also confirms the role of the chemical potential in the normalization of the particle distribution to their total number N.

4

Crystal Structures

4.1 Crystals

Crystals are solid bodies[1] where atoms are arranged in regular patterns, as shown in Fig. 4.1, where, for clarity, a two-dimensional arrangement is shown. Ideal crystals have a translational periodicity defined by three noncoplanar vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$, such that the crystal remains identical if translated by a vector

$$\boldsymbol{T} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3, \tag{4.1}$$

where n_1, n_2, and n_3 are any three integers. From this definition, it is clear that, ideally, crystals should be infinite in all directions. In practice, crystals of macroscopic dimensions are formed by an extremely large number of atoms so that their properties in the interior, or *bulk properties*, are those expected for the ideal crystal, and only near the edges the physical properties are influenced by *surface effects*. On the other hand, it is today possible to fabricate and study "nano-crystals" of very small dimensions, where the translational symmetry is completely lost, but we still call them crystals as long as the positions of their atoms are those predicted by the regular arrangement of the corresponding ideal crystal. There are, of course, intermediate cases, nowadays very frequent and important, where some of the bulk properties of the crystal can still be used, but the finite dimensions of the system play a crucial role. A typical example is given by quantum dots, where the energy levels of the electrons can be evaluated with good approximation by the solution of the Schrödinger equation for the finite system, using the electron effective mass obtained by the bulk bands of the material (see Chap. 7).

[1] Liquid crystals have properties between those of a liquid and of a conventional crystal. They are of great technological importance for their application in displays, but lie outside of the scope of this book and will not be considered further.

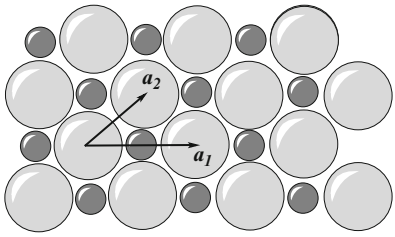


Fig. 4.1. A crystal is formed by a regular arrangement of atoms. In this two-dimensional example, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are the unit symmetry translations

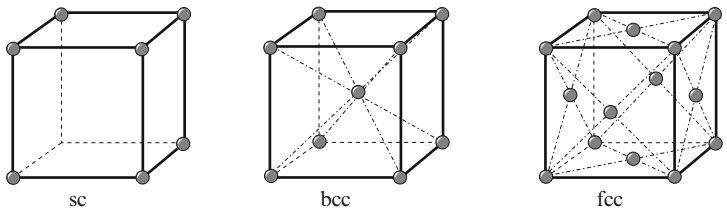


Fig. 4.2. Cubic lattices. In the simple cubic (sc), lattice points are situated at the corners of the cube. In the body centered cubic (bcc), lattice points are situated at the corners and at the center of the cube. In the face centered cubic (fcc), lattice points are situated at the corners of the cube and at the centers of the faces of the cube

A crystal has also other symmetry operations, besides translations, which bring the crystal over itself. They can be rotations, reflections, rotary-reflections, and inversion. The total set of symmetry operations of a crystal is its *space group*. The group of operations obtained by a space group setting to zero all translations is the *point group* of the crystal. Group theory indicates how to use the symmetry of the crystal, and in general of a physical system, to classify the eigenstates, evaluate degeneracies, selection rules, etc. For reasons of space this subject, albeit very important, cannot be treated in this book, and we refer the interested reader to specialized books, such as [447].

4.2 Lattices

The set of points defined by the translation vectors in (4.1) is called a *Bravais lattice* or simply a *lattice*. The parallelepiped formed by the vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ is a *primitive unit cell*. The lattice in (4.1) is sometimes called a *direct lattice*, to distinguish it from the *reciprocal lattice* that we shall see in Sect. 4.4.

In Fig. 4.2, the three possible cubic lattices are shown. The simple cubic (sc), the body-centered cubic (bcc), and the face-centered cubic (fcc). A primitive cell of the sc lattice coincides with the cube. As any primitive cell, it contains, inside the cube, only one atom, more precisely 1/8 of atom at each

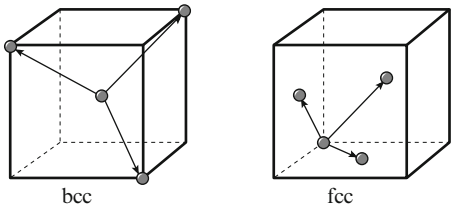


Fig. 4.3. Primitive cells of body-centered and face-centered cubic lattices. For the simple cubic, the cubic cell shown in Fig. 4.2 is also a primitive cell

of the 8 corners. In the bcc lattice, the cube is not a primitive cell. It contains, besides the eight eighths of atom at the corners, a second atom at the center of the cube. A primitive cell of the bcc lattice is shown in Fig. 4.3. The three primitive elementary translations connect an atom at a center of the cube with three atoms at three nonadjacent corners. In the fcc lattice, the cube contains four atoms: eight eighths at the corners and six halves at the centers of the faces. A primitive unit cell, shown in Fig. 4.3, is formed with elementary translations connecting an atom at a corner of the cube with three atoms at the centers of three faces. Both the bcc and fcc primitive cells indicated above are formed by rhombohedrons where atoms are located at the eight corners and shared by eight primitive cells, as for the sc primitive cells.

Often the directions of the primitive translation vectors are used as *crystallographic axes*. This, however, is not always the most convenient choice. For the most common cubic semiconductors of diamond and zincblende structures, shown in Fig. 4.4, the conventional cubic cell is considered, and cubic axes are always used as crystallographic axes, more convenient than the primitive translations shown in Fig. 4.3. The lengths of the elementary translations along the crystallographic axes are called *lattice constants*.

Wigner–Seitz Primitive Cell

The *Wigner–Seitz primitive cell* of a lattice is formed by all points closer to one of the lattice points than to any other. In Fig. 4.5, it is shown how to construct it in a simple two-dimensional case: a lattice point is connected with line segments to the nearby lattice points; then the perpendicular lines (or planes in the three-dimensional case) are drawn through the middle points of these segments. The Wigner–Seitz cell is the polygon (polyhedron) of smallest area (volume) enclosed by these lines (planes).

Points, Lines, and Planes in Crystals

Points in a crystal are identified by their coordinates in the crystallographic axes in units of the lattice vectors. For example, the point $(1/4, 1/4, 1/2)$ is the point given by

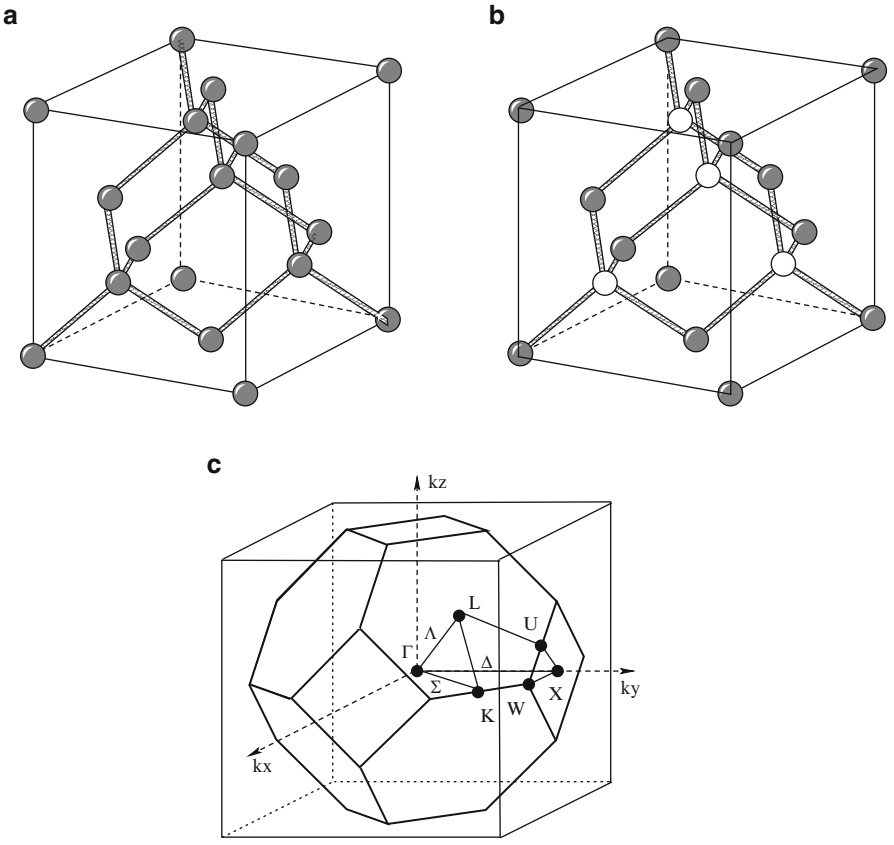


Fig. 4.4. Diamond (**a**) and zincblende (**b**) structures. Not all atoms and bonds of the structures are shown in the figure to make more evident the tetrahedral structure of the bonds. Part (**c**) of the figure shows the Brillouin zone of the fcc lattice with indication of the points and lines of high symmetry. The irreducible wedge is the polyhedron with vertices Γ, K, W, X, U, L

$$\frac{1}{4}\,\boldsymbol{a}_1 + \frac{1}{4}\,\boldsymbol{a}_2 + \frac{1}{2}\,\boldsymbol{a}_3,$$

where $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ are the conventional unit translations, not necessarily coincident with the primitive translations, as said above.

The direction of an oriented straight line is indicated by the smallest integers proportional to the components, along the crystallographic axes, of a vector oriented in the considered direction. They are usually given inside square brackets: $[n_1, n_2, n_3]$. A negative component is indicated by a minus sign above the index: $[n_1, \overline{n}_2, n_3]$. If all directions equivalent by symmetry are to be indicated, they are usually put in angular brackets: $\langle n_1, n_2, n_3 \rangle$.

The orientation of a plane in a crystal is indicated by its *Miller indices*: if x_1, x_2, and x_3 are the intercepts of a plane with the wanted orientation with

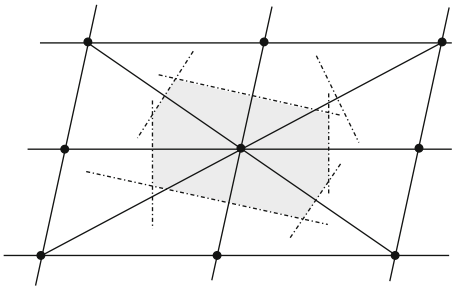


Fig. 4.5. The Wigner–Seitz primitive cell of a lattice is formed by all points closer to one of the lattice points than to any other

the crystallographic axes, in units of $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$, the smallest integers proportional to the inverse of such intersects are the Miller indices. They are usually given inside round brackets: (i, j, k). The inverse are used to avoid infinities if a plane is parallel to a crystallographic direction. If all plane orientations equivalent by symmetry are to be indicated, they are often put in curly brackets: $\{i, j, k\}$.

In a crystal with a cubic unit cell the direction $[ijk]$ is orthogonal to the plane (ijk).

Diamond and Zincblende Structures

In the two-dimensional example shown in Fig. 4.1, it is clear that there are two different atoms in each unit cell. The atomic species in the unit cell and the vectors defining their positions form the *basis*. Thus, a crystal is characterized by its lattice and basis.

The diamond structure, common to carbon (diamond), silicon, germanium, and gray tin, is formed by an fcc lattice with a basis formed by two identical atoms, one at the corners of the cube and a second one along the diagonal of the cube at 1/4 of its length. In the cubic crystallographic axes, the coordinates of the atoms of the basis are $(0, 0, 0)$ and $(1/4, 1/4, 1/4)$. Thus, the crystal is formed by two fcc interpenetrated structures at the distance of one fourth of the diagonal of the cube. The diamond structure is shown in part (a) of Fig. 4.4.

The zincblende (ZnS) structure is common to many III-V semiconductors. The structure is similar to that of diamond, but in the zincblende structure the two atoms of the basis are different. This structure is shown in part (b) of Fig. 4.4.

In Table 4.1, the crystal structures and lattice constants of some important semiconductors are given.

Table 4.1. Crystallographic data of some semiconductors

Material		Structure	Cube side (Å) at 300 K
Silicon	Si	Diamond	5.431
Germanium	Ge	Diamond	5.646
Gallium arsenide	GaAs	Zincblende	5.653
Gallium phosphide	GaP	Zincblende	5.451
Aluminum arsenide	AlAs	Zincblende	5.661
Indium arsenide	InAs	Zincblende	6.058
Indium phosphide	InP	Zincblende	5.869

4.3 Crystal Bonding

Different types of bonds hold atoms and molecules together in crystals. All of them are ultimately due to the electron charge distribution around the nuclei. Electrons in deep levels do not contribute to the binding, which is instead determined by the charge distribution in the most external states, or *valence states*. Thus, the forces responsible for the structure of matter are the electric forces that keep electrons bound to their nuclei and different atoms bound to each other in condensed matter, with a dynamics perfectly described by quantum mechanics. Before the discovery of quantum mechanics neither the structure of the atoms, nor the chemical bonds, nor the properties of condensed matter could have a plausible explanation.

Although due all to electric forces, crystal bonds, as said above, may have different forms, briefly described in the following.

Electrostatic Interaction, Ionic Crystals

Typical ionic crystals are the salts formed by an element of the first group and one of the seventh group, such as NaCl. The isolated electron in the external shell of the former is transferred to the external shell of the latter, that misses just one electron to be complete. This transfer is energetically favored, and the result is the formation of charged ions that are held together by the electrostatic force. In ionic crystals, an ion of one sign is surrounded by ions of opposite sign. If the ions get too close to each other the electronic wavefunctions tend to overlap. Since, however, electrons obey Pauli exclusion principle, this overlap generates a repulsive force, and the stable structure is determined by the minimum free energy due to the competitive actions of these forces.

Homopolar Bond, Covalent Crystals

When the atoms that form a crystal are all identical, there is no reason for the electrons to move from one atom to another. The crystal binding cannot

be of ionic nature. In the case of elements of group IV, a binding interaction arises, instead, by the sharing of electrons between two adjacent atoms. As for the hydrogen molecule, this *covalent bond* is present when two electrons of opposite spins have wavefunctions that overlap in the region between the atoms. This *bonding state* correspond to a higher charge density in the region of low potential energy between the two atomic cores.

For III-V semiconductors, such as GaAs, the bond is partially ionic and partially covalent.

Other Types of Bonds

Other types of crystal bonds are less important for semiconductors.

In crystals formed by noble-gas elements the electronic shells are completely filled, no charge rearrangement occurs, and the atoms in the solid are very similar to the isolated atoms. In a static picture, the spherical distribution of the electronic charge is such that no electrostatic force is acting between atoms. Nevertheless, dipole fluctuations generate the so-called *Van der Waals forces* that hold the crystal together with little cohesive energy. A similar situation occurs in the so-called *molecular crystals*. The combination of the Van der Waals potential and the repulsion due to the overlap of the wavefunctions yields the empirical *Lennard–Jones potential* [18].

In metal crystals, the electrons of the outer shell leave the parent atoms and move freely inside the solid. The resulting electron gas interacts electrostatically with the positive ions of the lattice, and this interaction determines the *metallic bond* of the crystal. Its quantitative evaluation is not simple and requires the use of many-body techniques in second-quantization formalism [145].

Another type of bond, important mainly in organic crystals and in ice, is the *hydrogen bond*. When a hydrogen atom is in the proximity of a strongly electronegative atom, it transfers its electron to this atom, forming two ions of opposite signs that are bonded as in ionic crystals. The hydrogen bond is responsible, for example, for the link between the two helices of the DNA.

A detailed description of the various types of crystal bonding and of the various crystal structures and symmetries can be found in standard books of solid state, such as [18,196], or [241].

4.4 Reciprocal Lattice

The *reciprocal lattice* is a fundamental concept in the whole theory of solid state. Given a direct lattice with unit vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, the corresponding reciprocal lattice is defined in the space of wavevectors by the three unit vectors

$$\boldsymbol{b}_1 = 2\pi \frac{\boldsymbol{a}_2 \times \boldsymbol{a}_3}{V_c}, \quad \boldsymbol{b}_2 = 2\pi \frac{\boldsymbol{a}_3 \times \boldsymbol{a}_1}{V_c}, \quad \boldsymbol{b}_3 = 2\pi \frac{\boldsymbol{a}_1 \times \boldsymbol{a}_2}{V_c},$$

where $V_c = \boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3$ is the volume of the unit cell of the direct lattice. It is immediate to verify that between the unit vectors of the direct lattice and those of the reciprocal lattice the following relation holds:

$$\boxed{\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi\delta_{ij}} \tag{4.2}$$

Thus, given a vector $\boldsymbol{T}$ of the direct lattice and a vector $\boldsymbol{G}$ of the reciprocal lattice,

$$\boldsymbol{T} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3, \qquad \boldsymbol{G} = m_1\boldsymbol{b}_1 + m_2\boldsymbol{b}_2 + m_3\boldsymbol{b}_3, \tag{4.3}$$

we have

$$\boldsymbol{T} \cdot \boldsymbol{G} = 2\pi\,(n_1m_1 + n_2m_2 + n_3m_3),$$

so that

$$\mathrm{e}^{\mathrm{i}\boldsymbol{T}\boldsymbol{G}} = 1.$$

For the scalar product of a generic vector $\boldsymbol{r}$ of the direct space and a generic vector $\boldsymbol{q}$ of the reciprocal space

$$\boldsymbol{r} = r_1\boldsymbol{a}_1 + r_2\boldsymbol{a}_2 + r_3\boldsymbol{a}_3, \qquad \boldsymbol{q} = q_1\boldsymbol{b}_1 + q_2\boldsymbol{b}_2 + q_3\boldsymbol{b}_3,$$

we have, except for the factor 2π, the traditional sum of coordinate products:

$$\boldsymbol{r} \cdot \boldsymbol{q} = 2\pi\,(r_1q_1 + r_2q_2 + r_3q_3)\ .$$

The volume, in the reciprocal space, of the unit cell of the reciprocal lattice, is easily found with the application of the formula for the vector triple product:

$$\boldsymbol{b}_1 \cdot \boldsymbol{b}_2 \times \boldsymbol{b}_3 = \left(\frac{2\pi}{V_c}\right)^3 (\boldsymbol{a}_2 \times \boldsymbol{a}_3) \cdot [(\boldsymbol{a}_3 \times \boldsymbol{a}_1) \times (\boldsymbol{a}_1 \times \boldsymbol{a}_2)] = \frac{(2\pi)^3}{V_c}. \tag{4.4}$$

It turns out that the reciprocal of an sc lattice is an sc lattice; the reciprocal of an fcc lattice is a bcc, and of a bcc is an fcc [241].

The *first Brillouin zone* (BZ) of a lattice is formed by all points of the reciprocal space closer to one of the points of the reciprocal lattice than to any other. It is the analogous, in the reciprocal space, of the Wigner–Seitz cell of direct space. The BZ of the fcc direct lattice, which is a unit cell of a bcc lattice, is shown in Fig. 4.4.

The symmetry operations of the crystal bring points inside the BZ to coincide with other points inside the BZ. The whole BZ can be generated, through symmetry operations, by the points inside a polyhedron, called the *irreducible wedge*. For the fcc lattice, the irreducible wedge is the polyhedron whose vertices are the points Γ, K, W, X, U, L in Fig. 4.4. Its volume is 1/48 of the volume of the entire BZ.

5

Phonons

In the previous chapter, a crystal was defined as a periodic arrangement of atoms. The equilibrium positions of the atoms in such arrangement correspond to the minimum of the potential energy of the system. Around the equilibrium positions, the atoms of the crystal vibrate with their thermal energy.

The physics of lattice vibrations is one of the most important and best established branches of modern solid-state physics. Not only electron transport, of direct interest in this book, is heavily influenced by the crystal vibrations, but also many other important physical characteristics of solids depend critically upon their vibrational properties, from specific heat to superconductivity, to mention only some of the most important.

Crystals are made of atoms, and atoms are relatively complex systems with nuclei and electrons around them. Nevertheless, many features of crystal vibrations are reasonably well described considering atoms simply as single particles located at the lattice points. This is the result of the *adiabatic approximation*, introduced by Born and Oppenheimer in 1927 [58] and well verified in most cases. According to this approximation, since the electrons are much lighter than the nuclei, they can follow their nuclei, at least the most internal ones closer to the nuclei, without internal energy change, i.e., adiabatically. We may then solve first the Schrödinger equation for the electrons with a given set of positions $\boldsymbol{R}_i$, and then use the energy found in this way in the dynamics of the crystal vibrations, with the nuclear coordinates as parameters. For more external electrons, and in particular for conduction electrons in metals, the above considerations do not apply, and refined calculations do take them into account in the study of crystal vibrations.

If small enough, lattice vibrations are described as independent harmonic oscillations of normal coordinates, as indicated in Sect. 1.5. From quantum mechanics of the harmonic oscillator (see appendix C), discrete energies are associated with each of such oscillation modes:

$$\epsilon_n = (n + 1/2)\hbar\omega \qquad n = 0, 1, 2, \ldots,$$

where ω is the frequency associated with the harmonic oscillator. When the energy is ϵ_n, we say that there are n *phonons* of that particular mode. If, owing to some interaction, the harmonic oscillator makes a transition from the state with energy ϵ_n to the state with energy ϵ_{n-1} we say that a phonon has been absorbed. A phonon is emitted if a transition occurs from the state with energy ϵ_n to that of energy ϵ_{n+1}.

Let us now develop the above ideas in detail, starting from the simplest case of a classical vibrating string.

5.1 The Vibrating String

Consider a homogeneous string of length L with linear density ρ. Let the string be fixed at both ends in $x = 0$ and $x = L$, and subject to a tension T. We consider transverse vibrations of the string and assume that the vibrations are small, so that the angle θ formed by the string with the direction given by its equilibrium position is small, and $\sin\theta \approx \theta \approx \tan\theta$. Consider the element of string between the longitudinal positions x and $x + \mathrm{d}x$, as indicated in Fig. 5.1. The forces acting on the extrema of the element along the direction y, orthogonal to the direction of the string in equilibrium, are

$$F_1 = T\sin\theta_1 \approx T\tan\theta_1 = -T\left(\frac{\partial Y}{\partial x}\right)_x,$$

$$F_2 = T\sin\theta_2 \approx T\tan\theta_2 = T\left(\frac{\partial Y}{\partial x}\right)_{x+\mathrm{d}x} = T\left[\left(\frac{\partial Y}{\partial x}\right)_x + \left(\frac{\partial^2 Y}{\partial x^2}\right)_x \mathrm{d}x\right],$$

where Y(x) is the transverse displacement of the string at x. Thus, the total force acting on the string element is

$$F_2 + F_1 = T\,\frac{\partial^2 Y}{\partial x^2}\mathrm{d}x = \rho\,\mathrm{d}x\,\frac{\partial^2 Y}{\partial t^2},$$

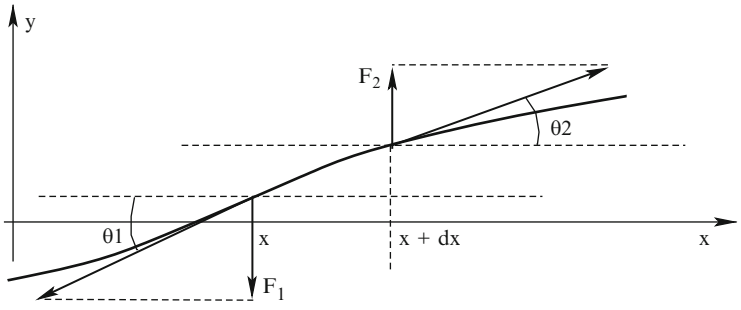


Fig. 5.1. For the dynamics of a vibrating string (see text)

where Newton's law has been applied in the last step. The equation of motion of the vibrating string is therefore

$$\frac{\partial^2 Y}{\partial t^2} = \frac{T}{\rho}\frac{\partial^2 Y}{\partial x^2}.$$

The general solution is given by a superposition of waves of the type

$$Y(x,t) = \sum_k \left[A_k \mathrm{e}^{\mathrm{i}(kx-\omega(k)t)} + B_k \mathrm{e}^{\mathrm{i}(-kx-\omega(k)t)}\right], \quad k > 0, \tag{5.1}$$

with

$$\omega(k) = \sqrt{T/\rho}\, k. \tag{5.2}$$

The above linear dependence of ω upon wavevector indicates that in the homogeneous vibrating string there is no dispersion: all components of a wavepacket travel at the same velocity so that group velocity and phase velocity (see Appendix B) are the same:

$$v_\varphi = v_g = \sqrt{T/\rho}.$$

We assume that the string has fixed extrema at $x = 0$ and $x = L$. The first condition requires

$$Y(0,t) = \sum_k \left[A_k \mathrm{e}^{-\mathrm{i}\omega(k)t} + B_k \mathrm{e}^{-\mathrm{i}\omega(k)t}\right] = 0$$

at any t. For this to be true, it is necessary that $A_k = -B_k$, so that the general solution (5.1) may be written as

$$Y_k(x,t) = \sum_k C_k \sin(kx)\mathrm{e}^{-\mathrm{i}\omega(k)t}. \tag{5.3}$$

The condition of vanishing Y also at $x = L$ for any t requires that $\sin(kL) = 0$, i.e., that L contains an integer number of half wavelengths:

$$n\frac{\lambda_n}{2} = L\ , \quad k_n = \frac{2\pi}{\lambda_n} = n\frac{\pi}{L} \qquad n = 1, 2, \ldots. \tag{5.4}$$

These wavevectors define the *modes of oscillation* of the string, shown in Fig. 5.2. The corresponding frequencies, from (5.2) and (5.4), are

$$\omega(k_n) = n\frac{\pi}{L}\sqrt{T/\rho} \quad n = 1, 2, \ldots.$$

From (5.4), the distance in k-space between two successive wavevectors is $\delta k = \pi/L$. Thus, the number of modes per unit k-length, called the *density of states* is

$$g(k) = L/\pi.$$

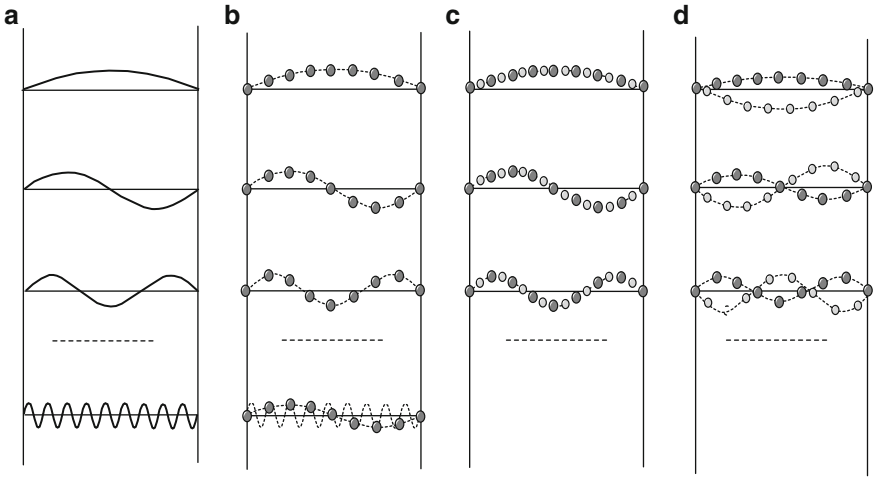


Fig. 5.2. Vibration modes. (**a**) Vibrations of a string. (**b**) Vibrations of a monatomic linear chain; in the lowest part of (**b**) it is shown that when the wavelength is shorter than the interparticle distance the displacements of the particles correspond to a longer wavelength. (**c**) Acoustic and (**d**) optical vibrations of a diatomic linear chain

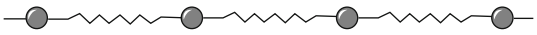


Fig. 5.3. Simple linear chain

The above results have been obtained with fixed boundary conditions $Y(0) = Y(L) = 0$. Similar results would have been obtained with periodic boundary conditions $Y(0) = Y(L)$, $(\mathrm{d}Y/\mathrm{d}x)_{x=0} = (\mathrm{d}Y/\mathrm{d}x)_{x=L}$. In such a case traveling waves would have been obtained, instead of the standing waves in (5.3). The density of states would be half of the previous one: $g = L/2\pi$. However, both positive and negative wavevectors have to be considered in this case, corresponding to waves traveling along the positive and negative directions.

5.2 The Simplest Linear Chain

Let us now consider a chain of N equal particles of mass M, connected by equal massless springs with elastic constant C, as shown in Fig. 5.3. Their equilibrium separation is the lattice constant a.

Traveling Waves

For simplicity we assume, for the time being, that the particles can vibrate only along the direction of the chain (longitudinal waves). Let x_j be the displacement of the j-th particle from its equilibrium position. To handle the

boundary conditions for a chain of finite length, we assume that the chain is formed by a very large ring, so that we can use *cyclic boundary conditions*:

$$x_{j+N}(t) = x_j(t), \tag{5.5}$$

for any j and t. In this way, we keep the translational symmetry of the ideal chain. The vibrational properties that will be found are appropriate for regions of the real chain far from its edges. The same will be true for a three-dimensional crystal, where the *bulk phonons* will be obtained with periodic boundary conditions: *surface modes* of vibration can be found by considering explicitly the surface of the crystal.

The classical equations of motion of the particles in the simple linear chain are

$$M\ddot{x}_j = C(x_{j+1} + x_{j-1} - 2x_j). \tag{5.6}$$

With the experience of the vibrating string, we look for solutions like

$$x_j = \xi \mathrm{e}^{\mathrm{i}(qja - \omega(q)t)}. \tag{5.7}$$

Note that ja is the equilibrium position of the j-th particle. Substitution into the equations of motion (5.6) yields

$$M(-\omega^2(q))\xi \mathrm{e}^{\mathrm{i}(qja-\omega(q)t)} = C\xi\left[\mathrm{e}^{\mathrm{i}q(j+1)a} + \mathrm{e}^{\mathrm{i}q(j-1)a} - 2\mathrm{e}^{\mathrm{i}qja}\right]\mathrm{e}^{-\mathrm{i}\omega(q)t},$$

or

$$-\omega^2(q)M = C\left[\mathrm{e}^{\mathrm{i}qa} + \mathrm{e}^{-\mathrm{i}qa} - 2\right] = C[2\cos(qa) - 2] = -4C\sin^2\left(\frac{qa}{2}\right).$$

We thus obtain that (5.7) is a solution of (5.6) with the dispersion relation

$$\omega(q) = 2\sqrt{\frac{C}{M}}\left|\sin\left(\frac{qa}{2}\right)\right|. \tag{5.8}$$

Let us now apply the cyclic conditions (5.5):

$$\xi \mathrm{e}^{\mathrm{i}(q(j+N)a-\omega(q)t)} = \xi \mathrm{e}^{\mathrm{i}(qja-\omega(q)t)},$$

or

$$\mathrm{e}^{\mathrm{i}qNa=1}, \qquad qNa = 2\pi n \quad n = 0, \pm 1, \pm 2, \ldots. \tag{5.9}$$

The allowed wavevectors are therefore

$$q_n = \frac{2\pi}{L}n \quad n = 0, \pm 1, \pm 2, \ldots, \tag{5.10}$$

where $L = Na$ is the length of the chain. Thus, as in the vibrating string, the normal modes are discrete, the closer the modes the longer the chain. The density of states is now

$$g(q) = \frac{L}{2\pi}. \tag{5.11}$$

This result is coherent with the discussion at the end of the section on the vibrating string: now, with the cyclic conditions, equivalent to the periodic boundary conditions, we obtained traveling waves and a density of states half of that obtained with fixed edges and standing waves, but both positive and negative wavevectors are now to be considered.

Periodicity of $\omega(q)$, Brillouin Zone

The dispersion relation (5.8) is periodic with period $2\pi/a$: the value of ω_q does not change if q is substituted by

$$q' = q + \frac{2\pi}{a} r, \qquad r \ \text{ integer}, \tag{5.12}$$

as shown in Fig. 5.4. This periodicity is due to the finite distance between the particles of the chain. In fact, the atom displacements in (5.7) do not change if $2\pi/a$ is added to the wavevector q, as also shown in the bottom of part (b) of Fig. 5.2. It is therefore sufficient to consider wavevectors between $-\pi/a$ and $+\pi/a$. This interval constitutes the first unidimensional *Brillouin zone*, (BZ), a concept already encountered in Sect. 4.4, which will be used over and over again in this book. The total number of normal modes is the width of the BZ multiplied by the density of states (5.11):

$$\frac{2\pi}{a} \times \frac{L}{2\pi} = N,$$

equal to the number of particles, according to the theory of small oscillations, as reported in Sect. 1.5.

The dispersion relation (5.8) is not linear. This means that for the linear chain there is dispersion, and phase and group velocities are different. For $qa \ll 1$, however, when the wavelength is much longer than the interatomic distance, the frequency in (5.8) is approximately given by the linear relation

$$\omega(q) \approx a\sqrt{C/M}q \quad \text{(small } q\text{)}.$$

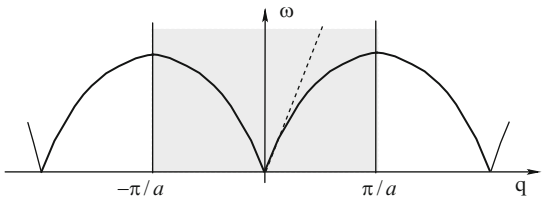


Fig. 5.4. Dispersion of the simple linear chain. The dashed line shows, for comparison, the limit of a continuous string. The shaded area identifies the Brillouin zone

This result coincides with the result for the vibrating string, if we identify C with T/a and ρ with M/a, as shown in Fig. 5.4: for long wavelengths the granularity of the chain has no influence.

In the above, we have considered only longitudinal waves, but there are also two independent transverse modes of oscillation that can be derived in a similar way.[1] The returning action of the spring for transverse displacements is weaker, so that we expect a lower frequency for the same q, i.e., a lower dispersion relation, as is found in the real dispersion relations of the crystals. The most general vibration of the chain is a linear superposition of all possible modes, each with a given wavevector q, polarization $\mathbf{e}_{q\ell}$ (direction of the displacements, identified with the label ℓ), and amplitude $\xi_{q,\ell}$:

$$\boldsymbol{x}_j = \sum_{q,\ell} \mathbf{e}_{q,\ell} \left\{ \xi_{q\ell} \mathrm{e}^{\mathrm{i}(qja - \omega_\ell(q)t)} + \xi^*_{q\ell} \mathrm{e}^{-\mathrm{i}(qja - \omega_\ell(q)t)} \right\}, \tag{5.13}$$

where, according to the use of exponentials for oscillating quantities, the complex conjugate has been added to get a physical quantity given by a real number.

5.3 Monatomic Linear Chain with Multiple Coupling

The simple dispersion relation (5.8) has been obtained under the hypothesis that only the nearest particles interact through the spring between them. If, still considering for simplicity longitudinal waves in a linear chain, we refine the model by assuming that the particle can interact with more distant particles, the total force acting on the j-th particle is

$$F_j = \sum_{p=1}^{\tilde{n}} C_p (x_{j+p} + x_{j-p} - 2x_j),$$

where C_p is the constant of the spring connecting particles p positions away, as indicated in Fig. 5.5, and $\tilde{n}$ is the number of considered interactions.

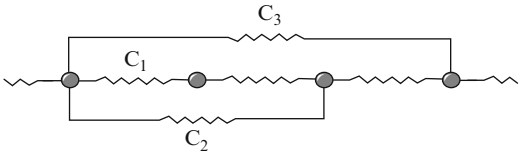


Fig. 5.5. Linear chain with multiple coupling

[1] In real crystals for arbitrary directions, the vibrational modes are neither purely longitudinal nor purely transverse. There are always, however, three independent polarizations.

With the same procedure followed in the previous section, we obtain, for the dispersion relation

$$\omega_q^2 = \frac{2}{M} \sum_p C_p[1 - \cos(pqa)]. \tag{5.14}$$

This reduces to the simpler (5.8) when only C_1 is different from zero, as it should. All considerations about the possible modes q, the density of states, the periodicity of the dispersion relation, and the BZ remain valid. The spring constants can be obtained from an experimental determination of the dispersion relation, since (5.14) can be easily inverted with Fourier analysis.

5.4 Diatomic Linear Chain

Let us now consider a linear chain composed of two types of different particles, with alternate masses M_1 and M_2: a simple model of a one-dimensional crystal with a base of two atoms, as shown in Fig. 5.6. For simplicity, we assume that the two adjacent interatomic distances at equilibrium are equal to b. The lattice constant is $a = 2b$. Also the two spring constants are assumed to be equal to C. The results would not be essentially different for the more general case of different interatomic distances and spring constants. We also consider the simple case of interactions between only nearest particles.

The equations of motions are now

$$\begin{cases} M_1 \ddot{x}_{2j+1} = C(x_{2j+2} + x_{2j} - 2x_{2j+1}) \\ \\ M_2 \ddot{x}_{2j} = C(x_{2j+1} + x_{2j-1} - 2x_{2j}) \end{cases}$$

where odd and even indices are used for positions of masses M_1 and M_2, respectively. We look for solutions like

$$\begin{cases} x_{2j+1} = \xi \mathrm{e}^{\mathrm{i}[(2j+1)bq - \omega(q)t]} \\ \\ x_{2j} = \eta \mathrm{e}^{\mathrm{i}[2jbq - \omega(q)t]} \end{cases}$$

Substituting into the equations of motion and dividing by the time exponentials, we obtain

$$\begin{cases} -M_1\omega^2(q)\xi \mathrm{e}^{\mathrm{i}(2j+1)bq} = C\left[\eta \mathrm{e}^{\mathrm{i}(2j+2)bq} + \eta \mathrm{e}^{\mathrm{i}2jbq} - 2\xi \mathrm{e}^{\mathrm{i}(2j+1)bq}\right] \\ \\ -M_2\omega^2(q)\eta \mathrm{e}^{\mathrm{i}2jbq} = C\left[\xi \mathrm{e}^{\mathrm{i}(2j+1)bq} + \xi \mathrm{e}^{\mathrm{i}(2j-1)bq} - 2\eta \mathrm{e}^{\mathrm{i}2jbq}\right] \end{cases}$$

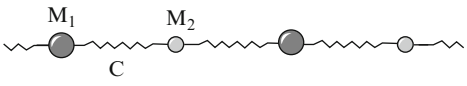


Fig. 5.6. Diatomic linear chain

Now we divide by $e^{i(2j+1)bq}$ the first equation and by e^{i2jbq} the second:

$$\begin{cases} -M_1\omega^2(q)\xi = C\left[\eta e^{ibq} + \eta e^{-ibq} - 2\xi\right] \\ -M_2\omega^2(q)\eta = C\left[\xi e^{ibq} + \xi e^{-ibq} - 2\eta\right] \end{cases}$$

or

$$\begin{cases} (2C - M_1\omega^2(q))\xi - C2\cos(bq)\eta = 0 \\ -C2\cos(bq)\xi + (2C - M_2\omega^2(q))\eta = 0 \end{cases} \tag{5.15}$$

For nonzero solutions of this homogeneous system, its determinant must be zero:

$$(2C - M_1\omega^2(q))(2C - M_2\omega^2(q)) - 4C^2\cos(bq)\cos(bq) = 0.$$

This equation can be solved for the square of the frequency with straightforward algebra, leading to

$$\omega^2(q) = C\left(\frac{1}{M_1} + \frac{1}{M_2}\right)\left\{1 \mp \sqrt{1 - \frac{4M_1M_2}{(M_1+M_2)^2}\sin^2(bq)}\right\}. \tag{5.16}$$

The matrix of the elements containing the force constant in the system (5.15) is called the *dynamical matrix* of the crystal. In real three-dimensional crystals and more realistic models, it is much more complicated. The dynamical matrix is the basic quantity for the calculations of the dispersion relations of the vibrational modes of any crystal. It yields the potential energy of the crystal for any configuration of the atomic displacements, and its eigenvalues are strictly related to the frequency of each mode.

Acoustic and Optical Modes

Equation (5.16) shows that for a diatomic chain the dispersion relation has two branches. They are called *acoustic modes* and *optical modes*, for a reason that will be seen shortly. The branches above have been found for longitudinal modes and are indicated by LA and LO, respectively. There are of course four other transverse modes, two acoustic, called TA, and two optical, called TO.

To analyze the form of the dispersions in (5.16), let us first consider their small-q limits. For $qb \ll 1$, we have, using $\sqrt{1+\varepsilon} \approx 1 + \varepsilon/2$,

$$\omega^2(q) \approx \begin{cases} \frac{2C}{M_1+M_2}b^2q^2 & \text{ac}: \quad \omega_{ac}(q) \propto q \\ 2C\left(\frac{1}{M_1} + \frac{1}{M_2}\right) & \text{op}: \quad \omega_{op}(q) \text{ indep. of } q \end{cases} \qquad q \to 0 \tag{5.17}$$

Thus, the acoustic branch starts, at small q, with a linear relation as the vibrating string. The optical modes, instead, start with a constant frequency,

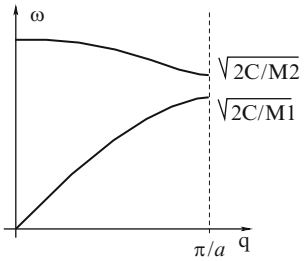


Fig. 5.7. Acoustic (lower) and optical (upper) branches of the vibrations of a diatomic linear chain

a completely different behavior. Optical modes are intrinsically related to the discrete nature of the chain composed by different atoms. The consideration of the reduction of the allowed modes inside the first BZ is still valid, and we can analyze the values of the two branches at the zone edge. For $q = \pm\frac{\pi}{a} = \pm\frac{\pi}{2b}$, after straightforward calculations, we find, for $M_1 > M_2$,

$$\omega_{ac}\left(\pm\frac{\pi}{a}\right) = \sqrt{\frac{2C}{M_1}}, \qquad \omega_{op}\left(\pm\frac{\pi}{a}\right) = \sqrt{\frac{2C}{M_2}}.$$

The full dispersion curves for acoustic and optical modes are schematically shown in Fig. 5.7.

To understand the nature of the two branches, let us go back to (5.15) for the displacements of the atoms. From those equations, it is immediately found the ratio between the displacements of the two types of atoms:

$$\frac{\xi}{\eta} = \frac{2C\cos(qb)}{2C - M_1\omega^2},$$

or, using (5.17) for small q,

$$\frac{\xi}{\eta} \sim \begin{cases} 1 & \text{for acoustic modes} \\ -\frac{M_2}{M_1} & \text{for optical modes} \end{cases} \qquad q \to 0$$

Thus, while the acoustic modes behave like the simple chain, in the optical modes the two different particles vibrate in opposite directions, as shown in part (d) of Fig. 5.2, keeping their center of mass fixed.

If the crystal has p atoms per unit cell, there are three acoustic branches, and $3(p-1)$ optical branches. The density of states is the same for each branch, and the total number of modes is again equal to the number of degrees of freedom.

When the two atoms carry opposite charges, as in ionic crystals, optical modes are associated with electromagnetic waves (in the infrared range at room temperature) as effect of the oscillating dipoles. This is the origin of the name of optical modes.

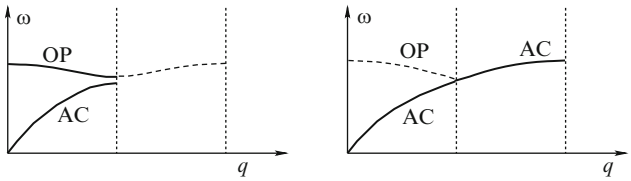


Fig. 5.8. Reduction of the two branches of acoustic and optical modes to the simple acoustic branch when the two masses of the chain become equal, with the doubling of the BZ

From the above discussion, it may seem that an unphysical discontinuity should occur when the mass M_2 is changed continuously until it becomes equal to M_1 since the lattice constant changes from the value $a = 2b$, as long as the two masses are different, to b, when M_2 becomes equal to M_1. The BZ should discontinuously double its size and the optical branch suddenly disappear. This, in a sense, is true. But, taking into account the periodicity of the dispersion relations, we may realize that the optical branch is present also outside of BZ, and when the masses become equal, the separation between the two branches at the zone edge vanishes (in a continuous way) and the optical branch becomes the continuation of the acoustic (and now only) branch in the double BZ, as illustrated in Fig. 5.8.

5.5 Three-Dimensional Lattice Vibrations

If we consider a monatomic three-dimensional crystal, three integer numbers are necessary to identify the equilibrium position of a particular atom, given by a vector of the direct lattice:

$$\boldsymbol{r_j} = j_1\boldsymbol{a}_1 + j_2\boldsymbol{a}_2 + j_3\boldsymbol{a}_3.$$

By a generalization of the result in (5.13), the displacements of the particles are given by

$$\boldsymbol{x_j} = \sum_{\boldsymbol{q},\ell} \mathbf{e}_{\boldsymbol{q},\ell} \left\{ \xi_{\boldsymbol{q}\ell} \mathrm{e}^{\mathrm{i}(\boldsymbol{q}\boldsymbol{r_j} - \omega_\ell(\boldsymbol{q})t)} + \xi^*_{\boldsymbol{q}\ell} \mathrm{e}^{-\mathrm{i}(\boldsymbol{q}\boldsymbol{r_j} - \omega_\ell(\boldsymbol{q})t)} \right\}. \tag{5.18}$$

The one-dimensional restrictions (5.9) become, in the three-dimensional case,

$$\begin{cases} N_1\boldsymbol{q}\boldsymbol{a}_1 = 2\pi n_1 \\ N_2\boldsymbol{q}\boldsymbol{a}_2 = 2\pi n_2 \\ N_3\boldsymbol{q}\boldsymbol{a}_3 = 2\pi n_3 \end{cases} \qquad n_1, n_2, n_3 = 0, \pm 1, \pm 2, \ldots, \tag{5.19}$$

where N_1, N_2, N_3 are the numbers of cells in the three directions. If the wavevectors $\boldsymbol{q}$ are written in terms of the unit vectors of the reciprocal lattice

as defined in Sect. 4.4,

$$\boldsymbol{q} = q_1\boldsymbol{b}_1 + q_2\boldsymbol{b}_2 + q_3\boldsymbol{b}_3,$$

the above relation (5.19) yield

$$q_1 = \frac{n_1}{N_1}, \quad q_2 = \frac{n_2}{N_2}, \quad q_3 = \frac{n_3}{N_3},$$

corresponding to (5.10) of the one-dimensional chain. The allowed wavevectors are then

$$\boldsymbol{q} = \frac{n_1}{N_1}\boldsymbol{b}_1 + \frac{n_2}{N_2}\boldsymbol{b}_2 + \frac{n_3}{N_3}\boldsymbol{b}_3, \qquad n_1, n_2, n_3 = 0, \pm 1, \pm 2, \ldots. \tag{5.20}$$

Let us note finally that if we add to an allowed vector in (5.20) a vector of the reciprocal lattice,

$$\boldsymbol{q}' = \boldsymbol{q} + s_1\boldsymbol{b}_1 + s_2\boldsymbol{b}_2 + s_3\boldsymbol{b}_3,$$

with s_i integers, the atom displacements in (5.18) do not change for the same reason illustrated in part (b) of Fig. 5.2 for the one-dimensional case. Thus, the only essential wavevectors $\boldsymbol{q}$ are those contained in a unit cell of the reciprocal lattice. For this purpose, it is convenient to consider $\boldsymbol{q}$ within the first BZ.

Density of States

From the foregoing analysis, it is clear that the allowed wavevectors $\boldsymbol{q}$ are those at the vertices of a parallelepiped in the reciprocal space with sides

$$\frac{1}{N_1}\boldsymbol{b}_1, \quad \frac{1}{N_2}\boldsymbol{b}_2, \quad \frac{1}{N_3}\boldsymbol{b}_3.$$

The volume δV_q of this parallelepiped is the volume of the unit cell, $\boldsymbol{b}_1 \cdot \boldsymbol{b}_2 \times \boldsymbol{b}_3$, given by (4.4), divided by $N = N_1N_2N_3$, the number of unit cells in the crystal. Thus, the density of allowed modes $\boldsymbol{q}$ in the reciprocal space, for each branch, is

$$g(\boldsymbol{q}) = \frac{1}{\delta V_q} = \frac{N}{(2\pi)^3/V_c} = \frac{V}{(2\pi)^3}, \tag{5.21}$$

where $V = NV_c$ is the total volume of the crystal. The total number of modes $\boldsymbol{q}$ is thus N per branch. There are three possible polarizations for the acoustic modes and three for the optical mode. If the basis of the unit cell of the crystal is formed by n atoms, the optical modes are $3(n-1)N$ and the acoustic modes are $3N$. Thus, the total number of normal modes of vibrations of a three-dimensional crystal are again equal to the total number $3N$ of the degrees of freedom of the lattice, as expected. The dispersion curves $\omega(\boldsymbol{q})$ are functions of the three-dimensional vector $\boldsymbol{q}$. The modes are purely longitudinal or purely transverse only for $\boldsymbol{q}$ oriented along directions of high symmetry.

The density of states in the reciprocal space, as given by (5.21), is constant. We are often interested, however, in the density of states in the space of frequency, that is, the number of modes with frequency between ω and $\omega + \mathrm{d}\omega$. This depends upon the dispersion relation $\omega(\boldsymbol{q})$. It is thus useful to start again from the simple atomic chain, for which we have a simple analytical dispersion, given by (5.8). The density of states in the reciprocal space for the one-dimensional case is given by (5.11). For symmetry reasons, the frequencies of the modes q and $-q$ coincide, as confirmed by (5.8). The number of allowed wavevectors between q and $q + \mathrm{d}q$ is given by

$$\mathrm{d}N = \frac{L}{2\pi}\mathrm{d}q.$$

The corresponding frequency range is

$$\mathrm{d}\omega = \frac{\mathrm{d}\omega}{\mathrm{d}q}\mathrm{d}q.$$

Thus, the number of modes in the frequency interval $\mathrm{d}\omega$, taking into account positive and negative q, is

$$\mathrm{d}N = 2 \times \frac{L}{2\pi}\mathrm{d}q = 2 \times \frac{L}{2\pi}\frac{1}{\mathrm{d}\omega/\mathrm{d}q}\mathrm{d}\omega.$$

The desired density of states $D(\omega)$ is then

$$D(\omega) = \frac{\mathrm{d}N}{\mathrm{d}\omega} = 2 \times \frac{L}{2\pi}\frac{1}{\mathrm{d}\omega/\mathrm{d}q}.$$

For the simple linear chain, using (5.8), this becomes

$$D_{lc} = 2 \times \frac{L}{2\pi}\frac{1}{a\sqrt{\frac{C}{M}}cos\left(\frac{qa}{2}\right)} = \frac{L}{\pi a\sqrt{\frac{C}{M}\left[1-\sin^2\left(\frac{qa}{2}\right)\right]}},$$

or, taking into account the dispersion relation in (5.8),

$$\mathcal{D}(\omega) = \frac{2L}{\pi a}\frac{1}{\sqrt{\omega_M^2 - \omega^2}}, \tag{5.22}$$

where ω_M is the maximum frequency $\omega_M = 2\sqrt{C/M}$. The behavior of the density of states in (5.8) is shown in Fig. 5.9. The divergence in $\omega = \omega_M$ is related to the zero value of the derivative of the dispersion curve at $\omega = \omega_M$, and is called a *van Hove singularity*.

If transverse modes are also considered, the density of states in the wavevector space (5.21) is three times larger (one per polarization) and the density of states for the transverse modes must be added to the expression in (5.22).

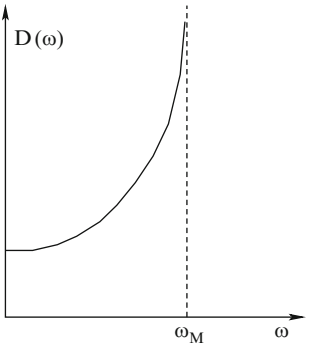


Fig. 5.9. Density of states per branch of a simple linear chain, with the van Hove singularity where the dispersion relation has zero derivative

In the case of the diatomic chain, also optical modes must be considered. Since their dispersion is almost constant, their density of states in the frequency space is strongly peaked.

In the three-dimensional case, ω is a function of the vector $\boldsymbol{q}$, not necessarily spherically symmetric as in isotropic materials. In the spherical case, the number of states per branch in $\mathrm{d}\omega$ is found considering the volume in $\boldsymbol{q}$ space between two spheres of radius q and $q + \mathrm{d}q$:

$$\mathrm{d}n = 4\pi q^2 \mathrm{d}q \frac{V}{(2\pi)^3} = 4\pi q^2 \frac{V}{(2\pi)^3} \frac{1}{\mathrm{d}\omega/\mathrm{d}q} \mathrm{d}\omega,$$

so that

$$D(\omega) = \frac{V q^2(\omega)}{2\pi^2 v_g},$$

where v_g is the group velocity $\mathrm{d}\omega/\mathrm{d}q$. This expression must then be summed over the different branches, acoustic and optical, longitudinal and transverse.

In the more general case, the surface $S(\omega)$ of constant $\omega(\boldsymbol{q})$ has a more complex shape, and the number of states between ω and $\omega + \mathrm{d}\omega$ is

$$\mathrm{d}n = \int_{S(\omega)} \frac{V}{(2\pi)^3} \, \mathrm{d}S \, \mathrm{d}q = \frac{V}{(2\pi)^3} \int_{S(\omega)} \frac{\mathrm{d}S}{|v_g|} \, \mathrm{d}\omega.$$

Thus,

$$D(\omega) = \frac{V}{(2\pi)^3} \int_{S(\omega)} \frac{\mathrm{d}S}{|v_g|}.$$

The modulus of v_g takes into account that the volume in $\boldsymbol{q}$ space is positive for frequencies both increasing and decreasing at increasing q. The density of states obtained in this way must be added for the different branches, and features van Hove singularities, as seen in the linear chain, every time ω has an extremum, corresponding to vanishing group velocity.

5.6 Normal Coordinates and Quantization – Phonons

Having in mind the expression (5.18) for the particle vibrations of a three-dimensional lattice, let us define:

$$\alpha_{\boldsymbol{q}\ell}(t) = \left(\frac{2MN}{\hbar}\omega_\ell(\boldsymbol{q})\right)^{\frac{1}{2}} \xi_{\boldsymbol{q}\ell} \mathrm{e}^{-\mathrm{i}\omega_\ell(\boldsymbol{q})t}. \tag{5.23}$$

In terms of these new coordinates, the displacements are given by

$$\boldsymbol{x_j} = \sum_{\boldsymbol{q},\ell} \mathbf{e}_{q,\ell} \left(\frac{\hbar}{2MN\omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}} \left\{\alpha_{\boldsymbol{q}\ell} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r_j}} + \alpha^*_{\boldsymbol{q}\ell} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\boldsymbol{r_j}}\right\}. \tag{5.24}$$

In the sum of the second terms, we may substitute $\boldsymbol{q}$ with $-\boldsymbol{q}$, and the sum runs over the same values since for each $\boldsymbol{q}$ there is a $-\boldsymbol{q}$:

$$\mathbf{x_j} = \sum_{\boldsymbol{q},\ell} \mathbf{e}_{q,\ell} \left(\frac{\hbar}{2MN\omega_\ell(\mathbf{q})}\right)^{\frac{1}{2}} \left\{\alpha_{\boldsymbol{q}\ell} + \alpha^*_{-\boldsymbol{q}\ell}\right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r_j}}. \tag{5.25}$$

The time derivatives of (5.23) are

$$\dot{\alpha}_{\boldsymbol{q}\ell} = -\mathrm{i}\omega_\ell(\boldsymbol{q})\alpha_{\boldsymbol{q}\ell}, \qquad \ddot{\alpha}_{\boldsymbol{q}\ell} = -\omega^2_\ell(\boldsymbol{q})\alpha_{\boldsymbol{q}\ell}. \tag{5.26}$$

These expressions show that the new variables satisfy the equation of motion of independent harmonic oscillators. They are the normal coordinates of our small-oscillation problem (see Sect. 1.5).

To find the Hamiltonian in terms of the new variables, we must write the energy associated with a generic vibration. To simplify the notation, we consider again the simple linear chain with only one polarization. The kinetic energy is

$$T = \frac{1}{2}M\sum_n \dot{x}_n^2$$

$$= \frac{1}{2}M\sum_{jqq'} \frac{\hbar}{2MN\sqrt{\omega(q)\omega(q')}} \left\{\dot{\alpha}_q + \dot{\alpha}^*_{-q}\right\} \left\{\dot{\alpha}_{q'} + \dot{\alpha}^*_{-q'}\right\} \mathrm{e}^{\mathrm{i}(q+q')ja}.$$

Using (5.26), the product of the α becomes

$$-\omega(q)\omega(q') \left\{\alpha_q\alpha_{q'} - \alpha_q\alpha^*_{-q'} - \alpha^*_{-q}\alpha_{q'} + \alpha^*_{-q}\alpha^*_{-q'}\right\}.$$

Furthermore, the sum over j is the sum of terms of a geometric series that yields

$$\sum_{j=0}^{N} \mathrm{e}^{\mathrm{i}(q+q')ja} = \frac{1-\mathrm{e}^{\mathrm{i}(q+q')Na}}{1-\mathrm{e}^{\mathrm{i}(q+q')a}} = \frac{1-\mathrm{e}^{\mathrm{i}2\pi(n+n')}}{1-\mathrm{e}^{\mathrm{i}\frac{2\pi}{N}(n+n')}}, \tag{5.27}$$

where, in the last step, we have taken into account the allowed values of q given in (5.10), and that $Na = L$. Since $n + n'$ is an integer, the numerator

is always zero so that the sum vanishes, unless also the denominator is zero. This happens only when $n + n'$ is zero,[2] i.e., $q = -q'$. In such case each term in the sum in (5.27) is unity, and the value of the sum is N. Collecting the above results, the kinetic energy takes the form

$$T = \sum_q \frac{\hbar\omega(q)}{4} \left\{ -\alpha_q \alpha_{-q} + \alpha_q \alpha_q^* + \alpha_{-q}^* \alpha_{-q} - \alpha_{-q}^* \alpha_q^* \right\}. \tag{5.28}$$

As it regards to the potential energy, if we limit ourselves to interactions between neighboring atoms, it is given by

$$V = \frac{1}{2} C \sum_j (x_{j+1} - x_j)^2$$

$$= \frac{1}{2} C \sum_{qq'} \left(\frac{\hbar}{2N\sqrt{\omega(q)\omega(q')}} \right) \left(\alpha_q + \alpha_{-q}^* \right) \left(\alpha_{q'} + \alpha_{-q'}^* \right) \sum_j \mathrm{e}^{\mathrm{i}(q+q')ja} \times$$

$$\times \left(\mathrm{e}^{\mathrm{i}(q+q')a} - \mathrm{e}^{\mathrm{i}qa} - \mathrm{e}^{\mathrm{i}q'a} + 1 \right).$$

The sum over j is evaluated as before; only $q = -q'$ contributes, and the last bracket becomes

$$\left(\mathrm{e}^{\mathrm{i}(q+q')a} - \mathrm{e}^{\mathrm{i}qa} - \mathrm{e}^{\mathrm{i}q'a} + 1 \right) = 2\left(1 - \cos(qa)\right) = 4\sin^2 \frac{qa}{2}.$$

The potential energy is then

$$V = \frac{1}{2} C \sum_q \frac{\hbar N}{2N\omega(q)} \left(\alpha_q \alpha_{-q} + \alpha_q \alpha_q^* + \alpha_{-q}^* \alpha_{-q} + \alpha_{-q}^* \alpha_q^* \right) 4\sin^2 \frac{qa}{2}.$$

The frequency is given by (5.8), so that

$$V = \sum_q \frac{\hbar\omega(q)}{4} \left(\alpha_q \alpha_{-q} + \alpha_q \alpha_q^* + \alpha_{-q}^* \alpha_{-q} + \alpha_{-q}^* \alpha_q^* \right).$$

Summing this to the kinetic energy in (5.28), we obtain the Hamiltonian in terms of the new variables:

$$H = T + V = \sum_q \frac{\hbar\omega(q)}{4} \left[\left\{ -\alpha_q \alpha_{-q} + \alpha_q \alpha_q^* + \alpha_{-q}^* \alpha_{-q} - \alpha_{-q}^* \alpha_q^* \right\} \right.$$

$$\left. + \left(\alpha_q \alpha_{-q} + \alpha_q \alpha_q^* + \alpha_{-q}^* \alpha_{-q} + \alpha_{-q}^* \alpha_q^* \right) \right],$$

[2] The case of $n + n'$ equal to a multiple of N, corresponds to $q + q' = G$, with G a vector of the reciprocal lattice, already included in $q + q' = 0$ with the substitution of q' with the equivalent $q' - G$.

or

$$H = \sum_q \frac{\hbar\omega(q)}{4} \left\{ +2\alpha_q \alpha_q^* + 2\alpha_{-q}^* \alpha_{-q} \right\} = \sum_q \hbar\omega(q) \left(\alpha_q \alpha_q^* \right),$$

where we have taken into account that the sums over q or $-q$ are the same. This Hamiltonian is separated for the different normal coordinates α_q, as wanted. It is still the classical Hamiltonian. To move to its quantum equivalent, first we must put it in the symmetric form:

$$H = \sum_q \hbar\omega(q) \frac{1}{2} \left(\alpha_q \alpha_q^* + \alpha_q^* \alpha_q \right).$$

Comparing this expression with the Hamiltonian for the harmonic oscillator in Appendix C, we substitute the variables α_q and α_q^* with the corresponding operators a_q and $\mathsf{a}_q^\dagger$, with commutator

$$\left[\mathsf{a}_q, \mathsf{a}_q^\dagger \right] = 1. \tag{5.29}$$

The Hamiltonian becomes the sum of Hamiltonians of independent harmonic oscillators:

$$\mathcal{H} = \sum_q \hbar\omega(q) \left(\mathsf{a}_q^\dagger \mathsf{a}_q + \frac{1}{2} \right). \tag{5.30}$$

In quantum theory, the vibrations of a crystal lattice are therefore quantized. The excitations, with energy

$$\epsilon(q) = \hbar\omega(q),$$

are called *phonons*. q indicates the phonon wavevector; $\mathsf{a}_q^\dagger$ and a_q are the creation and annihilation operators for the phonon mode q. The operator $\mathcal{N}_q = \mathsf{a}_q^\dagger \mathsf{a}_q$ is the *number operator*, whose eigenvalues are the integers $N_q = 0, 1, 2, \ldots$. The eigenvalues of the total energy are therefore

$$\epsilon_{tot} = \sum_q \hbar\omega(q) \left(N_q + \frac{1}{2} \right).$$

The above result is easily extended to longitudinal and transverse modes, to three-dimensional crystals and to crystals with acoustic and optical branches.

The quantum theory of lattice vibrations has a very important result: in the ground state, that is, at T = 0, the energy associated with each mode is not zero, but is $1/2(\hbar\omega(\boldsymbol{q}))$. This means that even at zero absolute temperature the crystal vibrates. This *zero point vibration* is well detectable experimentally and has important physical consequences.

5.7 Phonon Momentum and Crystal Momentum

For any nonzero wavevector $\boldsymbol{q}$, along a wavelength there are atoms of equal masses which vibrate with velocities that are opposite at any time. Thus, the total momentum of a phonon is zero. This result can be derived analytically by evaluating the velocity, and therefore the momentum, of the particles as time derivatives of the atom displacements given in (5.18), and then summing over all particles. For $q = 0$, this is not true because all atoms moves with equal velocities, but this case corresponds to a translation of the crystal as a whole without an internal restoring force.[3] This case, therefore, should not be really considered as a phonon mode.

On the other hand, we shall see in the following chapters that it is convenient to consider a *crystal momentum* of the phonons, defined as $\hbar\boldsymbol{q}$, since a conservation law holds in crystals for such quantity. The continuous translational symmetry of the Hamiltonian in vacuum implies the momentum conservation law. Similarly, the discrete translational symmetry of the Hamiltonian in crystals implies a conservation law for the crystal momentum, to within a vector $\boldsymbol{G}$ of the reciprocal lattice. This is coherent with the fact that the $\boldsymbol{q}$ themselves are defined to within a vector $\boldsymbol{G}$. The energy of the phonons, on the contrary, is well defined owing to the periodicity of the dispersion relations. The fact that the crystal momentum is a good quantum number and the real momentum is not, is related to the virtually infinite mass of the whole crystal, which can exchange any amount of real momentum with electrons and phonons.

5.8 Experimental Determination of Phonon Dispersions

When radiation, for example light or neutrons, is scattered by a crystal, it can emit or absorb phonons, so that the analysis of the frequency and momentum of the scattered radiation yields information on the phonon dispersion curves. The scattering of electromagnetic radiation from crystals takes different names according to the different processes involved: if light is scattered by acoustic phonons we have *Brillouin scattering*. The spectrum of the light detected after the collision with the crystal shows a central peak with the same frequency of the original beam and two lateral peaks corresponding to the light scattered with a phonon emission or a phonon absorption, respectively. If the light is scattered by optical phonons, we have *Raman scattering*, and the two lateral lines are called *Stokes* for phonon absorption and *anti-Stokes* for phonon emission. This is not the place to discuss this phenomenon and the relative experimental techniques. We simply mention that these experiments, in particular neutron scattering, are nowadays a standard technique to determine the vibrational properties of the crystalline materials.

[3] In optical modes, the two sublattices vibrate in opposite directions keeping the center of mass fixed.

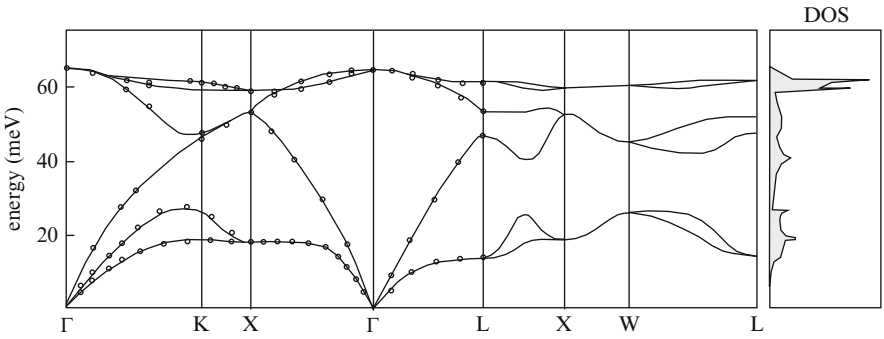


Fig. 5.10. Theoretical (*lines*) and experimental (*circles*) phonon dispersions, and density of states, in silicon, from [163]. Letters indicate the points of high symmetry in the Brillouin zone defined in Fig. 4.4

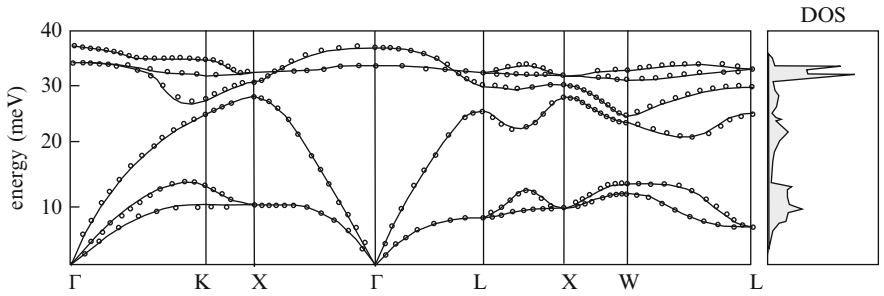


Fig. 5.11. Theoretical (*lines*) and experimental (*circles*) phonon dispersions, and density of states, in gallium arsenide, from [163]. Letters indicate the points of high symmetry in the Brillouin zone defined in Fig. 4.4

The results are in good agreement with theoretical predictions obtained in [163] by first-principle calculations applying a *density-functional theory* (DFT).[4] In [163], the energy of a crystal and its gradient are obtained, via

[4] *Density functional theory.* Given an ensemble of electrons subject to an external potential $V(\boldsymbol{r})$ and to their mutual interaction, Hohenberg and Kohn [191] have shown that the expression that yields the electron density as a functional of $V(\boldsymbol{r})$ can be inverted. This means that two equal density functions cannot be the result of different potentials and that all quantities of interest of a many-body electron system may be obtained if the electron density is known. This is an enormous simplification of the theoretical problem of finding the properties of a many-body system, since the density is a function of only one $\boldsymbol{r}$, and not of as many positions as particles in the system. DFT has become particularly effective after Kohn and Sham theorem [249]. According to this theorem, the solution of a many-body problem, via DFT, can be obtained solving a one-particle Schrödinger equation with a Hartree term and an (unknown) exchange-correlation term. The correctness of the solution depends on the quality of the guess for this last term,

DFT, as function of the configuration of the atoms, thus obtaining the dynamical matrix whose diagonalization leads to the phonon spectra. A detailed description of the application of DFT to lattice dynamics calculations is given in [29].

The phonon spectra for the two main semiconductors, silicon and gallium arsenide, are shown in Figs. 5.10 and 5.11, respectively.

for which several approaches and approximations have been given. For a general treatment of DFT, see, for example, [121].

6

Bloch States and Band Theory

The topic of central interest in this book, i.e., electron transport in semiconductors, deals with the response to external forces of electrons that are able to move inside a semiconductor. Therefore, the determination of the states available to the electrons in the crystal and their energies is of fundamental importance. The knowledge of electron states in solids is essential also for the analysis of their optical properties. It is not surprising, therefore, that a great effort has been devoted in the past to develop both experimental and theoretical techniques able to yield information on this subject. The problem is not an easy one, however, owing to its many-body nature. Here, we shall review very briefly the main ideas and techniques, starting from the fundamental theorem, known in solid-state physics as *Bloch theorem* and in mathematical analysis as *Floquet theorem*. Demonstrated by mathematicians at the end of the nineteenth century, it leads to the definition of *Bloch states*, the basic bricks of any theoretical study of electronic properties of solids.

6.1 Bloch Theorem

A correct analysis of the electron dynamics in crystals requires the application of advanced many-body techniques. For the present purpose, however, we may limit ourselves to consider a single electron subject to the potential due to the nuclei, to the core electrons and to the average interaction with all the other external electrons. Let us then consider the Schrödinger equation for the Hamiltonian eigenfunctions of a particle of mass m subject to a potential $V_{cr}(\boldsymbol{r})$ with the periodicity of the direct lattice:

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\boldsymbol{r}) + V_{cr}(\boldsymbol{r})\psi(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}). \tag{6.1}$$

In terms of the unit vectors $\boldsymbol{a}_i$ of the lattice in (4.3), each position $\boldsymbol{r}$ can be written as

$$\boldsymbol{r} = \xi_1 \boldsymbol{a}_1 + \xi_2 \boldsymbol{a}_2 + \xi_3 \boldsymbol{a}_3.$$

The potential energy V_{cr} is therefore a periodic function of the three variables (ξ_1, ξ_2, ξ_3) and can be expanded in Fourier series as

$$V_{cr}(\xi_1, \xi_2, \xi_3) = \sum_{m_1 m_2 m_3} V_{m_1 m_2 m_3} \mathrm{e}^{\mathrm{i} m_1 2\pi \xi_1 / a_1} \mathrm{e}^{\mathrm{i} m_2 2\pi \xi_2 / a_2} \mathrm{e}^{\mathrm{i} m_3 2\pi \xi_3 / a_3}.$$

If we now consider the vector $\boldsymbol{G}$ of the reciprocal lattice defined in (4.3), we have, taking into account the relations (4.2) between the vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$,

$$\boldsymbol{G} \cdot \boldsymbol{r} = (m_1 \boldsymbol{b}_1 + m_2 \boldsymbol{b}_2 + m_3 \boldsymbol{b}_3) \cdot (\xi_1 \boldsymbol{a}_1 + \xi_2 \boldsymbol{a}_2 + \xi_3 \boldsymbol{a}_3)$$

$$= m_1 \frac{\boldsymbol{b}_1 \boldsymbol{a}_1}{a_1} \xi_1 + m_2 \frac{\boldsymbol{b}_2 \boldsymbol{a}_2}{a_2} \xi_2 + m_3 \frac{\boldsymbol{b}_3 \boldsymbol{a}_3}{a_3} \xi_3 = \frac{2\pi}{a_1} m_1 \xi_1 + \frac{2\pi}{a_2} m_2 \xi_2 + \frac{2\pi}{a_3} m_3 \xi_3.$$

Therefore, putting $V_{m_1 m_2 m_3} = V(\boldsymbol{G})$,

$$\boxed{V_{cr}(\boldsymbol{r}) = \textstyle\sum_{\boldsymbol{G}} V(\boldsymbol{G}) \mathrm{e}^{\mathrm{i} \boldsymbol{G} \boldsymbol{r}}} \tag{6.2}$$

More generally, a function with the periodicity of the direct lattice can be expanded in Fourier series with wavevectors of the reciprocal lattice. Inversely, if the Fourier expansion of a function contains only wavevectors of the reciprocal lattice, the function is periodic with the periodicity of the direct lattice. This result is of paramount importance and is used very often in solid-state physics.

If we put (6.2) into the Schrödinger equation (6.1), we obtain

$$-\frac{\hbar^2}{2m} \nabla^2 \psi(\boldsymbol{r}) + \sum_{\boldsymbol{G}} V(\boldsymbol{G}) \mathrm{e}^{\mathrm{i} \boldsymbol{G} \boldsymbol{r}} \psi(\boldsymbol{r}) = \epsilon \psi(\boldsymbol{r}).$$

Let us now expand the wavefunction in Fourier series,

$$\psi(\boldsymbol{r}) = \sum_{\boldsymbol{k}'} C(\boldsymbol{k}') \mathrm{e}^{\mathrm{i} \boldsymbol{k}' \boldsymbol{r}} \tag{6.3}$$

and obtain

$$\sum_{\boldsymbol{k}'} \frac{\hbar^2 k'^2}{2m} C(\boldsymbol{k}') \mathrm{e}^{\mathrm{i} \boldsymbol{k}' \boldsymbol{r}} + \sum_{\boldsymbol{G} \boldsymbol{k}'} V(\boldsymbol{G}) C(\boldsymbol{k}') \mathrm{e}^{\mathrm{i} (\boldsymbol{k}' + \boldsymbol{G}) \boldsymbol{r}} = \epsilon \sum_{\boldsymbol{k}'} C(\boldsymbol{k}') \mathrm{e}^{\mathrm{i} \boldsymbol{k}' \boldsymbol{r}}.$$

In the second sum, we may substitute $\boldsymbol{k}'$ with $\boldsymbol{k}'' = \boldsymbol{k}' + \boldsymbol{G}$ and the sum over $\boldsymbol{k}''$ runs over the reciprocal space exactly as $\boldsymbol{k}'$ so that we may keep the same

symbol. The equation becomes

$$\sum_{\boldsymbol{k}'} \frac{\hbar^2 k'^2}{2m} C(\boldsymbol{k}') \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} + \sum_{\boldsymbol{G}\boldsymbol{k}'} V(\boldsymbol{G}) C(\boldsymbol{k}' - \boldsymbol{G}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} = \epsilon \sum_{\boldsymbol{k}'} C(\boldsymbol{k}') \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}}.$$

Equating the Fourier coefficients, we obtain

$$\left[\frac{\hbar^2 k'^2}{2m} - \epsilon \right] C(\boldsymbol{k}') + \sum_{\boldsymbol{G}} V(\boldsymbol{G}) C(\boldsymbol{k}' - \boldsymbol{G}) = 0. \tag{6.4}$$

Of all the Fourier coefficients in (6.3), this equation connects only those with wavevectors that differ of a vector $\boldsymbol{G}$ of the reciprocal lattice, as shown in Fig. 6.1. Thus, given a $\boldsymbol{k}$ in the first BZ, we may write an equation like (6.4) for each $\boldsymbol{k}'$ that differs from $\boldsymbol{k}$ for a vector $\boldsymbol{G}$. The differential Schrödinger equation has been therefore transformed into an infinite set of linear algebraic equations, one set for each $\boldsymbol{k}$. We may therefore consider wavefunctions whose Fourier series contains only one $\boldsymbol{k}$ and all the $\boldsymbol{k} + \boldsymbol{G}$. These are the *Bloch functions*, written as

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} C(\boldsymbol{k} + \boldsymbol{G}) \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G})\boldsymbol{r}},$$

or

$$\boxed{\psi_{\boldsymbol{k}}(\boldsymbol{r}) = u_{\boldsymbol{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}} \tag{6.5}$$

where

$$u_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} C(\boldsymbol{k} + \boldsymbol{G}) \mathrm{e}^{\mathrm{i}\boldsymbol{G}\boldsymbol{r}}. \tag{6.6}$$

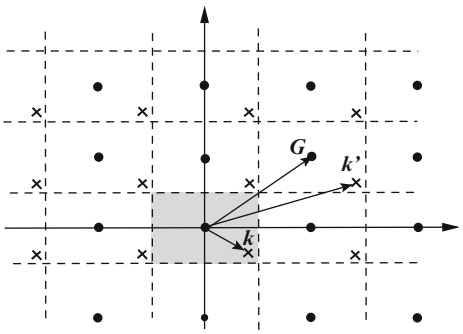


Fig. 6.1. Wavevectors of the plane waves that form a Bloch state. Dots indicate the vectors $\boldsymbol{G}$ of the reciprocal lattice; $\boldsymbol{k}$ in the first (*shaded*) BZ is the Bloch wavevector appearing in (6.5), and all $\boldsymbol{k}'$ (indicated by *crosses*), obtained by adding any $\boldsymbol{G}$ to $\boldsymbol{k}$, are the wavevectors involved in the Fourier expansion of $u_{\boldsymbol{k}}(\boldsymbol{r})$. Larger $\boldsymbol{G}$s yield faster oscillations of $u_{\boldsymbol{k}}(\boldsymbol{r})$ in space

Since the Fourier series in (6.6) contains only wavevectors of the reciprocal lattice, $u_{\boldsymbol{k}}(\boldsymbol{r})$ is a periodic function with the period of the direct lattice. $\boldsymbol{k}$ in (6.5), multiplied by $\hbar$, is called the *crystal momentum* of the Bloch state.

The Fourier expansion of $u_{\boldsymbol{k}}(\boldsymbol{r})$ may be truncated at values of $\boldsymbol{G}$ large enough for all the oscillations of the wavefunction to be correctly accounted for. The system in (6.4) is then reduced to a finite system, and the solutions of the secular equation are the energy eigenvalues $\epsilon_n(\boldsymbol{k})$ of our Hamiltonian, corresponding to the vector $\boldsymbol{k}$.

If all $\boldsymbol{k}$s are considered, the eigenvalues $\epsilon_n(\boldsymbol{k})$ distribute themselves in intervals of allowed values called *energy bands*, separated by gaps, called *energy band gaps*.

If we had started with a $\boldsymbol{k}$ outside of the BZ, we would have obtained the same eigenvalues and eigenfunctions. Thus, the $\boldsymbol{k}$s that label the wavefunctions are defined to within vectors of the reciprocal lattice, and the function $\epsilon_n(\boldsymbol{k})$ is periodic with the period of the reciprocal lattice. To take $\boldsymbol{k}$ inside, the first BZ is a question of opportunity. The first BZ for real crystals is of course three-dimensional, and part (c) of Fig. 4.4 shows the first BZ of fcc lattices (of most common semiconductors).

6.2 Density of States

If we assume periodic boundary conditions for our Bloch states, we have (in a one-dimensional case)

$$\psi_k(0) = u_k(0) = \psi_k(L) = u_k(L)\mathrm{e}^{\mathrm{i}kL}.$$

Owing to the periodicity of u, this is satisfied when

$$\mathrm{e}^{\mathrm{i}kL} = 1, \qquad or \qquad k = \frac{2\pi}{L}n, \quad n = 0, \pm 1, \pm 2, ...$$

As in the case of phonons, the distance between successive ks in the reciprocal axis is $\delta k = 2\pi/L$, which corresponds to a density of states given by

$$g(k) = \frac{1}{\delta k} = \frac{L}{2\pi}$$

for each band. To find the total number of states per band in the BZ, we remember that the length of the BZ in one dimension is $2\pi/a$, the density of states is $L/2\pi$ so that the wanted number is

$$\frac{2\pi/a}{(2\pi/Na)} = N.$$

This result is identical in three dimensions, where

$$g(\boldsymbol{k}) = \frac{V}{(2\pi)^3}. \tag{6.7}$$

We may thus conclude that the number of Bloch states in each band in the first BZ is equal to the number of atoms in the crystal. This is coherent with the tight-binding picture (see below), in which each atomic electron state contributes to the formation of a Bloch state in the band.

In addition to what seen above, we have a factor of 2 in the density of states for the spin degeneracy.

In a macroscopic crystal, the distance between two adjacent $\boldsymbol{k}$ wavevectors is very small, i.e., the density of states is very large. If we have to sum a regular function $f(\boldsymbol{k})$ over the states in the BZ, we may assume that in a small interval $\Delta\boldsymbol{k}$ the function is constant, and a number of states given by $g(\boldsymbol{k})\Delta\boldsymbol{k}$ contribute to the sum. When $\Delta\boldsymbol{k}$ becomes sufficiently small, the sum may be substituted by the Riemann integral, yielding the practical rule

$$\boxed{\sum_k \to \frac{V}{(2\pi)^3} \int \mathrm{d}\boldsymbol{k}} \tag{6.8}$$

The set of all Bloch states (all $\boldsymbol{k}$s in the BZ and all bands) form a complete base.

6.3 Tight-Binding Approach

We have seen that crystals are periodic arrangements of atoms. Electrons in the deepest atomic levels, forming the so-called atomic *cores*, can be described as occupying their atomic levels without big modifications, since their wavefunctions do not reach in an appreciable way the neighboring atoms. More external electron states, in particular those that participate in the chemical bond, cannot be described in this simple way and require a study that takes into account the presence of the nearby atoms.

There is a very intuitive approach to the problem of electron states, called *tight-binding* approach, which exploits the above idea: let us consider, as a starting point, a fictitious crystal that has the same structure of the real one, but with an arbitrarily large lattice constant. Such a system can be described as a set of isolated atoms, so that atomic wavefunctions represent exact states for all electrons.

We then let the lattice constant decrease. At a certain distance, the most external electrons start to feel the presence of the neighboring atoms. The atomic wavefunctions overlap and form states extended over the entire crystal. Equivalent atomic orbitals, which are degenerate when far apart, split into different levels as consequence of the interaction. Starting from the most outer states, as the atoms get closer, single levels become narrow energy bands that increase and eventually overlap, as shown in Fig. 6.2. The actual lattice constant is determined by the minimum of the total energy of the crystal.

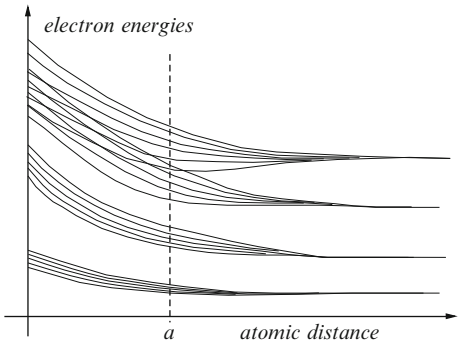


Fig. 6.2. Schematic representation of the formation of the energy bands, as the distance between atoms decreases, in a tight-binding approach. Electrons in deeper core levels are affected less by adjacent atoms and at smaller distances. The actual lattice constant a is determined by the minimum of the energy of the crystal due to occupied states

6.4 Band-Structure Calculations

As indicated above, the calculation of the eigenvalues and eigenstates of the Hamiltonian of the electrons in a crystal is the first basic step for the theoretical understanding of most properties of a solid. Many methods have been developed for the solution of such a problem. Here, we can only mention the most important of them since this subject is somewhat outside of our scope.

In general terms, the first problem to solve consists in the reduction of the many-body problem to a one-particle equation. For this purpose, some sort of *mean-field approximation* is to be performed, where the effect of all "other" electrons is embodied in an effective potential acting on each single electron. Self-consistency is obviously required, since the solution of the single-particle equation for one electron will affect the mean field acting on the others. In the *Hartree–Fock approximation* (see, e.g., [18, 145]), the mean potential acting on each electron is that given by the charge density of all other electrons, including the exchange effects due to antisymmetrization of the many-fermion wavefunction. More often, in recent times, the DFT approach is used, mentioned in the note at the end of Sect. 5.8. Often the potential to use in the one-particle Schrödinger equation is given a particular analytical form on the basis of general theoretical considerations, with some parameters fixed *a posteriori* by the comparison of the theoretical results with experimental data.

Many computer programs have been developed for band-structure calculations, some of which are commercially available.

6.4.1 LCAO Method

The method of *linear combination of atomic orbitals* (LCAO) is the immediate quantitative application of the idea of tight binding introduced above, and, in fact, the two phrases are often used as synonyms. This method was developed at the very beginning of the development of quantum physics and is particularly useful for the calculations of the electron states in molecules, the so-called *molecular orbitals* of quantum chemistry.

Since we are interested here only in the basic idea of the method, let us assume, for simplicity, that we are dealing with a simple solid with one atom per unit cell, and let $\phi_n(\boldsymbol{r})$ be the n-th atomic wavefunction centered in the origin. We may then generate a linear combination of such functions centered at the lattice sites $\boldsymbol{R}$:

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{R}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} \phi_n(\boldsymbol{r} - \boldsymbol{R}). \tag{6.9}$$

The reader is invited to show, as an exercise, that this expression is of the form required by Bloch theorem.

The development may proceed with perturbation theory (see Appendix E), considering as perturbation the difference between the crystal potential and the atomic potential. The resulting energy eigenvalues are expressed in terms of the matrix elements of the perturbation between various atomic orbitals. Such matrix elements may then be used to adjust the results to experimental data.[1]

For a better approximation, the atomic wavefunctions in (6.9) are substituted by a linear combination of different atomic wavefunctions with energies close to that of the considered level. This is important, in particular, in case of degenerate levels.

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{R}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} \phi(\boldsymbol{r} - \boldsymbol{R}) \ , \qquad \phi(\boldsymbol{r}) = \sum_n b_n \phi_n(\boldsymbol{r}). \tag{6.10}$$

Substituting the above wavefunction into the Schrödinger equation and expressing this equation on the basis of the atomic wavefunctions, a number of equations are obtained for the unknown coefficients b_n [18].

It can be shown that the exact Bloch wavefunction can be put in the form given by the first part of (6.10). This equation, in fact, may be inverted and the resulting functions $\phi(\boldsymbol{r})$ are called *Wannier functions* and form a

[1] As an example, in the simple cubic lattice, near $k = 0$, accounting for overlap of the atomic wavefunctions only between nearest-neighbor atoms, an isotropic parabolic band is obtained with an effective mass $m^* = \hbar^2/2\gamma a^2$ where $-\gamma$ is the *overlap integral*, i.e., the matrix element of the perturbation between two adjacent atoms and a the lattice constant. It is clear that the band becomes closer and closer to the atomic level (flat band, infinite effective mass) as the overlap of the atomic wavefunction decreases [241].

complete orthonormal basis. They are concentrated around the lattice sites and if the tight-binding is a good approximation, they are expected to be similar to the atomic wavefunctions, but they are not eigenfunctions of the crystal Hamiltonian, being linear combinations of Bloch states belonging to different energies.

6.4.2 $\boldsymbol{k} \cdot \boldsymbol{p}$ Method

If we substitute a Bloch function into the Schrödinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\boldsymbol{r}) + V_{cr}(\boldsymbol{r})\psi(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}),$$

we obtain the equation for the periodic part

$$\left(\frac{p^2}{2m} + \frac{\hbar}{m}\boldsymbol{k}\cdot\boldsymbol{p} + \frac{\hbar^2 k^2}{2m} + V_{cr}(\boldsymbol{r})\right) u_{n\boldsymbol{k}}(\boldsymbol{r}) = \epsilon_n(\boldsymbol{k}) u_{n\boldsymbol{k}}(\boldsymbol{r}), \tag{6.11}$$

where the band index n has been added. At $\boldsymbol{k} = 0$, the above equation becomes the same equation satisfied by the total wavefunction:

$$\left(\frac{p^2}{2m} + V_{cr}(\boldsymbol{r})\right) u_{n0}(\boldsymbol{r}) = \epsilon_n(\boldsymbol{k}) u_{n0}(\boldsymbol{r}). \tag{6.12}$$

If we are interested in the electronic states near $\boldsymbol{k} = 0$ (point Γ), the two terms containing k in (6.11) can be treated as a perturbation. For this, however, we need the eigenfunctions at Γ. On the other hand, (6.12) for such functions is much easier to solve than the general Schrödinger equation. As expected, the states obtained with this method are rather good for small $\boldsymbol{k}$, although the method can be extended to expand the band structure around any given value $\boldsymbol{k}_\circ$. In particular, the effective masses (i.e., the curvatures of the bands around their minima, as described in Sect. 6.6), obtained with this method, with simple approximations for the wavefunctions at Γ, are in good agreement with the experimental values [483].

The $\boldsymbol{k} \cdot \boldsymbol{p}$ method has been applied to obtain the band structures of Ge and Si in [93].

6.4.3 Pseudopotential Method

The concept of pseudopotential was introduced by Fermi in 1934 [139] for the analysis of atomic wavefunctions. When it was proposed for solids by Phillips and Kleinman in 1959 [339, 340], it generated a significant improvement in the calculations of band structures and it is today the approach most used for this purpose, often in connection with the density functional theory.

The main idea of the method is based on the fact that a Bloch state must have rapid oscillations near the crystal nuclei, where the electron kinetic

energy is large, while in the outer region between the nuclei is much smoother. If we expand the Bloch states in plane waves, a very large number of them will be necessary to reproduce correctly these oscillations. On the other hand, we are not particularly interested in these oscillations since valence and conduction properties are mainly due to electron wavefunctions in the outer regions of the atoms. Furthermore, for Pauli exclusion principle, outer electrons spend little time (small wavefunctions) in the region near the nuclei, already occupied by the core electrons. The idea is therefore to find smooth approximate wavefunctions that yield correct valence and conduction bands and correct electron densities in the regions far from the inner cores of the atoms as shown in Fig. 6.3.

To realize the above project, let us start from the expansion of the desired Bloch state in (6.5) in plane waves. Owing to the periodicity of the function $u_{\boldsymbol{k}}(\boldsymbol{r})$ this expansion has the form

$$|\psi_{\boldsymbol{k}}\rangle = \sum_{\boldsymbol{G}} C(\boldsymbol{G})|\phi_{\boldsymbol{k}+\boldsymbol{G}}\rangle,$$

where $|\phi_{\boldsymbol{k}+\boldsymbol{G}}\rangle$ is the state represented by the plane wave of wavevector $(\boldsymbol{k}+\boldsymbol{G})$. We then recall that the oscillations of an outer Bloch state make it orthogonal to the core Bloch states, and large values of $\boldsymbol{G}$ are necessary to obtain such oscillations. However, if we subtract from a plane wave its part "parallel" to a core Bloch state, the resulting function, called *orthogonalized plane wave* (OPW), should already contain at least part of these oscillations. Let us then define the OPWs as

$$|\phi^{opw}_{\boldsymbol{k}}\rangle = |\phi_{\boldsymbol{k}}\rangle - \sum_j a_j|c_{j\boldsymbol{k}}\rangle \ , \qquad a_j \equiv \langle c_{j\boldsymbol{k}}|\phi_{\boldsymbol{k}}\rangle, \tag{6.13}$$

where $|c_{j\boldsymbol{k}}\rangle$ is the j-th core Bloch state. Now we consider a trial wavefunction for the desired Bloch state given by a combination of the above OPWs.

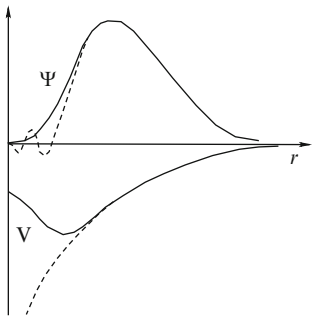


Fig. 6.3. Schematic representation of the comparison between pseudo wavefunction and pseudopotential (*continuous lines*), and true wavefunction and potential (*dashed lines*)

Since they are already orthogonal to the core states and already contain the oscillations near the nuclei, a few of them should be enough to give a good approximation of the correct Bloch state.

$$|\psi_{\boldsymbol{k}}\rangle = \sum_{\boldsymbol{G}} d(\boldsymbol{G})|\phi^{opw}_{\boldsymbol{k}+\boldsymbol{G}}\rangle. \tag{6.14}$$

If we now write the same combination of plane waves, we obtain the *pseudo wavefunction*,

$$|\tilde{\psi}_{\boldsymbol{k}}\rangle = \sum_{\boldsymbol{G}} d(\boldsymbol{G})|\phi_{\boldsymbol{k}+\boldsymbol{G}}\rangle. \tag{6.15}$$

Since the core Bloch states are significantly different from zero only near the core regions, these pseudo wavefunctions are good approximations of the true wavefunctions outside the core regions and do not have unnecessary oscillations inside them. If we substitute the definition (6.13) of the OPW into the trial wavefunction (6.14), we obtain

$$\begin{aligned}|\psi_{\boldsymbol{k}}\rangle &= \sum_{\boldsymbol{G}} d(\boldsymbol{G})\left[|\phi_{\boldsymbol{k}+\boldsymbol{G}}\rangle - \sum_j \langle c_{j\boldsymbol{k}}|\phi_{\boldsymbol{k}+\boldsymbol{G}}\rangle|c_{j\boldsymbol{k}}\rangle\right] \\ &= |\tilde{\psi}_{\boldsymbol{k}}\rangle - \sum_j |c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}| \sum_{\boldsymbol{G}} d(\boldsymbol{G})|\phi_{\boldsymbol{k}+\boldsymbol{G}}\rangle = \left[1 - \sum_j |c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}|\right]|\tilde{\psi}_{\boldsymbol{k}}\rangle.\end{aligned}$$

Now we substitute this expression into the Schrödinger equation:

$$\mathcal{H}|\psi_{\boldsymbol{k}}\rangle = \epsilon(\boldsymbol{k})|\psi_{\boldsymbol{k}}\rangle,$$

obtaining

$$\mathcal{H}|\tilde{\psi}_{\boldsymbol{k}}\rangle - \sum_j \mathcal{H}|c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}|\tilde{\psi}_{\boldsymbol{k}}\rangle = \epsilon(\boldsymbol{k})|\tilde{\psi}_{\boldsymbol{k}}\rangle - \sum_j \epsilon(\boldsymbol{k})|c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}|\tilde{\psi}_{\boldsymbol{k}}\rangle.$$

Take into account that the core Bloch states are eigenstates of the Hamiltonian with eigenvalues $\epsilon_j(\boldsymbol{k})$ and obtain

$$\mathcal{H}|\tilde{\psi}_{\boldsymbol{k}}\rangle + \sum_j \left(\epsilon(\boldsymbol{k}) - \epsilon_j(\boldsymbol{k})\right)|c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}|\tilde{\psi}_{\boldsymbol{k}}\rangle = \epsilon(\boldsymbol{k})|\tilde{\psi}_{\boldsymbol{k}}\rangle,$$

or

$$\left[\mathcal{H} + \sum_j \left(\epsilon(\boldsymbol{k}) - \epsilon_j(\boldsymbol{k})\right)|c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}|\right]|\tilde{\psi}_{\boldsymbol{k}}\rangle = \epsilon(\boldsymbol{k})|\tilde{\psi}_{\boldsymbol{k}}\rangle.$$

This equation shows that the pseudo wavefunctions obey a Schrödinger equation

$$\left[\mathcal{H} + \mathcal{V}_{\text{eff}}\right]|\tilde{\psi}_{\boldsymbol{k}}\rangle = \epsilon(\boldsymbol{k})|\tilde{\psi}_{\boldsymbol{k}}\rangle, \tag{6.16}$$

with exact eigenvalues and with an effective potential

$$\mathcal{V}_{\text{eff}} = \sum_j \left(\epsilon(\boldsymbol{k}) - \epsilon_j(\boldsymbol{k})\right) |c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}| \tag{6.17}$$

added to the crystal Hamiltonian. The pseudo wavefunctions are smooth also in the core region, but are pushed away from this region by the effective potential that results repulsive, owing to the energy differences in (6.17) that are all positive, and *nonlocal*. In fact, to see the effect of the operator $\mathcal{V}_{\text{eff}}$ on a function $f(\boldsymbol{r})$, we must perform the following calculation

$$\langle \boldsymbol{r}|\mathcal{V}_{\text{eff}}|f\rangle = \langle \boldsymbol{r}| \sum_j \left(\epsilon(\boldsymbol{k}) - \epsilon_j(\boldsymbol{k})\right) |c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}}|f\rangle.$$

Insert the identity $1 = \int |\boldsymbol{r}'\rangle \mathrm{d}\boldsymbol{r}'\langle \boldsymbol{r}'|$ before the last ket and obtain

$$\begin{aligned}\langle \boldsymbol{r}|\mathcal{V}_{\text{eff}}|f\rangle &= \langle \boldsymbol{r}| \sum_j \left(\epsilon(\boldsymbol{k}) - \epsilon_j(\boldsymbol{k})\right) |c_{j\boldsymbol{k}}\rangle\langle c_{j\boldsymbol{k}} \int |\boldsymbol{r}'\rangle \mathrm{d}\boldsymbol{r}'\langle \boldsymbol{r}'|f\rangle \\ &= \sum_j \left(\epsilon(\boldsymbol{k}) - \epsilon_j(\boldsymbol{k})\right) c_{j\boldsymbol{k}}(\boldsymbol{r}) \int c^*_{j\boldsymbol{k}}(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}' f(\boldsymbol{r}'),\end{aligned}$$

which shows that the effect of $\mathcal{V}_{\text{eff}}$ on the function in $\boldsymbol{r}$ depends upon the values of the same function in all points.

The full potential,

$$\mathcal{V}_p = \mathcal{V}_{cr} + \mathcal{V}_{\text{eff}},$$

sum of the true crystal potential and the effective potential in (6.16) is called *pseudopotential*. It is weakly attractive outside the cores and weakly repulsive inside them, as shown in Fig. 6.3.

It remains the problem to find good expressions for the pseudopotential. Several approximations are possible at this point. The simplest one is the *empirical local pseudopotential*: $\mathcal{V}_p$ is approximated by a local $V_p(\boldsymbol{r})$; then, owing to its periodicity, its Fourier expansion requires wavevectors of the reciprocal lattice, and it turns out that a small number of them is sufficient. From its very definition, it is clear that a sum must be performed over the atoms present in the unit cell. Thus, the Fourier expansion of V_p contains *structure factors*, which account for the positions of the atoms in the unit cell, and *form factors*, which depend on the chemical species of these atoms. The latter are then adjusted, in the empirical approach, to yield the correct values of the band structure in special points of high symmetry [106]. When the cores of the crystal contain high angular momentum states, such as d-states, the local approximation is not sufficient and *empirical nonlocal pseudopotential* methods have been developed, which yield better results [104, 341]. Finally, if the pseudopotential parameters are obtained by atomic wavefunctions, the method is called *ab initio*, as opposed to the *empirical*, where the parameters are used to fit experimental bands.

Once the pseudopotential has been determined, the diagonalization of the pseudo Hamiltonian yields the wanted eigenvalues, pseudo eigenfunctions, and charge density. We must be aware, however, that the pseudo wavefunctions give a realistic charge distribution only outside the core regions. In the core regions, they neglect oscillations and can be considered only as averages.

6.5 Band Structures of Most Important Semiconductors

Figures 6.4 and 6.5 show the electron energy bands and densities of states of silicon and gallium arsenide, respectively, calculated in [104] with an empirical nonlocal pseudopotential method.

6.6 Effective-Mass Approximation

Near a minimum of the energy of a band, the energy $\epsilon(\boldsymbol{k})$ can be approximated by a quadratic form:

$$\epsilon(\boldsymbol{k}) = \frac{1}{2}\hbar^2 \sum_{ij} \left(\frac{1}{m}\right)_{ij} k_i k_j, \tag{6.18}$$

where $(1/m)_{ij}$ is the *inverse effective-mass tensor*. In case of spherical symmetry, we have a simple effective mass m defined by

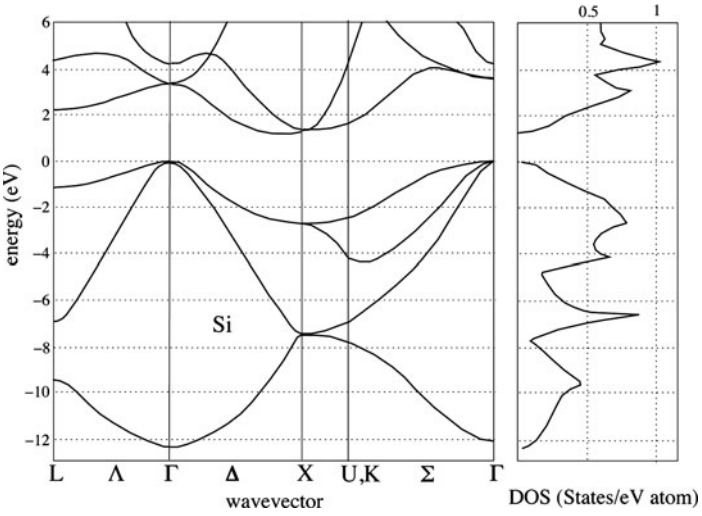


Fig. 6.4. Electron bands and density of states in silicon, calculated with an empirical nonlocal pseudopotential scheme [104]

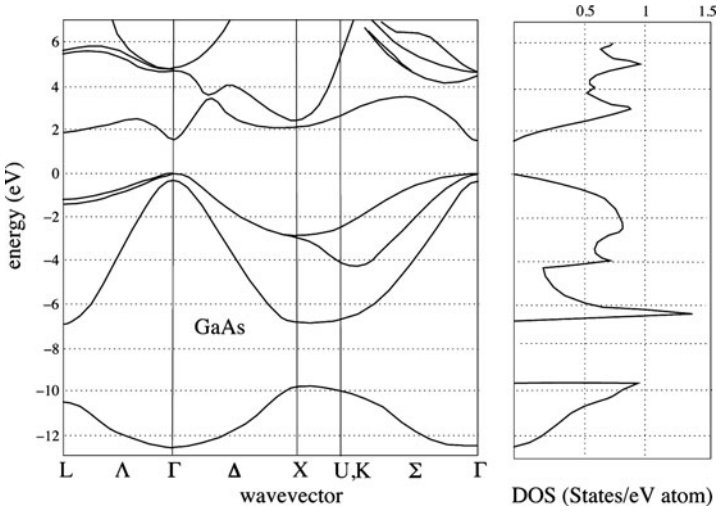


Fig. 6.5. Electron bands and density of states in gallium arsenide, calculated with an empirical nonlocal pseudopotential scheme [104]

$$\epsilon(\boldsymbol{k}) = \frac{\hbar^2 k^2}{2m}.$$

We shall return to this concept over and over again and in particular at the end of this chapter and in next chapter, where we shall discuss effective-mass theorems, not to be confused with the effective-mass approximation. We should note, however, that this approximation is extensively used in the theory of electron transport in semiconductors.

6.7 Bloch Wavepackets

Before leaving the subject of Bloch states, let us analyze the properties of wavepackets formed by superpositions of such states, representing electrons moving inside a crystal. For simplicity, we shall assume here that the states forming the wavepackets belong to one single band and shall omit the band index in the equations.

With the inclusion of the time dependence, a wavepacket formed by the superposition of Bloch states has the form

$$\psi_{\boldsymbol{k}_\circ}(\boldsymbol{r}, t) = \sum_{\boldsymbol{k}} a_{\boldsymbol{k}_\circ}(\boldsymbol{k}) u_{\boldsymbol{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}[\boldsymbol{k}\boldsymbol{r} - \epsilon(\boldsymbol{k})t/\hbar]}. \tag{6.19}$$

To have a well-defined wavepacket, we assume that the coefficients of the superposition are given by a function $a_{\boldsymbol{k}_\circ}(\boldsymbol{k})$ strongly peaked around a value $\boldsymbol{k}_\circ$.

At the same time, we wish the wavepacket itself to be significantly different from zero in a limited region of space.[2]

In (6.19), the sum is extended over all $\boldsymbol{k}$ values in the first BZ. To find the true momentum (not crystal momentum) components of such wave packet, we must perform its Fourier transform. To this purpose, we use the form in (6.6) for $u_{\boldsymbol{k}}(\boldsymbol{r})$ and obtain

$$\psi_{\boldsymbol{k}_\circ}(\boldsymbol{r},t) = \sum_{\boldsymbol{k},\boldsymbol{G}} C(\boldsymbol{k}+\boldsymbol{G}) a_{\boldsymbol{k}_\circ}(\boldsymbol{k}) \mathrm{e}^{\mathrm{i}[(\boldsymbol{k}+\boldsymbol{G})\boldsymbol{r}-\epsilon(\boldsymbol{k})t/\hbar]}. \tag{6.20}$$

Here $\boldsymbol{k}$ runs over the first BZ, and $\boldsymbol{G}$ over the reciprocal lattice, so that the double sum is equivalent to a single sum over the whole reciprocal space. In Sect. 6.1, we have also seen that the energy function $\epsilon(\boldsymbol{k})$ is a periodic function outside of the first Brillouin zone with the period of the reciprocal lattice. The last expression can then be written as

$$\psi_{\boldsymbol{k}_\circ}(\boldsymbol{r},t) = \sum_{\boldsymbol{k}'} c(\boldsymbol{k}') \mathrm{e}^{\mathrm{i}\left[\boldsymbol{k}'\boldsymbol{r}-\epsilon(\boldsymbol{k}')t/\hbar\right]}, \tag{6.21}$$

where $c(\boldsymbol{k}') = C(\boldsymbol{k}+\boldsymbol{G}) a_{\boldsymbol{k}_\circ}(\boldsymbol{k})$ is essentially different from zero in a small area in each cell of the reciprocal lattice centered around the wave vectors $\boldsymbol{k}_\circ + \boldsymbol{G}$, as shown in Fig. 6.6. Thus, the wave packet in (6.19) is in fact a superposition of a number of wave packets with the same crystal momenta but different true momenta.

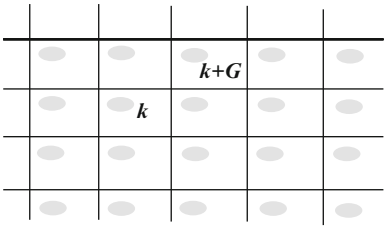


Fig. 6.6. Reciprocal lattice with a wave packet of Bloch states. The coefficients $c(\boldsymbol{k})$ in (6.21) are different from zero in the shaded areas

[2] If we assume a momentum uncertainty $\Delta p = m\Delta v$, a minimum necessary energy is $m(\Delta v)^2/2$; if this energy is taken as the thermal energy at room temperature, the corresponding position uncertainty is about 1.6 nm. In macroscopic systems, this kind of "thermal dimension" of electron wavepackets is not disturbing, and electrons may still be regarded as having well-defined positions in space, but in devices realized by modern technology that may have linear dimensions of a few nanometers, this situation requires special attention, as will be seen in later chapters.

6.7.1 Group Velocity

From the theory of Fourier analysis, we know (see Chap. B) that the group velocities of our wavepackets are given by

$$\boldsymbol{v}(\boldsymbol{k}_\circ + \boldsymbol{G}) = \frac{1}{\hbar}\, \nabla_{\boldsymbol{k}} \epsilon(\boldsymbol{k})|_{\boldsymbol{k}=\boldsymbol{k}_\circ+\boldsymbol{G}}\,. \tag{6.22}$$

Owing to the periodicity of the energy function, however, all these group velocities are equal, so that all the wave packets that contribute to (6.21) travel jointly with group velocity

$$\boxed{\boldsymbol{v}(\boldsymbol{k}_\circ) = \tfrac{1}{\hbar}\, \nabla_{\boldsymbol{k}} \epsilon(\boldsymbol{k})|_{\boldsymbol{k}=\boldsymbol{k}_\circ}} \tag{6.23}$$

which is therefore the velocity of our original wave packet in (6.19). Note that the above expression for the electron wave packet is formally identical to that for free electrons but, now, the momentum $\hbar\boldsymbol{k}$ is actually the crystal momentum of the electron, and $\epsilon(\boldsymbol{k})$ indicates its band energy.

In the effective-mass approximation (6.18), the group velocity becomes

$$v_i = \frac{1}{\hbar}\frac{\partial \epsilon}{\partial k_i} = \hbar \sum_j \left(\frac{1}{m}\right)_{ij} k_j, \tag{6.24}$$

and for a diagonal, isotropic, effective mass reduces to

$$\boldsymbol{v} = \frac{1}{m}\hbar\boldsymbol{k}.$$

7

Effective-Mass Theorems, Envelope Function, and Semiclassical Dynamics

The subject treated in this chapter is of extreme importance to understand electron transport properties in semiconductors and in solid state in general. In fact, it will be shown that, in spite of the crowded presence of atoms in solids, electrons can move freely, in absence of imperfections, owing to the wave nature of their dynamics combined with the periodicity of the atom positions. The periodic potential in the crystal has the effect that the momentum of the electron is no more a good quantum number and it must be substituted by the crystal momentum. Furthermore, the relation between kinetic energy and momentum, which for free electrons is given by the parabolic dependence $p^2/2m_\circ$, must be substituted by the relation between energy and crystal momentum, given by the band $\epsilon_n(\boldsymbol{k})$. With these changes, the electron dynamics is described as that of electrons in free space. In particular, when the potentials externally applied to the crystal are slowly varying, the approximations that in free space lead to classical dynamics, in crystals lead to the so-called *semiclassical dynamics*, extremely useful to understand the behavior of most semiconductor systems. It must be added, however, that in the last decades the technology became able to produce systems with such small dimensions that quantum effects must be considered, as it was necessary to consider quantum physics at the beginning of last century, to understand microscopic systems in free space.

7.1 Effective-Mass Theorem for Bloch States

We shall begin this chapter with an important theorem due to Wannier. Let us first recall the relation

$$\mathrm{e}^{\boldsymbol{R}\nabla} f(\boldsymbol{r}) = f(\boldsymbol{r} + \boldsymbol{R}), \tag{7.1}$$

which can be immediately obtained by expanding the exponential in power series. Now, given a band $\epsilon_n(\boldsymbol{k})$, owing to its periodicity in the reciprocal $\boldsymbol{k}$-space, if we expand it in Fourier series, we obtain

$$\epsilon_n(\boldsymbol{k}) = \sum_{\boldsymbol{R}} C_n(\boldsymbol{R}) \mathrm{e}^{\mathrm{i}\boldsymbol{R}\boldsymbol{k}}, \tag{7.2}$$

where $\boldsymbol{R}$ are the vectors of the direct lattice. In this equation, we may formally substitute $\boldsymbol{k}$ with $-\mathrm{i}\nabla$. The resulting expression can then be applied to a Bloch wavefunction, obtaining

$$\epsilon_n(-\mathrm{i}\nabla)\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{R}} C_n(\boldsymbol{R}) \mathrm{e}^{\boldsymbol{R}\nabla} \left[u_{n\boldsymbol{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}\right] = \sum_{\boldsymbol{R}} C_n(\boldsymbol{R}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{R}} u_{n\boldsymbol{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}},$$

where we have used (7.1) and the periodicity of the functions u_n. With (7.2), this becomes

$$\boxed{\epsilon_n(-\mathrm{i}\nabla)\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \epsilon_n(\boldsymbol{k})\psi_{n\boldsymbol{k}}(\boldsymbol{r})} \tag{7.3}$$

The above equation is the simplest form of the effective-mass theorem. It affirms that Bloch states are eigenfunctions of a Schrödinger equation for a free particle whose dispersion relation is given by the band function $\epsilon_n(p)$ instead of the one for particles in free space $p^2/2m$. The crystal periodic potential has disappeared from the equation and its effect is completely contained in the shape of the band. The theorem also confirms the role of the crystal momentum inside the crystal as the equivalent of the true momentum in free space.[1]

[1] The theorem, however, must be considered with care. In particular, if $\boldsymbol{k}$ lies in the vicinity of a central minimum of the band, such that (effective-mass approximation)

$$\epsilon(\boldsymbol{k}) \approx \frac{\hbar^2 \boldsymbol{k}^2}{2m}, \tag{7.4}$$

one could think, naively, that the Schrödinger equation satisfied by the Bloch state is

$$-\frac{\hbar^2}{2m}\nabla^2\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \epsilon(\boldsymbol{k})\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \frac{\hbar^2\boldsymbol{k}^2}{2m}\psi_{\boldsymbol{k}}(\boldsymbol{r}), \tag{7.5}$$

where the only difference with respect to the case of an electron in free space would be the substitution of the free electron mass $m_\circ$ with the electron effective mass m. This equation, however, has eigenfunctions given by plane waves, while we know that the Bloch functions are plane waves multiplied by the periodic part $u_{\boldsymbol{k}}(\boldsymbol{r})$. The solution of this apparent paradox lies in the fact that the band $\epsilon(\boldsymbol{k})$ is periodic in $\boldsymbol{k}$-space with the periodicity of the reciprocal lattice; the series expansion around the minimum must contain terms necessary to produce the periodicity. In physical terms, wavefunctions with small crystal momentum may have large components of real momentum due to oscillations of their periodic parts. Thus, the form in (7.4) is incomplete, and should be substituted by

$$\epsilon(\boldsymbol{k} + \boldsymbol{G}) \approx \frac{\hbar^2 \boldsymbol{k}^2}{2m},$$

where $\boldsymbol{k}$ is small in the first BZ, and $\boldsymbol{G}$ is any vector of the reciprocal lattice. This can be written, for $\boldsymbol{k}$ also large but close to a $\boldsymbol{G}$,

7.2 Effective-Mass Theorem in Presence of a Scalar Potential

Constant Potential

The above result can easily be generalized, under rather general conditions, when a constant scalar potential $\phi(\boldsymbol{r})$ is present, besides the crystal potential. The Schrödinger equation is

$$[\mathcal{H}_\circ + \phi(\boldsymbol{r})]\,\psi(\boldsymbol{r}) = \left\{\frac{\mathsf{p}^2}{2m_\circ} + V_{cr}(\boldsymbol{r}) + \phi(\boldsymbol{r})\right\}\psi(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}), \tag{7.6}$$

where

$$\mathcal{H}_\circ = \frac{\mathsf{p}^2}{2m_\circ} + V_{cr}(\boldsymbol{r}) \tag{7.7}$$

is the Hamiltonian of the perfect crystal with periodic potential $V_{cr}(\boldsymbol{r})$. We may write the eigenfunction as combination of Bloch states:

$$[\mathcal{H}_\circ + \phi(\boldsymbol{r})]\sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}}\,[\mathcal{H}_\circ + \phi(\boldsymbol{r})]\,\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}),$$

$$\epsilon(\boldsymbol{k}) \approx \frac{\hbar^2}{2m}(\boldsymbol{k}-\boldsymbol{G})^2,$$

where, now, $\boldsymbol{k}-\boldsymbol{G}$ is small in the first BZ, and $\boldsymbol{G}$ is any vector of the reciprocal lattice. Thus, we are not allowed to write

$$\epsilon(-\mathrm{i}\nabla) \approx -\frac{\hbar^2}{2m}\nabla^2\,;$$

we should rather write

$$\epsilon(-\mathrm{i}\nabla) \approx \frac{\hbar^2}{2m}(-\mathrm{i}\nabla-\boldsymbol{G})^2,$$

and the effective-mass theorem tells us that the Bloch function obeys the equation

$$\frac{\hbar^2}{2m}(-\mathrm{i}\nabla-\boldsymbol{G})^2\psi_{\boldsymbol{k},\boldsymbol{G}}(\boldsymbol{r}) = \epsilon(\boldsymbol{k})\psi_{\boldsymbol{k},\boldsymbol{G}}(\boldsymbol{r}) = \frac{\hbar^2\boldsymbol{k}^2}{2m}\psi_{\boldsymbol{k},\boldsymbol{G}}(\boldsymbol{r}),$$

with solutions

$$\psi_{\boldsymbol{k},\boldsymbol{G}}(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}+\boldsymbol{G})\boldsymbol{r}}.$$

One such eigenfunction exists for each $\boldsymbol{G}$, and they are degenerate because of the periodicity of $\epsilon(\boldsymbol{k})$. The Bloch state is given by a linear combination of such functions:

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} b_{\boldsymbol{k}}(\boldsymbol{G})\psi_{\boldsymbol{k},\boldsymbol{G}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} b_k(\boldsymbol{G})\mathrm{e}^{\mathrm{i}\boldsymbol{G}\boldsymbol{r}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} = u_{\boldsymbol{k}}(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}},$$

with the right periodicity of u.

or

$$\sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}} \left[\epsilon_n(\boldsymbol{k}) + \phi(\boldsymbol{r})\right] \psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}).$$

Then, using the result in (7.3),

$$\sum_{n} \left[\epsilon_n(-\mathrm{i}\nabla) + \phi(\boldsymbol{r})\right] \sum_{\boldsymbol{k}} c_{n\boldsymbol{k}} \psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}). \tag{7.8}$$

If only Bloch states belonging to a single band n are required in the expansion, the eigenfunction can be recomposed, yielding

$$\left[\epsilon_n(-\mathrm{i}\nabla) + \phi(\boldsymbol{r})\right] \psi(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}).$$

We shall shortly see that the above condition is not always fulfilled: if the potential $\phi(\boldsymbol{r})$ is not slowly varying with $\boldsymbol{r}$, it may mix Bloch states of different bands.

Time-Dependent Potential

In case the scalar potential depends slowly upon time, a similar result can be obtained for the time-dependent Schrödinger equation:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = \left[\mathcal{H}_\circ + \phi(\boldsymbol{r},t)\right]\Psi(\boldsymbol{r},t).$$

The electron wavefunction can still be written as a combination of Bloch states with time-dependent coefficients. Using the above result (7.8), we obtain

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = \sum_{n} \left[\epsilon_n(-\mathrm{i}\nabla) + \phi(\boldsymbol{r},t)\right] \sum_{\boldsymbol{k}} c_{n\boldsymbol{k}}(t)\psi_{n\boldsymbol{k}}(\boldsymbol{r}).$$

In the present case, however, the possibility to expand the wavefunction over Bloch states belonging to a single band must also be confronted with the energy quanta associated with the time-dependent potential. If the frequencies contained in $\phi(\boldsymbol{r},t)$ are high enough, they can induce interband transitions (see also the following sections). If the applied potential is slowly varying in space and time, a single band approximation can be used, and the above equation reduces to

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi_n(\boldsymbol{r},t) = \left[\epsilon_n(-\mathrm{i}\nabla) + \phi(\boldsymbol{r},t)\right]\Psi_n(\boldsymbol{r},t), \tag{7.9}$$

where, again, the effect of the crystal potential is absorbed by the band function.

The case of a general electromagnetic field, with the inclusion, therefore, of a magnetic field, will be treated in Sect. 7.5 within the envelope-function approximation. Before then, however, let us analyze in great detail the case of a constant and uniform electric field since it presents very interesting features.

7.3 Accelerated Waves

Let us assume that a constant and uniform electric field $\boldsymbol{E}$ is present inside a crystal. The classical force $\boldsymbol{F}$ and the Hamiltonian are

$$\boldsymbol{F} = q\boldsymbol{E}, \quad H = \frac{\boldsymbol{p}^2}{2m} - \boldsymbol{F}\boldsymbol{r}, \tag{7.10}$$

where q is the charge of the carrier.

7.3.1 Accelerated Classical Electrons in Free Space

In classical mechanics, the particle performs an accelerated motion. In terms of Hamilton equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r} = \frac{\partial}{\partial \boldsymbol{p}}H = \frac{\boldsymbol{p}}{m}, \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p} = -\frac{\partial}{\partial \boldsymbol{r}}H = \boldsymbol{F},$$

with solution

$$\boldsymbol{p}(t) = \boldsymbol{p}_\circ + \boldsymbol{F}t, \tag{7.11}$$

if at time $t = 0$ the particle momentum was $\boldsymbol{p}_\circ$.

7.3.2 Accelerated Quantum Electrons in Free Space

The eigenfunctions of momentum are plane waves. When the field $\boldsymbol{E}$ is applied, we may solve the time-independent Schrödinger equation with the Hamiltonian in (7.10). We would obtain as eigenfunctions the Airy functions with a continuum, nondegenerate, energy spectrum [261].

It is more interesting for us here to consider an *accelerated plane wave*, given by [115, 192]:

$$\Psi(\boldsymbol{r}, t) = C\mathrm{e}^{\mathrm{i}\left[\boldsymbol{k}(t)\boldsymbol{r} - \int^t \omega(\boldsymbol{k}(t'))\mathrm{d}t'\right]},$$

where the wavevector changes with time following the classical law (7.11)

$$\boldsymbol{k}(t) = \boldsymbol{k}_\circ + \frac{\boldsymbol{F}}{\hbar}t, \quad \omega(\boldsymbol{k}(t)) = \frac{1}{\hbar}\epsilon(\boldsymbol{k}(t)). \tag{7.12}$$

This function is a solution of the time-dependent Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r}, t) = \left[\frac{\mathsf{p}^2}{2m} - \boldsymbol{F}\boldsymbol{r}\right]\Psi(\boldsymbol{r}, t), \tag{7.13}$$

as can be immediately verified.

7.3.3 Accelerated Bloch States

It will now be shown that the above result holds true also for Bloch functions [192], as long as the field is not so intense to generate interband transitions, known as Zener effect [484].

The accelerated Bloch function is

$$\Psi_{n\boldsymbol{k}(t)}(\boldsymbol{r},t) = \psi_{n\boldsymbol{k}(t)}(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}\int^t \omega_n(\boldsymbol{k}(t'))\mathrm{d}t'}, \tag{7.14}$$

where $\boldsymbol{k}(t)$ is given by (7.12). Our purpose, now, is to see if and when it is a solution of the time-dependent Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = [\mathcal{H}_\circ - \boldsymbol{F}\boldsymbol{r}]\,\Psi(\boldsymbol{r},t), \tag{7.15}$$

where $\mathcal{H}_\circ$ is the crystal Hamiltonian in (7.7), without the applied field. Substituting (7.14) into the l.h.s. of (7.15) yields

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi_{n\boldsymbol{k}(t)}(\boldsymbol{r},t) = \mathrm{i}\hbar\nabla_{\boldsymbol{k}}\psi_{n\boldsymbol{k}(t)}\dot{\boldsymbol{k}}\mathrm{e}^{-\mathrm{i}\int^t \omega_n(\boldsymbol{k}(t'))\mathrm{d}t'} + \hbar\omega_n(\boldsymbol{k}(t))\Psi_{n\boldsymbol{k}(t)}(\boldsymbol{r},t). \tag{7.16}$$

The first term of the r.h.s. of (7.15) yields (in this term t is just a parameter)

$$\mathcal{H}_\circ\Psi_{n\boldsymbol{k}(t)} = \hbar\omega_n(\boldsymbol{k}(t))\Psi_{n\boldsymbol{k}(t)}.$$

To evaluate the second term, we observe that

$$\nabla_{\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \nabla_{\boldsymbol{k}}[u_{n\boldsymbol{k}}(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}] = [\nabla_{\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r})]\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} + \mathrm{i}\boldsymbol{r}\psi_{n\boldsymbol{k}}(\boldsymbol{r}).$$

Thus,

$$\boldsymbol{r}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = -\mathrm{i}\nabla_{\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) + \mathrm{i}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}\nabla_{\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r}).$$

The r.h.s. of (7.15) becomes

$$\hbar\omega_n(\boldsymbol{k}(t))\Psi_{nk(t)} - \boldsymbol{F}\cdot\left[-\mathrm{i}\nabla_{\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{r}) + \mathrm{i}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}\nabla_{\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r})\right]\mathrm{e}^{-\mathrm{i}\int^t \omega_n(\boldsymbol{k}(t'))\mathrm{d}t'}. \tag{7.17}$$

The first two terms coincide with those of the time derivative in (7.16). Thus, the accelerated Bloch state (7.14) is a solution of the time-dependent Schrödinger equation if last term in (7.17) is negligible. To understand the effect of this term, we note that it is a contribution to the time derivative of the state induced by the Hamiltonian. Therefore, its projection on an arbitrary Bloch state $\Psi_{n'\boldsymbol{k}'}$ gives the probability amplitude of inducing a transition to that state. This projection is proportional to

$$\int u^*_{n'k'}(\boldsymbol{r})\mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\boldsymbol{r}}\nabla_k u_{nk}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}.$$

The function $u^*\nabla u$ is periodic with the period of the direct lattice and can be expanded in the Fourier series with wavevector of the reciprocal lattice, so that our integral can be written as

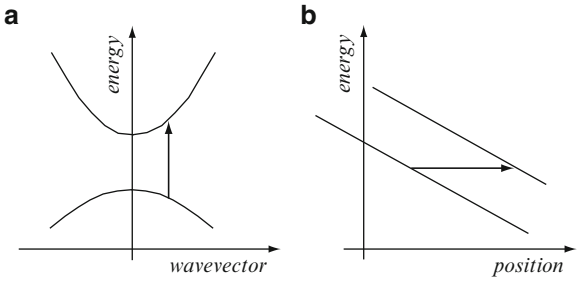


Fig. 7.1. (**a**) A carrier performs a vertical optical transition to a different band by exchanging energy with the applied field. (**b**) In Zener tunneling, the carrier performs a transition to an upper band by tunneling through the band gap and maintaining the same energy. The tilted lines indicate the energy of two bands at $\boldsymbol{k}$ as functions of position

$$\sum_{G} C(\boldsymbol{G}) \int \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\boldsymbol{r}} \mathrm{e}^{\mathrm{i}\boldsymbol{G}\boldsymbol{r}} \, \mathrm{d}\boldsymbol{r}.$$

The integral is the plane-wave representation of the δ function (see Appendix A) and requires $\boldsymbol{k} - \boldsymbol{k}' + \boldsymbol{G} = 0$. Since, however, $\boldsymbol{k}$ and $\boldsymbol{k}'$, as labels of Bloch functions, are defined inside the first BZ, the matrix element above can be different from zero only for $\boldsymbol{k} = \boldsymbol{k}'$. The term under inspection can therefore induce transitions only between states with the same crystal momentum. A vertical transition to an upper band can occur only if the field can supply the necessary energy, i.e., if it has a time dependence with the proper frequency (see Appendix E). This is the case of *optical transitions*, as shown in part (a) of Fig. 7.1, and cannot occur in our situation since we assumed a constant field. A second possibility is that of *Zener tunneling* [484] where an electron tunnels through the band gap, as illustrated in part (b) of the figure.

In conclusion, the term of the Hamiltonian proportional to $\boldsymbol{r}$ due to a uniform and constant field has two effects on a Bloch state $(n\boldsymbol{k})$: the first changes continuously the crystal wavevector $\boldsymbol{k}$ according to the semiclassical law $\dot{\boldsymbol{k}} = \boldsymbol{F}/\hbar$ (semiclassical and not classical because $\hbar\boldsymbol{k}$ is the crystal momentum and not the real momentum); the second produces Zener tunneling if the field is sufficiently intense.

7.4 Envelope Function for Steady States

In the previous sections, we have seen that the crystal momentum $\hbar\boldsymbol{k}$ has many dynamical properties that make its function inside a crystal very similar to that of the real momentum in free space. In the following pages, we shall push this idea further, showing that, under suitable hypotheses, a wavepacket of Bloch states may be described ignoring the periodic part of the

Bloch wavefunction. This is the *envelope-function approximation* used very frequently is solid-state theory.

Let us start from the determination of electron states when a static field $\phi(\boldsymbol{r})$ is present inside a crystal, a case already considered in previous sections, where it was shown that an effective-mass theorem holds, which does not mix bands if the field is not too intense.

Our purpose is, therefore, to find solutions of the Schrödinger equation (7.6). Let us develop the wanted eigenstate in Bloch states, and project the equation on the Bloch state $|\psi_{n'k'}\rangle$:

$$\left\langle \psi_{n'\boldsymbol{k}'} \middle| [\mathcal{H}_\circ + \phi(\boldsymbol{r})] \sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}} \middle| \psi_{n\boldsymbol{k}} \right\rangle = \epsilon \left\langle \psi_{n'\boldsymbol{k}'} \middle| \sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}} \middle| \psi_{n\boldsymbol{k}} \right\rangle,$$

$$\epsilon_{n'}(\boldsymbol{k}')c_{n'\boldsymbol{k}'} + \sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}} \langle \psi_{n'\boldsymbol{k}'} | \phi(\boldsymbol{r}) | \psi_{n\boldsymbol{k}} \rangle = \epsilon c_{n'\boldsymbol{k}'}. \tag{7.18}$$

Let us now examine the matrix elements of ϕ, developing it in Fourier series:

$$\langle \psi_{n'\boldsymbol{k}'} | \phi(\boldsymbol{r}) | \psi_{n\boldsymbol{k}} \rangle = \left\langle \psi_{n'\boldsymbol{k}'} \middle| \sum_{\boldsymbol{k}''} \varphi(\boldsymbol{k}'') \mathrm{e}^{\mathrm{i}\boldsymbol{k}''\boldsymbol{r}} \middle| \psi_{n\boldsymbol{k}} \right\rangle$$

$$= \sum_{\boldsymbol{k}''} \varphi(\boldsymbol{k}'') \int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}''\boldsymbol{r}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} u^*_{n'\boldsymbol{k}'}(\boldsymbol{r}) u_{n\boldsymbol{k}}(\boldsymbol{r}) \mathrm{d}\boldsymbol{r}.$$

The integral must be performed over the crystal. Let us separate into the sum of integrals over the unit cells centered at $\boldsymbol{R}_j$; then define $\boldsymbol{r} = \boldsymbol{R}_j + \boldsymbol{r}'$ in each of such integrals and take into account the periodicity of the functions $u(\boldsymbol{r})$. The matrix elements become

$$\sum_{k''} \varphi(\boldsymbol{k}'') \left[\sum_j \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{k}'')\boldsymbol{R}_j} \right] \int_{\text{cell}} \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{k}'')\boldsymbol{r}'} u^*_{n'k'}(\boldsymbol{r}') u_{nk}(\boldsymbol{r}') \mathrm{d}\boldsymbol{r}'. \tag{7.19}$$

The sum over j can be performed remembering that the crystal momenta of the Bloch states are given by

$$\frac{n_1}{N_1}\boldsymbol{b}_1 + \frac{n_2}{N_2}\boldsymbol{b}_2 + \frac{n_3}{N_3}\boldsymbol{b}_3 \quad n_i = 0, \pm 1, \pm 2, \ldots,$$

where N_1, N_2, and N_3 are the number of unit cells in the three directions of the crystal, and the vector $\boldsymbol{b}$ are the unit vectors of the reciprocal lattice; the vectors of the direct lattice are given by

$$\boldsymbol{R}_j = m_1\boldsymbol{a}_1 + m_2\boldsymbol{a}_2 + m_3\boldsymbol{a}_3, \quad m_i = 0, 1, 2, \ldots, N_i.$$

Thus, if $\boldsymbol{g} = \boldsymbol{k} - \boldsymbol{k}' + \boldsymbol{k}''$,

$$\sum_{\boldsymbol{R}} \mathrm{e}^{\mathrm{i}\boldsymbol{g}\boldsymbol{R}} = \sum_{m_1 m_2 m_3} \mathrm{e}^{\mathrm{i}\left[\frac{n_1}{N_1}\boldsymbol{b}_1 + \frac{n_2}{N_2}\boldsymbol{b}_2 + \frac{n_3}{N_3}\boldsymbol{b}_3\right][m_1\boldsymbol{a}_1 + m_2\boldsymbol{a}_2 + m_3\boldsymbol{a}_3]}$$

$$= \sum_{m_1=0}^{N_1} \mathrm{e}^{2\pi\mathrm{i}\frac{n_1}{N_1}m_1} \sum_{m_2=0}^{N_2} \mathrm{e}^{2\pi\mathrm{i}\frac{n_2}{N_2}m_2} \sum_{m_3=0}^{N_3} \mathrm{e}^{2\pi\mathrm{i}\frac{n_3}{N_3}m_3},$$

where the relation (4.2) between the unit vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ has been used. Each sum contains terms in geometric progression. The result is

$$\sum_{m_1=0}^{N_1} \mathrm{e}^{2\pi\mathrm{i}\frac{n_1}{N_1}m_1} = \frac{1 - \mathrm{e}^{2\pi\mathrm{i}\frac{n_1}{N_1}N_1}}{1 - \mathrm{e}^{2\pi\mathrm{i}\frac{n_1}{N_1}}},$$

and similarly for the other two sums. The exponential in the numerator is unity so that the entire expression is zero unless also the denominator vanishes. This happens when n_1 is a multiple of N_1, and in this case all terms in the sum have value 1. In conclusion (N very large $\approx N + 1$), we have the important result

$$\boxed{\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{g}\boldsymbol{R}_j} = \begin{cases} N & \text{if } \boldsymbol{g} = \boldsymbol{G} \\ 0 & \text{otherwise} \end{cases}} \tag{7.20}$$

where $\boldsymbol{G}$ is a vector of the reciprocal lattice. This is an important result, very useful each time a matrix element between Bloch states is to be evaluated.

Going back to our matrix element in (7.19), from the above result it is zero unless

$$\boldsymbol{k}'' = \boldsymbol{k}' - \boldsymbol{k} + \boldsymbol{G}. \tag{7.21}$$

This means that the Fourier component $\boldsymbol{k}''$ of the potential $\phi(\boldsymbol{r})$ couples Bloch states with crystal wavevectors that differ by $\boldsymbol{k}''$, to within a vector of the reciprocal lattice. This is again identical (without $\boldsymbol{G}$) to what happens in free space in plane-wave scattering.

At this point, we have to make several assumptions. The first assumption is that the potential $\phi(\boldsymbol{r})$ does not vary appreciably within a unit cell of the direct lattice. If it is not so, the potential influences the behavior of the Bloch state within a unit cell, and we cannot hope to get rid of $u(\boldsymbol{r})$. If the potential does not contain significant variations within a distance of the order of the lattice constant, its Fourier transform does not contain wavevectors larger than the unit vector of the reciprocal lattice, and we may assume $\boldsymbol{G} = 0$ in (7.21). In such a case, the Schrödinger equation in the form (7.18) becomes

$$\epsilon_{n'}(\boldsymbol{k}')c_{n'\boldsymbol{k}'} + \sum_{n\boldsymbol{k}} c_{n\boldsymbol{k}}\varphi(\boldsymbol{k}' - \boldsymbol{k})\Delta^{n'n}_{\boldsymbol{k}'\boldsymbol{k}} = \epsilon c_{n'\boldsymbol{k}'}, \tag{7.22}$$

where

$$\Delta_{\boldsymbol{k}'\boldsymbol{k}}^{n'n} = N \int_{\text{cell}} u_{n'\boldsymbol{k}'}^*(\boldsymbol{r}') u_{n\boldsymbol{k}}(\boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}$$

is the overlap integral, and N the number of unit cell in the crystal.

The second assumption is still related to the variations of the potential and restricts them even more: not only the potential does not mix Bloch states in different BZ, but also it connects only Bloch states with crystal momentum so close that the periodic parts of the wavefunction can be assumed to be identical. Then, the orthogonality of the eigenstates ensures that $\Delta_{k'k}^{n'n} = \Delta_{kk}^{n'n} = \delta_{n'n}$. As a consequence of this assumption, the potential does not mix states of different bands, a result that may be connected to the discussion at the end of last section. Equation (7.22) is further simplified:

$$\epsilon_n(\boldsymbol{k}') c_{n\boldsymbol{k}'} + \sum_{\boldsymbol{k}} \varphi(\boldsymbol{k}' - \boldsymbol{k}) c_{n\boldsymbol{k}} = \epsilon c_{n\boldsymbol{k}'}.$$

Now we have separate equations for the different bands. At this point, if we want to antitransform to go back to the original (approximated) wavefunction, we should multiply this equation by the corresponding Bloch wavefunction and sum them all. Instead, we take advantage of the fact that the periodic part has disappeared from the equation and multiply by the plane wave, and sum:

$$\sum_{\boldsymbol{k}'} \epsilon_n(\boldsymbol{k}') c_{n\boldsymbol{k}'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} + \sum_{\boldsymbol{k}\boldsymbol{k}'} \varphi(\boldsymbol{k}' - \boldsymbol{k}) c_{n\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} = \sum_{\boldsymbol{k}'} \epsilon c_{n\boldsymbol{k}'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}}. \tag{7.23}$$

If we apply the operator $\epsilon_n(-\mathrm{i}\nabla)$ to the exponential in the first term of the above equation, we obtain the same function ϵ_{nk}:

$$\sum_{\boldsymbol{k}'} \epsilon_n(\boldsymbol{k}') c_{n\boldsymbol{k}'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} = \epsilon_n(-\mathrm{i}\nabla) \sum_{\boldsymbol{k}'} c_{n\boldsymbol{k}'} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} = \epsilon_n(-\mathrm{i}\nabla) F_n(\boldsymbol{r}),$$

where we have defined the *envelope function*

$$F_n(\boldsymbol{r}) = \sum_{\boldsymbol{k}} c_{n\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}. \tag{7.24}$$

To have a better insight into the above definition, let us compare it with the development of the wavefunction in series of Bloch states:

$$\psi(\boldsymbol{r}) = \sum_n \left[\sum_{\boldsymbol{k}} c_{n\boldsymbol{k}} u_{n\boldsymbol{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} \right].$$

If the coefficients $c_{n\boldsymbol{k}}$ are peaked at a wavevector $\boldsymbol{k}_\circ$ so that the periodic parts can be approximated with only one of them, $u_{n\boldsymbol{k}}(\boldsymbol{r}) \approx u_{n\boldsymbol{k}_\circ}(\boldsymbol{r})$, then

$$\boxed{\psi(\boldsymbol{r}) \approx \textstyle\sum_n u_{n\boldsymbol{k}_\circ}(\boldsymbol{r}) F_n(\boldsymbol{r})} \tag{7.25}$$

Thus, the envelope function is the function that envelops the periodic part of the Bloch wavefunction that mainly contributes to $\psi(\boldsymbol{r})$. In particular, if the wavefunction reduces to a single Bloch wavefunction, the envelope function is the corresponding plane wave.

The second term in (7.23) is easily recognized to be

$$\sum_{\boldsymbol{k}\boldsymbol{k}'} \varphi(\boldsymbol{k}'-\boldsymbol{k}) c_{n\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}} = \sum_{\boldsymbol{k}\boldsymbol{k}'} \varphi(\boldsymbol{k}'-\boldsymbol{k}) \mathrm{e}^{\mathrm{i}(\boldsymbol{k}'-\boldsymbol{k})\boldsymbol{r}} c_{n\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} = \phi(\boldsymbol{r}) F_n(\boldsymbol{r}),$$

so that the entire (7.23) becomes the time-independent Schrödinger equation for the envelope function, identical to the Schrödinger equation in free space with the kinetic energy given by the band operator:

$$\boxed{\left[\epsilon_n(-\mathrm{i}\nabla) + \phi(\boldsymbol{r})\right] F_n = \epsilon F_n} \tag{7.26}$$

As regards the normalization of the envelope function, we recall that it must vary slowly in the distance of the unit cell, so that, if the entire wavefunction is normalized to one in the crystal volume V, we have

$$1 = \int_V \psi^* \psi \, \mathrm{d}\boldsymbol{r} = \int_V |F|^2 |u|^2 \, \mathrm{d}\boldsymbol{r} \approx \sum_j |F(\boldsymbol{R}_j)|^2 \int_{\text{cell}} |u|^2 \, \mathrm{d}\boldsymbol{r} = \sum_j |F(\boldsymbol{R}_j)|^2 \frac{1}{N}$$
$$= \sum_j |F(\boldsymbol{R}_j)|^2 \frac{V_c}{N V_c} = \frac{1}{V} \int_V |F(\boldsymbol{r})|^2 \, \mathrm{d}\boldsymbol{r}.$$

Therefore, the envelope function must be normalized to the volume of the crystal:

$$\int |F(\boldsymbol{r})|^2 \mathrm{d}\boldsymbol{r} = V. \tag{7.27}$$

The envelope function technique is very often used to evaluate localized electron states inside a semiconductor. It is done so, often, also in heterostructures, where the applied potential is due to a sudden change in the nature of the crystal (see Chap. 19). In such cases, the hypotheses at the basis of the envelope function approximation are not fulfilled and better calculations are needed.

7.5 Effective-Mass Theorem for a Wavepacket in Slow-Varying Electric and Magnetic Fields

Let us now turn to the problem of the effect of the contemporary presence of both an electric and a magnetic field on the electron dynamics in a crystal. We shall consider here the case of an electron wavepacket of small enough dimension, such that the electromagnetic potentials can be considered constant where the electron wavefunction is appreciably different from zero. We

can do so because in this section we are interested in obtaining the so-called semiclassical dynamics, neglecting special quantum effects, such as interference and tunneling. These quantum effects will be treated in the chapters devoted to quantum transport.

The procedure we are going to apply is the following: first, using the approximations indicated above, we find a gauge in which the vector potential does not appear; then, in absence of the vector potential we may use the result obtained in Sect. 7.2; finally we may go back to a generic gauge and find the equation satisfied by the time-dependent electron wavefunction.

The Hamiltonian of an electron in a crystal in presence of an electromagnetic field is

$$\mathcal{H} = \frac{1}{2m_\circ}\left[\boldsymbol{p} - q\boldsymbol{A}(\boldsymbol{r},t)\right]^2 + V_{cr}(\boldsymbol{r}) + q\phi(\boldsymbol{r},t),$$

where $\boldsymbol{A}$ and ϕ are the vector and scalar potential, respectively. The electromagnetic potentials can be changed through gauge transformations (see Sect. 1.7)

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \nabla\Lambda(\boldsymbol{r},t), \quad \phi \to \phi' = \phi - \frac{\partial\Lambda}{\partial t}, \tag{7.28}$$

where Λ is an arbitrary scalar function of space and time.

If we assume, as indicated above, that the vector potential is varying slowly enough to be considered constant in the region where the wavepacket is different from zero, then we may write

$$\boldsymbol{A}(\boldsymbol{r},t) \approx \boldsymbol{A}(\langle\boldsymbol{r}\rangle,t), \tag{7.29}$$

where $\langle\boldsymbol{r}\rangle$ indicates the mean position of the electron wavepacket. Now we can search for a new gauge in which

$$\boldsymbol{A}'(\langle\boldsymbol{r}\rangle,t) = 0. \tag{7.30}$$

This can be achieved taking in (7.28)

$$\Lambda = -\boldsymbol{r}\cdot\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t). \tag{7.31}$$

In fact,

$$\boldsymbol{A}' = \boldsymbol{A} + \nabla\Lambda = \boldsymbol{A}(\boldsymbol{r},t) - \boldsymbol{A}(\langle\boldsymbol{r}\rangle,t) \approx 0.$$

In this new gauge, the Hamiltonian does not contain the vector potential, and the time-dependent Schrödinger equation is:

$$\mathrm{i}\hbar\frac{\partial\Psi'}{\partial t} = \mathcal{H}\Psi' = \left[\frac{\mathsf{p}^2}{2m_\circ} + V_c(\boldsymbol{r}) + q\phi'(\boldsymbol{r},t)\right]\Psi'.$$

Since the vector potential is absent, we may apply the effective-mass theorem in the form (7.9), and write the Schrödinger equation as

$$\mathrm{i}\hbar\frac{\partial\Psi'}{\partial t} = [\epsilon(-\mathrm{i}\nabla) + q\phi'(\boldsymbol{r},t)]\,\Psi', \tag{7.32}$$

where we assume that the electron wavepacket is formed with Bloch states of only one band. The effect of the magnetic field is contained in the gauge, and we shall find it again when we shall write the wavefunction in the original gauge. In Sect. 2.5, it is shown that when a gauge transformation is performed, the wavefunction is also transformed according to

$$\Psi \to \Psi' = \mathrm{e}^{\mathrm{i}q\Lambda/\hbar}\Psi. \tag{7.33}$$

Thus, to find the Schrödinger equation obeyed by the original wavefunction Ψ, let us multiply (7.32) by the exponential factor:

$$\mathrm{i}\hbar\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\frac{\partial\Psi'}{\partial t} = \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\,[\epsilon(-\mathrm{i}\nabla) + q\phi'(\boldsymbol{r},t)]\,\Psi'. \tag{7.34}$$

Now we analyze the various terms of this equation. The l.h.s. is, using (7.28),

$$\mathrm{i}\hbar\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\frac{\partial\Psi'}{\partial t} = \mathrm{i}\hbar\frac{\partial}{\partial t}\left(\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\Psi'\right) - \mathrm{i}\hbar\frac{-\mathrm{i}q}{\hbar}\frac{\partial\Lambda}{\partial t}\Psi = \mathrm{i}\hbar\frac{\partial\Psi}{\partial t} - q(\phi-\phi')\Psi.$$

In the first term in the r.h.s. of (7.34) we Fourier expand the band ϵ as in (7.2) and use the translation operator in (7.1). It becomes

$$\mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{\boldsymbol{R}\nabla}\Psi'(\boldsymbol{r}) = \sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{\boldsymbol{R}\nabla}\left[\mathrm{e}^{-\mathrm{i}q\Lambda(\boldsymbol{r}-\boldsymbol{R})/\hbar}\Psi'(\boldsymbol{r})\right].$$

Now, from the definition of Λ in (7.31), $\Lambda(\boldsymbol{r}-\boldsymbol{R}) = -(\boldsymbol{r}-\boldsymbol{R})\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t) = \Lambda(\boldsymbol{r}) + \boldsymbol{R}\boldsymbol{A}(\langle\boldsymbol{r}\rangle),t)$, and the above can be further transformed as

$$\sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{\mathrm{i}\boldsymbol{R}(-\mathrm{i}\nabla)}\left[\mathrm{e}^{(-\mathrm{i}q/\hbar)\boldsymbol{R}\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)}\Psi(\boldsymbol{r})\right].$$

Since the exponential in the square brackets does not depend on $\boldsymbol{r}$, it commutes with the gradient operator and we can apply the rule of the product of exponentials:

$$= \sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{\mathrm{i}\boldsymbol{R}\cdot[-\mathrm{i}\nabla-(q/\hbar)\boldsymbol{A}(\langle\boldsymbol{R}\rangle,t)]}\Psi(\boldsymbol{R}) = \epsilon[-\mathrm{i}\nabla - (q/\hbar)\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)]\Psi.$$

Finally, the last term in the r.h.s. of (7.34) is $q\phi'(\boldsymbol{r},t)\Psi$.

Collecting the above results our Schrödinger equation (7.34) becomes, remembering also (7.29),

$$\boxed{\mathrm{i}\hbar\tfrac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = \left\{\epsilon\left(-\mathrm{i}\nabla - \tfrac{q}{\hbar}\boldsymbol{A}(\boldsymbol{r},t)\right) + q\phi\right\}\Psi} \tag{7.35}$$

We have again an effective-mass theorem: the Schrödinger equation is obtained by inserting the operator $[-\mathrm{i}\nabla - (q/\hbar)\boldsymbol{A}(\boldsymbol{r},t)]$ in the band function to account for the periodic crystal potential.

7.6 Time-Dependent Envelope Function

Let us consider the wavefunction with the explicit gauge dependence, as given in (7.33) and expand Ψ' in Bloch states of a single band around $\boldsymbol{k}_\circ$:

$$\Psi(\boldsymbol{r},t) = \mathrm{e}^{-\mathrm{i}q\Lambda/\hbar}\Psi' = \mathrm{e}^{-\mathrm{i}q\Lambda(\boldsymbol{r},t)/\hbar}\sum_{\boldsymbol{k}} a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k},t)u_{\boldsymbol{k}}(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}}.$$

Here, again, $a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k},t)$ is a function strongly peaked around $\boldsymbol{k}_\circ$.[2] Thus, since the functions u_k are usually slowly varying with respect to $\boldsymbol{k}$, we may write

$$\Psi(\boldsymbol{r},t) = \mathrm{e}^{-\mathrm{i}q\Lambda(\boldsymbol{r},t)/\hbar}u_{\boldsymbol{k}_\circ(t)}(\boldsymbol{r})\sum_{\boldsymbol{k}} a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k},t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}},$$

or

$$\boxed{\Psi(\boldsymbol{r},t) = F(\boldsymbol{r},t)u_{\boldsymbol{k}_\circ(t)}(\boldsymbol{r})} \tag{7.36}$$

where

$$F(\boldsymbol{r},t) = \mathrm{e}^{-\mathrm{i}q\Lambda(\boldsymbol{r},t)/\hbar}\sum_{\boldsymbol{k}} a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k},t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} \tag{7.37}$$

is again the envelope function that envelops the periodic part of the Bloch function to form the wavepacket. Now it is time dependent.

To find the equation satisfied by the envelope function, let us insert (7.36) into (7.35), with the vector potential evaluated again in the mean position of the wavepacket:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\left[F(\boldsymbol{r},t)u_{\boldsymbol{k}_\circ(t)}(\boldsymbol{r})\right] = \left\{\epsilon\left[-\mathrm{i}\nabla - (q/\hbar)\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)\right] + q\phi(\boldsymbol{r},t)\right\}\left[Fu_{\boldsymbol{k}_\circ(t)}(\boldsymbol{r})\right].$$

Now we apply the usual expansion of the band function:

$$\mathrm{i}\hbar\left[\frac{\partial F}{\partial t}u + F\frac{\partial u}{\partial \boldsymbol{k}_\circ}\dot{\boldsymbol{k}}_\circ\right] = \sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{\mathrm{i}\boldsymbol{R}\cdot(-\mathrm{i}\nabla-(q/\hbar)\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t))}[Fu] + q\phi Fu$$

$$\mathrm{i}\hbar u\left[\frac{\partial F}{\partial t} + F\frac{\partial \ln u}{\partial \boldsymbol{k}_\circ}\dot{\boldsymbol{k}}_\circ\right] = \sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{-\frac{\mathrm{i}q}{\hbar}\boldsymbol{R}\cdot\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)}F(\boldsymbol{r}+\boldsymbol{R})u_{\boldsymbol{k}_\circ(t)}(\boldsymbol{r}+\boldsymbol{R}) + q\phi Fu.$$

The last term in the l.h.s. can be neglected since the functions u are slowly varying with respect to $\boldsymbol{k}$. We should, however, remember that the results we will find do depend upon this assumption that in the time variation of the wavepacket the variation of the envelope function dominates with respect to the time variation of the main enveloped function. Then, taking also into account the periodicity of u, the above equation becomes, after division by u,

[2] This condition is to be reconciled with the one, considered above, of a wavepacket small enough to consider the potentials constant in the region where the wavefunction is significantly different from zero.

$$\mathrm{i}\hbar\frac{\partial F}{\partial t} = \sum_{\boldsymbol{R}} C(\boldsymbol{R})\mathrm{e}^{-\frac{\mathrm{i}q}{\hbar}\boldsymbol{R}\cdot\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)}\mathrm{e}^{\boldsymbol{R}\nabla}F(\boldsymbol{r}) + q\phi F.$$

Finally, remembering again (7.29),

$$\boxed{\mathrm{i}\hbar\tfrac{\partial}{\partial t}F(\boldsymbol{r},t) = \left\{\epsilon\left(-\mathrm{i}\nabla - \tfrac{q}{\hbar}\boldsymbol{A}(\boldsymbol{r},t)\right) + q\phi\right\}F} \tag{7.38}$$

Thus, under the above assumptions, the envelope function F satisfies the same Schrödinger equation (7.35) as the wavefunction Ψ, with Hamiltonian

$$\mathcal{H} = \epsilon\left(\frac{\mathsf{p} - q\boldsymbol{A}}{\hbar}\right) + q\phi, \tag{7.39}$$

where p is the momentum operator $-\mathrm{i}\hbar\nabla$, and $\epsilon(\boldsymbol{k})$ is the function that gives the considered band.

7.7 Semiclassical Dynamics

We are now in the condition to study how an electron described by a wavepacket of small enough dimension around $\langle\boldsymbol{r}\rangle$ with a reasonably well-defined crystal momentum around $\boldsymbol{k}_\circ$, changes its position and its crystal momentum as effect of the applied fields.

From the discussion in Sect. 6.7, we have seen that the crystal momentum of a Bloch state is given by the wavevector appearing in the exponential of the Bloch wavefunction, while higher components of the momentum are given by the oscillations of the periodic part $u_{\boldsymbol{k}}(\boldsymbol{r})$. Since the envelope function keeps the exponentials of the wavepacket and envelops its main periodic part, it is useful to evaluate the mean value of the momentum with the envelope function:

$$\langle\boldsymbol{p}\rangle_F = \frac{1}{V}\int F^*(\boldsymbol{r},t)(-\mathrm{i}\hbar\nabla)F(\boldsymbol{r},t)\,\mathrm{d}\boldsymbol{r}, \tag{7.40}$$

where the normalization in (7.27) has been taken into account. Inserting (7.37), we obtain

$$\begin{aligned}\langle\boldsymbol{p}\rangle_F = \frac{1}{V}\sum_{\boldsymbol{k},\boldsymbol{k}'}\int a^*_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k},t)\mathrm{e}^{\frac{\mathrm{i}q}{\hbar}\Lambda(\boldsymbol{r},t)}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{r}}(-\mathrm{i}\hbar)\left(-\frac{\mathrm{i}q}{\hbar}\nabla\Lambda(\boldsymbol{r},t) + \mathrm{i}\boldsymbol{k}'\right)\\ \times a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k}',t)\mathrm{e}^{-\frac{\mathrm{i}q}{\hbar}\Lambda(\boldsymbol{r},t)}\mathrm{e}^{\mathrm{i}\boldsymbol{k}'\boldsymbol{r}}\mathrm{d}\boldsymbol{r}.\end{aligned}$$

Now, according to (7.31) $\nabla\Lambda = -\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)$, and, remembering that $a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k},t) \approx a_{\boldsymbol{k}_\circ(t)}(\boldsymbol{k}_\circ,t)$, the above yields

$$\langle\boldsymbol{p}\rangle_F = \left[q\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t) + \hbar\boldsymbol{k}_\circ\right]\frac{1}{V}\int|F|^2\,\mathrm{d}\boldsymbol{r} = \left[\hbar\boldsymbol{k}_\circ(t) + q\boldsymbol{A}(\langle\boldsymbol{r}\rangle,t)\right]. \tag{7.41}$$

From the time derivative of the above equation, we obtain

$$\hbar\dot{\boldsymbol{k}}_\circ(t) = \frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{p}\rangle_F - q\frac{\partial}{\partial t}\boldsymbol{A}(\langle \boldsymbol{r}\rangle, t) - q\boldsymbol{v}\nabla\boldsymbol{A}(\langle \boldsymbol{r}\rangle, t). \tag{7.42}$$

To evaluate the first term on the r.h.s., we take the time derivative of (7.40):

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{p}\rangle_F = \frac{1}{V}\int \left\{ \frac{\partial F^*(\boldsymbol{r},t)}{\partial t}\mathsf{p}F(\boldsymbol{r},t) + F^*(\boldsymbol{r},t)\mathsf{p}\frac{\partial F(\boldsymbol{r},t)}{\partial t} \right\} \mathrm{d}\boldsymbol{r},$$

and from (7.38) and (7.39)

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{p}\rangle_F = \frac{1}{V}\int \frac{1}{\mathrm{i}\hbar}\left\{(-\mathcal{H}F)^*\mathsf{p}F(\boldsymbol{r},t) + F^*\mathsf{p}\mathcal{H}F(\boldsymbol{r},t)\right\} \mathrm{d}\boldsymbol{r}.$$

Now, since $\mathcal{H}$ is hermitian, the above equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{p}\rangle_F = \frac{1}{V}\frac{1}{\mathrm{i}\hbar}\int F^*[\mathsf{p},\mathcal{H}]F\,\mathrm{d}\boldsymbol{r}.$$

We know from quantum mechanics that

$$[\mathsf{p},\mathcal{H}] = -\mathrm{i}\hbar\nabla\mathcal{H}.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{p}\rangle_F = -\langle\nabla\mathcal{H}\rangle. \tag{7.43}$$

Now, with the form (7.39) of the Hamiltonian, we obtain

$$\nabla\mathcal{H} = q\nabla\phi - q(\boldsymbol{v}_g\cdot\nabla)\boldsymbol{A} - q\boldsymbol{v}_g\times(\nabla\times\boldsymbol{A}).$$

This expression can be easily obtained in cartesian coordinates, keeping in mind that the derivative of the band with respect to its argument gives the group velocity. Thus, always remembering that we are dealing with a wavepacket with reasonably well-defined position and crystal momentum,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \boldsymbol{p}\rangle_E = -q\nabla\phi(\langle \boldsymbol{r}\rangle) + q(\boldsymbol{v}_g\cdot\nabla)\boldsymbol{A} + q\boldsymbol{v}_g\times\boldsymbol{B}.$$

Substituting this result into (7.42), we obtain

$$\hbar\dot{\boldsymbol{k}}_\circ(t) = -q\nabla\phi(\langle \boldsymbol{r}\rangle) + q\boldsymbol{v}_g\times(\boldsymbol{B}) - q\frac{\partial}{\partial t}\boldsymbol{A}(\langle \boldsymbol{r}\rangle, t).$$

Finally, taking into account the definition of the scalar potential in (1.24), this becomes

$$\boxed{\hbar\dot{\boldsymbol{k}}_\circ(t) = q\left(\boldsymbol{E} + \boldsymbol{v}_g\times\boldsymbol{B}\right)} \tag{7.44}$$

This is the fundamental law of semiclassical dynamics that governs the dynamics of electrons in solids: when the applied fields have negligible variations within the region occupied by the wavepacket the mean crystal momentum of the electron follows the Newton dynamics with the classical force acting on a charge in an electromagnetic field. The difference with the case of a charged particle in free space, in the classical limit, is that $\hbar \boldsymbol{k}_\circ$ *is not the momentum of the particle but its mean crystal momentum, and the relation between the energy and the crystal momentum is not the classical parabola* $\epsilon = p^2/2m$*, but the function* $\epsilon(\boldsymbol{k})$ *that gives the band in which the electron is located.*

8

Semiconductors

8.1 Free Dynamics of Bloch Electrons

To understand the basic properties of different materials from the point of view of their electrical conduction, let us start from the final result of last chapter: under rather general conditions, the crystal momentum of a Bloch state subject to an external force $\boldsymbol{F}$ changes according to the semiclassical law

$$\frac{\mathrm{d}(\hbar \boldsymbol{k})}{\mathrm{d}t} = \boldsymbol{F}. \tag{8.1}$$

This result is to some extent amazing if we consider that the dynamics of the electron wave is continuously affected by the interaction with the atoms of the crystal. The effect of the crystal periodic potential lies in the fact that $\hbar \boldsymbol{k}$ appearing in (8.1) is the crystal momentum, not the real momentum, and its relation with the energy is given by the band function $\epsilon(\boldsymbol{k})$.

If an electric field is applied, an electron will continuously change its $\boldsymbol{k}$, according to (8.1), in absence of collisions, as shown in Fig. 8.1. Its velocity, given by the derivative of the band, will also change continuously, increasing as long as the crystal momentum lies in the lower, concave, part of the band. When the electron reaches the upper, convex, part of the band, the velocity starts to decrease, and this is a crucial effect of the crystal potential. When $\boldsymbol{k}$ reaches the energy maximum, the electron group velocity vanishes. In a simple band, this happens at the BZ edge, as shown in the figure. As the effect of the force continues, $\boldsymbol{k}$ surpasses the zone edge, and its velocity is reversed. Since $\boldsymbol{k}$ is defined to within a vector of the reciprocal lattice, it may be considered to reenter the BZ from the opposite side. Thus, the crystal momentum, in presence of a constant and homogeneous electric field, performs oscillations in the BZ, corresponding to oscillations of the wavepacket in real space, called

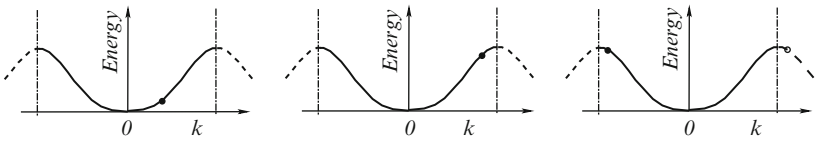


Fig. 8.1. Under the action of a constant and homogeneous electric field, and in absence of collisions, the crystal momentum of an electron performs oscillations inside the Brillouin zone. The three parts of the figure show the positions of the crystal momentum in the band at three different, successive times. When it exits from the BZ at one edge, it reenters from the opposite edge

Bloch oscillations. It is important to remember that these oscillations occur in absence of collisions, a situation practically unrealizable in bulk materials.[1]

8.2 A Fully Occupied Band Cannot Carry Current

In any given band, the time-reversal symmetry requires that for each state $\boldsymbol{k}$ there exist a state $-\boldsymbol{k}$ such that

$$\epsilon(\boldsymbol{k}) = \epsilon(-\boldsymbol{k}). \tag{8.2}$$

Furthermore, since ϵ is an even function of $\boldsymbol{k}$, its gradient must be odd, so that

$$\boldsymbol{v}(\boldsymbol{k}) = \frac{1}{\hbar}\nabla_{\boldsymbol{k}}\,\epsilon(\boldsymbol{k}) = -\frac{1}{\hbar}\nabla_{\boldsymbol{k}}\,\epsilon(-\boldsymbol{k}) = -\boldsymbol{v}(-\boldsymbol{k}). \tag{8.3}$$

Thus, if all states in the band are occupied by electrons, the total crystal momentum and the total current of the electrons are both zero:

$$\sum_{k\in\mathrm{BZ}} \boldsymbol{k} = 0, \quad \sum_{k\in\mathrm{BZ}} \boldsymbol{v}(\boldsymbol{k}) = 0.$$

If now a constant, uniform, electric field is applied to the crystal, all electrons in a completely occupied band move their wavevector as indicated in the previous section. Electrons that exit from one side of the BZ reenter from the opposite side; the occupation of the band is unaltered, as shown in the left part of Fig. 8.2, and no charge current is generated: *a fully occupied band does not conduct.*

The situation is quite different if the band is only partially occupied, as happens in metals. Also in this case $\boldsymbol{k}$ changes according to (8.1), but collisions occur which tend to restore the equilibrium situation. The electron distribution in $\boldsymbol{k}$-space becomes asymmetric and a net current results, as shown in the right part of Fig. 8.2.

[1] Superlattices have been conceived [132] to obtain crystals with a very large lattice constant in one direction and therefore a very narrow BZ. In this way, Bloch oscillations can be observed. See Chap. 19.

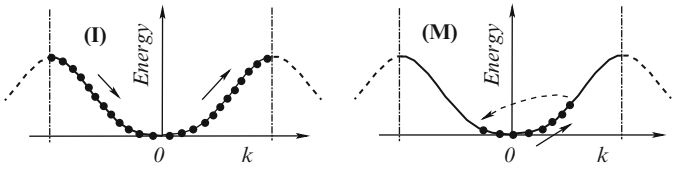


Fig. 8.2. Under the action of an electric field, a completely full band (I) does not change its status since all Bloch states move at the same rate (*continuous arrows*) and collisions are not effective since there are no empty states available as possible final states of collisions. In a partially filled band (M), as in metals, electrons change continuously their crystal momenta, and collisions (*broken arrow*) tend to restore the equilibrium situation. The electron distribution in $\boldsymbol{k}$-space becomes asymmetric and a net current results

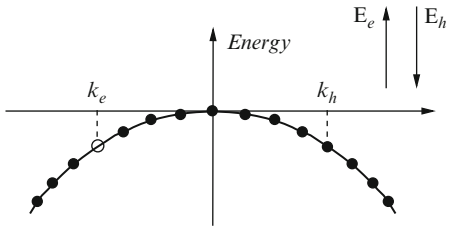


Fig. 8.3. The concept of hole. See text

8.3 Holes

Let us consider a band entirely occupied by electrons, except for one single Bloch state $\boldsymbol{k}_\mathrm{e}$, as shown in Fig. 8.3. In what follows we shall show that the electrical property of the band in such situation can conveniently be described in terms of the presence of a single particle, called *hole*. This description provides a language much more intuitive than that in terms of all electrons actually present.

We have just seen that the sum of all possible $\boldsymbol{k}$s in the BZ is zero since opposite $\boldsymbol{k}$s cancel in pairs. Thus, the total crystal momentum of all electrons in the band with one missing electron in $\boldsymbol{k}_\mathrm{e}$ is equal to

$$\sum_{k\neq k_\mathrm{e}} \boldsymbol{k} = \sum_{k\in \mathrm{BZ}} \boldsymbol{k} - \boldsymbol{k}_\mathrm{e} = -\boldsymbol{k}_\mathrm{e},$$

where the second sum is extended to all states in the BZ, while the first is over the occupied states. Thus, the total momentum[2] of all present electrons is $-\boldsymbol{k}_\mathrm{e}$. This situation can be described as the presence of a single particle, a hole, with crystal momentum opposite to that of the missing electron:

$$\boldsymbol{k}_h = -\boldsymbol{k}_\mathrm{e}. \tag{8.4}$$

[2] Often, in the following, we shall simply call "momentum" the crystal momentum, where no ambiguity can arise.

Let us now consider the energy of the electrons that occupy all the states of a band with the exception of $\boldsymbol{k}_\mathrm{e}$. If, for simplicity, we set to zero the total energy of the band where all states are filled by electrons, by repeating the argument above we find that the energy of the hole is the opposite of the energy of the missing electron.

$$\epsilon_h(\boldsymbol{k}_h) = -\epsilon_\mathrm{e}(\boldsymbol{k}_\mathrm{e}).$$

This result is very intuitive if we observe that the present electrons tend to occupy, at equilibrium, the states with lower energies so that the hole tends to stay in a state with higher (electron) energy.

If we consider a wavepacket of Bloch states localized in space, its center will move with the group velocity (6.23). If the wavepacket is formed with empty states, the missing electron, i.e., the hole, will move with the same velocity. This suggests that

$$\boldsymbol{v}_h = \boldsymbol{v}_\mathrm{e}.$$

This relation is coherent with the formal expression that is obtained by the definitions of energy and momentum of the hole:

$$\boldsymbol{v}_\mathrm{e} = \frac{1}{\hbar}\nabla_{\boldsymbol{k}_\mathrm{e}}\epsilon_\mathrm{e} = \frac{1}{\hbar}\nabla_{-\boldsymbol{k}_\mathrm{e}}(-\epsilon_\mathrm{e}) = \frac{1}{\hbar}\nabla_{\boldsymbol{k}_h}\epsilon_h = \boldsymbol{v}_h.$$

If an electric field $\boldsymbol{E}$ is applied to the crystal, from (8.1) we have

$$\hbar\frac{\mathrm{d}\boldsymbol{k}_\mathrm{e}}{\mathrm{d}t} = (-\mathrm{e})\boldsymbol{E},$$

where $(-\mathrm{e})$ is the electron charge. Applying (8.4), this relation may be read as

$$\hbar\frac{\mathrm{d}\boldsymbol{k}_h}{\mathrm{d}t} = (+\mathrm{e})\boldsymbol{E},$$

showing that the hole must be considered as positively charged with charge $(+\mathrm{e})$, coherently with the charge neutrality of the fully occupied band. In presence of an additional magnetic field $\boldsymbol{B}$, an analogous argument leads to

$$\hbar\frac{\mathrm{d}\boldsymbol{k}_h}{\mathrm{d}t} = (+\mathrm{e})\left(\boldsymbol{E} + \boldsymbol{v}_h \times \boldsymbol{B}\right).$$

8.4 Insulators, Conductors, Semiconductors

We know that the energy eigenvalues of electrons in perfect crystals fall in intervals, the energy bands, separated by intervals where the energy eigenvalues are absent, called band gaps. Quantum statistics tells us (see Chap. 3) that an electron state can be occupied only by one electron, or two if we account for spin degeneracy.

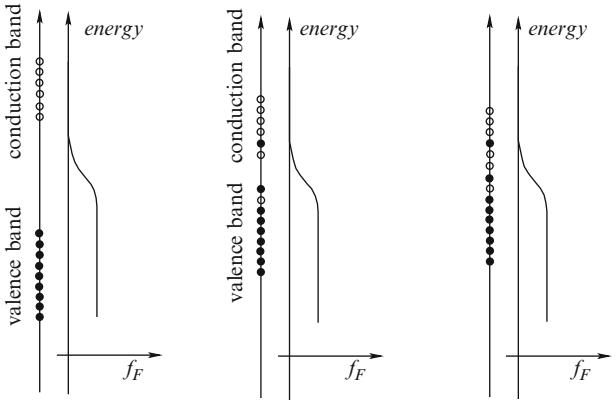


Fig. 8.4. Bands and Fermi–Dirac distribution functions in insulators, intrinsic semiconductors, and metallic conductors. Closed circles represent states occupied by electrons, and open circles represent empty states

When at zero absolute temperature the last occupied band, in the order of increasing energies, is totally occupied by electrons, it is called *valence band*, since the corresponding electron states are the main responsible for the bonds of the atoms that form the crystal. The next higher band is called *conduction band*, since the corresponding electron states are the main responsible for electrical conduction (Fig. 8.4).

What follows is a basic classification of the different materials from the point of view of their electrical properties.

- If at zero absolute temperature the highest band occupied by electrons is entirely occupied (valence band) and the energy gap between this band and the next higher band (conduction band) is much higher than $K_{\mathrm{B}}T$, where K_{B} is the Boltzmann constant and T the room temperature, the conduction band remains empty even at room temperature since the thermal energy is not sufficient to promote electrons from the last occupied band to the next empty band. In absence of impurities, the material is an *insulator*. In fact, the totally occupied valence band does not conduct as shown in Sect. 8.2, and the conduction band does not conduct because it is void of electrons. We shall see shortly that the presence of impurities changes drastically this scenario.
- If the energy gap between the valence band, entirely occupied at zero temperature, and the conduction band, empty at zero temperature, is comparable with $K_{\mathrm{B}}T$ at room temperature, some electrons are promoted to the conduction band by the thermal energy. The conduction band contains some electrons, the valence band contains an equal number of holes, and in absence of impurities the material is an *intrinsic semiconductor*. It is

obvious that at zero temperature any intrinsic semiconductor becomes an insulator.
– If at zero absolute temperature the last occupied band is only partially occupied, the material is a *metallic conductor*, as described in Sect. 8.2.

8.5 Intrinsic and Doped Semiconductors

8.5.1 Donors and Acceptors

One of the most peculiar characteristics of semiconductors is that it is possible to change their conductivity by many orders of magnitude with the introduction of a very little quantity, even one part in ten millions, of an appropriate material, called *dopant*. Let us see how this happens.

Consider a perfect crystal of, for example, silicon. The four electrons of the outer atomic shell, together with the analogous electrons of the neighboring atoms, form the tetrahedral bonds shown in part (a) of Fig. 4.4. In terms of Bloch states they form the valence band, totally occupied by electrons. If an Si atom is substituted by, say, a phosphorous atom, the latter, called *impurity atom*, has five electrons in the outer shell. Four of them contribute to the bonds, and the fifth is kept around the atom by the Coulomb force of the nucleus. The whole impurity forms a system with properties similar to a hydrogen atom, and for this reason is called a *hydrogenic impurity*. With respect to a true hydrogen atom, there are two important differences: owing to the effect of the periodic potential of the host crystal, the dynamics of the electron is modified with respect to that of electrons in isolated atoms, and it is possible to account for this difference with a reasonable accuracy by using the effective-mass approximation described in Sect. 6.6. The second main difference is that the dielectric constant of the material must be inserted in the Coulomb interaction between the nucleus charge and the electron. These two features of the theoretical approach are justified by the fact that the resulting electron wavefunction is large compared with the lattice constant. The resulting energy level is slightly negative, i.e., the electron is weakly bound to the atom. As a consequence, the impurity is described by a bound localized state with energy just below the bottom of the conduction band, as shown in part (d) of Fig. 8.5. A small thermal energy is then sufficient to ionize the impurity with the effect of leaving a fixed positive charge in the crystal (the *ionized impurity*) and an electron free to move in the conduction band. For such a reason, this kind of impurity is called a *donor*.

The hydrogenic model for impurities of group V in silicon is a good approximation, also for its excited states above the ground level, but is not exact. One of the main differences between the energy levels obtained with this approximation and the experimental values is the so-called *chemical shift*. It is due to the interaction of the "extra" electron with the *core electrons* of the impurity atom, i.e., the electrons closer to the nucleus than the valence electrons. This

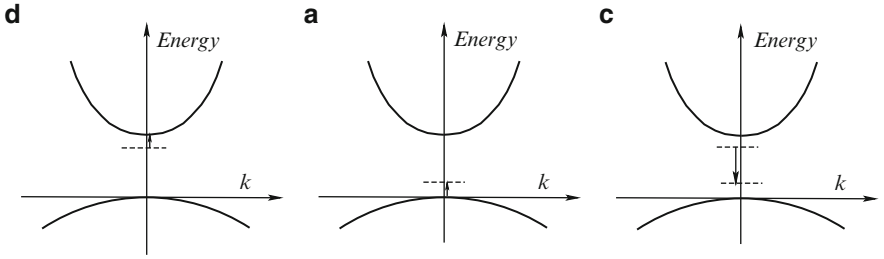


Fig. 8.5. (**d**) Donor levels, (**a**) Acceptor levels, and (**c**) compensated donors and acceptors

interaction is such that the potential deviates from the pure Coulombic nature and depends on the chemical species of the impurity. Furthermore, the impurity distorts the crystal potential in its vicinity. The part of this distortion inside the unit cell of the impurity is called *central cell correction*. It may be a strongly varying potential so that the treatment of the electron dynamics in terms of effective mass may break down. *Deep levels* (closer to the center of the energy gap) may also result, which are much more difficult to calculate than the hydrogenic impurity levels.

In a similar and symmetric way, if an impurity of the third group of the periodic table, such as aluminum, is substitutionally inserted into a crystal of silicon or germanium, an electron is missing for the formation of the tetrahedral bonds with the neighboring atoms. An electron that would come around and complete the bonds would have a somewhat higher energy with respect to the other binding electrons because the corresponding nuclear attraction is missing. An electron state is generated by the impurity that is empty at zero temperature, but may easily be occupied, with thermal excitation, by an electron coming from the valence band. The impurity is thus called an *acceptor*, whose energy level is near, and above, the top of the valence band, as shown in part (a) of Fig. 8.5. When an electron leaves the valence band to occupy an acceptor level, it generates a hole; the impurity is negatively charged, since there is an extra electron with respect to the nuclear charge, and the hole tends to stay close to the "mother" impurity. Thus, the description given above may be reformulated in terms of a hole state with energy somewhat lower (remember that the energy of a hole is opposite of that of the missing electron) of the top of the valence band, as long as it remains bound to the impurity in a sort of hydrogenic atom with reverse charges. The hole can easily be excited to the valence band and move freely around the crystal. This situation is perfectly symmetric to that of the donor impurity. An acceptor impurity "donates" holes for the conduction mechanism.

In a III–V semiconductor, such as GaAs, the tetrahedral bonds are formed by three electrons furnished by the Ga atoms and five electrons furnished by the As atoms. From the above discussion, it should be clear that an atom of, say, silicon, behaves as a donor impurity if it is located in a position that in

the pure crystal belongs to gallium. In this case, the impurity is indicated as Si_{Ga}. It behaves as an acceptor impurity if it takes an arsenic position (Si_{As}).

8.5.2 n-Doped and p-Doped Semiconductors

At zero temperature, the valence band of a perfect semiconductor is entirely occupied by electrons, and the conduction band is empty. Thus, as seen above, a semiconductor at $T = 0$ is actually an insulator. For some electrons to be "promoted" from the valence band to the conduction band, an energy at least equal to the band gap is necessary. If we consider that the band gap of Si is about 1.1 eV, corresponding to 1.3×10^4 K and that of GaAs is 1.4 eV (1.7×10^4 K), we understand that the conductivity of the intrinsic semiconductors of most technological interest is negligible. If, however, a number of donor impurities are present in the material, a significant fraction of them can be ionized at lower temperature, and free electrons are available in the conduction band. The conductivity of the semiconductor is increased by orders of magnitude, as anticipated at the beginning of Sect. 8.5.1. Since the current is carried in this case by electrons with negative charge, the material is called a *doped semiconductor of n type*. If the impurities present in the material are acceptors, the free carriers are positive holes and the doped semiconductor is of *p type*.

When both types of impurities are present, electrons may fall from donor states into acceptor states; both types of impurities become ionized, but the free charge carriers available for the conduction mechanism are reduced with respect to the case when only one type is present. The semiconductor is said to be *compensated*. This situation is illustrated in part (c) of Fig. 8.5. The density of free charge carriers in the conduction or valence band at equilibrium is determined by the densities of dopants and by statistics, as discussed in next section.

8.6 Charge-Carrier Statistics

The probability that an electron state in a semiconductor with energy ϵ is occupied by an electron is given by the Fermi–Dirac distribution in (3.19), repeated here for convenience:

$$f_F(\epsilon) = \frac{1}{e^{(\epsilon-\mu)/K_B T} + 1},$$

where μ is the electrochemical potential.

In the limit of zero temperature, the Fermi distribution has a typical rectangular shape as indicated in part (a) of Fig. 8.6; At $T > 0$, the energy interval in which the distribution is appreciably less than 1 and greater than zero is

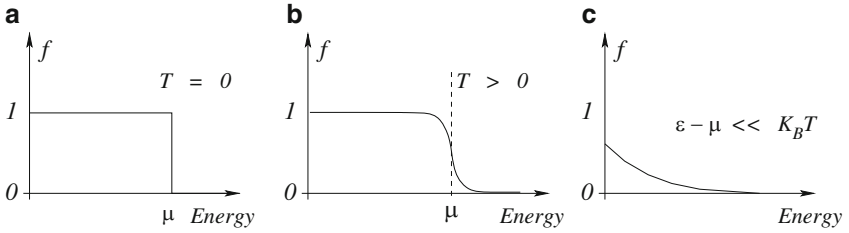


Fig. 8.6. Fermi distributions at (**a**) zero temperature, (**b**) positive temperature, and (**c**) in the classical limit of $\epsilon - \mu \gg K_{\mathrm{B}}T$

of the order of $K_{\mathrm{B}}T$; When $\epsilon - \mu \gg K_{\mathrm{B}}T$ the Fermi distribution becomes the classical Boltzmann distribution:

$$f(\epsilon) \propto \mathrm{e}^{-\epsilon/K_{\mathrm{B}}T},$$

as shown in Fig. 8.6.

In analyzing the distribution of electrons in energy, it is useful to consider also the case of metals, both for comparison with semiconductor materials, and because semiconductor devices have metal contacts.

8.6.1 Metals

Let us consider the electrons in a metallic band and define the origin of the energy ϵ at the bottom of this band. If $g_\epsilon(\epsilon)$ is the density of states in energy, the total number N of electrons in the band is given by

$$N = \int_0^\infty g_\epsilon(\epsilon) f_F(\epsilon) \mathrm{d}\epsilon. \tag{8.5}$$

This normalization equation determines the electrochemical potential μ. In fact, if we define

$$G(\epsilon) = \int_0^\epsilon g_\epsilon(\epsilon') \mathrm{d}\epsilon'$$

and integrate (8.5) by parts, we obtain

$$N = G(\epsilon) f_F(\epsilon)\big|_0^\infty - \int_0^\infty G(\epsilon) \frac{\mathrm{d}f_F}{\mathrm{d}\epsilon} \mathrm{d}\epsilon = \int_0^\infty G(\epsilon) \left(-\frac{\mathrm{d}f_F}{\mathrm{d}\epsilon} \right) \mathrm{d}\epsilon, \tag{8.6}$$

where we have taken into account that $G(0) = 0$ and $f_F(\infty) = 0$. At $T = 0$ the Fermi distribution is a perfect step at $\mu = \epsilon_F$; its derivative is $-\delta(\epsilon - \epsilon_F)$, and (8.6) yields

$$N = G(\epsilon_F) = \int_0^{\epsilon_F} g_\epsilon(\epsilon)\, \mathrm{d}\epsilon. \tag{8.7}$$

This result is obvious since at $T = 0$ all states below ϵ_F are occupied, and all states above are empty.

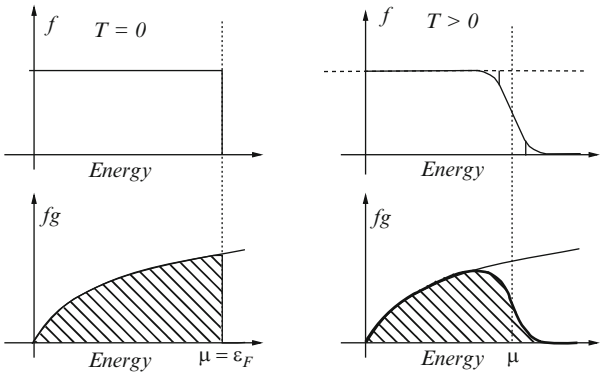


Fig. 8.7. *Top*: Fermi distributions at $T = 0$ and at $T > 0$. *Bottom*: the same distributions multiplied by the density of states. The shaded areas indicate the total number of electrons

At increasing T above zero, we may easily realize that μ must decrease, with respect to ϵ_F if, as happens most of the times, the density of states g is an increasing function of energy. In fact, we note that the Fermi distribution has the symmetry property indicated by the relation

$$1 - f_F(\mu - \delta) = f_F(\mu + \delta),$$

that can immediately be verified from the definition. This symmetry implies that the two small vertical lines on the top right parts of Fig. 8.7 are equal. At $T = 0$, all states up to ϵ_F are occupied, as indicated in the lower left part of Fig. 8.7, where the shaded area represents the total number of electrons. At increasing T above zero, the distribution spreads around μ in an interval of energies of the order of $K_B T$. Should the electrochemical potential remain constant, the fraction of empty states below μ would be equal to the fraction of occupied states above μ for the symmetry just seen. For an increasing density of states, this would imply a larger total number of electrons, that, instead, must remain the same. Thus, μ must decrease at increasing temperature to compensate the increasing density of states.

It is possible to translate the above considerations into a quantitative, analytical form. Let us expand $G(\epsilon)$ in powers of ϵ around μ inside the integral in (8.6), obtaining:

$$N = \int_0^\infty \left[G(\mu) + G'(\mu)(\epsilon - \mu) + \frac{1}{2} G''(\mu)(\epsilon - \mu)^2 + \ldots \right] \left(-\frac{\mathrm{d}f_F}{\mathrm{d}\epsilon} \right) \mathrm{d}\epsilon. \quad (8.8)$$

The first term yields immediately $G(\mu)$ assuming that at our temperature $f_F(0)$ is still one, as shown in Fig. 8.8.

For the second term, let us consider that

$$-\frac{\mathrm{d}f_F}{\mathrm{d}\epsilon} = \frac{1}{K_B T} \frac{\mathrm{e}^{(\epsilon-\mu)/K_B T}}{\left(\mathrm{e}^{(\epsilon-\mu)/K_B T} + 1\right)^2} = \frac{1}{K_B T} \frac{1}{\left(1 + \mathrm{e}^{(\epsilon-\mu)/K_B T}\right)\left(1 + \mathrm{e}^{-(\epsilon-\mu)/K_B T}\right)}.$$

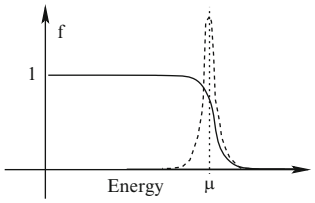


Fig. 8.8. Fermi distribution (*continuous line*) and its derivative (*broken line*)

Thus, the second term in (8.8) is

$$G'(\mu)K_{\mathrm{B}}T\int_{-\infty}^{\infty}\frac{z}{(1+\mathrm{e}^{z})\,(1+\mathrm{e}^{-z})}\mathrm{d}z=0,$$

which vanishes because the integrand is odd. The first limit of integration could be shifted to $-\infty$ as the derivative of the delta function limits the non-vanishing values of the integrand function to a small region around μ (see Fig. 8.8). This result is to be connected with the qualitative discussion above, taking into account that the linear term of G corresponds to a constant density of states g_ϵ.

The third term in (8.8) can be calculated using tabulated integrals (or evaluated as summable series):

$$\frac{1}{2}G''(\mu)(K_{\mathrm{B}}T)^2\int_{-\infty}^{\infty}\frac{z^2}{(1+\mathrm{e}^{z})\,(1+\mathrm{e}^{-z})}\mathrm{d}z=G''(\mu)(K_{\mathrm{B}}T)^2\frac{1}{6}\pi^2.$$

Collecting our results, (8.8), together with (8.7), yields

$$N=\int_0^{\epsilon_F}g_\epsilon(\epsilon)\mathrm{d}\epsilon=\int_0^{\mu}g_\epsilon(\epsilon)\mathrm{d}\epsilon+\frac{\pi^2}{6}(K_{\mathrm{B}}T)^2\frac{\mathrm{d}g_\epsilon(\epsilon)}{\mathrm{d}\epsilon}\bigg|_{\mu}+\ldots \tag{8.9}$$

If the density of states is known as function of energy, the above expression, neglecting higher-order terms, yields the electrochemical potential μ as a function of temperature. An analytical expression can be obtained as long as μ remains close to the Fermi energy, its value at zero temperature. In such a case, in fact, (8.9) becomes

$$\int_0^{\epsilon_F}g_\epsilon(\epsilon)\mathrm{d}\epsilon-\int_0^{\mu}g_\epsilon(\epsilon)\mathrm{d}\epsilon=\int_{\mu}^{\epsilon_F}g_\epsilon(\epsilon)\mathrm{d}\epsilon\approx g_\epsilon(\epsilon_F)(\epsilon_F-\mu)=\frac{\pi^2}{6}(K_{\mathrm{B}}T)^2\frac{\mathrm{d}g_\epsilon(\epsilon)}{\mathrm{d}\epsilon}\bigg|_{\mu}.$$

From this, we obtain

$$\mu\approx\epsilon_F-\frac{\pi^2}{6}(K_{\mathrm{B}}T)^2\frac{\mathrm{d}}{\mathrm{d}\epsilon}\log(g_\epsilon(\epsilon))\bigg|_{\epsilon_F}. \tag{8.10}$$

This result confirms that, as expected, the electrochemical potential decreases with increasing temperature above zero, if the density of states is an increasing function of energy.

The result in (8.10) may be given a more explicit form if we know the analytical form of the density of states $g_\epsilon(\epsilon)$. For a parabolic band, it is easy to find g_ϵ and the result is very important since it is useful in many physical situations. The density of states in $\boldsymbol{k}$-space is known from the theory of Bloch states and it is given by (6.7). Accordingly, the number of states in a shell of radius k and thickness $\mathrm{d}k$ is, including spin degeneracy,

$$g(k)\mathrm{d}k = 2\frac{V}{(2\pi)^3}4\pi k^2\mathrm{d}k. \tag{8.11}$$

For a parabolic band with effective mass m:

$$\epsilon = \frac{\hbar^2k^2}{2m}, \quad \mathrm{d}\epsilon = \frac{\hbar^2k}{m}\mathrm{d}k, \quad k = \frac{\sqrt{2m\epsilon}}{\hbar}, \quad \mathrm{d}k = \frac{\sqrt{2m}}{\hbar}\frac{1}{2\sqrt{\epsilon}}\,\mathrm{d}\epsilon.$$

Thus,

$$\boxed{g_\epsilon(\epsilon)\,\mathrm{d}\epsilon = \tfrac{V}{2\pi^2}\left(\tfrac{2m}{\hbar^2}\right)^{3/2}\sqrt{\epsilon}\;\mathrm{d}\epsilon} \tag{8.12}$$

With this result, (8.10) becomes

$$\mu \approx \epsilon_F - \frac{\pi^2}{6}(K_\mathrm{B}T)^2\frac{1}{2\epsilon_F} = \epsilon_F\left[1 - \frac{\pi^2}{12}\left(\frac{K_\mathrm{B}T}{\epsilon_F}\right)^2\right].$$

8.6.2 Semiconductors

We have seen that semiconductors have a concentration of free charge carriers, electrons and/or holes, strongly influenced by doping. It is convenient to introduce first some notation that is independent of the particular situation. Then we shall consider a few particular cases, but the general problem of carrier statistics in semiconductors is very complex and diversified, and we refer the interested reader to the classic text of Blackemore [47].

If ϵ_v is the top of the valence band and ϵ_c the bottom of the conduction band, so that the band gap $\Delta\epsilon_g$ is $\epsilon_c - \epsilon_v$, we define

$$\epsilon_\mathrm{e} = \epsilon - \epsilon_c, \quad \epsilon_h = \epsilon_v - \epsilon, \tag{8.13}$$

as electron and hole energies, respectively, as shown in Fig. 8.9. Note that the hole energy is defined in the opposite direction with respect to the electron energy, as discussed above. Then the electron distribution becomes

$$f_\mathrm{e}(\epsilon_\mathrm{e}) = f_F(\epsilon_c + \epsilon_\mathrm{e}) = \frac{1}{\mathrm{e}^{(\epsilon_c+\epsilon_\mathrm{e}-\mu)/K_\mathrm{B}T}+1} = \frac{1}{\mathrm{e}^{(\epsilon_\mathrm{e}-\mu_\mathrm{e})/K_\mathrm{B}T}+1},$$

where we have put, coherently with the above definitions,

$$\mu_\mathrm{e} = \mu - \epsilon_c.$$

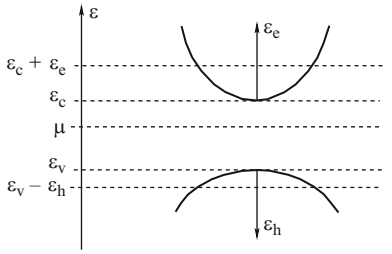


Fig. 8.9. Electron and hole energies

Accordingly, since the probability to find a hole in a state with energy ϵ_h, is the probability that an electron is absent in the state with corresponding energy $\epsilon = \epsilon_v - \epsilon_h$, we have

$$f_h(\epsilon_h) = 1 - f_F(\epsilon_v - \epsilon_h) = 1 - \frac{1}{\mathrm{e}^{\frac{\epsilon_v - \epsilon_h - \mu}{K_\mathrm{B}T}} + 1} = \frac{1}{\mathrm{e}^{\frac{-\epsilon_v + \epsilon_h + \mu}{K_\mathrm{B}T}} + 1} = \frac{1}{\mathrm{e}^{\frac{\epsilon_h - \mu_h}{K_\mathrm{B}T}} + 1},$$

where

$$\mu_h = \epsilon_v - \mu \; [= -(\mu - \epsilon_v)],$$

coherently with (8.13): electrochemical potential measured from ϵ_v with opposite sign.

Nondegenerate Semiconductors – Law of Mass Action

If μ is far from both band edges, such that

$$\epsilon_c - \mu \gg K_\mathrm{B}T \text{ and } \mu - \epsilon_v \gg K_\mathrm{B}T, \tag{8.14}$$

the following inequalities are also true:

$$\epsilon_\mathrm{e} - \mu_\mathrm{e} = \epsilon - \epsilon_c - \mu + \epsilon_c \gg K_\mathrm{B}T \text{ and } \epsilon_h - \mu_h = \epsilon_v - \epsilon - \epsilon_v + \mu \gg K_\mathrm{B}T,$$

where it has been taken into account that for electrons $\epsilon > \epsilon_c$ and for holes $\epsilon < \epsilon_v$. The following approximations are then correct, which define *nondegenerate semiconductors*:

$$f_\mathrm{e}(\epsilon_\mathrm{e}) \approx \mathrm{e}^{-\frac{\epsilon_\mathrm{e} - \mu_\mathrm{e}}{K_\mathrm{B}T}} \propto \mathrm{e}^{-\frac{\epsilon_\mathrm{e}}{K_\mathrm{B}T}}, \quad f_h(\epsilon_h) \approx \mathrm{e}^{-\frac{\epsilon_h - \mu_h}{K_\mathrm{B}T}} \propto \mathrm{e}^{-\frac{\epsilon_h}{K_\mathrm{B}T}}. \tag{8.15}$$

Thus, the number of electrons can be obtained as:

$$N_\mathrm{e} = \int_{\epsilon_c}^{\infty} g_c(\epsilon)\mathrm{e}^{-\frac{\epsilon - \mu}{K_\mathrm{B}T}}\mathrm{d}\epsilon = \mathrm{e}^{-\frac{\epsilon_c - \mu}{K_\mathrm{B}T}} \int_{\epsilon_c}^{\infty} g_c(\epsilon)\mathrm{e}^{-\frac{\epsilon - \epsilon_c}{K_\mathrm{B}T}}\mathrm{d}\epsilon = N_c(T)\mathrm{e}^{-\frac{\epsilon_c - \mu}{K_\mathrm{B}T}},$$

where g_c is the density of states in the conduction band,

$$N_c(T) = \int_{\epsilon_c}^{\infty} g_c(\epsilon) \mathrm{e}^{-\frac{\epsilon-\epsilon_c}{K_\mathrm{B}T}} \mathrm{d}\epsilon = \int_0^{\infty} g_\mathrm{e}(\epsilon_\mathrm{e}) \mathrm{e}^{-\frac{\epsilon_\mathrm{e}}{K_\mathrm{B}T}} \mathrm{d}\epsilon_\mathrm{e},$$

and $g_\mathrm{e}(\epsilon_\mathrm{e}) = g_c(\epsilon_c + \epsilon_\mathrm{e})$. Similarly, the number of holes is found as

$$N_h = \int_{-\infty}^{\epsilon_v} g_v(\epsilon) \mathrm{e}^{-\frac{\mu-\epsilon}{K_\mathrm{B}T}} \mathrm{d}\epsilon = N_v(T) \mathrm{e}^{-\frac{\mu-\epsilon_v}{K_\mathrm{B}T}},$$

where g_v is the density of states in the valence band,

$$N_v(T) = \int_{-\infty}^{\epsilon_v} g_v(\epsilon) \mathrm{e}^{-\frac{\epsilon_v-\epsilon}{K_\mathrm{B}T}} \mathrm{d}\epsilon = \int_0^{\infty} g_h(\epsilon_h) \mathrm{e}^{-\frac{\epsilon_h}{K_\mathrm{B}T}} \mathrm{d}\epsilon_h,$$

and $g_h(\epsilon_h) = g_v(\epsilon_v - \epsilon_h)$. Note that the above results have been obtained in the hypothesis of non degeneracy (8.14) without any assumption about the presence of doping.

If we assume that the bands are parabolic in the region of energies of interest, with effective masses m_e and m_h, the densities of states are those given by (8.12), and the results take the more explicit forms:

$$N_c(T) = \int_0^{\infty} \frac{V}{2\pi^2} \left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2} \sqrt{\epsilon_\mathrm{e}}\ \mathrm{e}^{-\frac{\epsilon_\mathrm{e}}{K_\mathrm{B}T}}\ \mathrm{d}\epsilon_\mathrm{e} = \frac{V}{4} \left(\frac{2m_\mathrm{e}K_\mathrm{B}T}{\pi\hbar^2}\right)^{3/2},$$

where we have used the relation $\int_0^{\infty} \sqrt{z}\mathrm{e}^{-z}\mathrm{d}z = \sqrt{\pi}/2$. The electron concentration is then given by

$$n_\mathrm{e} = \frac{N_\mathrm{e}}{V} = \frac{1}{4} \left(\frac{2m_\mathrm{e}K_\mathrm{B}T}{\pi\hbar^2}\right)^{3/2} \mathrm{e}^{-\frac{\epsilon_c-\mu}{K_\mathrm{B}T}}. \tag{8.16}$$

Similarly for the holes:

$$n_h = \frac{N_h}{V} = \frac{1}{4} \left(\frac{2m_h K_\mathrm{B}T}{\pi\hbar^2}\right)^{3/2} \mathrm{e}^{-\frac{\mu-\epsilon_v}{K_\mathrm{B}T}}.$$

From the above a very important result follows for the product of the electron and hole concentrations:

$$\boxed{n_\mathrm{e} n_h = \tfrac{1}{2} \left(\tfrac{K_\mathrm{B}T}{\pi\hbar^2}\right)^3 (m_\mathrm{e} m_h)^{3/2}\, \mathrm{e}^{-\frac{\Delta\epsilon_g}{K_\mathrm{B}T}}} \tag{8.17}$$

where $\Delta\epsilon_g$ is the band gap. This is the *law of mass action* for the charge carriers at equilibrium in nondegenerate semiconductors. Its importance derives from the fact that it holds independently of the position of μ and therefore of the type of doping. If the amount of one type of carriers increases, part of the carriers of opposite sign will recombine in such a way that the product of the concentrations remains unchanged.

If the semiconductor is degenerate, i.e., when the conditions (8.14) are not fulfilled, and therefore the approximations (8.15) do not hold, the statistics of the carriers is more complex. A detailed treatment can be found in [47].

Intrinsic, Nondegenerate Semiconductors

If the nondegenerate semiconductor treated in the previous section is also intrinsic, the concentrations of electrons and holes must be the same, and the law of mass action (8.17) yields

$$n_{\mathrm{e}} = n_h = \frac{1}{4}\left(\frac{2K_{\mathrm{B}}T}{\pi\hbar^2}\right)^{3/2}(m_{\mathrm{e}}m_h)^{3/4}\,\mathrm{e}^{-\frac{\Delta\epsilon_g}{2K_{\mathrm{B}}T}}.$$

Comparing with (8.16) yields

$$\mathrm{e}^{-\frac{\epsilon_c-\mu}{K_{\mathrm{B}}T}} = \left(\frac{m_h}{m_{\mathrm{e}}}\right)^{\frac{3}{4}}\mathrm{e}^{-\frac{\Delta\epsilon_g}{2K_{\mathrm{B}}T}},$$

or

$$-\epsilon_c + \mu + \frac{\Delta\epsilon_g}{2} = K_{\mathrm{B}}T\frac{3}{4}\log\left(\frac{m_h}{m_{\mathrm{e}}}\right).$$

Since $\epsilon_c = \epsilon_v + \Delta\epsilon_g$, the above yields the position of the electrochemical potential as function of temperature for intrinsic, nondegenerate semiconductors:

$$\mu = \epsilon_v + \frac{1}{2}\Delta\epsilon_g + \frac{3}{4}K_{\mathrm{B}}T\log\left(\frac{m_h}{m_{\mathrm{e}}}\right).$$

From this expression, we learn that at $T = 0$ the Fermi level is in the middle of the gap and that at increasing temperatures it moves toward the band with lower effective mass. This result is coherent with what we have seen in metals, since a lower effective mass means a lower density of states, as indicated in (8.12).

Again for more complex situations of doped and/or degenerate materials, we refer to [47].

8.7 General Models of Bands for Cubic Semiconductors

From the foregoing, it is clear that the bands of interest for electrical transport are the conduction and the valence bands for electrons and holes, respectively. For cubic semiconductors with diamond and zinc-blende structure, a general model for the band structure may be considered [213], as shown in Fig. 8.10. It consists of one conduction band, with three sets of minima, and three valence bands. The minima of the conduction bands are located at the Γ point ($k = 0$, see Fig. 4.4), at the L points ($\boldsymbol{k} = (\pi/a, \pi/a, \pi/a)$, a being the lattice constant), and along the Δ lines ($\boldsymbol{k} = k, 0, 0$). The tops of the valence bands are located at Γ. Two of these bands are degenerate at this point, while the third one is split off by spin–orbit interaction.

We shall see in the following chapters that the particular form of the energy-wavevector relationship $\epsilon(\boldsymbol{k})$ of charge carriers is of great importance

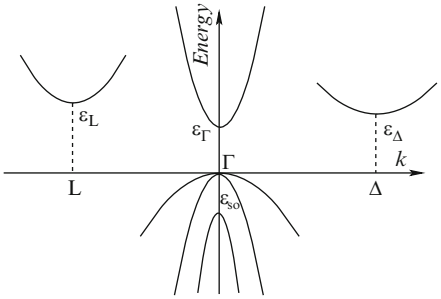


Fig. 8.10. A general model for band structure of cubic semiconductors

for the transport properties, because it determines both the velocity of each electron state and the density of states, of central importance for electron scattering.

In the regions around the minima of the conduction band, called *valleys* or around the maximum of the valence band, the function $\epsilon(\boldsymbol{k})$ may be approximated by a quadratic function of $\boldsymbol{k}$ (parabolic bands, see Sect. 6.6), which may assume one of the following forms:

$$\epsilon(\boldsymbol{k}) = \frac{\hbar^2 k^2}{2m} \quad \text{(spherical bands)}; \tag{8.18}$$

$$\epsilon(\boldsymbol{k}) = \frac{\hbar^2}{2}\left[\frac{k_\ell^2}{m_\ell} + \frac{k_t^2}{m_t}\right] \quad \text{(ellipsoidal bands)}; \tag{8.19}$$

$$\epsilon(\boldsymbol{k}) = ak^2[1 \mp g(\vartheta, \psi)] \quad \text{(warped bands)}. \tag{8.20}$$

In the above equations, $\boldsymbol{k}$ is measured from the value of the wavevector where the band has its relative minimum, that is from the center of the valley.

Equation (8.18) represents a band with spherical equienergetic surfaces as shown in Fig. 8.11, with a single scalar effective mass m, and it is appropriate for the minimum of the conduction band located at Γ and for the maximum of the split-off band. This is the simplest case and it is generally adopted as a simple model for any material when rough estimates of transport properties are sought.

Equation (8.19) represents a band with ellipsoidal equienergetic surfaces (see Fig. 8.11) with a diagonal inverse effective-mass tensor. $1/m_\ell$, and $1/m_t$ are the longitudinal and transverse components, respectively, of the inverse effective mass (see (6.18)). The ellipsoids have rotational symmetry around the crystallographic directions which contain the center of the valleys. This case is appropriate for the minima of the conduction bands located at L and along Δ; for symmetry reasons several equivalent valleys are present. In the *many-valley model*, it is assumed that an electron cannot move from one valley to another with continuous variation of its momentum as effect of the applied fields, because of the existence of intermediate regions of $\boldsymbol{k}$ space with too high

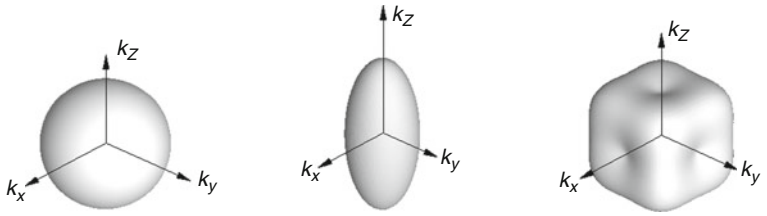


Fig. 8.11. Different types of equienergy surfaces in cubic semiconductors: spherical, for electron valleys at Γ, ellipsoidal for electron valleys along high symmetry directions such as Δ or Λ, and warped, for heavy holes

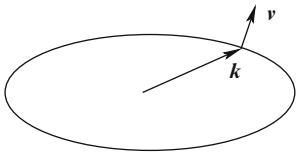


Fig. 8.12. Crystal momentum and velocity are not parallel, in general, in ellipsoidal bands. The velocity is always orthogonal to the equienergetic surfaces

energies. This assumption is removed in the full-band models, as discussed in Sect. 14.2.8. It is important to note that in ellipsoidal valleys the group velocity $\boldsymbol{v}(\boldsymbol{k})$ of charge carriers, always orthogonal to the equienergetic surfaces, is not parallel, in general, to their wavevector $\boldsymbol{k}$. From (6.24), in fact, we have, for a diagonal inverse effective-mass tensor,

$$v_i = \frac{\hbar}{m_i} k_i. \tag{8.21}$$

Since the constant of proportionality between momentum and velocity is different along the longitudinal and transverse directions, the crystal momentum $\hbar\boldsymbol{k}$ and the velocity are no more parallel in general (they are still parallel along the symmetry directions), as shown in Fig. 8.12.

Equation (8.20) represents a band with warped equienergetic surfaces (see Fig. 8.11) and is appropriate for the two degenerate maxima of the valence bands. Here, the signs $\mp$ refer to heavy (minus sign) and light (plus sign) holes. The symbols ϑ and ψ indicate the polar and azimuthal angles of $\boldsymbol{k}$ with respect to crystallographic axis where the valley is located so that $g(\vartheta, \psi)$ contains the angular dependence of the effective mass and is given by [122]

$$g(\vartheta, \psi) = [b^2 + c^2(\sin^4\vartheta\cos^2\psi\sin^2\psi + \sin^2\vartheta\cos^2\vartheta)]^{1/2}, \tag{8.22}$$

with

$$a = \frac{\hbar^2|A|}{2m_\circ}, \quad b = \frac{|B|}{|A|}, \quad c = \frac{|C|}{|A|},$$

where A, B, and C are the inverse valence-band parameters, [122, 231, 363] and $m_\circ$ the free-electron mass. An analysis of the warped bands of holes in silicon can be found in [333].

8.7.1 Different Types of Effective Masses

The standard definition of inverse effective-mass tensor is given in (6.18) of Sect. 6.6 and is related to the quadratic shape of the band as approximation around the extrema of the band function $\epsilon(\boldsymbol{k})$. However, it is often useful to consider specific different definitions of the effective mass, also for nonparabolic bands, which depend on its specific different use.

Acceleration Effective Mass

The *acceleration effective mass* is defined in connection with the fundamental law of (semiclassical) dynamics. The acceleration of a Bloch wavepacket is given by

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\hbar}\frac{\partial \epsilon}{\partial k_i} = \frac{1}{\hbar^2}\sum_j \frac{\partial^2 \epsilon(\boldsymbol{k})}{\partial k_i \partial k_j}\frac{\mathrm{d}(\hbar k_j)}{\mathrm{d}t} = \frac{1}{\hbar^2}\sum_j \frac{\partial^2 \epsilon(\boldsymbol{k})}{\partial k_i \partial k_j} F_j,$$

where use has been made of (8.1). If this expression is compared with fundamental law of dynamics $\boldsymbol{F} = m\boldsymbol{a}$, then the inverse acceleration effective mass is defined by the second derivative

$$\left(\frac{1}{m_a}\right)_{ij} = \frac{1}{\hbar^2}\frac{\partial^2 \epsilon(\boldsymbol{k})}{\partial k_i \partial k_j}. \tag{8.23}$$

In the parabolic case, the above acceleration effective mass coincides with the standard definition in (6.18).

Conductivity Effective Mass

As we shall see in the following chapters, when an electric field is applied to a semiconductor, the balance between the acceleration effect of the field and the equilibrating effect of the collisions produces an average crystal momentum $\hbar\boldsymbol{k}_\circ$. The current is related to the average velocity, so that a *conductivity effective mass* m_c is defined by means of the ratio $\hbar\boldsymbol{k}/m_c = \boldsymbol{v}$. In a general band, this reads

$$v_i = \frac{1}{\hbar}\frac{\partial \epsilon}{\partial k_i} = \sum_j \left(\frac{1}{m_c}\right)_{ij} \hbar k_j.$$

In a reference frame in which the tensor $1/m_c$ is diagonal, the above definition yields

$$\left(\frac{1}{m_c}\right)_{ii} = \frac{1}{\hbar^2}\frac{1}{k_i}\frac{\partial\epsilon}{\partial k_i}.$$

In the parabolic case, for a single ellipsoid the result is $m_c = m_\ell$ or $m_c = m_t$ according to the direction considered. However, we know that for symmetry reasons there are several ellipsoidal valleys and the ratio between current and applied field in regime of linear response[3] is independent of the direction of the applied field. Thus, if in silicon we apply a field along a [100] direction, two valleys have a drift momentum along the longitudinal direction of the ellipsoid, and four along a transverse direction. Thus, a simple average of the contributions of the different valleys (the contributions of the different valleys add up, but the electrons divide themselves among the six valleys) leads to

$$\frac{1}{m_c} = \frac{1}{6}\left(\frac{2}{m_\ell} + \frac{4}{m_t}\right) = \frac{1}{3m_\ell} + \frac{2}{3m_t}.$$

Density-of-States Effective Mass

The *density-of-states effective mass* relates to the mass that in the spherical and parabolic bands appears in the expression (8.12) for the density of states. Since in $\boldsymbol{k}$-space the density of states has the constant value $g(\boldsymbol{k}) = 2V/(2\pi)^3$, as given by (6.7) with the addition of the spin multiplicity, the number of electron states in the energy interval between ϵ and $\epsilon + \delta\epsilon$ is given by

$$\delta N = g(\boldsymbol{k})\delta V_k = g(\boldsymbol{k})\int_{S(\epsilon)} \mathrm{d}\sigma\ \mathrm{d}k_\perp = g(\boldsymbol{k})\int_{S(\epsilon)} \mathrm{d}\sigma \frac{\mathrm{d}\epsilon}{|\nabla_k \epsilon|},$$

where δV_k is the volume in $\boldsymbol{k}$-space containing states with energy between ϵ and $\epsilon + \delta\epsilon$; $S(\epsilon)$ is the surface in $\boldsymbol{k}$-space of the points with energy ϵ; $\mathrm{d}k_\perp$ is the increment of crystal wavevector orthogonal to S corresponding to the energy increment $\delta\epsilon$. The density of states in energy is thus given by

$$g_\epsilon(\epsilon) = g(\boldsymbol{k})\int_{S(\epsilon)} \frac{\mathrm{d}\sigma}{|\nabla_k \epsilon|}. \tag{8.24}$$

If the band function $\epsilon(\boldsymbol{k})$ is known, the above integral yields the density of states, and the comparison of the result with (8.12) leads to the definition of the density-of-state effective mass.

In the case of parabolic bands with ellipsoidal equienergetic surfaces, we may apply the procedure indicated above, but there is a simpler approach. Let

[3] In the linear response regime, by definition, the current density $\boldsymbol{j}$ produced by an applied field $\boldsymbol{E}$ is proportional to $\boldsymbol{E}$ so that in the equation $\boldsymbol{j} = \sigma\boldsymbol{E}$, the conductivity tensor σ must be a property of the material, independent of the applied field. In a semiconductor with cubic symmetry, this tensor reduces to a scalar, i.e., must be diagonal with equal diagonal elements.

us consider the transformation of variables (the Herring–Vogt transformation discussed below):

$$k_\ell^* = k_\ell \sqrt{\frac{m_\circ}{m_\ell}}, \quad k_t^* = k_t \sqrt{\frac{m_\circ}{m_t}},$$

where $m_\circ$ is the free electron mass. In the new variables, the band in (8.19) assumes a spherical form

$$\epsilon(\boldsymbol{k}) = \frac{\hbar^2 k^{*2}}{2m_\circ}. \tag{8.25}$$

The density of states in the starred space $g^*(\boldsymbol{k}^*)$ is such that

$$g^*(\boldsymbol{k}^*)\mathrm{d}k_x^*\mathrm{d}k_y^*\mathrm{d}k_z^* = g(\boldsymbol{k})\mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z = g(\boldsymbol{k})\mathrm{d}k_x^*\mathrm{d}k_y^*\mathrm{d}k_z^* \sqrt{\frac{m_t^2 m_\ell}{m_\circ^3}},$$

or

$$g^*(\boldsymbol{k}^*) = g(\boldsymbol{k}) \sqrt{\frac{m_t^2 m_\ell}{m_\circ^3}}.$$

For the spherical band in (8.25), and the density of states above, following the same argument that led to (8.12), we obtain

$$g_\epsilon(\epsilon)\mathrm{d}\epsilon = \frac{V}{2\pi^2} \sqrt{\frac{m_t^2 m_\ell}{m_\circ^3}} \left(\frac{2m_\circ}{\hbar^2}\right)^{3/2} \sqrt{\epsilon}\ \mathrm{d}\epsilon,$$

and the comparison with (8.12) leads to the density-of-states effective mass for an ellipsoidal valley given by

$$m_d = (m_\ell m_t^2)^{1/3}. \tag{8.26}$$

To obtain the total density of states, this result must be multiplied by the number of equivalent valleys.

8.7.2 Herring–Vogt Transformation

When considering the ellipsoidal bands in (8.19), to simplify the analytical calculations, it is often useful to introduce the *Herring–Vogt transformation* [185], which reduces the ellipsoidal equienergetic surfaces to spheres and is defined by

$$k_i^{*(m)} = \sum_j T_{ij} k_j^{(m)}, \tag{8.27}$$

where $\boldsymbol{k}^{*(m)}$ is the transformed wavevector of an electron with wavevector $\boldsymbol{k}^{(m)}$ in the m-th valley. In the valley frame of reference, i.e., in the frame centered at the center of the valley, with the z axis along its symmetry axis, the transformation matrix takes the form

$$T_{ij} = \begin{pmatrix} \left(\frac{m_\circ}{m_t}\right)^{1/2} & 0 & 0 \\ 0 & \left(\frac{m_\circ}{m_t}\right)^{1/2} & 0 \\ 0 & 0 & \left(\frac{m_\circ}{m_\ell}\right)^{1/2} \end{pmatrix}.$$

Since this matrix is diagonal, the inverse transformation is again diagonal with the inverse matrix elements. The energy–wavevector relationship in the starred space becomes of spherical type, as given in (8.25).

The electron velocity in (8.21) becomes

$$v_i = \frac{\hbar}{\sqrt{m_i m_\circ}} k_i^*.$$

To preserve vector equations, the transformation (8.27) must be applied to other vector quantities involved, such as driving forces and phonon wavevectors. Thus, the equation of motion for an electron under the influence of an external force $\boldsymbol{F}$ becomes

$$\hbar \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{k}^* = \boldsymbol{F}^*.$$

8.7.3 Nonparabolicity

For values of $\boldsymbol{k}$ far from the minima of the conduction bands or from the maxima of the valence bands, the energy deviates from the simple quadratic expression seen above, and nonparabolicity occurs.

For the conduction band, a simple analytical way of introducing nonparabolicity is to consider an energy–wavevector relation of the type [108]

$$\epsilon(1 + \alpha\epsilon) = \gamma(\boldsymbol{k}), \tag{8.28}$$

or

$$\epsilon(\boldsymbol{k}) = \frac{-1 + \sqrt{1 + 4\alpha\gamma}}{2\alpha},$$

where the r.h.s. $\gamma(\boldsymbol{k})$ of (8.28) must be replaced by the appropriate r.h.s. of (8.18) and (8.19). The factor α is the *nonparabolicity parameter*, which can be related to other band quantities [137,213] but is often used as a free parameter to fit experimental data.

For the valence band, nonparabolicity cannot be parameterized in a simple form like that in (8.28). In this case, nonparabolicity has two main features [231]: (a) it is more pronounced along the $\langle 110 \rangle$ and $\langle 111 \rangle$ directions for heavy and light holes, respectively; (b) if ϵ_{so} is the split-off energy of the lowest valence band, nonparabolicity is stronger near the energy $\epsilon_{so}/3$ and in the limits of $\epsilon \ll \epsilon_{so}$ and $\epsilon \gg \epsilon_{so}$ the bands are parabolic.

For a nonparabolic band of the type described in (8.28), the velocity associated with a state $\boldsymbol{k}$ is

$$\boldsymbol{v} = \frac{1}{\hbar} \nabla_{\boldsymbol{k}} \epsilon = \frac{\hbar \boldsymbol{k}}{m(1 + 2\alpha\epsilon)}.$$

Part II

Semiclassical Transport in Bulk Semiconductors

9

Electronic Interactions

Together with the band shape and semiclassical dynamics, electronic interactions in crystals form the basic ingredients for any semiclassical theory of electron transport in semiconductors. In general terms, to write down the Boltzmann transport equation (BE), we need the transition rate, or transition probability per unit time, from a state with crystal momentum $\boldsymbol{k}$ to a state with crystal momentum $\boldsymbol{k}'$, due to the various possible interaction mechanisms. Most of the times the latter are considered independent from each other, and the corresponding transition rates are calculated by means of the Fermi golden rule, described in Appendix E. This formula contains a δ function of energy conservation. However, this result is correct in the limit of long times between collisions, i.e., in the *completed-collision approximation*. When this condition is not satisfied, we need a more precise theory that accounts for *collisional broadening*, as will be discussed later in this book.

In the following sections of this chapter, we shall first present a classification of the various scattering processes, then we shall develop a general theory of electron scattering in crystals, in the completed-collision approximation, followed by an analysis of the different possible scattering mechanisms.

9.1 Classification

There are several ways to classify the electronic scattering processes relevant for transport in semiconductors. In a many-valley model, the transitions may be classified according to the valleys of the initial and final states: a transition can be

- *Intravalley*, if the initial and final states lie in the same valley;
- *Intervalley*, if the initial and final states lie in different valleys.

More detailed distinctions can then be made in both cases, as illustrated if Fig. 9.1. For example, in silicon intervalley transitions may occur

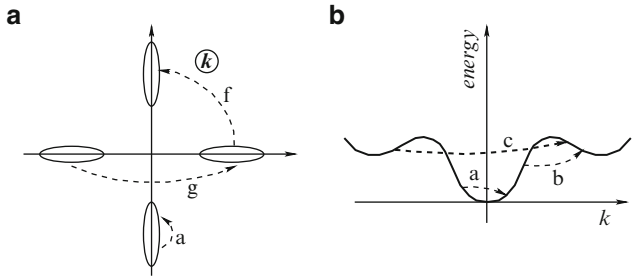


Fig. 9.1. Classification of electronic transitions in (**a**) silicon, and (**b**) gallium arsenide. a: intravalley transitions; b: intervalley transitions between central and upper valleys; c: between upper valleys; f: between perpendicular valleys; g: between parallel valleys

between parallel (g-transitions) or perpendicular valleys (f-transitions); in GaAs intravalley transitions may occur in the central, lower, valley or in the upper valleys; intervalley transitions may occur between central valley and upper valleys, or between upper valleys.

As regards the mechanisms that induce the transitions, we have again several possibilities:

– **Phonons**. Electron scattering may consist in an absorption or emission of a phonon, that, in turn, may be an acoustic or an optical phonon. The interaction mechanism may also be distinguished in electrostatic, as we have for polar optical or piezoelectric acoustic phonons in compound semiconductors, or due to the variation of the band edge produced by the deformation of the lattice. In the different cases we talk of *polar interaction*, *piezoelectric interaction*, and *deformation-potential interaction*, respectively. From the point of view of energy exchange, we shall see that optical–phonon scattering is anelastic, while acoustic phonons carry a very small amount of energy, and the scattering can be considered elastic at room temperature.

– **Impurities**. They can be *ionized impurities*, which interact with electrons through a long-range Coulomb field, or *neutral impurities*, with a short-range interaction and much weaker effect in bulk transport properties. Owing to the large mass of the impurities, this kind of scattering is always considered elastic. The effect of the impurities becomes more important at lower temperatures, when phonons become less effective. Impurities are always present in semiconductors, even when not introduced on purpose as dopants. Today, however, the technological perfection reaches extraordinary levels, so that in some cases the effect of impurities may not be present down to few Kelvin.

– **Alloy scattering**. The model of an alloy most frequently used is that of the *virtual crystal*, introduced by Nordheim in 1931 [328], with physical features intermediate between the two components of the alloy. The model requires, of course, that the alloy is perfectly homogeneous, and in general this

is not true. The variation in space of the composition of the alloy produces a perturbation that generates transitions between Bloch states of the virtual crystal.

– **Other electrons**. This type of Coulomb interaction depends on the probability of an electron to encounter another electron with given momentum. Therefore, it depends on the electron distribution function $f(\boldsymbol{p}, t)$ and, for this reason, makes the Boltzmann equation nonlinear in the unknown function f. In a collision between two electrons, the total momentum and the total energy of the electron gas do not change. As a consequence, this type of interaction is not dissipative *per se*. It changes, however, the shape of the distribution function and, as a consequence, influences the effect of the other scattering mechanisms. Its effect is relevant at high electron concentrations ($\geq 10^{17}$–10^{18} cm^{-3}).

Surface-roughness scattering is not included in the list above since it does not contribute to bulk transport properties. It will be treated in the more appropriate Chap. 19.

9.2 Fundamentals of Scattering – Crystal-Momentum Conservation

To study the transitions of an electron between different Bloch states in a crystal, due to a perturbation, one starts with the assumption that the system can be separated into the electron of interest and the rest of the crystal. The Hamiltonian may be written as

$$\mathcal{H} = \mathcal{H}_e + \mathcal{H}_c + \mathcal{H}', \tag{9.1}$$

where $\mathcal{H}_e$ represents the Hamiltonian of the electron in the perfect crystal, $\mathcal{H}_c$ the Hamiltonian of the rest of the crystal, and $\mathcal{H}'$ the perturbation Hamiltonian that describes the interaction between the two subsystems. The Hamiltonian of the unperturbed system is given by the first two terms in (9.1), and its eigenstates are separable. They can be written as the direct product

$$|\boldsymbol{k}, c\rangle = |\boldsymbol{k}\rangle|c\rangle,$$

where $|\boldsymbol{k}\rangle$ and $|c\rangle$ represent the unperturbed states of the electron and the crystal, respectively.

The transition probability per unit time from a state $|\boldsymbol{k}, c\rangle$ to a state $|\boldsymbol{k}', c'\rangle$ induced by the perturbation Hamiltonian $\mathcal{H}'$ is given, to the lowest order, by the Fermi golden rule reported in Appendix E:

$$P(\boldsymbol{k}, c; \boldsymbol{k}', c') = \frac{2\pi}{\hbar} |\langle \boldsymbol{k}', c' | \mathcal{H}' | \boldsymbol{k}, c \rangle|^2 \delta(\epsilon(\boldsymbol{k}', c') - \epsilon(\boldsymbol{k}, c)), \tag{9.2}$$

where $\epsilon(\boldsymbol{k}, c)$ is the unperturbed energy of the state $|\boldsymbol{k}, c\rangle$. $\mathcal{H}'$ acts, in general, on the coordinate $\boldsymbol{r}$ of the electron and on the variables, say $\boldsymbol{y}$, that describe

the state of the crystal, e.g., atom or impurity positions. We already mentioned in the introduction to this chapter that the application of this rule is restricted to weak interactions, when collisional broadening can be neglected, a condition usually satisfied in the linear-response regime at room temperature.

It is now convenient to expand $\mathcal{H}'$ in a Fourier series

$$\mathcal{H}'(\boldsymbol{r},\boldsymbol{y}) = \frac{1}{\sqrt{V}} \sum_{\boldsymbol{q}} \left[\mathcal{A}(\boldsymbol{q},\boldsymbol{y}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} + \mathcal{A}^{\dagger}(\boldsymbol{q},\boldsymbol{y}) \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \right], \tag{9.3}$$

where the sum of the two terms has been inserted to ensure the hermiticity of $\mathcal{H}'$. Keeping for the moment the first term in (9.3), the matrix element in (9.2) takes the form:

$$\langle \boldsymbol{k}', c' | \mathcal{H}' | \boldsymbol{k}, c \rangle = \frac{1}{\sqrt{V}} \sum_{\boldsymbol{q}} \langle c' | \mathcal{A}(\boldsymbol{q},\boldsymbol{y}) | c \rangle \int_V \psi^*_{\boldsymbol{k}'}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \psi_{\boldsymbol{k}}(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r}, \tag{9.4}$$

where $\psi_k(\boldsymbol{r})$ are the Bloch wavefunctions, eigenfunctions of $\mathcal{H}_e$. The band index has been omitted for simplicity.

Let us now focus on the integral I on the r.h.s. of (9.4). A similar integral has been considered in Sect. 7.4. Following the same procedure, let us split the integral as a sum of integrals over the crystal cells labeled by the direct-lattice vectors $\boldsymbol{R}_j$. Then putting $\boldsymbol{r} = \boldsymbol{R} + \boldsymbol{r}'$ and remembering the form (6.5) of the Bloch functions, we obtain

$$I = \sum_j \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{q})\cdot\boldsymbol{R}_j} \int_{\mathrm{cell}} u^*_{\boldsymbol{k}}(\boldsymbol{r}') u_{\boldsymbol{k}}(\boldsymbol{r}') \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{q})\cdot\boldsymbol{r}'} \, \mathrm{d}\boldsymbol{r}'. \tag{9.5}$$

The sum on the left has been calculated in Sect. 7.4 and is given by (7.20) The integral I reduces to

$$I = \delta_{\boldsymbol{G},\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{q}} N \int_{\mathrm{cell}} u^*_{\mathbf{k}'}(\boldsymbol{r}) u_{\mathbf{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\mathbf{G}\mathbf{r}} \mathrm{d}\boldsymbol{r},$$

where N is the number of unit cells in the crystal. The Kronecker δ in this equation describes the conservation of crystal momentum in electronic scattering in crystals, verified to within a vector $\boldsymbol{G}$ of the reciprocal lattice:

$$\boxed{\boldsymbol{k}' = \boldsymbol{k} \pm \boldsymbol{q} + \boldsymbol{G}} \tag{9.6}$$

where the double sign refers to the two terms in (9.3). Collecting the above results, the transition rate in (9.2) becomes

$$P(\boldsymbol{k}, c; \boldsymbol{k}', c') = \frac{2\pi}{\hbar} \frac{1}{V} \left| \sum_{\boldsymbol{q}} \langle c' | \mathcal{A}(\boldsymbol{q},\boldsymbol{y}) | c \rangle \right|^2 \mathcal{G}(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{G}) \delta(\epsilon(\boldsymbol{k}', c') - \epsilon(\boldsymbol{k}, c)), \tag{9.7}$$

where the sum contains all $\boldsymbol{q}$s that verify the crystal-momentum conservation (9.6), and $\mathcal{G}$ is the *overlap integral*:

$$\mathcal{G}(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{G}) = \left| N \int_{\text{cell}} u^*_{\boldsymbol{k}'}(\boldsymbol{r}) u_{\boldsymbol{k}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}\boldsymbol{G}\boldsymbol{r}} \, \mathrm{d}\boldsymbol{r} \right|^2 . \tag{9.8}$$

The factor N appearing here is due to the normalization of the Bloch wavefunction to unity in the volume of the crystal, so that the periodic parts u are normalized to $1/N$.

The scattering processes in which $\boldsymbol{G}$ in (9.6) is different from zero are called *umklapp*, or *U-processes*; if $\boldsymbol{G} = 0$ the process is called *normal* or *non-umklapp* or *N-process*. To proceed further, we need explicit forms for $\mathcal{H}'$, which depend on the particular scattering mechanism considered.

Overlap Integral

Expressions for the overlap integral $\mathcal{G}$ in (9.8) as a function of $\boldsymbol{k}$ and $\boldsymbol{k}'$ have been given in the literature for various cases.

Intravalley transitions are, in general, N-processes, because the distance between $\boldsymbol{k}$ and $\boldsymbol{k}'$ is small compared to the dimensions of the Brillouin zone. For N-processes, $\mathcal{G}$ is equal to unity for exact plane waves or for wavefunctions formed with pure s states. When lower symmetries are involved in the Bloch wavefunctions, an overlap integral less than unity is obtained, which depends mainly upon the angle θ between initial and final states $\boldsymbol{k}$ and $\boldsymbol{k}'$, measured from the center of the Brillouin zone. In the case of the minimum of the conduction band in Γ, $\mathcal{G}$ is related to the nonparabolicity parameter α (see Sect. 8.7.3), since both $\mathcal{G}$ and α come from the presence of p terms in the electron wavefunctions. Fawcett et al. [137] gave for $\mathcal{G}$ the expression

$$\mathcal{G}(\boldsymbol{k}, \boldsymbol{k}') = \frac{[(1+\alpha\epsilon)^{1/2}(1+\alpha\epsilon')^{1/2} + \alpha(\epsilon\epsilon')^{1/2}\cos\theta]^2}{(1+2\alpha\epsilon)(1+2\alpha\epsilon')},$$

where $\epsilon = \epsilon(\boldsymbol{k})$ and $\epsilon' = \epsilon(\boldsymbol{k}')$.

In the many-valley model, defined in Sect. 8.7, intervalley transitions can be U-processes, because of the large values associated with $\boldsymbol{k}$ and $\boldsymbol{k}'$. For both intravalley (in silicon or germanium) and intervalley transitions, the angle θ between initial and final states depends mostly on the valleys involved in the transition, and $\mathcal{G}$ is thus almost constant within each type and may be included in the coupling constants.

For transitions of holes within heavy or light bands, Wiley [474] found the simple expression

$$\mathcal{G}(\boldsymbol{k}, \boldsymbol{k}') = \frac{1}{4}\left(1 + 3\cos^2\theta\right), \tag{9.9}$$

while for hole interband transitions he found

$$\mathcal{G}(\boldsymbol{k}, \boldsymbol{k}') = \frac{3}{4}\sin^2\theta. \tag{9.10}$$

Since $\mathcal{G}$ is either almost constant or averaged over many scattering angles, it is often included in the coupling constant of the interaction mechanism at hand.

9.3 Electron–Phonon Scattering Rates – Deformation Potential

The general theory reviewed above will now be applied to the case of electronic interactions with phonons. In covalent crystals, the distortion of the lattice due to thermal vibrations is not accompanied by polarization fields and the interaction of the vibrations with the electrons is simply due to the deformation of the lattice. If the displacement field of the atoms $\boldsymbol{y}(\boldsymbol{r})$ were uniform, the lattice would not be deformed, but shifted. Thus, in the *deformation-potential* theory it is assumed that the interaction Hamiltonian is proportional to the variation in space of the displacement field:

$$\mathcal{H}' = \sum_{ij} E_{ij} \frac{\partial y_i}{\partial r_j}, \tag{9.11}$$

where E_{ij} is the deformation-potential tensor constant.

Values of the deformation-potential constants in several materials can be found, for example, in [46, 213, 364].

According to the theory of lattice vibrations developed in Chap. 5, the displacement $\boldsymbol{y}(\boldsymbol{r})$ of the ion whose equilibrium position is $\boldsymbol{r}$ (in the limit of a continuous displacement field) is given by (cf. 5.25)

$$\boldsymbol{y}(\boldsymbol{r}) = \sum_{\boldsymbol{q},\ell} \mathbf{e}_{\boldsymbol{q}\ell} \left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})} \right)^{\frac{1}{2}} \left\{ \mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell} \right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}, \tag{9.12}$$

where each mode is characterized by wavevector $\boldsymbol{q}$, polarization $\mathbf{e}_{\boldsymbol{q}\ell}$, and angular frequency $\omega_\ell(\boldsymbol{q})$; ρ is the density of the crystal; $\mathsf{a}_{\boldsymbol{q}\ell}$ and $\mathsf{a}^\dagger_{\boldsymbol{q}\ell}$ are annihilation and creation phonon operators. By using the above equation, we obtain for the interaction Hamiltonian (9.11)

$$\mathcal{H}' = \sum_{ij} E_{ij} \sum_{\boldsymbol{q},\ell} [\mathbf{e}_{\boldsymbol{q}\ell}]_i \mathrm{i} q_j \left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})} \right)^{\frac{1}{2}} \left\{ \mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell} \right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}. \tag{9.13}$$

This equation is an explicit form of the Fourier transform indicated in (9.3) with

$$\mathcal{A}(\boldsymbol{q}, \boldsymbol{y}) = \sum_{ij} E_{ij} \sum_{\ell} [\mathbf{e}_{\boldsymbol{q}\ell}]_i \mathrm{i} q_j \left(\frac{\hbar}{2\rho \omega_\ell(\boldsymbol{q})} \right)^{\frac{1}{2}} \mathsf{a}_{\boldsymbol{q}\ell}.$$

Remember that $\boldsymbol{y}$ appears as an argument of $\mathcal{A}$ to indicate that this operator acts on the crystal subsystem. Taking into account the effect of the creation and annihilation operators (see Appendix C), we note that they appear in

(9.13) in such a way that only two terms in the sum over $\boldsymbol{q}$ in (9.7) are different from zero. They correspond to vibrational states of the crystal $|c'\rangle$ with the occupation number of one mode $\boldsymbol{q}\ell$ changed by one unit with respect to $|c\rangle$. They correspond to the emission and the absorption of a phonon with wavevector

$$\boldsymbol{q} = \boldsymbol{k}' - \boldsymbol{k} + \boldsymbol{G} \quad \text{or} \quad \boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}' + \boldsymbol{G}$$

for absorption, or emission, respectively. More precisely,

$$|\langle c'|a_{\boldsymbol{q}\ell}|c\rangle|^2 = N_{\boldsymbol{q}\ell} \quad \text{and} \quad |\langle c'|a^{\dagger}_{-\boldsymbol{q}\ell}|c\rangle|^2 = N_{\boldsymbol{q}\ell} + 1,$$

respectively, where $N_{\boldsymbol{q}\ell}$ is the number of phonons $\boldsymbol{q}\ell$ in the state $|c\rangle$.

By collecting the foregoing results, we obtain the following expression for the transition probability per unit time of an electron from state $\boldsymbol{k}$ to state $\boldsymbol{k}'$:

$$P^{(d)}(\boldsymbol{k},\boldsymbol{k}') = \frac{\pi}{\rho V \omega_\ell(\boldsymbol{q})} \begin{bmatrix} N_{\boldsymbol{q}\ell} \\ N_{\boldsymbol{q}\ell}+1 \end{bmatrix} \mathcal{G} \left| \sum_{ij} E_{ij} q_j [\mathbf{e}_{\boldsymbol{q}\ell}]_i \right|^2 \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k}) \mp \hbar\omega_\ell(\boldsymbol{q})], \tag{9.14}$$

where the upper and lower symbols refer to absorption and emission, respectively. The same will always be true in the following of this chapter, when a double sign is present. The superscript d stands for "deformation potential". The term $(+1)$ added to $N_{\boldsymbol{q}\ell}$ in the emission case is due to the so-called *spontaneous emission*, a quantum phenomenon not present in classical physics, which describes the possibility of electron interaction with the phonon field (as well as with the electromagnetic field) even when phonons are not present. In a way, it may be attributed to the *zero-point vibrations* described in the harmonic-oscillator quantum theory in Appendix C. The term proportional to $N_{\boldsymbol{q}\ell}$ gives the transition probability induced by the phonons already present in the crystal and is called *stimulated emission*.

In bulk transport, electrons interact many times with lattice vibrations, emitting and absorbing phonons. In the evaluation of the macroscopic current, a statistical average is performed over many situations where the phonon gas appears in different states with different probabilities. Since the scattering probabilities appear linearly in the transport equation (see Sect. 10.3), the net result is that in the above equation the actual phonon number may be replaced by its average value, as long as the boson gas itself is maintained in equilibrium. Deviations from such situation lead to the problem of *hot phonons*, which will be briefly discussed in Sect. 13.7.

9.3.1 Electron Intravalley Scattering by Acoustic Phonons

Quasi Elasticity

For the qualitative considerations of this section, we may consider a spherical and parabolic band with effective mass m. Intravalley transitions induced by acoustic phonons involve a phonon wavevector that is a small fraction of the

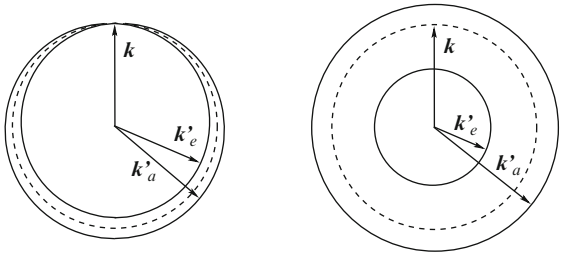


Fig. 9.2. Possible final states $\boldsymbol{k}'_a$ and $\boldsymbol{k}'_e$ after an absorption or emission of an acoustic phonon (*left*) or an optical phonon (*right*), given an initial state $\boldsymbol{k}$. Dashed circles indicate equi-energy states. The complete set of possible final states is given by the rotation of the figures around the initial state $\boldsymbol{k}$

Brillouin-zone size, so that the linear part of the acoustic dispersion branch (5.17) can be used. Then the phonon energy can be taken as $\hbar q v_s$, where v_s is the sound velocity. Energy and crystal-momentum conservations for an N-process require

$$\begin{cases} \hbar^2 k'^2/2m = \hbar^2 k^2/2m \pm \hbar q v_s \\ \boldsymbol{k}' = \boldsymbol{k} \pm \boldsymbol{q} \end{cases} . \tag{9.15}$$

If the second equation above is inserted into the first one, after simple algebraical steps, we obtain

$$q = \mp 2\,(k\cos\theta - m v_s/\hbar), \tag{9.16}$$

where θ is the angle between $\boldsymbol{k}$ and $\boldsymbol{q}$. The maximum value of q is obtained for absorption with backward scattering ($\theta = \pi$), as shown in the left part of Fig. 9.2, and is given by

$$q_{\max} = 2k + 2m v_s/\hbar. \tag{9.17}$$

If the scattering mechanism were elastic, a backward collision would involve $q_{\max} = 2k$; thus, the second term in the above equation is the correction due to the energy of the phonon involved in the transition. The relative contribution of this term to the electron wavevector can be seen in (9.17) to be given by v_s/v, where v is the velocity of the electron, which, in general, is much larger than v_s ($v \sim 10^5$ m/s, $v_s \sim 10^3$ m/s). The maximum q involved in these transitions is therefore very close to $2k$; the corresponding maximum energy transfer is

$$\hbar q_{\max} v_s \approx 2\hbar k v_s = 2mvv_s,$$

which, again, is in general much smaller than the electron kinetic energy $mv^2/2$. For this reason, acoustic scattering is very often considered an elastic process, mainly when dealing with transport at room temperature.[1] At very

[1] This approximation may cause problems in Monte Carlo simulations, as discussed in Chap. 14.

low temperatures, however, the energy dissipated via acoustic phonons may be relevant [161].

In the following sections, acoustic-phonon scattering will be treated first in the elastic approximation, then more detailed calculations will be developed using the correct energy exchange.

A - Elastic, Energy Equipartition, Spherical and Parabolic Bands

When we deal with acoustic scattering in the elastic approximation, the phonon population N_q is usually approximated by the equipartition expression

$$N_q = \frac{1}{\mathrm{e}^{\hbar q v_s / K_\mathrm{B} T} - 1} \approx \frac{K_\mathrm{B} T}{\hbar q v_s} - \frac{1}{2}. \tag{9.18}$$

This approximation is closely related to the elastic approximation discussed above. In fact, (9.18) is valid when $\hbar q v_s \ll K_\mathrm{B} T$, i.e., when the thermal energy is much larger than the energy of the phonon involved in the transition.[2]

Furthermore, for a nondegenerate parabolic band at the center of the Brillouin zone of a cubic semiconductor, the deformation-potential constant E_{ij} in (9.14) is a second-rank tensor with cubic symmetry. Such a tensor has a diagonal form with equal diagonal elements and therefore can be treated as a scalar quantity E_1. The squared factor that appears in the same equation reduces to $E_1^2 q^2$ for longitudinal phonons, while it vanishes for transverse modes. The expression for the scattering probability per unit time from a state $\boldsymbol{k}$ to a state $\boldsymbol{k}'$ given by (9.14) in the elastic and energy-equipartition approximation, becomes

$$P_{ae}^{(d)}(\boldsymbol{k}, \boldsymbol{k}') = \frac{\pi q E_1^2}{\rho V v_l} \left[\frac{K_\mathrm{B} T}{\hbar q v_l} \mp \frac{1}{2} \right] \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})], \tag{9.19}$$

where v_l is the sound velocity for longitudinal modes. The suffix ae stands for "acoustic elastic". The overlap integral $\mathcal{G}$ has been taken as equal to one, according to the comment at the end of Sect. 9.2. Since in the elastic approximation no distinction is made between final states attained by means of absorption or emission processes, we can consider the sum of the transition

[2] This result can be obtained with the Laurent expansion of the Bose distribution, but in more elementary terms it can be obtained with the power expansion of its denominator as

$$\frac{1}{\mathrm{e}^x - 1} \approx \frac{1}{1 + x + x^2/2 - 1} \approx \frac{1}{x(1 + x/2)} \approx \frac{1}{x}(1 - x/2) = \frac{1}{x} - \frac{1}{2}.$$

probabilities per unit time to be given by

$$P_{ae}^{(d)}(\boldsymbol{k},\boldsymbol{k}') = \frac{2\pi K_{\mathrm{B}}T}{\hbar\rho V v_l^2} E_1^2\, \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})]. \tag{9.20}$$

From this equation, it results that, in the present approximations, acoustic scattering becomes isotropic: the probability distribution of the direction of the state after scattering $\boldsymbol{k}'$ is independent of the direction of the initial state $\boldsymbol{k}$.

To obtain the total scattering probability per unit time, we have to sum the above expression over the possible final states $\boldsymbol{k}'$. First we transform the sum into an integral, using the proper density of states, as indicated in (6.8) of Chap. 6. On this respect no spin multiplicity is introduced since phonon interaction does not produce spin flip, and therefore only the state with the same spin orientation as that of the state before scattering is available:

$$P_{ae}^{(d)}(\boldsymbol{k}) = \frac{V}{(2\pi)^3} \int \frac{2\pi K_{\mathrm{B}}T}{\hbar\rho V v_l^2} E_1^2\, \delta\left[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})\right] \mathrm{d}\boldsymbol{k}'.$$

Next, we write the integral in polar coordinates of $\boldsymbol{k}'$ with $\boldsymbol{k}$ as polar axis. Since the integrand is isotropic, the angular variables are immediately integrated yielding the total solid angle 4π:

$$P_{ae}^{(d)}(\boldsymbol{k}) = \frac{1}{(2\pi)^2} \frac{K_{\mathrm{B}}T}{\hbar\rho v_l^2} E_1^2 4\pi \int k'^2\, \delta\left[\frac{\hbar^2 k'^2}{2m} - \frac{\hbar^2 k^2}{2m}\right] \mathrm{d}k'.$$

Then, we use the rule (A.22) in Appendix A for using a $\delta(f(x))$ within an integral:

$$P_{ae}^{(d)}(\boldsymbol{k}) = \frac{1}{\pi} \frac{K_{\mathrm{B}}T}{\hbar\rho v_l^2} E_1^2 \frac{k^2}{\hbar^2 k/m}.$$

This rate is a function of k only, therefore of energy, and can be written as

$$P_{ae}^{(d)}(\epsilon) = \frac{\sqrt{2} m^{3/2} K_{\mathrm{B}}T E_1^2}{\pi\hbar^4 \rho v_l^2} \sqrt{\epsilon}. \tag{9.21}$$

Equation (9.21) accounts for both absorption and emission processes. Its energy dependence is a simple proportionality to $\sqrt{\epsilon}$ and is shown in Fig. 9.3.

B - Elastic, Energy Equipartition, Ellipsoidal, Nonparabolic Bands

If we move from the simple spherical and parabolic case to more realistic band structures, several complications arise. For an ellipsoidal, nonparabolic model the energy-wavevector relationship is, with the simple nonparabolicity model of Sect. 8.7.3,

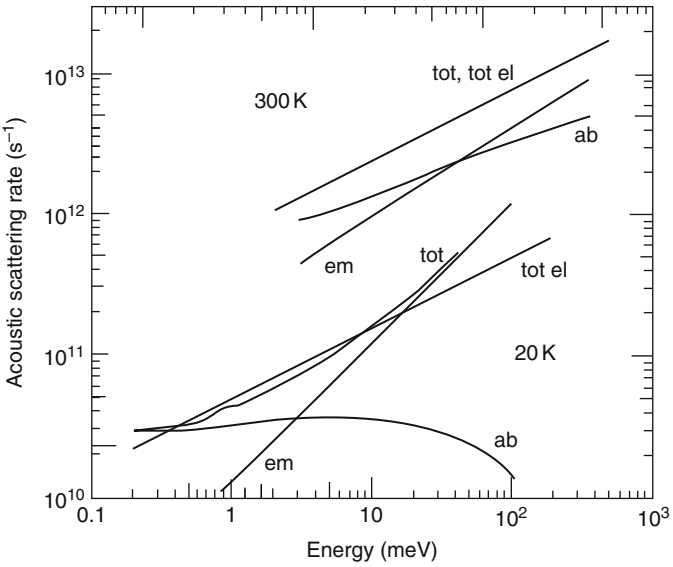


Fig. 9.3. Integrated scattering rate for electron scattering by acoustic modes as a function of energy at the indicated temperatures. The parameters used refer to silicon with a parabolic band. ab = absorption; em = emission; tot el = total in elastic approximation; tot = total when acoustic energy dissipation and exact phonon population are accounted for [213]

$$\gamma(\boldsymbol{k}) = \epsilon(1+\alpha\epsilon) = \frac{\hbar^2}{2}\left(\frac{1}{m}\right)_{ij} k_i k_j \tag{9.22}$$

and is usually treated by means of the Herring–Vogt transformation [185], as discussed in Chap. 8. In a many-valley model with valleys centered along the $\langle 100\rangle$ and/or $\langle 111\rangle$ directions, for symmetry reasons we have two independent components Ξ_d and Ξ_u of the deformation-potential tensor [185], and interaction of electrons with acoustic phonons is allowed with transverse as well as longitudinal modes. In this case, (9.20) becomes

$$P_{ae}^{dl}(\boldsymbol{k},\boldsymbol{k}') = \frac{2\pi K_{\mathrm{B}}T}{\hbar\rho V v_l^2}\left(\Xi_d + \Xi_u\cos^2\theta\right)^2\ \delta\left[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})\right] \tag{9.23}$$

for longitudinal modes, and

$$P_{ae}^{dt}(\boldsymbol{k},\boldsymbol{k}') = \frac{2\pi K_{\mathrm{B}}T}{\hbar\rho V v_t^2}\left(\Xi_u\sin\theta\cos\theta\right)^2\ \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})] \tag{9.24}$$

for transverse modes. Here v_t is the transverse sound velocity, and θ is the angle between $\boldsymbol{q}$ and the longitudinal axis of the valley.

The effect of anisotropy is not large [107], so that it is, in general, neglected by replacing v_l and v_t with an average value $v_s = (2v_t + v_l)/3$, and the

expression in brackets with mean values E_1^2 over the angle θ, often determined from the interpretation of transport measurements. Thus, the same expression for the scattering rate as (9.20) is obtained, with v_s in place of v_l and the Herring–Vogt transformed $\boldsymbol{k}^*$ and $\boldsymbol{k}^{*\prime}$ in place of $\boldsymbol{k}$ and $\boldsymbol{k}'$. The resulting expression is again independent of the direction of $\boldsymbol{k}^*$ and $\boldsymbol{k}^{*\prime}$, so that the scattering is still isotropic in the starred space. Integration over all possible final states $\boldsymbol{k}^{*\prime}$ is performed following the same method as in the previous case. Nonparabolicity is taken into account in the form of the energy in the argument of the δ function. The integrated scattering rate, which accounts for both absorption and emission processes, results

$$P_{ae}^{(d)}(\epsilon) = \frac{\sqrt{2} m_d^{3/2} K_\mathrm{B} T \, E_1^2}{\pi \hbar^4 \rho v_s^2} \sqrt{\epsilon}(1 + 2\alpha\epsilon)(1 + \alpha\epsilon)^{1/2}, \tag{9.25}$$

where m_d is the density-of-states effective mass of (8.26).

From the above equation, it is also possible to determine the expression for "intermediate" models, i.e., spherical nonparabolic [$m_d = m$ and $v = v_l$] or ellipsoidal parabolic [$\alpha = 0$].

C - Inelastic Acoustic Scattering, Spherical, Parabolic Bands

To treat correctly energy dissipation via acoustic phonons, we must take into account the energy of the phonon involved in the energy balance of the collision:

$$\epsilon(\boldsymbol{k}') = \epsilon(\boldsymbol{k}) \pm \hbar q v_s.$$

In a simple spherical and parabolic band, only longitudinal modes contribute to the scattering, and energy and momentum conservation imply [see (9.16)]

$$\cos\theta = \pm \frac{q}{2k} + \frac{m v_l}{\hbar k}.$$

The condition $-1 \leq \cos\theta \leq 1$ determines the range of phonon wavevectors involved in a collision with an electron in state $\boldsymbol{k}$. The results are reported in Table 9.1 where, instead of q, use has been made of the more convenient dimensionless variable $x = (\hbar q v_l / K_\mathrm{B} T)$, and where $\epsilon_s = m v_l^2/2$ is the kinetic energy of an electron with velocity equal to the longitudinal sound velocity.

To be consistent with the correct treatment of energy exchange in acoustic–phonon scattering, the exact expression of the phonon number N_q must be included in the calculations. The transition probability (9.14) for the case of interest becomes

$$P_a^{(d)}(\boldsymbol{k}, \boldsymbol{k}') = \frac{\pi q E_1^2}{\rho V v_l} \begin{bmatrix} N_q \\ N_q + 1 \end{bmatrix} \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k}) \mp \hbar q v_l]. \tag{9.26}$$

To perform the integration over all possible final states, it is convenient, this time, to consider integration over $\boldsymbol{q}$ and to use polar coordinates with the polar

Table 9.1. Limits of values of q for inelastic acoustic scattering determined by energy and momentum conservation. $x = (\hbar q v_l)/(K_B T)$, $\epsilon_s = m v_s^2/2$

Absorption	Emission	
$x_{1,a} = \frac{4\epsilon_s^{1/2}}{K_B T}(\epsilon_s^{1/2} - \epsilon^{1/2})$	Absent	$\epsilon < \epsilon_s$
$x_{2,a} = \frac{4\epsilon_s^{1/2}}{K_B T}(\epsilon_s^{1/2} + \epsilon^{1/2})$		
$x_{1,a} = 0$	$x_{1,e} = 0$	$\epsilon < \epsilon_s$
$x_{2,a} = \frac{4\epsilon_s^{1/2}}{K_B T}(\epsilon^{1/2} + \epsilon_s^{1/2})$	$x_{2,e} = \frac{4\epsilon_s^{1/2}}{K_B T}(\epsilon^{1/2} - \epsilon_s^{1/2})$	

axis along $\boldsymbol{k}$. The δ function of energy conservation can be used to integrate over the polar angle θ between $\boldsymbol{q}$ and $\boldsymbol{k}$. This leads to

$$P_a^{(d)}(k,q)\mathrm{d}q = \frac{mE_1^2}{4\pi\rho v_l \hbar^2 k}\begin{bmatrix} N_q \\ N_q + 1 \end{bmatrix} q^2 \mathrm{d}q. \tag{9.27}$$

Final integration over q leads to the integrated scattering probability per unit time as a function of energy:

$$P_a^{(d)}(\epsilon) = \frac{m^{1/2}(K_B T)^3 E_1^2}{2^{5/2}\pi\rho v_l^4 \hbar^4}\epsilon^{-1/2}\begin{bmatrix} F_1(x_{2,a}) - F_1(x_{1,a}) \\ G_1(x_{2,e}) - G_1(x_{1,e}) \end{bmatrix}, \tag{9.28}$$

where

$$F_1(x) = \int_0^x N_q(x') x'^2 \mathrm{d}x',$$

$$G_1(x) = \int_0^x [N_q(x') + 1] x'^2 \mathrm{d}x',$$

where $x_{1,a}$, $x_{2,a}$, $x_{1,e}$, and $x_{2,e}$ are those given in Table 9.1. For practical calculations, it is useful to use a Laurent expansion of the phonon occupation numbers [88].

In Fig. 9.3, the integrated acoustic scattering rates, given by (9.28), are shown and compared with the results of the elastic approximation.

D - Inelastic Acoustic Scattering, Ellipsoidal, Nonparabolic Bands

When all details of ellipsoidal valleys and acoustic energy relaxation are considered, the transition rate $P_a^{(d)}(\boldsymbol{k}, \boldsymbol{k}')$ has the form given in (9.23) and (9.24), where, in addition, the argument of the δ function contains the energy of the phonon, as in (9.26), and nonparabolicity is accounted for by assuming, for the energy-wavevector relationship, the expression in (9.22). The isotropic approximation is still used for the deformation-potential coupling and, in applying Herring–Vogt transformation for $\boldsymbol{k}$ and $\boldsymbol{q}$ vectors, the magnitude

of $\boldsymbol{q}$ is approximated by

$$q = q^* \left[\frac{m_l}{m_\circ} \cos^2 \theta^* + \frac{m_t}{m_\circ} \sin^2 \theta^* \right]^{1/2} \approx q^* \left[\frac{m_d}{m_\circ} \right]^{1/2}, \tag{9.29}$$

where θ^* is the angle between $\boldsymbol{k}^*$ and the principal axis of the valley. This last approximation is rather poor when the valley is strongly anisotropic, as for the case of Ge. The calculation of the total scattering rate may proceed in analogy with the previous cases [213]. However, when it is necessary to consider details of the band, a full-band model is more advisable [148, 187] (see Sect. 14.2.8).

9.3.2 Electron Intravalley Scattering by Optical Phonons

The theory of electron scattering by acoustic phonons developed in the previous section made used of the continuous approximation, valid when the phonon wavelength is much larger than the lattice constant. This approximation may be good for acoustic phonons that induce intravalley transitions where small momentum transfers are involved. In the case of optical phonons, however, even when the wavevector is small, adjacent atoms vibrate in opposite directions, and the deformation of the crystal is relevant not only at distances comparable with the wavelength. A naïve extension of the deformation theory of acoustic modes to the optical case suggests to assume a fixed wavelength equal to the lattice constant, and this approach is justified by more rigorous theories [181, 267, 403, 463]. The scattering rate can then be written, starting from (9.14), by replacing $E_1^2 q^2$ with a squared optical coupling constant $(D_t K)^2$, which can also include the overlap integral.

The energy associated with optical phonons in intravalley transitions can be assumed constant, given by $\hbar\omega_{op}$, since the dispersion curve of such phonons is quite flat for the q-values involved in electronic intravalley transitions. For the same reason, N_q becomes q-independent: $N_q = N_{op}$. The resulting scattering rate is

$$P_{op}^{(d)}(\boldsymbol{k}, \boldsymbol{k}') = \frac{\pi (D_t K)^2}{\rho V \omega_{op}} \begin{bmatrix} N_{op} \\ N_{op} + 1 \end{bmatrix} \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k}) \mp \hbar\omega_{op}]. \tag{9.30}$$

No angular dependence upon the direction of $\boldsymbol{k}'$ is present, and the scattering is therefore isotropic. The possible final states after a transition induced by optical phonons is shown in the right part of Fig. 9.2.

A - Spherical, Parabolic Bands

For spherical, parabolic bands, integration of (9.30) over k' yields, with standard calculations, the scattering rate

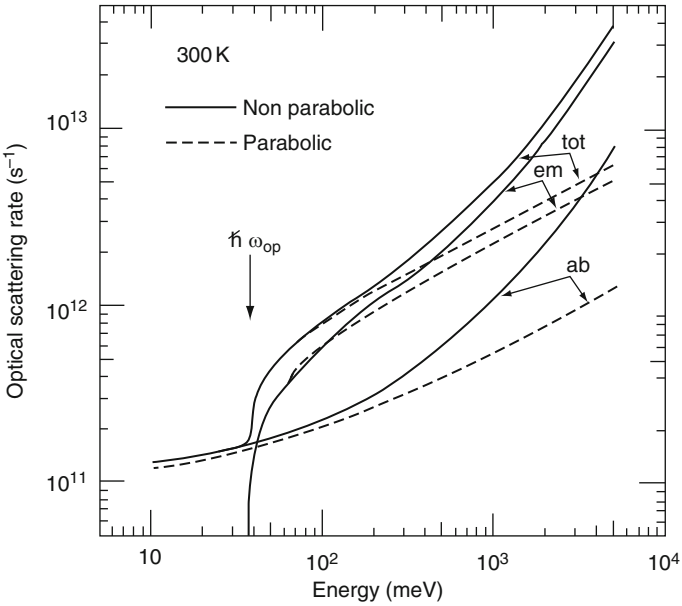


Fig. 9.4. Integrated optical scattering rates as a function of energy at room temperature. The model used refers to electrons in germanium. ab = absorption; em = emission; tot = total [213]

$$P_{op}^{(d)}(\epsilon) = \frac{m^{3/2}(D_t K)^2}{\sqrt{2}\pi\hbar^3\rho\omega_{op}} \begin{bmatrix} N_{op} \\ N_{op}+1 \end{bmatrix} (\epsilon \pm \hbar\omega_{op})^{1/2}. \tag{9.31}$$

The probability of emission is obviously zero when $\epsilon < \hbar\omega_{op}$, since the electron does not have enough energy to emit the phonon. Figure 9.4 shows the scattering rate by optical phonons as a function of energy, as given in (9.31). A simple proportionality to $\epsilon^{1/2}$, as for acoustic modes, is attained in the parabolic case when the carrier energy is much larger than the phonon energy, so that the scattering becomes approximately elastic.

B - Ellipsoidal Nonparabolic Bands

In dealing with ellipsoidal bands, the Herring–Vogt transformation is again applied, as for the scattering by acoustic phonons. The transition rate can be derived from the general expression in (9.14), as indicated above for simple bands. Since the transition probability is independent of $\boldsymbol{q}$, the approximation in (9.29) is not necessary. The resulting transition rate is the same as in (9.30), and the integrated scattering rate is the same as in (9.31), with m replaced by m_d.

When nonparabolicity is included, integration over $\boldsymbol{k}^{*\prime}$ yields an extra factor $[1 + 2\alpha(\epsilon \pm \hbar\omega_{op})]$, so that the integrated scattering rate for optical phonons is given by

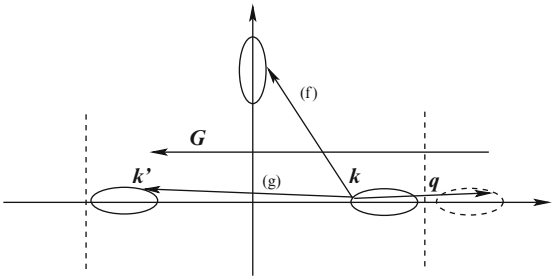


Fig. 9.5. Scattering of g type between parallel valleys in silicon is an umklapp process ($\boldsymbol{k}' = \boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G}$) and involves a phonon wavevector smaller than in scattering of f type between perpendicular valleys

$$P_{op}^{(d)}(\epsilon) = \frac{m_d^{3/2}(D_t K)^2}{\sqrt{2}\pi\hbar^3\rho\omega_{op}} \begin{bmatrix} N_{op} \\ N_{op}+1 \end{bmatrix} \gamma^{1/2}(\epsilon \pm \hbar\omega_{op})[1+2\alpha(\epsilon \pm \hbar\omega_{op})], \quad (9.32)$$

where γ is defined in (9.22).

In Fig. 9.4 the integrated optical intravalley scattering rates for the parabolic and nonparabolic cases ((9.31) and (9.32), respectively) are shown as a function of energy.

9.3.3 Electron Intervalley Scattering

Electron transitions between states in two different equivalent valleys can be induced by electron scattering with both acoustic and optical modes.

The phonon wavevector $\boldsymbol{q}$ involved in a transition remains very close to the distance between the minima of the initial and final valleys, even for high-energy electrons. Consequently, $\boldsymbol{q} = \Delta\boldsymbol{k}$ is almost constant, and, for a given branch of phonons, the energy $\hbar\omega_i$ involved in the scattering process is also about constant, as in the case of optical intravalley scattering, and depends upon the valleys involved in the transition, as shown in Fig. 9.5. Thus, intervalley scattering is usually treated, formally, in the same way as intravalley scattering by optical phonons [107, 181].

The squared coupling constant $(D_t K)_i^2$ depends upon the kinds of valleys (initial and final) and the branch of phonons involved in the transition, and may include, as in the other cases, an overlap integral. Furthermore, a factor Z_f must be included, equal to the number of possible equivalent valleys for the final state of the transition, when the method used for the solution of the transport problem requires to account for this multiplicity.

When the electron energy is high enough, electrons can be scattered into valleys higher, in energy, than the lowest-energy valleys. In this case, the appropriate $\Delta\boldsymbol{k}$ in the Brillouin zone must of course be considered together with the variation of the electron kinetic energy due to the energy difference between the minima of the initial and final valleys (see Fig. 9.1). The integrated scattering rate for intervalley transitions due to phonons, in spherical,

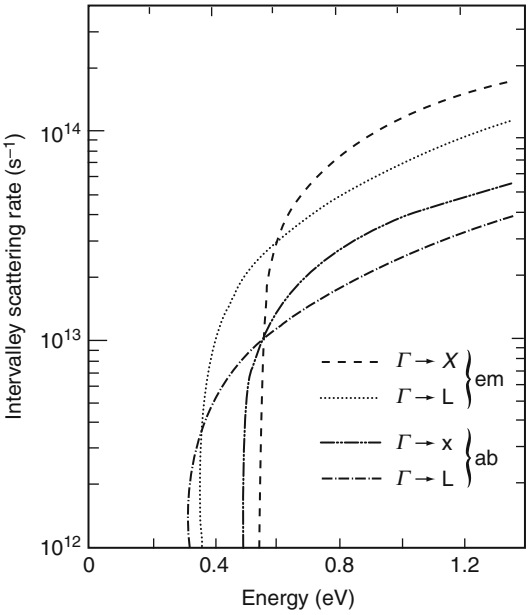


Fig. 9.6. Intervalley scattering rates as a function of energy in GaAs at room temperature [209]

parabolic valleys, is therefore given by

$$P_{iv}^{(d)}(\epsilon) = \frac{m^{3/2}(D_t K)_i^2 Z_f}{\sqrt{2}\pi\hbar^3\rho\omega_i} \begin{bmatrix} N_i \\ N_i + 1 \end{bmatrix} (\epsilon \pm \hbar\omega_i - \Delta\epsilon_{fi})^{1/2}, \tag{9.33}$$

where Z_f is the number of possible equivalent final valleys for the type of intervalley scattering under consideration, N_i is the number of phonons involved in the transition, and $\Delta\epsilon_{fi}$ is the difference between the energies of the bottoms of the final and initial valleys. Figure 9.6 shows the integrated scattering rates for intervalley transitions to nonequivalent valleys.

For equivalent valleys, when $\Delta\epsilon_{fi} = 0$, the behavior of $P_{iv}^{(d)}$ is the same as that of $P_{op}^{(d)}$, shown in Fig. 9.4.

When ellipsoidal and/or nonparabolic valleys are considered, the same corrections must be introduced relative to m_d and to the extra factor $[1 + 2\alpha(\epsilon \pm \hbar\omega_i - \Delta\epsilon_{fi})]$ as in the case of optical intravalley scattering.

9.3.4 Hole Intraband Scattering by Acoustic Phonons

Owing to the complexity of the valence-band structure, with its peculiarities of degeneracy and warping, the description of acoustic–phonon interaction requires three deformation potential parameters [41, 444, 445]. Furthermore,

the characteristic p-like symmetry of hole wavefunctions introduces an overlap integral as given in (9.9) and (9.10).

To overcome analytical difficulties, a single coupling constant E_1 and a single warped parabolic band can be used, as given in (8.20) and (8.22). Then, the treatment of hole intraband scattering rate by acoustic phonons closely follows that of electrons, and the result is [213, 365]

$$P_a^{(d)}(\boldsymbol{k}, \boldsymbol{k}') = \frac{\pi q E_1^2}{\rho V v_s} \begin{bmatrix} N_q \\ N_q + 1 \end{bmatrix} \frac{1}{4} \left(1 + 3\cos^2\theta\right) \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k}) \mp \hbar q v_s], \quad (9.34)$$

where θ is the angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$. In the integration to obtain the total scattering rate, the complex structure of the band plays its role in the argument of the δ function of energy conservation that carries the density of states. Depending on the type of approximation, different degrees of complexity are achieved [213, 365].

Often the transport of holes is much simplified by assuming one spherical and parabolic band (the heavy-hole band that has a higher density of states) or two such bands (heavy holes and light holes) with different effective masses. In the latter case, interband as well as intraband scattering must be considered.

9.3.5 Hole Intraband Scattering by Optical Phonons

Optical scattering of holes has been found to be isotropic even when the symmetry of the hole wavefunctions is considered [41, 266]. The transition rate from $\boldsymbol{k}$ to $\boldsymbol{k}'$ is therefore described in terms of one deformation-potential parameter and is similar to that of electrons. Once again the warped shape of the band enters the calculation of the integrated scattering rate in the δ of energy conservation [213, 365].

9.3.6 Hole Interband Scattering

When a two-band model, with heavy and light holes, is considered, interband scattering must be accounted for. The scattering probabilities have been studied for different mechanisms [61, 112, 401]. However, owing to the complexity of the calculations and the limited importance in practical cases, this subject will not be further examined in this textbook.

9.4 Electron–Phonon Scattering Rates – Electrostatic Interaction

In compound semiconductors, where atoms have some ionic charge, optical phonons are associated with vibrating dipoles. Furthermore, when a crystal lacks inversion symmetry, the strain associated with acoustic phonons may

generate a polarization field, accompanied by an electric field of piezoelectric nature. The result is a long-range electrostatic interaction of electrons with phonons, called *polar interaction* for optical phonons and *piezoelectric interaction* for acoustic phonons, besides the deformation potential interaction studied in the previous sections.

Following Ridley [372], we start with the expression for the interaction energy of an electron with a potential field $\phi(\boldsymbol{r})$:

$$H_{ep} = \int \rho(\boldsymbol{r}')\phi(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}',$$

where $\rho(\boldsymbol{r})$ is the electron charge density. Then, using the third Maxwell equation in (1.14), we transform this integral as

$$H_{ep} = \int \varepsilon_\circ \nabla \cdot \boldsymbol{E}_e(\boldsymbol{r}')\phi(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}',$$

where $\boldsymbol{E}_e(\boldsymbol{r})$ is the electric field associated with the electron charge. Integration by parts yields

$$H_{ep} = -\varepsilon_\circ \int \boldsymbol{E}_e(\boldsymbol{r}') \cdot \nabla\phi(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}' = \varepsilon_\circ \int \boldsymbol{E}_e(\boldsymbol{r}') \cdot \boldsymbol{E}_p(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}', \tag{9.35}$$

where $\boldsymbol{E}_p(\boldsymbol{r})$ is the electric field produced by the phonon. Now, an electron in $\boldsymbol{r}$ produces a screened Coulomb field in $\boldsymbol{r}'$ given by

$$\boldsymbol{E}_e(\boldsymbol{r}') = -\frac{(-\mathrm{e})}{4\pi\varepsilon_\circ}\nabla_{\mathrm{r}'}\left[\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{e}^{-q_\circ|\boldsymbol{r}-\boldsymbol{r}'|}\right], \tag{9.36}$$

where $q_\circ$ is the inverse screening length. In the Debye formulation, for nondegenerate statistics

$$q_\circ = \left[\frac{\mathrm{e}^2 n}{\varepsilon K_\mathrm{B} T}\right]^{1/2}, \tag{9.37}$$

where n is the electron density.

As regards the field produced by the phonon, we must distinguish between the phonon modes.

9.4.1 Acoustic Phonons – Piezoelectric Interaction

In the piezoelectric effect, the polarization field in terms of the strain is given by

$$P_i(\boldsymbol{r}) = \sum_{jk} \mathrm{e}_{ijk}\frac{\partial y_j(\boldsymbol{r})}{\partial r_k},$$

where, as in (9.11), $\boldsymbol{y}(\boldsymbol{r})$ is the displacement field of the atoms associated with the phonon, and e_{ijk} is the piezoelectric constant, a third-rank tensor.

From the displacement field given in (9.12), the strain tensor is

$$\frac{\partial y_j(\boldsymbol{r})}{\partial r_k} = \sum_{\boldsymbol{q},\ell} (\mathbf{e}_{\boldsymbol{q}\ell})_j \, \mathrm{i}q_k \left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}} \left\{\mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}. \tag{9.38}$$

In a zincblende structure, such as GaAs, there is only one independent constant. Nevertheless, several elements of e_{ijk} are nonzero: $\mathrm{e}_{123} = \mathrm{e}_{132} = \mathrm{e}_{213} = \mathrm{e}_{231} = \mathrm{e}_{312} = \mathrm{e}_{321}$. Thus, the first component of the polarization field associated with phonons is

$$\begin{aligned} P_1(\boldsymbol{r}) &= \mathrm{e}_{123} \left[\frac{\partial y_2(\boldsymbol{r})}{\partial r_3} + \frac{\partial y_3(\boldsymbol{r})}{\partial r_2}\right] \\ &= \mathrm{e}_{123} \, \mathrm{i} \sum_{\boldsymbol{q},\ell} \left[(\mathbf{e}_{\boldsymbol{q}\ell})_2 \, q_3 + (\mathbf{e}_{\boldsymbol{q}\ell})_3 \, q_2\right] \left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}} \left\{\mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}, \end{aligned} \tag{9.39}$$

and similar expressions for the other two components of $\boldsymbol{P}$.

Since the electric-displacement field $\boldsymbol{D}$ associated with the phonon is zero, the relation (1.17) yields the electric field associated with the phonon:

$$\boldsymbol{E}_p(\boldsymbol{r}) = -\frac{1}{\varepsilon_\circ} \boldsymbol{P}(\boldsymbol{r}). \tag{9.40}$$

After inserting (9.39) into (9.40), the result is used, together with (9.36), within the expression (9.35) for the Hamiltonian. This yields

$$\begin{aligned} H_{ep} = \mathrm{e}_{123} \, \mathrm{i} \int &\left\{\frac{(-\mathrm{e})}{4\pi\varepsilon_\circ} \frac{\partial}{\partial r'_1} \left[\frac{1}{|\boldsymbol{r} - \boldsymbol{r}'|} \mathrm{e}^{-q_\circ |\boldsymbol{r} - \boldsymbol{r}'|}\right]\right\} \\ &\times \left\{\sum_{\boldsymbol{q},\ell} \left[(\mathbf{e}_{\boldsymbol{q}\ell})_2 \, q_3 + (\mathbf{e}_{\boldsymbol{q}\ell})_3 \, q_2\right] \left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}} \left\{\mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}'}\right\} \mathrm{d}\boldsymbol{r}' + \ldots, \end{aligned}$$

where the dots indicate the products of the other components. Let us define a vector $\boldsymbol{a}(\boldsymbol{q}, \ell)$, whose first component is $\left[(\mathbf{e}_{\boldsymbol{q}\ell})_2 \, q_3 + (\mathbf{e}_{\boldsymbol{q}\ell})_3 \, q_2\right]$, and similarly for the other two components. Then

$$\begin{aligned} H_{ep} = \frac{(-\mathrm{e})}{4\pi} \frac{1}{\varepsilon_\circ} \mathrm{e}_{123} \, \mathrm{i} \sum_{\boldsymbol{q},\ell} &\left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}} \left\{\mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\} \\ &\times \boldsymbol{a}(\boldsymbol{q}, \ell) \cdot \int \nabla_{\mathsf{r}'} \left[\frac{1}{|\boldsymbol{r} - \boldsymbol{r}'|} \mathrm{e}^{-q_\circ |\boldsymbol{r} - \boldsymbol{r}'|}\right] \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}'} \mathrm{d}\boldsymbol{r}'. \end{aligned} \tag{9.41}$$

To evaluate the integral above, first we make the substitution $\boldsymbol{r}' - \boldsymbol{r} = \boldsymbol{s}$ and then integrate by parts (the screening exponential eliminates the contribution at infinity)

$$I = \int \nabla_{\mathsf{r}'} \left[\frac{1}{|\boldsymbol{r} - \boldsymbol{r}'|} \mathrm{e}^{-q_\circ |\boldsymbol{r} - \boldsymbol{r}'|} \right] \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}'} \, \mathrm{d}\boldsymbol{r}' = \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \mathrm{i}\boldsymbol{q} \int \frac{1}{|\boldsymbol{s}|} \mathrm{e}^{-q_\circ |\boldsymbol{s}|} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{s}} \, \mathrm{d}\boldsymbol{s}.$$

The integration can now be performed in polar coordinates, using the integral (860.80) in [124], and the result is

$$I = \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \mathrm{i}\boldsymbol{q} 4\pi \frac{1}{q} \frac{q}{q_\circ^2 + q^2}.$$

Equation (9.41) becomes

$$H_{ep} = -\frac{(-\mathrm{e})}{\varepsilon_\circ} \mathrm{e}_{123} \sum_{\boldsymbol{q},\ell} \boldsymbol{a}(\boldsymbol{q},\ell) \cdot \boldsymbol{q} \left(\frac{\hbar}{2\rho V \omega_\ell(\boldsymbol{q})} \right)^{\frac{1}{2}} \frac{1}{q_\circ^2 + q^2} \left\{ \mathsf{a}_{\boldsymbol{q}\ell} + \mathsf{a}^\dagger_{-\boldsymbol{q}\ell} \right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}.$$

This expression shows a complicated directional dependence of the Hamiltonian. An angular average over angles and phonon polarizations is convenient [372]. The Hamiltonian may then be written as

$$H_{ep} = -\frac{(-\mathrm{e})}{\varepsilon_\circ} p \sum_{\boldsymbol{q}\ell} \left(\frac{\hbar}{2\rho V q v_s} \right)^{\frac{1}{2}} \frac{q^2}{q_\circ^2 + q^2} \left\{ \mathsf{a}_{q\ell} + \mathsf{a}^\dagger_{-q\ell} \right\} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}},$$

where p and v_s are suitably averaged piezoelectric constant and sound velocity. Compared with the Hamiltonian in (9.13) for deformation potential interaction, the above Hamiltonian for piezoelectric interaction can be obtained from (9.13) with the substitution

$$E_1 \to p \frac{(-\mathrm{e})}{\varepsilon_\circ} \frac{q}{q_\circ^2 + q^2}.$$

Then, the transition rate in (9.14) becomes, for the piezoelectric interaction,

$$P_a^{(p)}(\boldsymbol{k}, \boldsymbol{k}') = \frac{\pi p^2 \mathrm{e}^2}{\rho V q v_s \varepsilon^2} \left(\frac{q^2}{q_\circ^2 + q^2} \right)^2 \begin{bmatrix} N_{\boldsymbol{q}} \\ N_{\boldsymbol{q}} + 1 \end{bmatrix} \mathcal{G} \; \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k}) \mp \hbar\omega_{\boldsymbol{q}}]. \quad (9.42)$$

The considerations regarding quasi-elasticity made in connection with deformation–potential interaction with acoustic phonons hold independently of the interaction mechanism. Thus, if elastic and equipartition approximations are made, and the overlap integral is taken to be unity, the scattering rate in (9.20) becomes, for the piezoelectric interaction,

$$P_{ae}^{(p)}(\boldsymbol{k}, \boldsymbol{k}') = \frac{2\pi p^2 \mathrm{e}^2 K_\mathrm{B} T}{\varepsilon^2 \hbar \rho V v_s^2} \left(\frac{q}{q_\circ^2 + q^2} \right)^2 \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})], \quad (9.43)$$

where both absorption and emission are included.

Integration over the possible final states yields again the total scattering rate:

$$P_{ae}^{(p)}(\boldsymbol{k}) = \frac{V}{(2\pi)^3} \int \frac{2\pi p^2 \mathrm{e}^2 K_\mathrm{B} T}{\varepsilon^2 \hbar \rho V v_s^2} \left(\frac{q}{q_\circ^2 + q^2} \right)^2 \delta[\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})] \, \mathrm{d}\boldsymbol{k}'.$$

It is convenient to integrate over the phonon wavevector. In a spherical and parabolic band:

$$P_{ae}^{(p)}(\boldsymbol{k}) = \frac{1}{(2\pi)^2} \frac{p^2 \mathrm{e}^2 K_\mathrm{B} T}{\varepsilon^2 \hbar \rho v_s^2} \int \left(\frac{q}{q_\circ^2 + q^2} \right)^2 \delta \left[\frac{\hbar^2 (\boldsymbol{k} \pm \boldsymbol{q})^2}{2m} - \frac{\hbar^2 k^2}{2m} \right] \mathrm{d}\boldsymbol{q}.$$

After straightforward calculations in polar coordinates with $\boldsymbol{k}$ as polar axis,

$$P_{ae}^{(p)}(\boldsymbol{k}) = \frac{1}{2\pi} \frac{p^2 \mathrm{e}^2 K_\mathrm{B} T m}{\varepsilon^2 \hbar^3 \rho v_s^2 k} \int_0^{2k} \frac{q^3}{\left(q_\circ^2 + q^2\right)^2} \, \mathrm{d}q,$$

where the integration limits have been fixed considering the discussion related to Fig. 9.2 of Sect. 9.3.1. From integral (123.2) of [124], we finally obtain the wanted transition rate as a function of electron energy [372]:

$$P_{ae}^{(p)}(\epsilon) = \frac{p^2 \sqrt{m} \mathrm{e}^2 K_\mathrm{B} T}{\sqrt{8} \pi \varepsilon^2 \hbar^2 \rho v_s^2} \frac{1}{\sqrt{\epsilon}} \left[\ln \left(1 + \frac{8m\epsilon}{\hbar^2 q_\circ^2} \right) - \frac{1}{\left(1 + \hbar^2 q_\circ^2 / 8m\epsilon\right)} \right]. \tag{9.44}$$

Figure 9.7 shows the scattering rate by piezoelectric phonons as a function of energy, as given by (9.44) in GaAs at room temperature.

9.4.2 Optical Phonons – Polar Interaction

The interaction of electrons with polar optical phonons is the dominant scattering mechanism at room temperature in polar semiconductors. It was

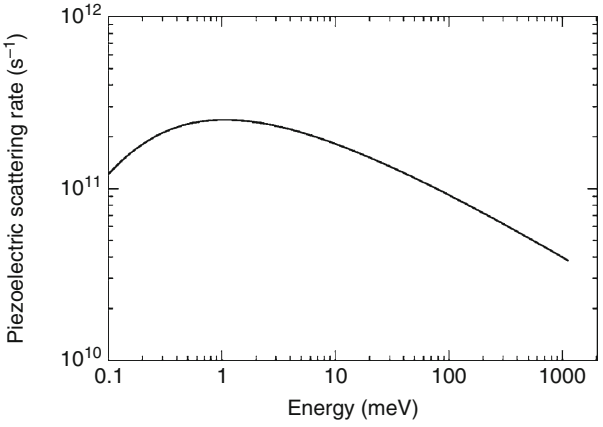


Fig. 9.7. Integrated scattering rates as a function of energy at room temperature for piezoelectric scattering. The model used refers to electrons in GaAs

originally studied by Frölich [156] and then by Callen [86] and Ehrenreich [128], and the corresponding interaction Hamiltonian carries Frölich's name.

For optical phonons, the polarization is given by

$$\boldsymbol{P}(\boldsymbol{r}) = \mathrm{e}^* \boldsymbol{y}(\boldsymbol{r})/V_c,$$

where V_c is the volume of the unit cell of the crystal, and e^* the effective charge on the atoms, to be determined later. The electric field produced by this polarization, with the displacement field given by (9.12), is

$$\begin{aligned}\boldsymbol{E}_p(\boldsymbol{r}) &= -\frac{1}{\varepsilon_\circ}\boldsymbol{P}(\boldsymbol{r}) = -\frac{1}{\varepsilon_\circ}\mathrm{e}^*\boldsymbol{y}(\boldsymbol{r})/V_c \\ &= -\frac{1}{\varepsilon_\circ}\frac{\mathrm{e}^*}{V_c}\sum_{\boldsymbol{q},\ell}\boldsymbol{e}_{\boldsymbol{q}\ell}\left(\frac{\hbar}{2\rho V\omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}}\left\{\mathsf{a}_{\boldsymbol{q}\ell}+\mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}.\end{aligned}$$

Putting this expression and (9.36) into (9.35) yields

$$\begin{aligned}H_{ep} = &-\frac{(-\mathrm{e})}{4\pi\varepsilon_\circ}\frac{\mathrm{e}^*}{V_c}\sum_{\boldsymbol{q},\ell}\mathbf{e}_{\boldsymbol{q}\ell}\left(\frac{\hbar}{2\rho V\omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}}\cdot\left\{\mathsf{a}_{\boldsymbol{q}\ell}+\mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\} \\ &\times\int\left\{\nabla_{\boldsymbol{r}'}\left[\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{e}^{-q_\circ|\boldsymbol{r}-\boldsymbol{r}'|}\right]\right\}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}'}\mathrm{d}\boldsymbol{r}'.\end{aligned}$$

The integral is the same calculated before, and the result is

$$H_{ep} = -\frac{(-\mathrm{e})}{4\pi\varepsilon_\circ}\frac{\mathrm{e}^*}{V_c}\sum_{\boldsymbol{q},\ell}\mathbf{e}_{\boldsymbol{q}\ell}\left(\frac{\hbar}{2\rho V\omega_\ell(\boldsymbol{q})}\right)^{\frac{1}{2}}\cdot\left\{\mathsf{a}_{\boldsymbol{q}\ell}+\mathsf{a}^\dagger_{-\boldsymbol{q}\ell}\right\}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}}\mathrm{i}\boldsymbol{q}4\pi\frac{1}{q_\circ^2+q^2}.$$

Again comparing with (9.13), this Hamiltonian is obtained with the substitution

$$E_1 \to \frac{(-\mathrm{e})\mathrm{e}^*}{\varepsilon_\circ V_c}\;\frac{1}{q^2+q_\circ^2},$$

so that the transition rate is

$$P^{(p)}_{op}(\boldsymbol{k},\boldsymbol{k}') = \frac{\pi}{\rho V\omega_{op}}\frac{\mathrm{e}^2\mathrm{e}^{*2}}{\varepsilon_\circ^2 V_c^2}\begin{bmatrix}N_{op}\\N_{op}+1\end{bmatrix}\mathcal{G}\left(\frac{q}{q^2+q_\circ^2}\right)^2\delta[\epsilon(\boldsymbol{k}')-\epsilon(\boldsymbol{k})\mp\hbar\omega_{op}]. \tag{9.45}$$

It remains to find a value for the effective charge e^* in the atoms. This quantity is not easy to determine directly, but it is possible to express its value in terms of quantities known experimentally. In fact, if a frequency-dependent electric field is applied to the material, at low frequency both charged atoms and electrons contribute to the polarization, while at high enough frequency the atoms, being heavier, do not contribute and only the electron contribution remains. By difference, the polarization of the atoms remains, that is related to their effective charge. Let us carry out this analysis, following [372].

In general, the relation between the polarization and the electric field is given by the definition of the electric induction $\boldsymbol{D} = \varepsilon \boldsymbol{E} = \varepsilon_\circ \boldsymbol{E} + \boldsymbol{P}$ in (1.17). From this, we have

$$\boldsymbol{P}(0) = \left(1 - \frac{\varepsilon_\circ}{\varepsilon(0)}\right) \boldsymbol{D}, \qquad \boldsymbol{P}(\infty) = \left(1 - \frac{\varepsilon_\circ}{\varepsilon(\infty)}\right) \boldsymbol{D},$$

where $\boldsymbol{P}(0)$, $\boldsymbol{P}(\infty)$, $\varepsilon(0)$, and $\varepsilon(\infty)$ are the polarizations and dielectric constant at low and high frequencies. From the difference, we obtain the polarization due to the atoms:

$$\boldsymbol{P}_{at} = \boldsymbol{P}(0) - \boldsymbol{P}(\infty) = \left(\frac{1}{\varepsilon_r(\infty)} - \frac{1}{\varepsilon_r(0)}\right) \boldsymbol{D}, \tag{9.46}$$

where $\varepsilon_r(0)$ and $\varepsilon_r(\infty)$ are the relative dielectric constants. As anticipated above, this polarization of the charged atoms can be obtained also from their dynamics. If ω_{op} is the frequency of the optical phonons, we may assume that they have a restoring force that tends to push them toward the equilibrium position given by $-\overline{M}\omega_{op}^2 \boldsymbol{y}$, where $\boldsymbol{y}$ is the relative position of the two atoms in the unit cell measured from its equilibrium value and $\overline{M}$ is their reduced mass. Thus, in presence of the microscopic electric field $\boldsymbol{E} = \boldsymbol{D}/\varepsilon_\circ$, the dynamics of the atom is[3]

$$\overline{M} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \boldsymbol{y} = -\overline{M}\omega_{op}^2 \boldsymbol{y} + \mathrm{e}^* \boldsymbol{D}/\varepsilon_\circ.$$

In the static case:

$$\boldsymbol{y} = \frac{\mathrm{e}^* \boldsymbol{D}}{\varepsilon_\circ \overline{M} \omega_{op}^2},$$

or

$$\boldsymbol{P}_{at} = \frac{\mathrm{e}^* \boldsymbol{y}}{V_c} = \frac{\mathrm{e}^*}{V_c} \frac{\mathrm{e}^* \boldsymbol{D}}{\varepsilon_\circ \overline{M} \omega_{op}^2}.$$

Comparing this with (9.46), we obtain

$$\mathrm{e}^{*2} = \varepsilon_\circ \overline{M} \omega_{op}^2 V_c \left(\frac{1}{\varepsilon_r(\infty)} - \frac{1}{\varepsilon_r(0)}\right).$$

Replacing this result into the transition rate (9.45) leads to[4]

[3] In this simplified derivation, the problem of the local effective field has been neglected. A more detailed derivation of the effective charge can be found, for example, in [57, 176].

[4] Note that the density ρ in (9.45) comes from the ratio between the atom mass and the volume of unit cell, and that in case of optical phonons the mass must be understood as the reduced mass of the two atoms as results from (5.16) and (5.17).

$$P_{op}^{(p)}(\boldsymbol{k},\boldsymbol{k}') = \frac{\pi e^2}{V}\omega_{op}\left[\frac{1}{\varepsilon(\infty)}-\frac{1}{\varepsilon(0)}\right]\begin{bmatrix} N_{op} \\ N_{op}+1\end{bmatrix}\mathcal{G}\left(\frac{q}{q^2+q_\circ^2}\right)^2 \times \delta[\epsilon(\boldsymbol{k}')-\epsilon(\boldsymbol{k})\mp\hbar\omega_{op}]. \tag{9.47}$$

Integration over the possible final states yields the total scattering rate:

$$P(\boldsymbol{k}) = \frac{V}{(2\pi)^3}\frac{\pi e^2}{V}\omega_{op}\left[\frac{1}{\varepsilon(\infty)}-\frac{1}{\varepsilon(0)}\right]\begin{bmatrix} N_{op} \\ N_{op}+1\end{bmatrix} \times \int \mathcal{G}\left(\frac{q}{q^2+q_\circ^2}\right)^2 \delta[\epsilon(\boldsymbol{k}')-\epsilon(\boldsymbol{k})\mp\hbar\omega_{op}]\mathrm{d}\boldsymbol{k}', \tag{9.48}$$

where the frequency of optical phonons is again assumed constant. The integration is easily performed for the simple case of spherical and parabolic bands and for an overlap integral equal to unity. As for piezoelectric scattering, since the scattering probability depends upon q it is convenient to perform the integration over $\boldsymbol{q}$. After the standard steps, the integral becomes:

$$\int\left(\frac{q}{q^2+q_\circ^2}\right)^2 \delta\left[\frac{\hbar^2(\boldsymbol{k}\pm\boldsymbol{q})^2}{2m}-\frac{\hbar^2k^2}{2m}\mp\hbar\omega_{op}\right]\mathrm{d}\boldsymbol{q}$$

$$= 2\pi\frac{m}{\hbar^2 k}\int_{q_{\min}}^{q_{\max}}\frac{q^3}{(q_\circ^2+q^2)^2}\,\mathrm{d}q.$$

From the right part of Fig. 9.2, it is clear that for absorption

$$q_{\min} = k'-k, \qquad q_{\max} = k'+k,$$

and for emission

$$q_{\min} = k-k', \qquad q_{\max} = k'+k,$$

where k and k' are the moduli of the wavevectors before and after scattering. This time the integral does not diverge, even in absence of screening, since the minimum q is not zero. Thus, for simplicity, we neglect screening, that is relevant only for highly doped materials. The value of the integral is then

$$2\pi\frac{m}{\hbar^2 k}\int_{q_{\min}}^{q_{\max}}\frac{1}{q}\,\mathrm{d}q = 2\pi\frac{m}{\hbar^2 k}\ln\frac{|q_{\max}|}{|q_{\min}|}. \tag{9.49}$$

The case of nonparabolic bands with also the inclusion of a realistic overlap integral is shown and discussed in [137]. Putting together our results, i.e., using the value (9.49) for the integral, with the limits of q just indicated, into (9.48), we obtain

$$P_{op}^{(p)}(\epsilon) = \frac{\sqrt{m}}{4\pi}\frac{e^2}{\hbar\sqrt{2\epsilon}}\omega_{op}\left[\frac{1}{\varepsilon(\infty)}-\frac{1}{\varepsilon(0)}\right]\begin{bmatrix} N_{op} \\ N_{op}+1\end{bmatrix}\ln\frac{\sqrt{\epsilon}+\sqrt{\epsilon'}}{|\sqrt{\epsilon}-\sqrt{\epsilon'}|}. \tag{9.50}$$

Figure 9.8 shows this scattering rate for polar optical phonons in GaAs.

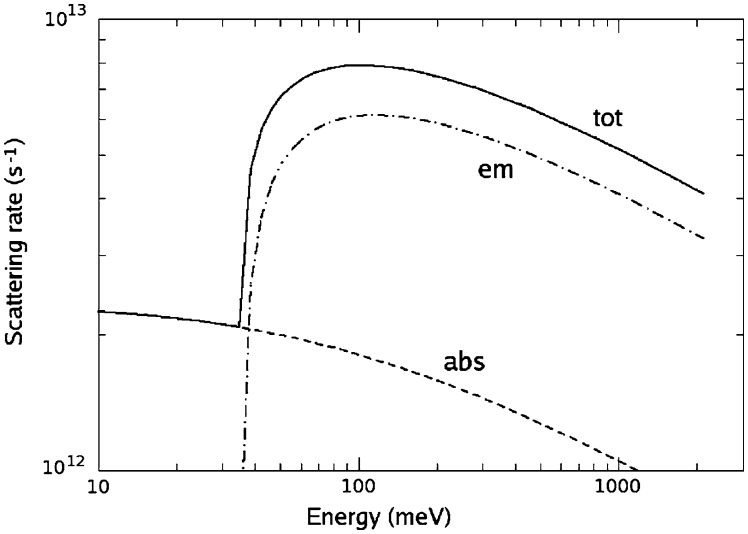


Fig. 9.8. Integrated scattering rates as a function of energy at room temperature for polar optical phonon scattering. The model used refers to electrons in GaAs

9.5 Selection Rules

The transition rates given by the Fermi golden rule in (9.2) contain the matrix elements of the interaction Hamiltonian between the final and initial states of the total system. In some cases, the symmetry properties of the electron wavefunctions and of the interaction Hamiltonian are such that the matrix elements are zero. In such cases, we say that the corresponding transitions are *forbidden by a selection rule*. These problems are best treated with the mathematical technique of group theory. Selection rules for electron–phonon scattering in semiconductors have been studied by many authors, for example in [41–44, 141, 181, 268, 403].

In cubic semiconductors, intravalley scattering by phonons is allowed, to the lowest (zero) order in the phonon wavevector, with LA modes for electrons in Γ, with acoustic modes for electrons in X, and with all modes for electrons in L. Furthermore, all zero-order transitions are allowed for holes at Γ [41,181].

Concerning intervalley scattering, group-theoretical analysis [44, 269] shows that g-scattering (between parallel valleys) is assisted by LO modes, and f-scattering (between perpendicular valleys) is assisted by LA and TO modes. From Γ to L, from L to L, and from L to X transitions are possible with longitudinal modes.

It is worth noting that, in practice, initial and final states never coincide exactly with high-symmetry points, and consequently the selection rules need not be strictly fulfilled, as confirmed in magnetophonon experiments [126]. From continuity, however, it is reasonable to expect that the "almost forbidden" transitions will remain weak when compared with allowed processes.

Ferry [141] has calculated the matrix elements for optical and intervalley scattering to first order in the phonon wavevector; they become significant when the zero-order transitions are forbidden by symmetry.

9.6 Impurity Scattering

Impurities are much heavier than electrons and their interactions may be described through a static potential $V(\boldsymbol{r})$. Thus, the collisions are elastic; they relax electron momentum but not energy. When linear transport is investigated the energy distribution of the carriers is assumed to be that of thermal equilibrium and impurities alone may control their transport properties. When, instead, nonequilibrium transport is studied, impurity scattering must be accompanied by some dissipative scattering mechanism, such as scattering by phonons, if the proper energy distribution of electrons is to be derived from theory. This is true, in particular, when Monte Carlo simulations are performed: in absence of dissipation, the energy of the carriers would increase indefinitely leading to meaningless results.

Electron–impurity interaction is described as scattering processes of an electron by the impurity field. A scattering event may be treated with the lowest order Born approximation when the electron mean wavelength is smaller than a characteristic dimension of the impurity field, otherwise a more rigorous quantum approach, such as the analysis of phase shifts of partial waves [306, 398] should be considered [103, 372]. Furthermore, the scattering of an electron by one impurity is treated as independent of all other impurities, an assumption that can be considered valid only at low impurity concentrations.

In the following, we shall present the simple semiclassical approach to impurity scattering that has been found sufficient for most cases of interest.

9.6.1 Ionized Impurities

For an ionized impurity, the scattering source is a screened Coulomb potential. The problem is generally treated with two different formulations: the Brooks and Herring (BH) approach [65] and the Conwell and Weisskopf (CW) approach [109]. The two approaches differ in the model used to screen the potential of the ion, and both of them use the Born approximation, equivalent to the perturbation theory as used in Sect. 9.2.

In the BH approach, an exponential screening is introduced, so that the scattering potential is given by

$$V_i(r) = \frac{Z\mathrm{e}(-\mathrm{e})}{4\pi\varepsilon r}\mathrm{e}^{-q_\circ r}, \tag{9.51}$$

where ε and $q_\circ$ are the dielectric constant of the materials and the inverse screening length already seen in the previous sections; Z the number of charge units of the impurity.

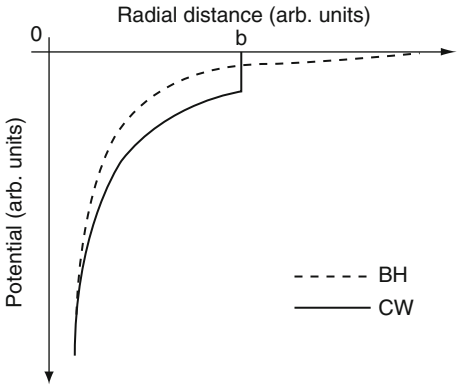


Fig. 9.9. Radial dependence of the ionized impurity potential in the BH and CW approaches [213]

In the CW approach, a bare potential is assumed, but an electron is supposed to be scattered only by the closest impurity. Thus, a cut-off of the potential is assumed at the mean distance b between the impurities:

$$b = \left[\frac{3}{4\pi n_I}\right]^{1/3},$$

where n_I is the impurity concentration.

Here, we shall unify the calculations by using the potential given in (9.51) also for the CW approach, taking for this case $q_\circ = 0$, and a maximum impact parameter b, or a minimum scattering angle θ_m.

When a high degree of compensation is present, very few free carriers are available to screen a much larger number of ionized (positive and negative) impurities. The latter, however, effectively screen each other, so that the CW approach seems to be more appropriate for this condition. In the opposite case, when each ionized impurity contributes one free carrier, n_I can be used as n in (9.37). Figure 9.9 shows, for this case, the space dependence of the scattering potential in the two approaches.

Also note that the screening length given in (9.37) is evaluated with the assumption that the electron has an equilibrium energy distribution. When at high fields the energy distribution deviates from equilibrium, (9.37) is no longer valid, and the screened potential depends on the carrier distribution, so that the screening problem makes the transport equation nonlinear in the distribution function, and a self-consistent solution is to be found.

With reference to (9.3), note that this time $\mathcal{H}'$ does not act on the crystal variables but only on the electron variable $\boldsymbol{r}$. In the position representation, it simply reduces to the function in (9.51). We can treat both terms in (9.3) together and write

$$V_i(\boldsymbol{r}) = \frac{1}{\sqrt{V}} \frac{V}{(2\pi)^3} \int A(\boldsymbol{q}) \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \, \mathrm{d}\boldsymbol{q}.$$

With standard technique, we may obtain the Fourier transform:

$$A(\boldsymbol{q}) = \frac{1}{\sqrt{V}} \int \mathrm{d}\boldsymbol{r} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\boldsymbol{r}} V_i(\boldsymbol{r}) = \frac{1}{\sqrt{V}} \int \mathrm{d}\boldsymbol{r} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \frac{Z\mathrm{e}(-\mathrm{e})}{4\pi\varepsilon r} \mathrm{e}^{-q_\circ r}.$$

The integral may be evaluated in polar coordinates, with $\boldsymbol{q}$ as polar axis, and using (860.80) of [124] for the final integration over q:

$$A(\boldsymbol{q}) = \frac{1}{\sqrt{V}} \frac{Z\mathrm{e}(-\mathrm{e})}{\varepsilon} \frac{1}{q_\circ^2 + q^2}.$$

Replacing this into (9.7), taking into account that the transition is elastic and that the state of the crystal does not change, we obtain

$$P_i(\boldsymbol{k}, \boldsymbol{k}') = \frac{2\pi}{\hbar} \frac{Z^2 \mathrm{e}^4}{\varepsilon^2 V^2} \frac{1}{[q_\circ^2 + q^2]^2} \mathcal{G} \delta(\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})).$$

Note that large momentum transfers are unlikely because of the term q^2 in the denominator. For this reason, U-processes and intervalley scattering by ionized impurities can be ignored. The dependence upon V^{-2} in the above equation has a physical origin worth to be underlined. On one side the density of final states that will enter the total scattering probability is proportional to the volume of the crystal and will cancel one of the V in the denominator. The second V depends upon the fact that we have assumed the presence of only one impurity, and the effect of this single scattering center on an extended Bloch state is of course inversely proportional to the volume of the crystal. However, we must assume that there are N_I impurities in the crystal corresponding to a density $n_I = N_I/V$. Furthermore, we assume that the impurities are located at random in space and that they are far away from each other enough to make the probability of multiple scattering negligible. Under these conditions, the transition probability for an electron to be scattered by an impurity becomes N_I times the one above, i.e.,

$$P_i(\boldsymbol{k}, \boldsymbol{k}') = \frac{2\pi}{\hbar} n_I \frac{Z^2 \mathrm{e}^4}{\varepsilon^2 V} \frac{1}{[q_\circ^2 + q^2]^2} \mathcal{G} \delta(\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})). \tag{9.52}$$

As always, to obtain the total scattering rate, we must multiply by the density of states $V/(2\pi)^3$ and integrate over the possible final states. For this purpose, we need information on the shape of the band and on the overlap integral.

Let us consider first the simplest case of a spherical parabolic band with an overlap integral equal to unity. The integration over $\boldsymbol{k}'$ is performed in polar coordinates using $\boldsymbol{k}$ as polar axis; the δ function is used for the integration over k, this yielding the expression

$$P_i(\boldsymbol{k}) = \frac{Z^2 \mathrm{e}^4 n_I}{2\pi\hbar\varepsilon^2} \int_{(\cos\theta)_{\min}}^{(\cos\theta)_{\max}} \frac{k^2 d(\cos\theta)}{[q_\circ^2 + 2k^2(1-\cos\theta)]^2} \frac{m}{\hbar^2 k}.$$

As regards the limits of integration for the angle, note that the maximum deflection is given by $\theta = \pi$, so that $(\cos\theta)_{\min} = -1$, while the minimum depends on the approach used. Hence,

$$P_i(\boldsymbol{k}) = \frac{Z^2 \mathrm{e}^4 n_I}{2\pi\hbar^3\varepsilon^2} mk \int_{-1}^{(\cos\theta)_{\max}} \frac{\mathrm{d}\xi}{[q_\circ^2 + 2k^2(1-\xi)]^2}. \tag{9.53}$$

The integral is easily evaluated by replacing the variable ξ with $x = q_\circ^2 + 2k^2(1-\xi)$, and the result is

$$P_i(\boldsymbol{k}) = \frac{Z^2 \mathrm{e}^4 n_I}{2\pi\hbar^3\varepsilon^2} mk \frac{1}{2k^2} \left\{ \frac{1}{q_\circ^2 + 2k^2(1-(\cos\theta)_{\max})} - \frac{1}{q_\circ^2 + 4k^2} \right\}. \tag{9.54}$$

At this point, it is necessary to separate the two approaches. In the BH approach, the angle of minimum deflection is zero, since the electron can be scattered from any distance, i.e., with any impact parameter. Then $(\cos\theta)_{\max} = 1$, and (9.54), after simple algebraic manipulations, reduces to

$$P_i^{(\mathrm{BH})}(\epsilon) = \frac{\sqrt{2} Z^2 \mathrm{e}^4 n_I}{4\pi\sqrt{m}\varepsilon^2 \epsilon_\circ^2} \sqrt{\epsilon} \frac{1}{1 + 4\epsilon/\epsilon_\circ}, \tag{9.55}$$

where $\epsilon_\circ = \hbar^2 q_\circ^2 / 2m$.

In the CW approach, a pure Coulombic potential ($q_\circ = 0$) is used with a maximum impact parameter b, corresponding to a minimum deflection. From the classical theory of Coulomb scattering [168], we know that the deflection angle θ is related to the impact parameter s by

$$\tan\left(\frac{\theta}{2}\right) = \frac{Z\mathrm{e}^2}{8\pi\varepsilon\epsilon s}.$$

The minimum deflection corresponds to the maximum impact parameter b. Thus,

$$\tan\left(\frac{\theta_{\min}}{2}\right) = \frac{Z\mathrm{e}^2}{8\pi\varepsilon\epsilon b} = \frac{\epsilon_b}{\epsilon},$$

where $\epsilon_b = (Z\mathrm{e}^2/8\varepsilon\pi b)$, and

$$1 - (\cos\theta)_{\max} = 2\sin^2\left(\frac{\theta_{\min}}{2}\right) = 2\frac{(\epsilon_b/\epsilon)^2}{1 + (\epsilon_b/\epsilon)^2} = \frac{2}{\epsilon^2/\epsilon_b^2 + 1}.$$

Equation (9.54) then becomes, after straightforward algebra,

$$P_i^{(CW)}(\epsilon) = n_I \frac{\hbar k}{m} \pi b^2 = n_I v(\epsilon) \sigma, \tag{9.56}$$

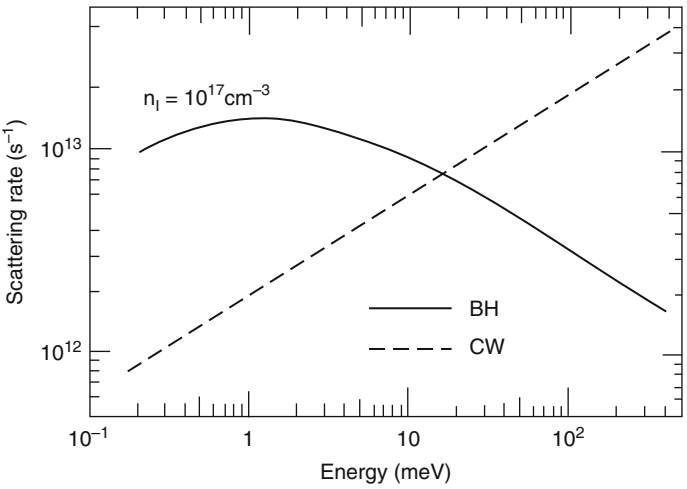


Fig. 9.10. Scattering rate for ionized impurities. The physical model refers to electrons in silicon with a parabolic band [213]

where $v(\epsilon)$ is the electron velocity and $\sigma = \pi b^2$ the scattering cross section. The above result has a very clear physical interpretation that does not require comments.

Figure 9.10 shows the energy dependence of the integrated scattering rate for ionized impurities in the two approaches. We should not be confused by the large difference between the two curves. In fact, they represent the total scattering rate, but the effect of each scattering event is, in average, very different in the two approaches, and the overall results are quite similar.

When ellipsoidal valleys are considered, the Herring–Vogt transformation is applied, as described above. Nonparabolicity may be accounted for in the integration over the magnitude of $\boldsymbol{k}^{*\prime}$, which is performed by means of the δ-function. The overlap integral $\mathcal{G}$ may be taken as unity to a good approximation. In fact, as mentioned before, owing to the presence of q^2 in the denominator of the transition rate (9.52), intervalley scattering by ionized impurities may be neglected, and, for intravalley scattering in the conduction band of cubic semiconductors and for small momentum transfer, $\mathcal{G}$ is unity to a good approximation. The calculations are straightforward, and the resulting scattering rates are [213]

$$P_i^{(\mathrm{BH})}(\epsilon) = \frac{\sqrt{2}Z^2 \mathrm{e}^4 n_I}{4\pi\sqrt{m_d}\varepsilon^2 \epsilon'^2_\circ}\sqrt{\gamma}\frac{1+2\alpha\epsilon}{1+4\gamma/\epsilon'_\circ}, \tag{9.57}$$

where γ and α are defined in (8.28), $\epsilon'_\circ = \hbar^2 q_\circ^2/2m_d$, and

$$P_i^{(CW)}(\epsilon) = n_I \pi b^2 \sqrt{\frac{2}{m_d}}\frac{\epsilon^2(1+2\alpha\epsilon)}{\gamma^{3/2}(\epsilon)}. \tag{9.58}$$

When holes in the valence bands are considered, the overlap integral $\mathcal{G}$ should be accounted for as well as warping. Complicated expressions result [112, 213, 362], often ignored in favor of simpler approximations.

In evaluating the effect of impurity scattering, it must be observed that at high energies Coulomb scattering is strongly peaked along the forward direction, owing to the presence of q^2 in the denominator of (9.52). Thus, a large number of scattering events may have a little effect on the electron path. In this respect, the BH approach seems to be more appropriate, since the integrated scattering rate itself decreases at sufficiently high energies.

9.6.2 Neutral Impurities

Most of the times, neutral impurities are neglected in electron transport in semiconductors. The interaction potential is of short range and is effective only for slow electrons. Two models have been developed for this type of scattering mechanism: hydrogenic models and spherically symmetrical square-well potentials. Standard references are the papers by Erginsoy [131] and by Sclar [399]. Useful discussions may be found also in, for example, [304, 308]. Ridley's book [372] covers the topic to a good extent.

9.7 Alloy Scattering

As indicated at the beginning of this chapter, the model of the virtual crystal with physical parameters intermediate between the two components, is most frequently used for the determination of the physical properties of an alloy. It assumes, however, that the alloy is perfectly homogeneous. The variation in space of the composition of the alloy, always present to some extent, produces a perturbation that generates transitions between Bloch states of the virtual crystal. Today, alloys are commonly used to engineer the band structure of materials and generate semiconductor structures with desired potential profiles (see Chap. 19). Thus, this type of electron scattering could play an important role in modern technology, strongly attenuated, however, by the great perfection reached by processing techniques to date.

Let us consider a material of the most common type in the physics of semiconductors, that is, an alloy $A_xB_{1-x}C$. If V_A and V_B are the atomic potentials of cations A and B, respectively, the corresponding virtual atomic potential can be described as

$$V_v = xV_A + (1-x)V_B.$$

Thus, in any site the actual potential is different from the virtual one: where an atom A is present the potential differs from the virtual one by

$$(\delta V)_A = V_A - V_v = (1-x)(V_A - V_B),$$

and where B is present

$$(\delta V)_B = V_B - V_v = x(V_B - V_A).$$

However, as long as the two types of atoms form a uniform periodic arrangement, we may consider that Bloch states exists with the corresponding bands, and they may be approximated by the solutions of the Hamiltonian of the virtual crystal. The perturbation that generates scattering arises when a site that in the perfect alloy should be occupied by, say, cation A is instead occupied by B or vice versa. The perturbation potential is thus given by the difference between the potential that would be present in the perfect alloy, and the potential that is present, instead, in a given site of the actual crystal. Thus, this scattering mechanism can be assimilated, to some extent, to the scattering due to impurities. If the two components A and B are not isoelectronic, the effect would be similar to that of ionized impurities. In case of isoelectronic components a neutral impurity model, such as the "square well" model, seems more appropriate.

As it regards the concentration of such scattering centers, it depends on the order parameter of the alloy. In the "random assumption," we assume that a fraction x of atoms A and a fraction $(1 - x)$ of atoms B are situated at random at the N cationic sites of the crystal. Thus, the probability that an "error" is present in a given site is the probability x that an atom A is required, times the probability $(1 - x)$ that an atom B is located there, plus the symmetric case. The density of "wrong atoms" is therefore

$$n_{ra} = \frac{N}{V} 2x(1 - x) = \frac{2}{V_c} x(1 - x),$$

where V_c is the volume of the unit cell. This concentration may of course be reduced in alloys with more perfect order. With this concentration of defects, and using a square well potential, the following total scattering rate for alloy scattering has been evaluated [182, 372]:

$$P_A(\epsilon) = \frac{\sqrt{2} V_c m^{3/2}}{\pi \hbar^4} x(1 - x)(V_B - V_A)^2 \sqrt{\epsilon}.$$

9.8 Carrier–Carrier Interaction

To include carrier–carrier (e–e) interaction in a transport theory, a knowledge of the electron distribution function $f(\boldsymbol{k})$ (defined in next chapter) is necessary to evaluate both the screening of the interaction potential between two carriers and the probability that a carrier with momentum $\boldsymbol{k}$ interacts with a second carrier with momentum $\boldsymbol{k}_1$.

The scattering problem between two identical charged particles can be solved in the center-of-mass frame of reference in the same way as the scattering from ionized impurities. If the inverse screening length is $q_\circ$, the integrated

scattering rate for an electron with momentum $\boldsymbol{k}$ that collides with an electron with momentum $\boldsymbol{k}_1$ is given by [302]

$$P_{e,e}^{\boldsymbol{k}_1}(\boldsymbol{k}) = \frac{\sqrt{2}\mathrm{e}^4}{16\pi\sqrt{m}\epsilon^2\epsilon_\circ^2}\sqrt{\epsilon_r}\frac{1}{1+\epsilon_r/\epsilon_\circ},$$

where

$$\epsilon_r = \frac{\hbar^2 k_r^2}{2m}, \qquad \epsilon_\circ = \frac{\hbar^2 q_\circ^2}{2m}, \qquad \boldsymbol{k}_r = \boldsymbol{k}_1 - \boldsymbol{k}.$$

To evaluate the total scattering rate, this probability must be integrated again over the distribution of the possible target electron wavevectors $\boldsymbol{k}_1$, in principle unknown, so that, as said above, some sort of self-consistent distribution must be obtained.

9.9 Relative Importance of the Different Scattering Mechanisms

In practice, it is useful to have some *a priori* knowledge of the type of mechanisms that may be important in given materials and given conditions of temperature and applied field, to set up the correct model for the special phenomenon under consideration. Without entering into details, a general picture can be obtained from the following considerations.

1. Phonon scattering is more effective at higher temperatures, when phonon populations are larger, and at higher fields, when the carrier energies are higher. While acoustic phonons can exchange arbitrarily small amounts of energy, optical and intervalley phonons have characteristic energies equivalent to a few hundreds Kelvin.
2. Ionized–impurity scattering becomes less effective as the carrier energy increases; it dissipates momentum but not energy.
3. Carrier–carrier scattering does not dissipate energy nor momentum, but it influences the shape of the distribution function, tending to make it Maxwellian.
4. Neutral impurities and alloy scattering are to be considered when special experimental conditions indicate that they may influence the electron transport under examination.

Therefore, at low temperature ($\leq\approx 100\,\mathrm{K}$) and low fields the following generalization can be made. Impurities are most probably important, unless the material is particularly pure. Acoustic phonons are certainly important, and their energy dissipation must be taken into account accurately. Optical phonons are generally not essential. Intervalley phonons must be considered if repopulation problems are to be investigated (see Sect. 13.3). The last point may be rather critical since, due to the characteristic energy of the intervalley phonons, intervalley transitions may be very rare at low temperature.

They will depend strongly upon the tail of the energy distribution function of the carriers, so that carrier–carrier interaction, which influences the shape of the electron distribution function, may also become relevant to anisotropy problems.

At low temperatures and high fields, owing to the higher electron energies (see Sect. 13.1), all kinds of phonon spontaneous emissions become important. Acoustic scattering may to some extent be approximated as an elastic process, and the relative importance of impurities becomes negligible.

At high temperatures, the situation is similar to that of low temperatures and high fields, except for the fact that also phonon absorption and stimulated emission play significant roles.

In Chap. 14, we shall see how it is possible to obtain from the Monte Carlo simulation information about the role of each scattering mechanism in dissipating the momentum and energy imparted to the carrier gas by the field under steady-state conditions.

10

Boltzmann Equation

10.1 The Distribution Function

Let us consider a gas of N classical particles, for example our semiclassical electrons. The *distribution function* $f(\boldsymbol{r}, \boldsymbol{v}, t)$ is defined in such a way that

$$f(\boldsymbol{r}, \boldsymbol{v}, t)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{v}$$

indicates the number of particles with positions in the volume $\mathrm{d}\boldsymbol{r}$ around $\boldsymbol{r}$ and velocities in $\mathrm{d}\boldsymbol{v}$ around $\boldsymbol{v}$, at time t. The normalization condition of the distribution function is then

$$\int \mathrm{d}\boldsymbol{r} \int \mathrm{d}\boldsymbol{v}\ f(\boldsymbol{r}, \boldsymbol{v}, t) = N.$$

If we integrate f over the velocity, we obtain the density $n(\boldsymbol{r}, t)$ in $\boldsymbol{r}$ at time t:

$$\int f(\boldsymbol{r}, \boldsymbol{v}, t)\ \mathrm{d}\boldsymbol{v} = n(\boldsymbol{r}, t).$$

In the case of electrons in crystals, we may work with wavepackets of "reasonably well defined" positions $\boldsymbol{r}$ and crystal momenta $\boldsymbol{k}$, as discussed in Sect. 6.7. In such a case

$$f(\boldsymbol{r}, \boldsymbol{k}, t)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{k}$$

is proportional to the number of electrons in $\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{k}$. As normalization condition we assume [1]

$$\boxed{\frac{2}{(2\pi)^3} \int \mathrm{d}\boldsymbol{r} \int \mathrm{d}\boldsymbol{k}\ f(\boldsymbol{r}, \boldsymbol{k}, t) = N, \quad \frac{2}{(2\pi)^3} \int \mathrm{d}\boldsymbol{k}\ f(\boldsymbol{r}, \boldsymbol{k}, t) = n(\boldsymbol{r})} \tag{10.1}$$

[1] Often, for simplicity, the limits of integration are not explicitly indicated. As it regards $\boldsymbol{r}$, they are given by the volume of the crystal; for $\boldsymbol{k}$ by the Brillouin zone.

This normalization has been chosen in such a way that for a homogeneous case

$$\frac{2V}{(2\pi)^3}\int \mathrm{d}\boldsymbol{k}\ f(\boldsymbol{k},t) = N. \tag{10.2}$$

Thus, since $2V/(2\pi)^3$ is the density of electron states, including spin, f becomes the occupation number of the state $\boldsymbol{k}$, and the condition $f \ll 1$ must be required for the electron gas to be non degenerate.

10.1.1 Mean Quantities

In general, the mean value $\langle A\rangle$ of a quantity $A(\boldsymbol{k})$ in a homogeneous system is obtained by means of the distribution function as

$$\langle A\rangle = \frac{\int A(\boldsymbol{k})f(\boldsymbol{k})\,\mathrm{d}\boldsymbol{k}}{\int f(\boldsymbol{k})\,\mathrm{d}\boldsymbol{k}}.$$

If the quantity of interest is a function of energy alone, and also the distribution function is a function of energy alone, as in the case of equilibrium, the above expression can be written as

$$\langle A\rangle = \frac{\int A(\epsilon)f(\epsilon)g(\epsilon)\mathrm{d}\epsilon}{\int g(\epsilon)f(\epsilon)\mathrm{d}\epsilon},$$

where $g(\epsilon)$ is the density of states in energy. If the considered band is assumed to be parabolic with spherical equienergetic surfaces, $g(\epsilon)$ is given by (8.12) of Chap. 8, so that, simplifying equal constants in numerator and denominator, the mean takes the form

$$\langle A\rangle = \frac{\int A(\epsilon)\epsilon^{\frac{1}{2}}f(\epsilon)\mathrm{d}\epsilon}{\int \epsilon^{\frac{1}{2}}f(\epsilon)\mathrm{d}\epsilon}.$$

Finally, if the particle gas can be described by a nondegenerate Maxwellian, we have

$$\langle A\rangle = \frac{\int A(\epsilon)\epsilon^{\frac{1}{2}}\mathrm{e}^{-\epsilon/K_\mathrm{B}T}\mathrm{d}\epsilon}{\int \epsilon^{\frac{1}{2}}\mathrm{e}^{-\epsilon/K_\mathrm{B}T}\mathrm{d}\epsilon}.$$

In the important example of the mean energy, the two integrals to be evaluated are

$$\int \epsilon^{\frac{1}{2}}\mathrm{e}^{-\frac{\epsilon}{K_\mathrm{B}T}}\mathrm{d}\epsilon = 2(K_\mathrm{B}T)^{\frac{3}{2}}\int_0^\infty x^2\mathrm{e}^{-x^2}\mathrm{d}x = \frac{1}{2}\sqrt{\pi}(K_\mathrm{B}T)^{\frac{3}{2}} \tag{10.3}$$

and

$$\int \epsilon^{\frac{3}{2}}\mathrm{e}^{-\frac{\epsilon}{K_\mathrm{B}T}}\mathrm{d}\epsilon = 2(K_\mathrm{B}T)^{\frac{5}{2}}\int_0^\infty x^4\mathrm{e}^{-x^2}\mathrm{d}x = 2(K_\mathrm{B}T)^{\frac{5}{2}}\frac{3}{8}\sqrt{\pi}. \tag{10.4}$$

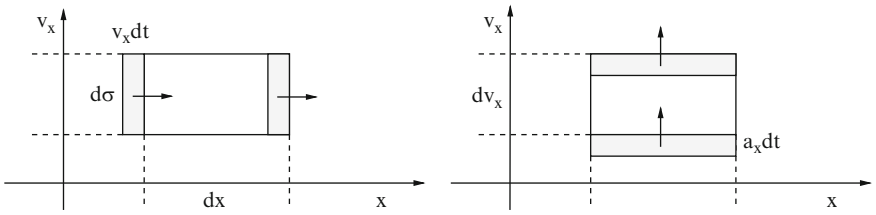


Fig. 10.1. For the derivation of Boltzmann equation

Thus,

$$\langle \epsilon \rangle = \frac{2(K_{\mathrm{B}}T)^{\frac{5}{2}}\frac{3}{8}\sqrt{\pi}}{\frac{1}{2}\sqrt{\pi}(K_{\mathrm{B}}T)^{\frac{3}{2}}} = \frac{3}{2}K_{\mathrm{B}}T. \tag{10.5}$$

This is a well-known result of classical statistics, a particular case of the *equipartition principle*: each quadratic term in the Hamiltonian contributes to the mean energy with a term $\frac{1}{2}K_{\mathrm{B}}T$. In our case, the Hamiltonian is simply given by the kinetic energy $p^2/2m$; each component of the momentum $\boldsymbol{p}$ brings a contribution $\frac{1}{2}K_{\mathrm{B}}T$, and the result is (10.5).

Remember that this result is found for the nondegenerate equilibrium distribution in a spherical, parabolic band.

10.2 Elementary Derivation of the Boltzmann Equation

Let us consider an infinitesimal parallelepiped in the six-dimensional space $(\boldsymbol{rv})$. The number of particles in this elementary volume at time t is $f(\boldsymbol{r}, \boldsymbol{v}, t)\, \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{v}$. Our purpose, now, is to evaluate the variation with time of such number. This variation is given by the balance between the numbers of particles that enter and exit from the faces of the parallelepiped. Let us examine first the plane (x, v_x) shown in Fig. 10.1. The number of particles that enter from the left face (accounting for the sign of v_x) during the interval of time $\mathrm{d}t$ is

$$f(x, v_x)\ v_x\, \mathrm{d}t\, \mathrm{d}\sigma,$$

where $\mathrm{d}\sigma$ is the area of the face of the parallelepiped orthogonal to the x axis. For brevity, the arguments of f not involved in the balance have been omitted. The number of particles that exit from the opposite side is (see the left part of Fig. 10.1)

$$f(x + \mathrm{d}x, v_x)\ v_x\, \mathrm{d}t\, \mathrm{d}\sigma.$$

The corresponding variation δN in the number of particles in the parallelepiped is then given by the difference:

$$(\delta N)_x = f(x, v_x)\, v_x\, \mathrm{d}t\, \mathrm{d}\sigma - f(x + \mathrm{d}x, v_x)\, v_x\, \mathrm{d}t\, \mathrm{d}\sigma,$$

or

$$(\delta f(x, v_x)\mathrm{d}\sigma\mathrm{d}x)_x = -\frac{f(x+\mathrm{d}x, v_x) - f(x, v_x)}{\mathrm{d}x}\,\mathrm{d}x\, v_x\,\mathrm{d}t\,\mathrm{d}\sigma.$$

Thus, the rate of change of f due to the motion of the particles along the x direction is

$$\left.\frac{\partial f}{\partial t}\right|_x = -\frac{\partial f}{\partial x} v_x.$$

Now we must consider the sides perpendicular to the v_x-axis, shown in the right part of Fig. 10.1. Particles cross these sides if they change their velocities, i.e., if they are accelerated by some external forces. Following the same argument used above, we reach the similar result:

$$\left.\frac{\partial f}{\partial t}\right|_{v_x} = -\frac{\partial f}{\partial v_x}\, a_x,$$

where a_x is the x-component of the acceleration of the particles in $(\boldsymbol{r}, \boldsymbol{v})$. By summing the two contributions and the analogous contributions along the other directions, we obtain the *Boltzmann equation* (BE) in absence of collisions, also called *Vlasov equation* in plasma physics:

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{r}} f + \boldsymbol{a} \cdot \nabla_{\boldsymbol{v}} f = 0.$$

It is important to note that this equation is identical to the Liouville theorem (3.5) seen in Sect. 3.2, where, however, ρ represents the density of representative points of the states of a statistical ensemble. This result shows that, while in our case the entire particle gas is the system under consideration, its individual particles can be considered as the ensemble replicas of a system formed by a single particle.

In the case of an electron gas in a crystal, since the external forces act on the crystal momentum, it is convenient to use the variables $\boldsymbol{r}$ and $\boldsymbol{k}$ and the collisionless BE reads

$$\frac{\partial}{\partial t} f(\boldsymbol{r}, \boldsymbol{k}, t) = -\boldsymbol{v} \cdot \nabla_{\boldsymbol{r}} f - \dot{\boldsymbol{k}} \cdot \nabla_{\boldsymbol{k}} f.$$

If we add the effect of the collisions, their contribution to the time variation of f must be added to the time variation above, due to the unperturbed dynamics. The result is the BE:

$$\boxed{\frac{\partial}{\partial t} f(\boldsymbol{r}, \boldsymbol{k}, t) + \boldsymbol{v} \cdot \nabla_{\boldsymbol{r}} f + \dot{\boldsymbol{k}} \cdot \nabla_{\boldsymbol{k}} f = \left.\frac{\partial f}{\partial t}\right|_{\mathrm{coll}}} \tag{10.6}$$

The l.h.s. represents the effect on the distribution function of the linear differential operator

$$\mathcal{L} = \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{r}} + \dot{\boldsymbol{k}} \cdot \nabla_{\boldsymbol{k}},$$

called *Liouvillian*. Thus, the BE may be written, in compact form, as

$$\mathcal{L} f(\boldsymbol{r}, \boldsymbol{k}, t) = \left.\frac{\partial f}{\partial t}\right|_{\text{coll}}. \tag{10.7}$$

The particular form of the contribution of the collisions, of fundamental importance in any transport theory, is analyzed in next section.

10.3 The Collision Integral – Detailed Balance

In an electron gas in a crystal, which for the moment will be considered homogeneous, collisions are due to imperfections of the crystal, as we have seen in the previous chapter, included, and of greatest importance, the lattice thermal vibrations, or phonons. They induce transitions between different Bloch states that form the wavepackets. $P(\boldsymbol{k}, \boldsymbol{k}')$ is the *scattering rate*, i.e., the probability per unit time that an electron in state $\boldsymbol{k}$ makes a transition to the state $\boldsymbol{k}'$, assumed empty, as effect of the perturbations. The probability per unit time that an electron in $\boldsymbol{k}$ undergoes a transition to a state in $\mathrm{d}\boldsymbol{k}'$ around $\boldsymbol{k}'$ is then

$$P(\boldsymbol{k}, \boldsymbol{k}') \frac{V}{(2\pi)^3} \left[1 - f(\boldsymbol{k}')\right] \mathrm{d}\boldsymbol{k}',$$

where the density of states has been taken into account, assuming that the collisions do not induce spin flip, so that the factor 2 due to the spin multiplicity is absent. The last factor accounts for the exclusion principle. The collision term in the BE (10.6) may then be written as the balance between the number of electrons that enter the element $\mathrm{d}\boldsymbol{k}$ (*scattering in*) and those that exit from the same element (*scattering out*) as effect of the collisions:

$$\boxed{\begin{aligned}\left.\frac{\partial f}{\partial t}\right|_{\text{coll}} = \frac{V}{(2\pi)^3} \int \big\{ & f(\boldsymbol{k}', t) P(\boldsymbol{k}', \boldsymbol{k}) \left[1 - f(\boldsymbol{k}, t)\right] \\ & - f(\boldsymbol{k}, t) P(\boldsymbol{k}, \boldsymbol{k}') \left[1 - f(\boldsymbol{k}', t)\right] \big\} \, \mathrm{d}\boldsymbol{k}'\end{aligned}} \tag{10.8}$$

It is clear that when this expression is inserted into the BE, the latter becomes an integro-differential equation, and even if the scattering rates do not depend on f, the integral term is not linear in f because of the exclusion principle.

Let us now consider an equilibrium situation, when collisions do not modify the Fermi–Dirac $f_F(\boldsymbol{k})$ distribution. This means that if we insert f_F into the collision integral, this must vanish:

$$\int \left\{ f_F(\boldsymbol{k}') P(\boldsymbol{k}', \boldsymbol{k}) \left[1 - f_F(\boldsymbol{k})\right] - f_F(\boldsymbol{k}) P(\boldsymbol{k}, \boldsymbol{k}') \left[1 - f_F(\boldsymbol{k}')\right] \right\} \mathrm{d}\boldsymbol{k}' = 0.$$

The *detailed-balance principle* states that this integral is zero because the integrand itself is zero, i.e., each transition occurs, at equilibrium, with the same frequency as the opposite transition:

$$f_F(\boldsymbol{k}')P(\boldsymbol{k}',\boldsymbol{k})\left[1-f_F(\boldsymbol{k})\right] = f_F(\boldsymbol{k})P(\boldsymbol{k},\boldsymbol{k}')\left[1-f_F(\boldsymbol{k}')\right].$$

With the explicit form of the Fermi distribution, this becomes

$$\frac{P(\boldsymbol{k},\boldsymbol{k}')}{P(\boldsymbol{k}',\boldsymbol{k})} = \frac{f_F(\boldsymbol{k}')\left[1-f_F(\boldsymbol{k})\right]}{f_F(\boldsymbol{k})\left[1-f_F(\boldsymbol{k}')\right]} = \frac{\mathrm{e}^{\frac{\epsilon(k)-\mu}{K_\mathrm{B}T}}+1}{\mathrm{e}^{\frac{\epsilon(k')-\mu}{K_\mathrm{B}T}}+1}\cdot\frac{\mathrm{e}^{-\frac{\epsilon(k')-\mu}{K_\mathrm{B}T}}+1}{\mathrm{e}^{-\frac{\epsilon(k)-\mu}{K_\mathrm{B}T}}+1},$$

or, after simple straightforward calculations (taking the first exponential as common factor in both the numerator and the denominator),

$$\boxed{\frac{P(\boldsymbol{k},\boldsymbol{k}')}{P(\boldsymbol{k}',\boldsymbol{k})} = \mathrm{e}^{-(\epsilon'-\epsilon)/K_\mathrm{B}T}} \tag{10.9}$$

For the transition rates to yield the detailed balance at equilibrium, it is necessary that the ratio between a transition probability and the probability of the opposite transition is given by the Boltzmann factor (10.9) of their energy difference.

We leave as an exercise to the reader to show that the transition rates found in the previous sections do verify the requirement in (10.9) if the phonon numbers that appear are given by their equilibrium values.

If the system is not homogeneous, the distribution function depends also upon $\boldsymbol{r}$. Also the transition rates $P(\boldsymbol{r},\boldsymbol{k},\boldsymbol{k}')$ will depend upon position, and they will be well defined only if the space variations due to nonhomogeneity are slow. In fact, the definition of crystal wavevectors requires a discrete translational symmetry. This must be achieved at least in the region occupied by the electron wavepacket. If this is not the case, the semiclassical theory of transport breaks down, and a more rigorous treatment, based on the Schrödinger equation, becomes necessary.

10.4 Moment Method

As we have just seen, the BE is an integro-differential equation, whose integral term may become very complicated, and there is no hope to find exact analytical solutions. Therefore, numerical methods have been developed that yield very satisfactory results. We shall see the most important of them, the Monte Carlo method, later in this book. However, the BE is also a fundamental theoretical tool to obtain general features of transport theory and approximate solutions of large applicability.

An important example is given by a set of macroscopic equations, called *drift-diffusion equations* and *hydrodynamic equations*, that can be derived from the BE with a method known as the *moment methods*.

Let us consider the BE in its compact form (10.7). The moment method consists in writing a certain number of equations, obtained by multiplying the BE by successive powers of the momentum or velocity, and then integrating over the momentum, leaving $\boldsymbol{r}$ as independent variable:

$$\frac{2}{(2\pi)^3}\int \boldsymbol{v}^\alpha(\boldsymbol{k})\ \mathcal{L}f(\boldsymbol{r},\boldsymbol{k},t)\mathrm{d}\boldsymbol{k} = \frac{2}{(2\pi)^3}\int \boldsymbol{v}^\alpha(\boldsymbol{k})\frac{\partial f}{\partial t}\bigg|_{\mathrm{coll}}\mathrm{d}\boldsymbol{k}, \quad \alpha = 0,1,2,\ldots. \tag{10.10}$$

For $\alpha > 1$, we actually mean tensor products. For example, for $\alpha = 2$ we mean the products $v_i v_j$. Keeping in mind that the BE describes the evolution in phase space of the particle distribution, it is reasonable to expect that (10.10) will lead to some continuity equations for physical quantities associated with the corresponding powers of $\boldsymbol{k}$, as we shall verify below. For this reason, the equations obtained by means of the moment method are also called *balance equations*.

10.4.1 Zero-Order Moment: Continuity Equation

In the case of the zero-order moment ($\alpha = 0$), (10.10) is simply the integral of the BE in the momentum space:

$$\frac{2}{(2\pi)^3}\int\left[\frac{\partial}{\partial t}f(\boldsymbol{r},\boldsymbol{k},t) + \boldsymbol{v}\cdot\nabla_{\boldsymbol{r}}f + \dot{\boldsymbol{k}}\cdot\nabla_{\boldsymbol{k}}f\right]\mathrm{d}\boldsymbol{k} = \frac{2}{(2\pi)^3}\int\frac{\partial f}{\partial t}\bigg|_{\mathrm{coll}}\mathrm{d}\boldsymbol{k}. \tag{10.11}$$

Let us consider, one by one, the terms in the l.h.s. The first term, remembering the normalization in (10.1) becomes

$$\frac{2}{(2\pi)^3}\int\frac{\partial}{\partial t}f(\boldsymbol{r},\boldsymbol{k},t)\ \mathrm{d}\boldsymbol{k} = \frac{\partial}{\partial t}n(\boldsymbol{r}). \tag{10.12}$$

As it regards the second term, we first note that, since $\boldsymbol{v}$ and $\boldsymbol{r}$ are independent variables,

$$\boldsymbol{v}\cdot\nabla_{\boldsymbol{r}}f = \nabla_{\boldsymbol{r}}\cdot(\boldsymbol{v}f). \tag{10.13}$$

Thus, the second term of (10.11) becomes

$$\frac{2}{(2\pi)^3}\int\boldsymbol{v}\cdot\nabla_{\boldsymbol{r}}f\ \mathrm{d}\boldsymbol{k} = \frac{2}{(2\pi)^3}\ \nabla_{\boldsymbol{r}}\cdot\int(\boldsymbol{v}f)\ \mathrm{d}\boldsymbol{k}.$$

The mean velocity of the particles in $\boldsymbol{r}$ is given by

$$\langle\boldsymbol{v}(\boldsymbol{r})\rangle = \frac{\int(\boldsymbol{v}f)\ \mathrm{d}\boldsymbol{k}}{\int f\ \mathrm{d}\boldsymbol{k}} = \frac{\frac{2}{(2\pi)^3}\int(\boldsymbol{v}f)\ \mathrm{d}\boldsymbol{k}}{n(\boldsymbol{r})},$$

so that

$$\frac{2}{(2\pi)^3}\int\boldsymbol{v}\cdot\nabla_{\boldsymbol{r}}f\ \mathrm{d}\boldsymbol{k} = \nabla_{\boldsymbol{r}}\cdot(n(\boldsymbol{r})\langle\boldsymbol{v}(\boldsymbol{r})\rangle) = \nabla_{\boldsymbol{r}}\cdot\boldsymbol{j}(\boldsymbol{r}).$$

Here, $\boldsymbol{j}$ indicates the particle current density (not the charge current density).

We thus understand that the first two terms of the zero-order moment yield the continuity equation, and expect that the other terms should vanish to maintain this result, as will be verified in what follows.

To analyze the third integral in the l.h.s. of (10.11), let us assume that both an electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ and a magnetic field $\boldsymbol{B}(\boldsymbol{r}, t)$ are present. The time variation of $\boldsymbol{k}$ due to such fields is given by the semiclassical dynamics and is proportional to $\boldsymbol{E}$ and $\boldsymbol{B}$. As it regards the electric field, its contribution to the integral is proportional to

$$\boldsymbol{E} \cdot \int \nabla_{\boldsymbol{k}} f \, \mathrm{d}\boldsymbol{k}. \tag{10.14}$$

This term vanishes since the integration results in the evaluation of the distribution function at infinity, or at the edge of the BZ, where f is supposed to be zero.

The contribution of the magnetic field is proportional to

$$\int \boldsymbol{v}(\boldsymbol{k}) \times \boldsymbol{B} \cdot \nabla_{\boldsymbol{k}} f \, \mathrm{d}\boldsymbol{k} = \boldsymbol{B} \cdot \int \nabla_{\boldsymbol{k}} f \times \boldsymbol{v}(\boldsymbol{k}) \, \mathrm{d}\boldsymbol{k}, \tag{10.15}$$

where use has been made of cyclic property of the vector mixed product. Now consider the vector equation

$$\nabla_{\boldsymbol{k}} \times (f\boldsymbol{v}) = \nabla_{\boldsymbol{k}} f \times \boldsymbol{v} + f \nabla_{\boldsymbol{k}} \times \boldsymbol{v}.$$

The last term is zero, being proportional to $\nabla_{\boldsymbol{k}} \times \nabla_{\boldsymbol{k}} \epsilon(\boldsymbol{k})$. The expression in (10.15) then becomes

$$\boldsymbol{B} \cdot \int \nabla_{\boldsymbol{k}} \times (f\boldsymbol{v}(\boldsymbol{k})) \mathrm{d}\boldsymbol{k}.$$

This is again zero for the same reason as the integral in (10.14).

It remains to consider the zero-order moment of the collision integral in (10.8). It is easy to understand that

$$\begin{aligned} &\int \int f(\boldsymbol{k}', t) P(\boldsymbol{k}', \boldsymbol{k}) \left[1 - f(\boldsymbol{k}, t)\right] \mathrm{d}\boldsymbol{k}' \mathrm{d}\boldsymbol{k} \\ &= \int \int f(\boldsymbol{k}, t) P(\boldsymbol{k}, \boldsymbol{k}') \left[1 - f(\boldsymbol{k}', t)\right] \mathrm{d}\boldsymbol{k}' \mathrm{d}\boldsymbol{k} \end{aligned}$$

because the integral on one side reduces to that on the other side by exchanging the names of the integration variables. The collision integral has therefore a zero-order moment equal to zero.

By collecting the above results, the zero-order moment of the BE yields the continuity equation

$$\boxed{\frac{\partial}{\partial t} n(\boldsymbol{r}) + \nabla \cdot \boldsymbol{j}(\boldsymbol{r}) = 0} \tag{10.16}$$

In the light of this derivation, we understand that the vanishing of the zero-order moment of the collision integral is equivalent to saying that collisions do not alter the local density of particles, since they modify the crystal momenta of the electrons, but not their positions.

Sometimes, in the analysis of particular semiconductor systems, such as electronic devices, a generation-recombination term must be added to the above continuity equation, not included in the BE (10.6).

10.4.2 First-Order Moment

The first-order moment of the BE is obtained by multiplying the equation by $\boldsymbol{v}$ and integrating over $\boldsymbol{k}$:

$$\frac{2}{(2\pi)^3}\int \boldsymbol{v}\left[\frac{\partial}{\partial t}f(\boldsymbol{r},\boldsymbol{k},t)+\boldsymbol{v}\cdot\nabla_{\boldsymbol{r}}f+\dot{\boldsymbol{k}}\cdot\nabla_{\boldsymbol{k}}f\right]\mathrm{d}\boldsymbol{k}=\frac{2}{(2\pi)^3}\int \boldsymbol{v}\frac{\partial f}{\partial t}\bigg|_{\mathrm{coll}}\mathrm{d}\boldsymbol{k}. \tag{10.17}$$

Let us again consider the various terms, one by one. The first term in the l.h.s. contains the integral that yields the current density, already seen in connection with the second term of the zero-order moment. The result is

$$\boldsymbol{M}_{1,t}=\frac{\partial}{\partial t}\frac{2}{(2\pi)^3}\int \boldsymbol{v}f(\boldsymbol{r},\boldsymbol{k},t)\mathrm{d}\boldsymbol{k}=\frac{\partial}{\partial t}\boldsymbol{j}. \tag{10.18}$$

To simplify the elaboration of the second term in (10.17), let us consider its x component first:

$$(M_{1,r})_x=\frac{2}{(2\pi)^3}\int v_x\left[\boldsymbol{v}\cdot\nabla_{\boldsymbol{r}}f\right]\mathrm{d}\boldsymbol{k}.$$

Now we make use again of (10.13):

$$\begin{aligned}(M_{1,r})_x&=\frac{2}{(2\pi)^3}\int v_x\left[\nabla_{\boldsymbol{r}}\cdot(f\boldsymbol{v})\right]\mathrm{d}\boldsymbol{k}=\frac{2}{(2\pi)^3}\nabla_{\boldsymbol{r}}\cdot\int(f\boldsymbol{v}v_x)\,\mathrm{d}\boldsymbol{k}\\&=\nabla_{\boldsymbol{r}}\cdot(n(\boldsymbol{r})\langle v_x\boldsymbol{v}\rangle)=\frac{\partial}{\partial x}(nv_xv_x)+\frac{\partial}{\partial y}(nv_yv_x)+\frac{\partial}{\partial z}(nv_zv_x)=(\nabla\cdot W)_x,\end{aligned}$$

where W is the tensor

$$W_{ij}(\boldsymbol{r})=n(\boldsymbol{r})\langle v_iv_j\rangle. \tag{10.19}$$

Thus, we can write

$$\boldsymbol{M}_{1,r}=\nabla\cdot W. \tag{10.20}$$

For the third term, we again assume that the force is given by an electric and a magnetic field and consider the x component of the moment. The electric-field contribution to the third term in (10.17) is given by

$$(M_{1,E})_x=\frac{2}{(2\pi)^3}\int v_x\left(\frac{q}{\hbar}\boldsymbol{E}(\boldsymbol{r})\cdot\nabla_{\boldsymbol{k}}f\right)\mathrm{d}\boldsymbol{k}=\frac{2}{(2\pi)^3}\frac{q}{\hbar}\boldsymbol{E}(\boldsymbol{r})\cdot\int(\nabla_{\boldsymbol{k}}f)v_x\mathrm{d}\boldsymbol{k}.$$

Let us use again the property of the ∇ of a product:

$$(\nabla_{\boldsymbol{k}} f)v_x = \nabla_{\boldsymbol{k}}(f v_x) - f(\nabla_{\boldsymbol{k}} v_x).$$

As before, the first term does not contribute to the integral, since, after integration, it involves the value of f at the boundaries, where it vanishes. Thus,

$$(M_{1,E})_x = -\frac{2}{(2\pi)^3}\frac{q}{\hbar}\boldsymbol{E}(\boldsymbol{r}) \cdot \int f(\nabla_{\boldsymbol{k}} v_x)\mathrm{d}\boldsymbol{k} = -n(\boldsymbol{r})\frac{q}{\hbar}\boldsymbol{E}(\boldsymbol{r}) \cdot \langle \nabla_{\boldsymbol{k}} v_x \rangle.$$

Similarly, for the magnetic field we have,

$$\begin{aligned}(M_{1,B})_x &= \frac{2}{(2\pi)^3}\int v_x \left(\frac{q}{\hbar}\boldsymbol{v} \times \boldsymbol{B}(\boldsymbol{r}) \cdot \nabla_{\boldsymbol{k}} f\right) \mathrm{d}\boldsymbol{k} \\ &= \frac{2}{(2\pi)^3}\frac{q}{\hbar}\boldsymbol{B}(\boldsymbol{r}) \cdot \int v_x(\nabla_{\boldsymbol{k}} f) \times \boldsymbol{v}\ \mathrm{d}\boldsymbol{k}.\end{aligned}$$

Use again the differential identity

$$[(\nabla_{\boldsymbol{k}} f) \times \boldsymbol{v}]v_x = \nabla_{\boldsymbol{k}} \times (f v_x \boldsymbol{v}) - f\nabla_{\boldsymbol{k}} \times (v_x \boldsymbol{v}).$$

After integration, the first term does not contribute because of the boundary conditions. Thus,

$$(M_{1,B})_x = -\frac{2}{(2\pi)^3}\frac{q}{\hbar}\boldsymbol{B}(\boldsymbol{r}) \cdot \int f\nabla_{\boldsymbol{k}} \times (v_x \boldsymbol{v})\ \mathrm{d}\boldsymbol{k}.$$

We already noticed in elaborating (10.15) that $\nabla_{\boldsymbol{k}} \times \boldsymbol{v}$ is zero. Therefore,

$$\begin{aligned}(M_{1,B})_x &= -\frac{2}{(2\pi)^3}\frac{q}{\hbar}\boldsymbol{B}(\boldsymbol{r}) \cdot \int f(\nabla_{\boldsymbol{k}} v_x) \times \boldsymbol{v}\ \mathrm{d}\boldsymbol{k} = -\frac{q}{\hbar}n(\boldsymbol{r})\langle \boldsymbol{B}(\boldsymbol{r}) \cdot (\nabla_{\boldsymbol{k}} v_x) \times \boldsymbol{v} \rangle \\ &= -\frac{q}{\hbar}n\langle (\nabla_{\boldsymbol{k}} v_x) \cdot \boldsymbol{v} \times \boldsymbol{B}(\boldsymbol{r}) \rangle.\end{aligned}$$

We can now collect the terms due to the electric and the magnetic fields and obtain the first-order moment of the term of the BE containing the external force $\boldsymbol{F}$ applied to the electrons:

$$(M_{1,F})_x = (M_{1,E})_x + (M_{1,B})_x = -n(\boldsymbol{r})\frac{1}{\hbar}\left\langle (\nabla_{\boldsymbol{k}} v_x) \cdot \boldsymbol{F} \right\rangle.$$

Note that the force cannot be taken out of the average since the magnetic force depends on the electron velocity. Now note that

$$\begin{aligned}(\nabla_{\boldsymbol{k}} v_x) \cdot \boldsymbol{F} &= \frac{\partial v_x}{\partial k_x}F_x + \frac{\partial v_x}{\partial k_y}F_y + \frac{\partial v_x}{\partial k_z}F_z \\ &= \frac{\partial}{\partial k_x}\frac{\partial}{\partial \hbar k_x}\epsilon(\boldsymbol{k})F_x + \frac{\partial}{\partial k_y}\frac{\partial}{\partial \hbar k_x}\epsilon(\boldsymbol{k})F_y + \frac{\partial}{\partial k_z}\frac{\partial}{\partial \hbar k_x}\epsilon(\boldsymbol{k})F_z.\end{aligned}$$

Changing the order of the derivatives, remembering the definition (8.23) of the inverse acceleration effective-mass tensor, and considering all three coordinates, our term becomes

$$\boldsymbol{M}_{1,F} = -n \left\langle \left(\frac{1}{m_a} \right) \boldsymbol{F} \right\rangle . \tag{10.21}$$

It remains to consider the first-order moment of the collision integral that contains the rate of change of the current density as effect of the collisions:

$$\boldsymbol{M}_{1,c} = \frac{2}{(2\pi)^3} \int \boldsymbol{v} \frac{\partial f}{\partial t}\bigg|_{\text{coll}} \mathrm{d}\boldsymbol{k} = \frac{\mathrm{d}}{\mathrm{d}t}(n(\boldsymbol{r})\langle \boldsymbol{v} \rangle)\bigg|_{\text{coll}} = \frac{\mathrm{d}\boldsymbol{j}}{\mathrm{d}t}\bigg|_{\text{coll}} .$$

In fact, collisions tend to destroy an average velocity of the particles. Let us define a tensor $1/\tilde{\tau}_v$ such that

$$\boldsymbol{M}_{1,c} = \frac{\mathrm{d}\boldsymbol{j}}{\mathrm{d}t}\bigg|_{\text{coll}} = -\frac{1}{\tilde{\tau}_v}\boldsymbol{j}. \tag{10.22}$$

This is not an approximation, but rather a definition of a characteristic time that depends on the distribution function, and therefore on the applied fields, temperature, concentration, etc. This definition is simply a way to indicate the first-order moment of the collision integral in more physical terms.

Combining the results (10.18), (10.20), (10.21), and (10.22) obtained above, for the first-order moment of the BE we obtain

$$\frac{\partial}{\partial t}\boldsymbol{j} + \nabla \cdot W - n \left\langle \frac{1}{m_a} \boldsymbol{F} \right\rangle = -\frac{1}{\tilde{\tau}_v}\boldsymbol{j}. \tag{10.23}$$

This is the velocity balance equation. According to this equation, the time variation of the particle flux $\boldsymbol{j}$ is given by a term due to variation in space of the distribution function, whose physical meaning will be clarified in the next section, a term due to the applied fields, and a term due to the relaxation induced by collisions.

10.4.3 Drift-Diffusion Equation

We will now see that under some simplifying assumptions the above balance equation (10.23) yields the well-known *drift-diffusion equation*, frequently used in the theoretical analysis of electronic devices. For this purpose, let us elaborate further the different terms one by one. The first term in (10.23) can be put in the form

$$\frac{\partial}{\partial t}\boldsymbol{j} = \frac{\partial}{\partial t}\left(n\langle \boldsymbol{v} \rangle\right) = \frac{\partial n}{\partial t}\langle \boldsymbol{v} \rangle + n\frac{\partial \langle \boldsymbol{v} \rangle}{\partial t}. \tag{10.24}$$

As regards the second term in (10.23), it is clear that the tensor W, defined in (10.19), is related to the kinetic energy of the electrons. Thus, $\nabla \cdot W$ in the balance equation (10.23) indicates a source of current density due to the

variation of kinetic energy in space. It is then useful to separate such kinetic energy into its ordered contribution, due to the drift of the carriers, and its fluctuating random contribution, related to the thermal excitation. For this purpose, let us consider the velocity of each particle as the sum of the mean value plus its fluctuation:

$$\boldsymbol{v} = \langle \boldsymbol{v} \rangle + \delta \boldsymbol{v}.$$

The tensor W becomes

$$W_{ij} = n\langle v_i v_j \rangle = n\langle (\langle v_i \rangle + \delta v_i)(\langle v_j \rangle + \delta v_j) \rangle = n\langle v_i \rangle \langle v_j \rangle + n\langle \delta v_i \delta v_j \rangle, \quad (10.25)$$

where we have taken into account that the mean value of the fluctuation is zero. We may define a tensor

$$w_{ij} = \langle \delta v_i \delta v_j \rangle$$

as the average product of the velocity fluctuations. It is clearly related to the particle temperature. The j-th component of $\nabla \cdot W$ due to the first term (drift energy) in (10.25) is

$$\begin{aligned} &\frac{\partial}{\partial x}(n\langle v_x \rangle \langle v_j \rangle) + \frac{\partial}{\partial y}(n\langle v_y \rangle \langle v_j \rangle) + \frac{\partial}{\partial z}(n\langle v_z \rangle \langle v_j \rangle) \\ &\quad = \langle v_j \rangle \nabla \cdot \boldsymbol{j} + n(\langle \boldsymbol{v} \rangle \cdot \nabla)\langle v_j \rangle. \end{aligned}$$

The second term in (10.25) yields a contribution to the j-th component of $\nabla \cdot W$ given by

$$\frac{\partial}{\partial x}(n\langle \delta v_x \delta v_j \rangle) + \frac{\partial}{\partial y}(n\langle \delta v_y \delta v_j \rangle) + \frac{\partial}{\partial z}(n\langle \delta v_z \delta v_j \rangle) = \langle (\nabla \cdot n\delta \boldsymbol{v})\, \delta v_j \rangle\,,$$

so that the total vector is

$$\nabla n \cdot w + n \nabla \cdot w.$$

Collecting the results above, the second term in (10.23) can be written in the form

$$\nabla \cdot W = \langle \boldsymbol{v} \rangle\ \nabla \cdot \boldsymbol{j} + n(\langle \boldsymbol{v} \rangle \cdot \nabla)\langle \boldsymbol{v} \rangle + \nabla n \cdot w + n \nabla \cdot w. \quad (10.26)$$

Substituting (10.24) and (10.26) into (10.23), we obtain

$$\frac{\partial n}{\partial t}\langle \boldsymbol{v} \rangle + n\frac{\partial \langle \boldsymbol{v} \rangle}{\partial t} + \langle \boldsymbol{v} \rangle \nabla \cdot \boldsymbol{j} + n(\langle \boldsymbol{v} \rangle \cdot \nabla)\langle \boldsymbol{v} \rangle + \nabla n \cdot w + n \nabla \cdot w - n \left\langle \frac{1}{m_a} \boldsymbol{F} \right\rangle = -\frac{1}{\tilde{\tau}_v} \boldsymbol{j}.$$

The first and third terms cancel because of the continuity equation (10.16). Furthermore, we make the physical assumption that the second term is negligible with respect to the right-hand side. In fact, in steady state the second term is zero; if the state is not stationary, because external forces or gradients change with time, we assume that the collisions are fast enough to let

the current density follow quasi-statically the variations of external fields and gradients, which is equivalent to the assumption just made.

Let us now multiply the above equation by $\tilde{\tau}_v$, the inverse tensor of $1/\tilde{\tau}_v$ defined in (10.22), and obtain

$$\boldsymbol{j} = -\tilde{\tau}_v n(\langle \boldsymbol{v} \rangle \cdot \nabla)\langle \boldsymbol{v} \rangle - \tilde{\tau}_v \nabla n \cdot w - \tilde{\tau}_v n \nabla \cdot w + \tilde{\tau}_v n \left\langle \frac{1}{m_a} \boldsymbol{F} \right\rangle . \tag{10.27}$$

The first term is a convective term, due to the variation in space of the average velocity. It is neglected in the drift-diffusion approach.

The last term in (10.27) is the drift term induced by the applied fields. If we assume that only an electric field is present, it takes the form

$$\tilde{\tau}_v n \frac{1}{m_a} q\boldsymbol{E} = n\mu \boldsymbol{E},$$

where the mobility μ is the tensor

$$\mu = \tilde{\tau}_v \frac{1}{m_a} q. \tag{10.28}$$

This expression is similar to what we will find in Chap. 11 in the relaxation-time approximation. Now, however, $\tilde{\tau}_v$ is not the relaxation time in the relaxation-time approximation, but is the relaxation characteristic time defined in (10.22).

The second term in (10.27), proportional to the gradient of the concentration n, is the diffusion term. Since w is a symmetric tensor, the x component of this term may be transformed as

$$(\tilde{\tau}_v \nabla n \cdot w)_x = \sum_i (\tilde{\tau}_v)_{xi} \sum_j (\nabla n)_j w_{ji} = \sum_{ij} (\tilde{\tau}_v)_{xi} w_{ij} (\nabla n)_j.$$

Thus,

$$\tilde{\tau}_v \nabla n \cdot w = D \nabla n,$$

which is the traditional form of the diffusion current, if we identify the diffusion coefficient D with the tensor [2]

$$D = \tilde{\tau}_v \, w. \tag{10.29}$$

[2] If we assume that the velocity fluctuations are those of thermal equilibrium and that the acceleration effective mass is the constant, scalar, effective mass m, then the tensor w reduces to the scalar $K_B T/m$, so that

$$D = \tilde{\tau}_v \frac{1}{m} K_B T,$$

and the comparison with (10.28) leads to the Einstein relation that we shall find in Sect. 12.2.

The third term in (10.27), proportional to the gradient of the particle temperature is responsible for the thermoelectric effect. This term too, as the convective term, is not considered in the drift-diffusion approach.

The result of this long elaboration of the first-order moment equation of the BE is the popular *drift-diffusion equation*, in absence of magnetic field:

$$\boxed{\boldsymbol{j} = n\mu\boldsymbol{E} - D\nabla n} \tag{10.30}$$

where $\boldsymbol{j}$ is here the particle current density, without charge. The current is formed by two terms: a drift term produced by the applied field and a diffusion term produced by the concentration gradient. The equation is very simple and physically intuitive, but, as we have just seen, its derivation from the BE is rather involved and shows that several approximations are necessary.

10.4.4 Higher-Order Moments: Hydrodynamic Equations

In the previous sections, we have seen that the zero-order moment of the BE yields the balance equation of the particle density and that the first-order moment yields the balance equation of the particle current density. The next moment equation contains the time derivative of the second-order velocity products $v_i v_j$ and leads to the balance of energy, i.e., the energy flux. This equation describes how power is imparted to the electron gas by the applied fields, it spreads around the gas and is dissipated by collisions. The resulting equations are called *hydrodynamic equations* [291, 385, 438] and are often used in the theoretical analysis of semiconductor electronic devices (see Sect. 18.7). Such analyses are more accurate than the ones based on the drift-diffusion equations, even though they do not account for specific details of the electron distribution function, such as high-energy tails that are of particular importance in specific features of modern devices.[3]

Balance equations are hierarchically entangled: the zero-order balance equation for the particles concentration contains the current density, a first order quantity. The first-order balance equation for the current contains the energy in the tensor W, and so on. A closure condition is required, usually obtained with suitable physical assumptions. Alternatively, we may notice that the equations that are derived with the moment method contain coefficients that are not furnished by the method itself, such as the mobility and the diffusion coefficient in the drift-diffusion equation, or the thermal conductivity in the hydrodynamical equations. These coefficients can be taken from experimental data or from numerical solutions of the BE.

Another way to use the moment method in transport problems, consists in assuming a reasonable analytical form for the carrier distribution function

[3] As we shall see, detailed information on the electron distribution function can be obtained by the Monte Carlo numerical solution of the BE, at the price of much longer computation times.

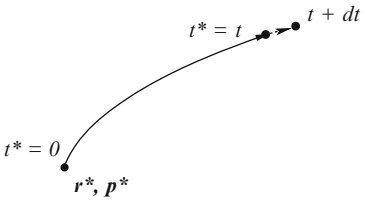


Fig. 10.2. Path variables

with some unknown parameters, and then obtaining such parameters from the moment equations. A typical example is that of the drifted and heated Maxwellian distribution (see Sect. 13.2). In this approach, the mean or drift velocity v_d of the electrons and their mean energy, or electron temperature T_e, are two parameters that can be obtained from the balance equations of first and second order.

10.5 Chambers' Integral Equation

10.5.1 Path Variables

In the phase-space formulation of the BE, $\boldsymbol{r}$, $\boldsymbol{p}$, and t are independent variables. We now make a change of variables from $\boldsymbol{r}$, $\boldsymbol{p}$, and t to three new variables $\boldsymbol{r}^*$, $\boldsymbol{p}^*$, and t^*. The last variable t^* is simply equal to t, while $\boldsymbol{r}^*$ and $\boldsymbol{p}^*$ are the position and momentum at time $t = 0$ of a particle that, in absence of collisions, has position $\boldsymbol{r}$ and momentum $\boldsymbol{p}$ at time t. As shown in Fig. 10.5.1, $\boldsymbol{r}^*$ and $\boldsymbol{p}^*$ identify a trajectory in phase space, and for this reason they are called *path variables*, and t^* gives the position of the particle on such trajectory.

Given a particle in $\boldsymbol{r}$, $\boldsymbol{p}$, at time t, let us follow it for an infinitesimal amount of time $\mathrm{d}t$, as shown in the figure. At this new time, the particle will be in $\boldsymbol{r}+\boldsymbol{v}\mathrm{d}t$ with momentum $\boldsymbol{p}+\boldsymbol{F}\mathrm{d}t$, while the path variables $\boldsymbol{r}^*$ e $\boldsymbol{p}^*$ do not change since they represent the initial position and momentum of the same particle. This means that

$$\boldsymbol{r}^*(\boldsymbol{r}+\boldsymbol{v}\mathrm{d}t, \boldsymbol{p}+\boldsymbol{F}\mathrm{d}t, t+\mathrm{d}t) = \boldsymbol{r}^*(\boldsymbol{r}, \boldsymbol{p}, t), \tag{10.31}$$

and

$$\boldsymbol{p}^*(\boldsymbol{r}+\boldsymbol{v}\mathrm{d}t, \boldsymbol{p}+\boldsymbol{F}\mathrm{d}t, t+\mathrm{d}t) = \boldsymbol{p}^*(\boldsymbol{r}, \boldsymbol{p}, t). \tag{10.32}$$

From the first one, we obtain

$$\boldsymbol{r}^*(\boldsymbol{r}, \boldsymbol{p}, t) + (\nabla_{\boldsymbol{r}}\boldsymbol{r}^*)\boldsymbol{v}\mathrm{d}t + (\nabla_{\boldsymbol{p}}\boldsymbol{r}^*)\boldsymbol{F}\mathrm{d}t + \frac{\partial \boldsymbol{r}^*}{\partial t}\mathrm{d}t = \boldsymbol{r}^*(\boldsymbol{r}, \boldsymbol{p}, t),$$

or

$$\boldsymbol{v}\ \nabla_{\boldsymbol{r}}\boldsymbol{r}^* + \boldsymbol{F}\ \nabla_{\boldsymbol{p}}\boldsymbol{r}^* + \frac{\partial \boldsymbol{r}^*}{\partial t} = 0. \tag{10.33}$$

In an identical way we obtain, for the momentum,

$$\boldsymbol{v}\ \nabla_{\boldsymbol{r}}\boldsymbol{p}^* + \boldsymbol{F}\ \nabla_{\boldsymbol{p}}\boldsymbol{p}^* + \frac{\partial \boldsymbol{p}^*}{\partial t} = 0. \tag{10.34}$$

Now we consider the distribution function as a function of the new variables and define

$$f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = f(\boldsymbol{r}, \boldsymbol{p}, t).$$

Consider one by one the various terms of the l.h.s. (Liouvillian) of the BE:

$$\frac{\partial f}{\partial t} = (\nabla_{\boldsymbol{r}^*} f^*)\frac{\partial \boldsymbol{r}^*}{\partial t} + (\nabla_{\boldsymbol{p}^*} f^*)\frac{\partial \boldsymbol{p}^*}{\partial t} + \frac{\partial f^*}{\partial t^*}\frac{\partial t^*}{\partial t},$$

$$\nabla_{\boldsymbol{r}} f = \nabla_{\boldsymbol{r}^*} f^* \nabla_{\boldsymbol{r}}\boldsymbol{r}^* + \nabla_{\boldsymbol{p}^*} f^* \nabla_{\boldsymbol{r}}\boldsymbol{p}^* + \frac{\partial f^*}{\partial t^*}\nabla_{\boldsymbol{r}} t^*,$$

$$\nabla_{\boldsymbol{p}} f = \nabla_{\boldsymbol{r}^*} f^* \nabla_{\boldsymbol{p}}\boldsymbol{r}^* + \nabla_{\boldsymbol{p}^*} f^* \nabla_{\boldsymbol{p}}\boldsymbol{p}^* + \frac{\partial f^*}{\partial t^*}\nabla_{\boldsymbol{p}} t^*.$$

Thus, the full Liouvillian becomes

$$\begin{aligned}\frac{\partial f}{\partial t} + \boldsymbol{v}\nabla_{\boldsymbol{r}} f + \boldsymbol{F}\nabla_{\boldsymbol{p}} f = {} & (\nabla_{\boldsymbol{r}^*} f^*)\frac{\partial \boldsymbol{r}^*}{\partial t} + (\nabla_{\boldsymbol{p}^*} f^*)\frac{\partial \boldsymbol{p}^*}{\partial t} + \frac{\partial f^*}{\partial t^*}\frac{\partial t^*}{\partial t} \\ & + \boldsymbol{v}\left[\nabla_{\boldsymbol{r}^*} f^* \nabla_{\boldsymbol{r}}\boldsymbol{r}^* + \nabla_{\boldsymbol{p}^*} f^* \nabla_{\boldsymbol{r}}\boldsymbol{p}^* + \frac{\partial f^*}{\partial t^*}\nabla_{\boldsymbol{r}} t^*\right] \\ & + \boldsymbol{F}\left[\nabla_{\boldsymbol{r}^*} f^* \nabla_{\boldsymbol{p}}\boldsymbol{r}^* + \nabla_{\boldsymbol{p}^*} f^* \nabla_{\boldsymbol{p}}\boldsymbol{p}^* + \frac{\partial f^*}{\partial t^*}\nabla_{\boldsymbol{p}} t^*\right].\end{aligned}$$

Rearranging the terms, taking into account that

$$\frac{\partial t^*}{\partial t} = 1, \quad \nabla_{\boldsymbol{r}} t^* = 0, \quad \nabla_{\boldsymbol{p}} t^* = 0,$$

and using (10.33) and (10.34), we finally obtain for the Liouvillian

$$\frac{\partial f}{\partial t} + \boldsymbol{v}\nabla_{\boldsymbol{r}} f + \boldsymbol{F}\nabla_{\boldsymbol{p}} f = \frac{\partial f^*}{\partial t^*}.$$

This result should not surprise: taking the derivative of f^* with respect to t^* means to look at the variation of the distribution function around the moving particle, and we know that according to Liouville theorem this variation, in absence of collisions, is zero.

10.5.2 Chambers' Integral Equation

Consider the full BE in the form

$$\frac{\partial}{\partial t}f(\boldsymbol{r},\boldsymbol{p},t)+\frac{\boldsymbol{p}}{m}\nabla_{\boldsymbol{r}}f+\boldsymbol{F}\nabla_{\boldsymbol{p}}f=\int f(\boldsymbol{r},\boldsymbol{p}',t)P(\boldsymbol{p}',\boldsymbol{p})\mathrm{d}\boldsymbol{p}'-\lambda(\boldsymbol{p})f, \quad (10.35)$$

where

$$\lambda(\boldsymbol{p})=\int P(\boldsymbol{p},\boldsymbol{p}')\mathrm{d}\boldsymbol{p}'.$$

The reader should be familiar, by now, with using indifferently $\boldsymbol{v}$, $\boldsymbol{p}$, or $\boldsymbol{k}$ as second argument of the distribution function. Furthermore, for simplicity, we assume here that the scattering rates do not depend on position. Moving to the path variables, this equation becomes

$$\frac{\partial}{\partial t^*}f^*(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*)=\int f^*(\boldsymbol{r}^*,\boldsymbol{p}',t^*)P(\boldsymbol{p}',\boldsymbol{p}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*))\mathrm{d}\boldsymbol{p}'-\lambda(\boldsymbol{p}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*))f^*. \quad (10.36)$$

Let us now consider the function

$$\tilde{f}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*)=\mathrm{e}^{\int_0^{t^*}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}f^*(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*), \quad (10.37)$$

where $\boldsymbol{p}(t')=\boldsymbol{p}(\boldsymbol{r}^*,\boldsymbol{p}^*,t')$ is the momentum of the particle in the trajectory $(\boldsymbol{r}^*,\boldsymbol{p}^*)$ at time t'. The time derivative of (10.37) is given by

$$\frac{\partial}{\partial t^*}\tilde{f}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*)=\lambda(\boldsymbol{p}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*))\tilde{f}+\mathrm{e}^{\int_0^{t^*}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}\frac{\partial}{\partial t^*}f^*(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*),$$

or, using (10.36),

$$\begin{aligned}\frac{\partial}{\partial t^*}\tilde{f}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*)&=\lambda(\boldsymbol{p}(t^*))\tilde{f}+\mathrm{e}^{\int_0^{t^*}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}\\&\qquad\left[\int f^*(\boldsymbol{r}^*,\boldsymbol{p}',t^*)P(\boldsymbol{p}',\boldsymbol{p}(t^*))\mathrm{d}\boldsymbol{p}'-\lambda(\boldsymbol{p}(t^*))f^*\right]\\&=\mathrm{e}^{\int_0^{t^*}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}\int f^*(\boldsymbol{r}^*,\boldsymbol{p}',t^*)P(\boldsymbol{p}',\boldsymbol{p}(t^*))\mathrm{d}\boldsymbol{p}'.\end{aligned}$$

Integration with respect to t^* yields

$$\begin{aligned}\tilde{f}(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*)&=\tilde{f}(\boldsymbol{r}^*,\boldsymbol{p}^*,0)\\&\quad+\int_0^{t^*}dt^{*\prime}\mathrm{e}^{\int_0^{t^{*\prime}}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}\int f^*(\boldsymbol{r}^*,\boldsymbol{p}',t^{*\prime})P(\boldsymbol{p}',\boldsymbol{p}(t^{*\prime}))\mathrm{d}\boldsymbol{p}'.\end{aligned}$$

Now we go back to the function f^* using (10.37) and noting that the two functions $\tilde{f}$ and f^* coincide for $t^*=0$:

$$\begin{aligned}f^*(\boldsymbol{r}^*,\boldsymbol{p}^*,t^*)&=f^*(\boldsymbol{r}^*,\boldsymbol{p}^*,0)\mathrm{e}^{-\int_0^{t^*}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}\\&\quad+\int_0^{t^*}dt^{*\prime}\mathrm{e}^{-\int_{t^{*\prime}}^{t^*}\lambda(\boldsymbol{p}(t'))\mathrm{d}t'}\int f^*(\boldsymbol{r}^*,\boldsymbol{p}',t^{*\prime})P(\boldsymbol{p}',\boldsymbol{p}(t^{*\prime}))\mathrm{d}\boldsymbol{p}'.\end{aligned}$$

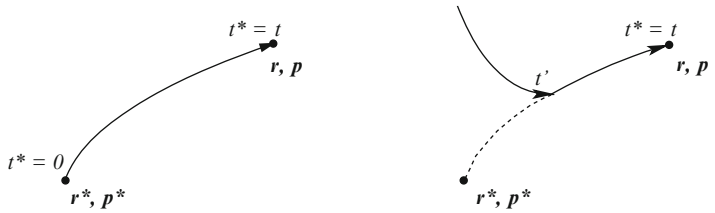


Fig. 10.3. Graphic representation of the two terms of Chambers equation: particles can reach the position $\boldsymbol{r}$ with momentum $\boldsymbol{p}$ at time t or from the initial position $\boldsymbol{r}^*$ with $\boldsymbol{p}^*$ at time $t = 0$ without suffering collisions (*left*), or being put in the right trajectory at time t' without another scattering event before t (*right*)

If we now return to the old variables, we obtain

$$\boxed{\begin{aligned} f(\boldsymbol{r},\boldsymbol{p},t) &= f(\boldsymbol{r}(0),\boldsymbol{p}(0),0)\mathrm{e}^{-\int_0^t \lambda(\boldsymbol{p}(t'))\mathrm{d}t'} \\ &+ \int_0^t \mathrm{d}t' \mathrm{e}^{-\int_{t'}^t \lambda(\boldsymbol{p}(t''))\mathrm{d}t''} \int f(\boldsymbol{r}(t'),\boldsymbol{p}',t')P(\boldsymbol{p}',\boldsymbol{p}(t'))\mathrm{d}\boldsymbol{p}' \end{aligned}} \tag{10.38}$$

where $\boldsymbol{r}(t') = \boldsymbol{r}(\boldsymbol{r}^*,\boldsymbol{p}^*,t')$ is the position of the particle in the trajectory $(\boldsymbol{r}^*,\boldsymbol{p}^*)$ at time t'. Equation (10.38) is Chambers' integral form of Boltzmann equation [99]. It can be interpreted in clear and simple physical terms graphically represented in Fig. 10.3. The distribution function in $\boldsymbol{r}$ and $\boldsymbol{p}$ at time t is given by two terms: the first one is the ballistic term $f(\boldsymbol{r}(0),\boldsymbol{p}(0),0)$, weighted by the probability that particles reach the final position, from the initial condition at $t = 0$, without being scattered off the trajectory before the observation time t; the second term is due to the contributions of the electrons that are put in the right trajectory at any time t' between $t = 0$ and time t and are not scattered off the trajectory before time t.

11

Linear Transport

The complexity of the BE is considerably reduced if only transport properties linear with respect to the applied fields are sought, as, for example, when we look for the conductivity of a material in Ohmic conditions. Before going into analytical elaborations, let us make some qualitative considerations that may provide a better insight into the phenomenon of electrical conduction.

When an electric field is applied to a semiconductor material, charge carriers are accelerated along the force direction, while, at the same time, collisions tend to restore the equilibrium momentum distribution. A stationary situation is set up where the distribution function in $\boldsymbol{k}$-space is shifted along the force direction, as qualitatively depicted in Fig. 11.1. It is clear in this figure that the number of electrons is increased in the front end at the expense of the rear region, and that the effect is appreciable where a gradient of the distribution function is present. Where the distribution is constant, its shift does not produce any effect, as in the central part of the distribution in (b).

Even in semiquantitative terms, we reach the same conclusion. If the distribution $f(k)$ is shifted of a quantity $\boldsymbol{k}_\circ$, to first order in the perturbation we may write

$$f(\boldsymbol{k}) = f_F(\boldsymbol{k} - \boldsymbol{k}_\circ) \approx f_F(\boldsymbol{k}) - \nabla_k f_F(\boldsymbol{k}) \cdot \boldsymbol{k}_\circ,$$

where f_F is the Fermi equilibrium distribution (possibly in its nondegenerate Maxwell limit). Since the equilibrium distribution is a function of the momentum only through the energy, the gradient in the above equation is proportional to the derivative of f_F with respect to energy. With these considerations in mind, let us now move to a more quantitative analysis.

11.1 Linearization of Boltzmann Equation

When the applied fields are sufficiently weak, as in the case under consideration, only effects linear with respect to the applied fields are sought, and we say we are in the *linear-response regime*. In such a case, we may write the

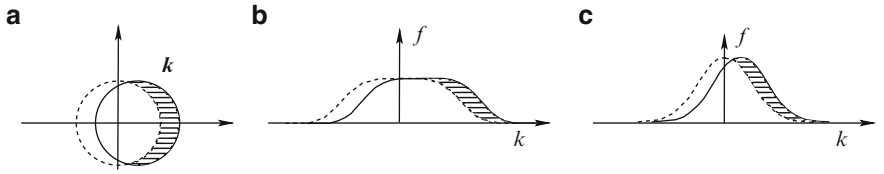


Fig. 11.1. Drifted distributions under the effects of an external field and scattering mechanisms. In part (**a**), a profile of the distribution is shown with a continuous line, while the dashed line indicates the same profile at equilibrium. In part (**b**), a one-dimensional distribution, somewhat degenerate, is shown for the two cases. The shaded areas show a portion of the distribution present because of the drift, while the symmetric part, occupied at equilibrium, is emptied by the drift. Part (**c**) shows the same situation as for the previous case for a nondegenerate electron gas

carrier distribution function as the sum of its equilibrium value $f_F(\boldsymbol{k})$ plus a term $f_1(\boldsymbol{k})$, supposed to be linear with the external fields:

$$f(\boldsymbol{k}) \approx f_F(\boldsymbol{k}) + f_1(\boldsymbol{k}). \tag{11.1}$$

More precisely, we assume $f_1(\boldsymbol{k})$ to be linear with respect to an applied electric field $\boldsymbol{E}$. We shall see, in fact, that the presence of a magnetic field alone, in a classical or semiclassical picture, leaves the distribution function unchanged, so that f_1 is zero if $\boldsymbol{E}$ is absent.

In this chapter, we shall study linear transport in a homogeneous and stationary situation, where the distribution function and the scattering mechanisms are independent of position and time. In such conditions the BE, with the position in (11.1), has the form

$$\begin{aligned}\dot{\boldsymbol{k}} \cdot \nabla_k \left[f_F(\boldsymbol{k}) + f_1(\boldsymbol{k})\right] = \frac{V}{(2\pi)^3} \int \big\{ &[f_F(\boldsymbol{k}') + f_1(\boldsymbol{k}')] P(\boldsymbol{k}', \boldsymbol{k}) \\ &\times [1 - f_F(\boldsymbol{k}) - f_1(\boldsymbol{k})] - [f_F(\boldsymbol{k}) + f_1(\boldsymbol{k})]\, P(\boldsymbol{k}, \boldsymbol{k}') \\ &\times \left[1 - f_F(\boldsymbol{k}') - f_1(\boldsymbol{k}'))\right] \big\}\, \mathrm{d}\boldsymbol{k}'.\end{aligned}$$

Terms of zero-order cancel for the equilibrium condition, and terms of second orders are neglected. The equation left for the first-order terms is

$$\begin{aligned}\dot{\boldsymbol{k}} \cdot \nabla_k f_F(\boldsymbol{k}) = \frac{V}{(2\pi)^3} \int \big\{ &f_1(\boldsymbol{k}') \left[P(\boldsymbol{k}', \boldsymbol{k})[1 - f_F(\boldsymbol{k})] + P(\boldsymbol{k}, \boldsymbol{k}') f_F(\boldsymbol{k})\right] \\ &- f_1(\boldsymbol{k}) \left[P(\boldsymbol{k}, \boldsymbol{k}')[1 - f_F(\boldsymbol{k}')] + P(\boldsymbol{k}', \boldsymbol{k}) f_F(\boldsymbol{k}')\right] \big\}\, \mathrm{d}\boldsymbol{k}'.\end{aligned} \tag{11.2}$$

Now, guided by the considerations at the beginning of this chapter, let us define a function $\phi(\boldsymbol{k})$ such that

$$f_1(\boldsymbol{k}) = -\phi(\boldsymbol{k}) \frac{\partial f_F}{\partial \epsilon} = \beta \phi(\boldsymbol{k}) f_F (1 - f_F),$$

where $\beta = 1/K_{\rm B}T$, and the derivative of the Fermi function has been evaluated explicitly. With this position, our linearized BE becomes

$$\dot{\boldsymbol{k}} \cdot \nabla_k f_1(\boldsymbol{k}) = \frac{V}{(2\pi)^3}\beta \int \left\{ \phi' f'_F(1-f'_F)\left[P(\boldsymbol{k}',\boldsymbol{k})[1-f_F] + P(\boldsymbol{k},\boldsymbol{k}')f_F\right] \right.$$
$$\left. - \phi f_F(1-f_F)\left[P(\boldsymbol{k},\boldsymbol{k}')[1-f'_F] + P(\boldsymbol{k}',\boldsymbol{k})f'_F\right] \right\} \mathrm{d}\boldsymbol{k}',$$

where for brevity we have written f for $f(\boldsymbol{k})$ and f' for $f(\boldsymbol{k}')$ and similarly for ϕ. Let $W(\boldsymbol{k},\boldsymbol{k}')$ be the transition frequencies at equilibrium:

$$W(\boldsymbol{k},\boldsymbol{k}') = f_F(\boldsymbol{k})P(\boldsymbol{k},\boldsymbol{k}')[1-f_F(\boldsymbol{k}')].$$

With this definition, our equation becomes

$$\dot{\boldsymbol{k}}\cdot\nabla_k f_1(\boldsymbol{k}) = \frac{V}{(2\pi)^3}\beta \int W(\boldsymbol{k},\boldsymbol{k}') \left\{\phi' \left[(1-f'_F)+f'_F\right] - \phi\left[(1-f_F)+f_F\right]\right\} \mathrm{d}\boldsymbol{k}',$$

where we have taken into account that, for the detailed balance,

$$W(\boldsymbol{k},\boldsymbol{k}') = W(\boldsymbol{k}',\boldsymbol{k}).$$

Thus, finally,

$$\dot{\boldsymbol{k}} \cdot \nabla_k f_1(\boldsymbol{k}) = \frac{V}{(2\pi)^3}\beta \int W(\boldsymbol{k},\boldsymbol{k}') \left\{\phi(\boldsymbol{k}') - \phi(\boldsymbol{k})\right\} \mathrm{d}\boldsymbol{k}'. \tag{11.3}$$

A variational method has been developed in [247, 248, 423] based on the properties of the integral operator in (11.3) [476, 489]. The method, initially derived for linear response, has been later extended to nonlinear transport [4]. We shall not pursue any further this topic here. The interested readers may refer to the cited literature. The variational principle is also related to the principle of maximum entropy production[1] [489].

11.2 Relaxation-Time Approximation

A great simplification of the BE is obtained by the assumption that the collision integral can be put in the form

$$\boxed{\left.\frac{\partial}{\partial t} f(\boldsymbol{k})\right|_{\rm coll} = -\frac{f-f_F}{\tau(\epsilon)} = -\frac{f_1(\boldsymbol{k})}{\tau(\epsilon)}} \tag{11.4}$$

where $\tau(\epsilon)$ is the *momentum relaxation time*, here assumed to be a function of the electron energy, as is often the case. We shall see later in this chapter

[1] A number of both maximum and minimum entropy-production principles exist, that hold in different conditions, as can be seen, for example, in [68, 118].

under what conditions we may justify this assumption, and how to evaluate the relaxation time starting from the scattering transition rates. However, the concept of relaxation-time is very powerful, in the linear response regime, also when not rigorously justified. It provides a language to describe nonequilibrium processes in many different contexts. The reason of the name "relaxation time" is due to the fact that if we drag the distribution function f out of equilibrium, and then, at $t = 0$ we remove any applied force and let f "relax" to its equilibrium value under the action of the collisions, according to (11.4), the relaxation process is governed by the equation[2]

$$\frac{\partial f_1}{\partial t} = -\frac{f_1(\boldsymbol{k})}{\tau(\epsilon)},$$

with immediate solution

$$f_1(\boldsymbol{k}, t) = f_1(\boldsymbol{k}, 0)\ \mathrm{e}^{-t/\tau}.$$

In linear response regime, τ relaxes only momentum, or velocity, because the energy distribution is still the equilibrium one. Far from equilibrium a different *energy relaxation time* may be defined, together with the momentum relaxation time, as discussed in [107].

11.3 Linear Transport Properties in a "Simple Semiconductor"

In this book we define, conventionally, a *simple semiconductor*, a virtual material with a spherical, parabolic band with effective mass m, with collisions described by a momentum relaxation time. Even though it is never rigorously correct, this model is very often used to obtain transport properties of a material, when we are not interested in fine details of physical quantities or in the analysis of phenomena that depend on specific features of the band structure or of the scattering mechanisms. The linear transport properties of such simple semiconductor are discussed in the following pages.

11.3.1 Ohmic Mobility

Let us apply a uniform and constant electric field $\boldsymbol{E}$ to our simple semiconductor. In steady state and linear-response conditions, our BE becomes

$$\nabla_{\boldsymbol{k}} f_F \cdot \dot{\boldsymbol{k}} = \frac{\partial f_F}{\partial \epsilon} \nabla_{\boldsymbol{k}} \epsilon \cdot \dot{\boldsymbol{k}} = \frac{\partial f_F}{\partial \epsilon} \boldsymbol{v} \cdot (-\mathrm{e}) \boldsymbol{E} = -\frac{f - f_F}{\tau(\epsilon)},$$

[2] While the collision integral always conserves the total number of particles, the relaxation-time approximation may violate this property.

where we have taken into account that the equilibrium distribution f_F depends upon $\boldsymbol{k}$ only through the energy. The above equation has the immediate solution

$$f = f_F + \mathrm{e}\tau \boldsymbol{v} \cdot \boldsymbol{E} \frac{\partial f_F}{\partial \epsilon}. \tag{11.5}$$

With the normalization of the distribution function in (10.2), the mean velocity is given by

$$\langle \boldsymbol{v} \rangle = \frac{1}{n} \frac{2}{(2\pi)^3} \int \boldsymbol{v}(\boldsymbol{k}) f(\boldsymbol{k}) \, \mathrm{d}\boldsymbol{k} = \frac{1}{n} \frac{2}{(2\pi)^3} \int \boldsymbol{v}(\boldsymbol{k}) \mathrm{e}\tau \boldsymbol{v} \cdot \boldsymbol{E} \frac{\partial f_F}{\partial \epsilon} \, \mathrm{d}\boldsymbol{k}, \tag{11.6}$$

where n is the carrier density, and, using the distribution function obtained in (11.5), we have taken into account that f_F does not contribute to the integral, being spherically symmetric. For symmetry reasons, the mean velocity is parallel to the field, which we may take along the z direction. The above becomes

$$\begin{aligned} \langle v \rangle &= \frac{1}{n} \frac{2}{(2\pi)^3} \mathrm{e}E \int v_z^2 \tau \frac{\partial f_F}{\partial \epsilon} \, \mathrm{d}\boldsymbol{k} = \frac{1}{n} \frac{2}{(2\pi)^3} \frac{1}{3} \mathrm{e}E \int v^2 \tau \frac{\partial f_F}{\partial \epsilon} \, \mathrm{d}\boldsymbol{k} \\ &= \frac{1}{n} \frac{2}{(2\pi)^3} \frac{2}{3} \frac{\mathrm{e}E}{m} \int \epsilon \tau \frac{\partial f_F}{\partial \epsilon} \, \mathrm{d}\boldsymbol{k}, \end{aligned} \tag{11.7}$$

where we used that the integral would have been the same for v_x^2 or v_y^2 in place of v_z^2

Until now, we have not made any assumption on the degeneracy of the equilibrium distribution function. If the semiconductor is totally degenerate (metal), the derivative of the distribution function may be approximated with a Dirac δ, but this case is not of interest for us in this section. In the opposite limit, when f_F is always much less than unity, the distribution may be approximated by a Maxwellian, as discussed in Sect. 3.7.4:

$$f_F \to f_M = C\mathrm{e}^{-\beta\epsilon} \,, \quad \frac{\partial f_F}{\partial \epsilon} = -\beta C \mathrm{e}^{-\beta\epsilon} = -\beta f_M.$$

In such a case

$$\langle v \rangle = -\frac{2}{3} \frac{\mathrm{e}E}{m} \beta \frac{1}{n} \frac{2}{(2\pi)^3} \int \epsilon \tau f_M \, \mathrm{d}\boldsymbol{k} = -\frac{\mathrm{e}E}{m} \frac{\langle \tau \epsilon \rangle}{\langle \epsilon \rangle}. \tag{11.8}$$

The *mobility* μ is defined as the ratio between mean velocity, or *drift velocity*, and electric field, and the conductivity σ as the ration between the current density and the applied field. Since the current density is $j = (-\mathrm{e})n\langle v \rangle$, we write

$$\boxed{\mu = \frac{\langle v \rangle}{E} = \frac{-\mathrm{e}}{m} \frac{\langle \tau \epsilon \rangle}{\langle \epsilon \rangle}, \quad \sigma = (-\mathrm{e}) n \mu} \tag{11.9}$$

For later use, it is convenient to find a more detailed expression for the mobility, as obtained from (11.8). First, we evaluate the normalization

constant C for the Maxwellian distribution:

$$n = \frac{N}{V} = \frac{1}{V}\frac{2V}{(2\pi)^3}\int C\mathrm{e}^{-\beta\epsilon}\,\mathrm{d}\boldsymbol{k} = C\frac{8\sqrt{2}\pi m^{3/2}}{\hbar^3\beta^{3/2}}\frac{2}{(2\pi)^3}\int_0^\infty x^2\mathrm{e}^{-x^2}\mathrm{d}x.$$

The above result has been obtained first moving to polar coordinates for $\boldsymbol{k}$, then using that $k^2\mathrm{d}k = \sqrt{2}m^{3/2}\sqrt{\epsilon}\mathrm{d}\epsilon/\hbar^3$, and finally putting $x^2 = \beta\epsilon$. Knowing that the integral has value $\sqrt{\pi}/4$, we obtain for C the value

$$C = \frac{\hbar^3}{\sqrt{2}}\left(\frac{\pi\beta}{m}\right)^{3/2} n.$$

With this result, repeating similar calculations, from (11.8) we obtain

$$\mu = \frac{4(-\mathrm{e})}{3\sqrt{\pi}m}\int \tau(x)x^{\frac{3}{2}}\mathrm{e}^{-x}\mathrm{d}x\ , \qquad x = \frac{\epsilon}{KT}. \tag{11.10}$$

A Simple, Intuitive Derivation

The relation

$$\mu = \frac{(-\mathrm{e})\tau}{m} \tag{11.11}$$

can be simply and approximately obtained with the following consideration, illustrated in Fig. 11.2. By definition, the mean velocity of the N particles in a gas is given, at any time t, by

$$\langle\boldsymbol{v}\rangle = \frac{1}{N}\sum_i \boldsymbol{v}^{(i)},$$

where $\boldsymbol{v}^{(i)}$ is the instantaneous velocity of the i-th particle at time t. It can be written as

$$\boldsymbol{v}^{(i)} = \boldsymbol{v}_\circ^{(i)} + \boldsymbol{a}\Delta t^{(i)}, \tag{11.12}$$

where $\boldsymbol{v}_\circ^{(i)}$ is the velocity of the i-th particle immediately after its last scattering event, $\boldsymbol{a}$ is the acceleration of the particles, here assumed to be the same for all particles and equal to $-\mathrm{e}\boldsymbol{E}/m$, and $\Delta t^{(i)}$ is the time interval

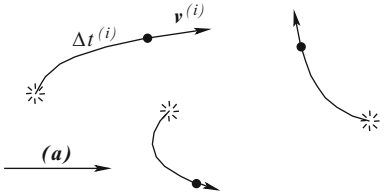


Fig. 11.2. For a simple interpretation of the mobility, see text

elapsed after the time of the last scattering event suffered by the i-th particle. Thus,

$$\langle \boldsymbol{v} \rangle = \langle \boldsymbol{v}_\circ \rangle + \boldsymbol{a} \langle \Delta t^{(i)} \rangle .$$

If we now assume that at each scattering event, a particle loses any memory of its previous velocity, and the velocity after scattering is distributed randomly with zero average, then $\langle \boldsymbol{v}_\circ \rangle = 0$. Furthermore, if the field is so weak that the energy gained during the free flight is negligible (linear response regime), then the mean time elapsed after the last collision, $\langle \Delta t^{(i)} \rangle$, is equal to the relaxation time τ and we obtain

$$\langle \boldsymbol{v} \rangle = \frac{(-\mathrm{e})\boldsymbol{E}}{m} \tau .$$

This yields the mobility in (11.11).

In the more exact expression (11.9) for the mobility, we see that the relaxation time must be weighted with the particle energy. From the derivation of this result, we realize that the larger weight at increasing energy is due in part to the larger distortion of the distribution function produced by the field at higher velocities (see (11.5)), and in part to the fact that faster electrons give a higher contribution to the mean velocity (see (11.6)).

11.3.2 Matthiessen Rule

The simple linear dependence of the electron mobility upon the relaxation time is on the basis of an empirical rule whose roots date back to the nineteenth century. If electrons are subject to several scattering mechanisms, and these may be considered independent from each other, the total scattering rate is the sum of the separate scattering rates due to the different sources:

$$P = P_1 + P_2 + \dots .$$

From this, since the relaxation time is linear with the inverse scattering probability, we may write

$$\frac{1}{\tau} = \frac{1}{\tau_1} + \frac{1}{\tau_2} + \dots ,$$

and therefore

$$\boxed{\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \dots} \tag{11.13}$$

This is Matthiessen rule, and we shall see some applications at the end of this chapter. In its derivation, it has been ignored that not always the scattering processes are independent from each other (not a serious problem in linear-response regime), that not all scattering mechanisms admit a rigorous definition of relaxation time, and that the averaging process of $\tau(\epsilon)$ is linear with respect to τ and not with respect to its inverse. Thus, this rule is

only approximately valid but it offers a very powerful tool to understand the complex mechanisms of transport in physical quantitative, albeit not exact, terms.

11.3.3 Magnetotransport

In presence of only a constant and uniform magnetic field $\boldsymbol{B}$,[3] the BE (10.6) for the steady state is written as

$$\nabla_{\boldsymbol{k}} f(\boldsymbol{k}) \cdot \left(-\frac{\mathrm{e}}{\hbar}\right) \cdot \boldsymbol{v} \times \boldsymbol{B} = \left.\frac{\partial f}{\partial t}\right|_{\mathrm{coll}}.$$

It is immediate to verify that the equilibrium distribution f_F is the solution of such equation. In fact, the collision integral with f_F vanishes, and the l.h.s. becomes

$$\frac{\partial f_F}{\partial \epsilon} \boldsymbol{v}\,(-\mathrm{e}) \cdot \boldsymbol{v} \times \boldsymbol{B} = 0. \tag{11.14}$$

This means that a magnetic field, alone, does not produce any effect on the distribution function. Figure 11.3 illustrates this fact, showing that a magnetic field, alone, induces the electrons to move along trajectories of constant energy (circles in our spherical band) where the equilibrium distribution function is constant. Thus, the effect of a magnetic field must be activated by the presence of an electric field. Let us consider, therefore, the linear response when both $\boldsymbol{E}$ and $\boldsymbol{B}$ are present. The BE for the steady state with the relaxation time is

$$\nabla_{\boldsymbol{k}} f \ \left(-\frac{\mathrm{e}}{\hbar}\right) \cdot (\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) = -\frac{f - f_F}{\tau} = -\frac{f_1}{\tau}.$$

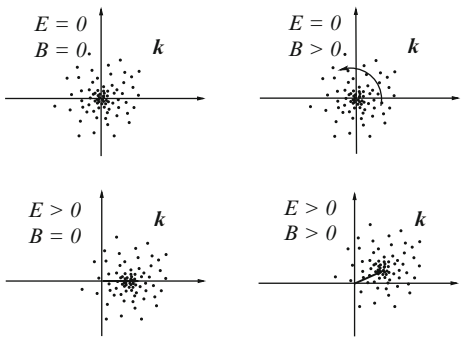


Fig. 11.3. Schematic representation of the electron distribution function with and without electric and magnetic fields

[3] Following a common practice in solid-state physics, we shall often write "magnetic field" also when we actually mean "magnetic induction field", if this misuse does not produce misunderstanding.

The term containing $\boldsymbol{E}$ can be linearized substituting f with f_F; the term with $\boldsymbol{B}$ keeps only f_1 because $\boldsymbol{B}$ has no effect on f_F, as just shown in (11.14). We thus obtain

$$\frac{\partial f_F}{\partial \epsilon}\ \mathrm{e}\boldsymbol{v}\cdot\boldsymbol{E}+\frac{\mathrm{e}}{\hbar}\nabla_{\boldsymbol{k}}f_1\ \cdot\boldsymbol{v}\times\boldsymbol{B}=\frac{f-f_F}{\tau}. \tag{11.15}$$

Guided by the previous experience with the electric field and by Fig. 11.3, we look for a solution of the form

$$f=f_F+\mathrm{e}\tau\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\frac{\partial f_F}{\partial\epsilon}. \tag{11.16}$$

In absence of $\boldsymbol{B}$, the vector $\boldsymbol{E}'$ coincides with $\boldsymbol{E}$ (see (11.5)). Now it is a vector to be determined, function of energy. If the trial position (11.16) is inserted into (11.15), we obtain

$$\frac{\partial f_F}{\partial\epsilon}\ \mathrm{e}\boldsymbol{v}\cdot\boldsymbol{E}+\frac{\mathrm{e}^2}{\hbar}\nabla_{\boldsymbol{k}}\left[\tau\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\frac{\partial f_F}{\partial\epsilon}\right]\ \cdot\boldsymbol{v}\times\boldsymbol{B}=\mathrm{e}\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\frac{\partial f_F}{\partial\epsilon}.$$

In the gradient of the square bracket, terms depending only upon energy yield results parallel to $\boldsymbol{v}$, which do not contribute. After simplification, it remains

$$\boldsymbol{v}\cdot\boldsymbol{E}+\frac{\mathrm{e}\tau}{m}\nabla_{\boldsymbol{v}}\left[\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\right]\ \cdot\boldsymbol{v}\times\boldsymbol{B}=\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon). \tag{11.17}$$

Let us consider separately the term with the $\nabla_{\boldsymbol{v}}$. Its x component is

$$\begin{aligned}\nabla_{\boldsymbol{v}}\left[\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\right]\big|_x&=\frac{\partial}{\partial v_x}\left[v_xE'_x(\epsilon)+v_yE'_y(\epsilon)+v_zE'_z(\epsilon)\right]\\&=E'_x(\epsilon)+v_x\frac{\partial E'_x}{\partial v_x}+v_y\frac{\partial E'_y}{\partial v_x}+v_z\frac{\partial E'_z}{\partial v_x}\pm v_x\frac{\partial E'_y}{\partial v_y}\pm v_x\frac{\partial E'_z}{\partial v_z},\end{aligned}$$

where the last terms have been added and subtracted. We then obtain

$$\nabla_{\boldsymbol{v}}\left[\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\right]\big|_x=E'_x(\epsilon)-\left[(\boldsymbol{v}\times\nabla_{\boldsymbol{v}})\times\boldsymbol{E}'\right]_x+\left(\nabla_{\boldsymbol{v}}\cdot\boldsymbol{E}'\right)v_x.$$

Collecting now the three components,

$$\nabla_{\boldsymbol{v}}\left[\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\right]=\boldsymbol{E}'(\epsilon)-(\boldsymbol{v}\times\nabla_{\boldsymbol{v}})\times\boldsymbol{E}'(\epsilon)+\left(\nabla_{\boldsymbol{v}}\cdot\boldsymbol{E}'(\epsilon)\right)\boldsymbol{v}. \tag{11.18}$$

The second term in (11.18) vanishes, since

$$(\boldsymbol{v}\times\nabla_{\boldsymbol{v}})\times\boldsymbol{E}'(\epsilon)=\left(\boldsymbol{v}\times m\boldsymbol{v}\frac{\partial}{\partial\epsilon}\right)\times\boldsymbol{E}'(\epsilon)=0.$$

Inserting (11.18) into (11.17), we obtain

$$\boldsymbol{v}\cdot\boldsymbol{E}+\frac{\mathrm{e}\tau}{m}\left[\boldsymbol{E}'(\epsilon)+\left(\nabla_{\boldsymbol{v}}\cdot\boldsymbol{E}'(\epsilon)\right)\boldsymbol{v}\right]\ \cdot\boldsymbol{v}\times\boldsymbol{B}=\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon).$$

The second term in the square brackets gives no contribution, being a mixed product with two parallel vectors. It remains, taking into account the cyclic property of the mixed product,

$$\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)=\boldsymbol{v}\cdot\boldsymbol{E}+\frac{\mathrm{e}\tau}{m}\boldsymbol{v}\cdot\boldsymbol{B}\times\boldsymbol{E}'(\epsilon).$$

For this to be true for any $\boldsymbol{v}$, it must be

$$\boldsymbol{E}'(\epsilon)=\boldsymbol{E}+\frac{\mathrm{e}\tau}{m}\boldsymbol{B}\times\boldsymbol{E}'(\epsilon). \tag{11.19}$$

To solve this equation, first substitute it into itself:

$$\boldsymbol{E}'(\epsilon)=\boldsymbol{E}+\frac{\mathrm{e}\tau}{m}\boldsymbol{B}\times\boldsymbol{E}+\left(\frac{\mathrm{e}\tau}{m}\right)^2\boldsymbol{B}\times\boldsymbol{B}\times\boldsymbol{E}'(\epsilon).$$

Then apply the property of the triple vector product:

$$\boldsymbol{E}'(\epsilon)=\boldsymbol{E}+\frac{\mathrm{e}\tau}{m}\boldsymbol{B}\times\boldsymbol{E}-\left(\frac{\mathrm{e}\tau B}{m}\right)^2\boldsymbol{E}'(\epsilon)+\left(\frac{\mathrm{e}\tau}{m}\right)^2(\boldsymbol{B}\cdot\boldsymbol{E}')\boldsymbol{B}.$$

From (11.19) above, it is clear that $\boldsymbol{B}\cdot\boldsymbol{E}=\boldsymbol{B}\cdot\boldsymbol{E}'$. Thus, solving for $\boldsymbol{E}'$,

$$\boldsymbol{E}'(\epsilon)=\frac{\boldsymbol{E}+\omega_c\tau\hat{\boldsymbol{B}}\times\boldsymbol{E}+(\omega_c\tau)^2(\hat{\boldsymbol{B}}\cdot\boldsymbol{E})\hat{\boldsymbol{B}}}{1+(\omega_c\tau)^2}, \tag{11.20}$$

where $\hat{\boldsymbol{B}}=\boldsymbol{B}/B$, and

$$\omega_c=\mathrm{e}B/m$$

is the cyclotron frequency. When this result is inserted into (11.16), the solution of the BE is obtained.

Now the distribution function must be inserted into the integral that yields the current $\boldsymbol{j}=(-\mathrm{e})n\langle\boldsymbol{v}\rangle$:

$$\begin{aligned}\boldsymbol{j}&=-\frac{2\mathrm{e}}{(2\pi)^3}\int\boldsymbol{v}f_1\mathrm{d}\boldsymbol{k}=-\frac{2\mathrm{e}}{(2\pi)^3}\int\boldsymbol{v}\mathrm{e}\tau\boldsymbol{v}\cdot\boldsymbol{E}'(\epsilon)\frac{\partial f_F}{\partial\epsilon}\mathrm{d}\boldsymbol{k}\\&=-\frac{2\mathrm{e}}{(2\pi)^3}\int\boldsymbol{v}\mathrm{e}\tau\frac{\boldsymbol{v}\cdot\boldsymbol{E}+\omega_c\tau\boldsymbol{v}\cdot\hat{\boldsymbol{B}}\times\boldsymbol{E}+(\omega_c\tau)^2(\hat{\boldsymbol{B}}\cdot\boldsymbol{E})\boldsymbol{v}\cdot\hat{\boldsymbol{B}}}{1+(\omega_c\tau)^2}\frac{\partial f_F}{\partial\epsilon}\mathrm{d}\boldsymbol{k}.\end{aligned} \tag{11.21}$$

According to the definition of conductivity, called in this case *magnetoconductivity*, this equation must be written in the form

$$\boldsymbol{j}=\sigma(\boldsymbol{B})\boldsymbol{E}.$$

From (11.21), we realize that now σ is a tensor (in fact the magnetic field breaks the spherical symmetry):

$$j_i=\sum_j\sigma_{ij}(\boldsymbol{B})E_j.$$

Using the property of the mixed vector product, the j-th component of the current density becomes

$$j_i = -\frac{2\mathrm{e}}{(2\pi)^3}\int v_i \mathrm{e}\tau \frac{v_j E_j + \omega_c\tau E_j(\boldsymbol{v}\times\hat{\boldsymbol{B}})_j + (\omega_c\tau)^2(\hat{B}_j E_j)(\boldsymbol{v}\cdot\hat{\boldsymbol{B}})}{1+(\omega_c\tau)^2}\frac{\partial f_F}{\partial\epsilon}\mathrm{d}\boldsymbol{k},$$

where the Einstein convention on the sum over repeated indices has been used. The magnetoconductivity tensor is then given by

$$\sigma_{ij} = \frac{-2\mathrm{e}^2}{(2\pi)^3}\int \frac{\partial f_F}{\partial\epsilon}\frac{\tau}{1+(\omega_c\tau)^2} v_i\left\{v_j + \omega_c\tau(\boldsymbol{v}\times\hat{B})_j + (\omega_c\tau)^2(\hat{B}_j)(\boldsymbol{v}\cdot\hat{\boldsymbol{B}})\right\}\mathrm{d}\boldsymbol{k}.$$

Defining the tensor product between two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$(\boldsymbol{A}\otimes\boldsymbol{B})_{ij} = A_i B_j,$$

we finally obtain the magnetoconductivity tensor:

$$\sigma(\boldsymbol{B}) = \frac{-2\mathrm{e}^2}{(2\pi)^3}\int \frac{\partial f_F}{\partial\epsilon}\frac{\tau}{1+(\omega_c\tau)^2}\boldsymbol{v}\otimes\left\{\boldsymbol{v}+\omega_c\tau(\boldsymbol{v}\times\hat{B}) + (\omega_c\tau)^2(\boldsymbol{v}\cdot\hat{\boldsymbol{B}})\hat{\boldsymbol{B}}\right\}\mathrm{d}\boldsymbol{k}. \tag{11.22}$$

Considering the symmetry properties of the integrand, we realize that the first term contains only diagonal elements, the second term contains only off-diagonal elements, and the third term contains elements connecting two directions along which $\boldsymbol{B}$ has nonzero components. If the z-axis is taken along $\boldsymbol{B}$, the magnetoconductivity takes the form

$$\sigma(\boldsymbol{B}\|z) = -\frac{2\mathrm{e}^2}{(2\pi)^3}\int \frac{\partial f_F}{\partial\epsilon}\frac{\tau}{1+(\omega_c\tau)^2}\begin{pmatrix} v_x^2 & -v_x^2\omega_c\tau & 0 \\ v_y^2\omega_c\tau & v_y^2 & 0 \\ 0 & 0 & v_z^2[1+(\omega_c\tau)^2]\end{pmatrix}\mathrm{d}\boldsymbol{k}. \tag{11.23}$$

It is easy to see that this tensor verifies the *Onsager relation* [255]

$$\sigma_{ij}(\boldsymbol{B}) = \sigma_{ji}(-\boldsymbol{B}). \tag{11.24}$$

Furthermore, if also $\boldsymbol{E}$ is parallel to the z direction, i.e., if electric and magnetic fields are parallel, the current is parallel to $\boldsymbol{E}$ and is given by the same quantity obtained in absence of magnetic field. This fact is usually expressed saying that *in a "simple semiconductor" no longitudinal magnetoresistance is present.*

If, instead, $\boldsymbol{E}$ is perpendicular to $\boldsymbol{B}$, say directed along y, the current has two components, one along $\boldsymbol{E}$ and one along x, i.e., perpendicular to both $\boldsymbol{E}$ and $\boldsymbol{B}$:

$$j_x = \sigma_{xy}E, \qquad j_y = \sigma_{yy}E,$$

with

$$\sigma_{xy} = \frac{2\mathrm{e}^2}{(2\pi)^3}\int \frac{\partial f_F}{\partial\epsilon}\frac{\omega_c\tau^2 v_x^2}{1+(\omega_c\tau)^2}\mathrm{d}\boldsymbol{k}$$

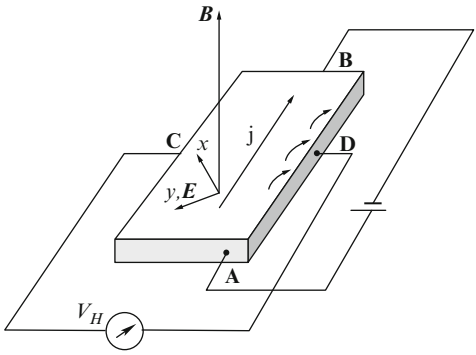


Fig. 11.4. Hall effect

and

$$\sigma_{yy} = -\frac{2\mathrm{e}^2}{(2\pi)^3}\int \frac{\partial f_F}{\partial \epsilon}\frac{\tau v_y^2}{1+(\omega_c\tau)^2}\mathrm{d}\boldsymbol{k}.$$

Since $\boldsymbol{B}\times\boldsymbol{E}$ is oriented along $-x$, we may write

$$\boldsymbol{j} = \sigma_E\boldsymbol{E} + \sigma_B\boldsymbol{B}\times\boldsymbol{E}, \tag{11.25}$$

where

$$\sigma_E = \sigma_{yy}, \qquad \sigma_B = -\frac{1}{B}\sigma_{xy}.$$

11.3.4 Hall Effect

The results obtained at the end of the previous section form the theoretical basis of the important Hall effect.[4] It consists in the appearance of a transverse voltage across a conducting bar carrying current in the presence of a perpendicular magnetic field. Figure 11.4 shows the geometry of the effect. A current is made to flow between the contacts A and B, while a magnetic

[4] Hall effect has repeatedly played a considerable role in the history of solid-state physics. It was discovered by the graduate student Edwin H. Hall in 1879 and it indicated that in metals the current was due to mobile negative charge much before the discovery of the electron. Later it became essential for the development of the idea of holes in semiconductors and a powerful tool for the characterization of semiconductor materials, in terms of sign and concentration of charge carriers. In 1980, a quantized Hall effect was discovered by K.V. Klitzing and coworkers in two-dimensional electron gases [244], where a quantized conductance appears in units of e^2/h, with such a precision that it is now used as a conductance standard. We shall discuss this effect in Chap. 21. But the Hall effect continued to offer surprises: in 1982 a fractional quantum Hall effect was discovered by D.C. Tsui, H.L. Störmer, and A.C. Gossard [452], whose explanation requires sophisticated many-body calculations.

field $\boldsymbol{B}$ is applied along the z direction, orthogonal to the conducting bar. A transverse voltage, called *Hall voltage* V_H is measured between the contacts in C and D.

Elementary Theory

We may assume that initially the field due to the applied voltage is parallel to the direction of the conducting sample, and carriers start moving along the same direction or the opposite one, depending on the sign of their charges. The Lorentz force, however, acting on the moving charges pushes them against the side of the bar. The charges accumulated in this way on the side generate a transverse Hall voltage, and the steady state is attained when the Hall field $\boldsymbol{E}_H$ balances exactly the Lorentz force. A crucial point in this effect is that the carriers are pushed toward the same side, independently of the sign of their charges. In fact, if this sign is changed, also the velocity of the carriers under the effect of the applied voltage is reversed, and the Lorentz force does not change. Thus, from the sign of the Hall voltage it is possible to determine the sign of the mobile charges.

If, in an elementary treatment, we assume that all carriers move with the drift velocity $\boldsymbol{v}_\mathrm{d}$, they are subject to the same Lorentz force, that must be counterbalanced by the force due to the Hall field:

$$q\boldsymbol{v}_\mathrm{d} \times \boldsymbol{B} = -q\boldsymbol{E}_H,$$

where this time the carrier charge is indicated with q to underline the possibility of positive or negative sign of its value. Since $\boldsymbol{B}$ is orthogonal to $\boldsymbol{v}_\mathrm{d}$, we have

$$E_H = v_\mathrm{d} B = \frac{jB}{qn}.$$

The Hall field is orthogonal to both $\boldsymbol{B}$ and $\boldsymbol{j}$, and the *Hall constant* R is defined by the relation

$$\boldsymbol{E}_H = R\ \boldsymbol{B} \times \boldsymbol{j}. \tag{11.26}$$

In the present simple case

$$R = \frac{1}{qn}. \tag{11.27}$$

Thus, the experimental measurement of R allows us to know the sign and the concentration of the carriers.

Kinetic Theory

In the elementary theory above, it has been assumed that all carriers move with the same velocity. We know, however, that there is a distribution of velocities, given by the distribution function f. To interpret the Hall effect in a more rigorous way, we must therefore use the kinetic theory developed in

the previous section, and in particular the result in (11.25). To obtain $\boldsymbol{E}$ from that equation, we use the same procedure applied to obtain $\boldsymbol{E}'$ from (11.19). First write (11.25) as

$$\boldsymbol{E} = \frac{1}{\sigma_E}\boldsymbol{j} - \frac{\sigma_B}{\sigma_E}\boldsymbol{B} \times \boldsymbol{E}.$$

Then substitute this equation into itself:

$$\boldsymbol{E} = \frac{1}{\sigma_E}\boldsymbol{j} - \frac{\sigma_B}{\sigma_E}\boldsymbol{B} \times \left[\frac{1}{\sigma_E}\boldsymbol{j} - \frac{\sigma_B}{\sigma_E}\boldsymbol{B} \times \boldsymbol{E}\right] = \frac{1}{\sigma_E}\boldsymbol{j} - \frac{\sigma_B}{\sigma_E^2}\boldsymbol{B} \times \boldsymbol{j} - \frac{\sigma_B^2}{\sigma_E^2}B^2\boldsymbol{E},$$

where the property of the triple vector product has been used. Thus,

$$\boldsymbol{E} = \frac{(1/\sigma_E)\boldsymbol{j} - (\sigma_B/\sigma_E^2)\boldsymbol{B} \times \boldsymbol{j}}{1 + (\sigma_B/\sigma_E)^2 B^2}.$$

In particular, the Hall component of the electric field, orthogonal to $\boldsymbol{j}$ is

$$\boldsymbol{E}_H = \boldsymbol{E}_\perp = -\frac{\sigma_B}{\sigma_E^2 + \sigma_B^2 B^2}\ \boldsymbol{B} \times \boldsymbol{j}, \tag{11.28}$$

while the component parallel to $\boldsymbol{j}$ is

$$\boldsymbol{E}_j = \boldsymbol{E}_\parallel = \frac{\sigma_E}{\sigma_E^2 + \sigma_B^2 B^2}\ \boldsymbol{j}.$$

If σ' is defined by the relation

$$\boldsymbol{j} = \sigma' \boldsymbol{E}_\parallel, \tag{11.29}$$

from (11.26), (11.28), and (11.29), we obtain

$$R = -\frac{\sigma_B}{\sigma_E^2 + \sigma_B^2 B^2}, \qquad \sigma' = \frac{\sigma_E^2 + \sigma_B^2 B^2}{\sigma_E}. \tag{11.30}$$

A simple relation holds between R and σ':

$$\sigma' = -\frac{\sigma_B}{\sigma_E R}. \tag{11.31}$$

Case of Constant τ

If the relaxation time may be approximated by a constant $\tau_\circ$, the magneto-conductivity tensor in (11.23) reduces to

$$\sigma(\boldsymbol{B}\|z) = \frac{\sigma_\circ}{1 + (\omega_c\tau_\circ)^2}\begin{pmatrix} 1 & -\omega_c\tau_\circ & 0 \\ \omega_c\tau_\circ & 1 & 0 \\ 0 & 0 & 1 + (\omega_c\tau_\circ)^2 \end{pmatrix},$$

where $\sigma_\circ$ is the conductivity in absence of magnetic field, given in this case by

$$\sigma_\circ = \frac{nq^2\tau_\circ}{m}.$$

In this approximation, the coefficients σ_E and σ_B, defined in (11.25), become

$$\sigma_E = \sigma_{yy} = \frac{\sigma_\circ}{1+(\omega_c\tau_\circ)^2}, \qquad \sigma_B = -\frac{1}{B}\sigma_{xy} = \frac{\omega_c\tau_\circ}{1+(\omega_c\tau_\circ)^2}\frac{\sigma_\circ}{B}. \tag{11.32}$$

Substitution into (11.30) leads, after a few simple algebraic steps, to

$$R = -\frac{\sigma_B}{\sigma_E^2+\sigma_B^2B^2} = -\frac{\omega_c\tau_\circ\frac{\sigma_\circ}{B}}{\sigma_\circ^2} = \frac{1}{nq}, \tag{11.33}$$

which is the same result (11.27) of the elementary theory. As mentioned above, this is the equation on which the application of the Hall effect for the characterization of a semiconductor material is based: R is determined experimentally from (11.26) with the measurements of $\boldsymbol{j}$, $\boldsymbol{B}$, and $\boldsymbol{E}_H$; then the sign of R yields the sign of the carrier charges, i.e., determines whether the semiconductor is of n-type or p-type, and its value determines the carrier concentration. It may be useful to underline the hypotheses that have been necessary to reach the above result: parabolic bands with spherical symmetry, weak electric field (linear-response theory), constant relaxation time, and nonquantizing magnetic fields.

In case of bipolar materials, where both positive and negative carriers are present, the results are somewhat more complicated and we refer the interested reader, for example, to [176] or [489].

From (11.31) and (11.32), we obtain σ' for a constant relaxation time:

$$\sigma' = -\frac{\frac{\omega_c\tau_\circ}{1+(\omega_c\tau_\circ)^2}\frac{\sigma_\circ}{B}}{\frac{\sigma_\circ}{1+(\omega_c\tau_\circ)^2}\frac{1}{nq}} = \sigma_\circ. \tag{11.34}$$

Again, no longitudinal magnetoresistance effect is present, with a different meaning with respect to the previous case, seen after (11.24).

Case of Weak Magnetic Field

A different approximation, alternative to the assumption of constant τ, is that of weak magnetic fields. It is preferable, since now τ is considered a function of energy, and the hypothesis of weak B may be controlled in the experimental conditions. Furthermore, the condition required by this approximation, that $\omega_c\tau \ll 1$, is always verified as long as quantization due to the magnetic field is avoided.

Following the same procedure used when only an electric field is present, assuming nondegenerate statistics and $\omega_c\tau \ll 1$, we obtain the following result:

$$\sigma(\boldsymbol{B}\|z) = \sigma_c \begin{pmatrix} 1 & -\omega_c \frac{\langle\tau^2\epsilon\rangle}{\langle\tau\epsilon\rangle} & 0 \\ \omega_c \frac{\langle\tau^2\epsilon\rangle}{\langle\tau\epsilon\rangle} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where, now,

$$\sigma_c = \frac{nq^2}{m}\frac{\langle\tau\epsilon\rangle}{\langle\epsilon\rangle}$$

is the conductivity in absence of magnetic field. It is easy to find, in the present case,

$$\sigma_E = \sigma_{yy} = \sigma_c, \qquad \sigma_B = -\frac{1}{B}\sigma_{xy} = \frac{\omega_c\sigma_c}{B}\frac{\langle\tau^2\epsilon\rangle}{\langle\tau\epsilon\rangle}, \tag{11.35}$$

$$R = -\frac{\sigma_B}{\sigma_E^2 + \sigma_B^2 B^2} = -\frac{\omega_c\sigma_\circ}{B}\frac{\langle\tau^2\epsilon\rangle}{\sigma_\circ^2\langle\tau\epsilon\rangle} = \frac{1}{nq}\frac{\langle\epsilon\rangle\langle\tau^2\epsilon\rangle}{\langle\tau\epsilon\rangle^2}, \tag{11.36}$$

$$\sigma' = \sigma_c.$$

Since $\sigma_E = \sigma_c$, as indicated in (11.35), also in this case no magnetoresistance effect is present. Such effect may arise, instead, when bands do not have spherical symmetry [489].

Hall Mobility and Hall Factor

In absence of magnetic field, conductivity, mobility, and concentration are related by $\sigma = nq\mu$. Once the conductivity is independently measured, the mobility can be determined if the concentration n is obtained with a Hall measurement. This operation, however, requires a correct expression of the Hall constant R. If the expression for R given by (11.33) for constant τ is used, the mobility is given by

$$\mu_H = R\sigma, \tag{11.37}$$

and this expression is called *Hall mobility*. However, if the energy dependence of the relaxation time is considered, and the expression (11.36) for R is used, the mobility, called in this case *drift mobility*, is given by

$$\mu_\mathrm{d} = \frac{1}{r_H}\mu_H, \qquad r_H = \frac{\langle\epsilon\rangle\langle\tau^2\epsilon\rangle}{\langle\tau\epsilon\rangle^2}. \tag{11.38}$$

The coefficient r_H is called *Hall factor*, or, sometimes *scattering factor*. It depends upon the scattering mechanisms that dominate the electron transport. The dependence of the relaxation time upon energy will be analyzed in Sect. 11.5 below. It yields a characteristic temperature dependence of the conductivity. Thus, from the analysis of the temperature dependence of the

conductivity an indication of the dominant scattering mechanisms can be obtained. The knowledge of the Hall factor may then be used for a better determination of the charge carrier concentration. In particular, for mobilities controlled by acoustic phonons with deformation potential interaction and by ionized impurities, the Hall factors result to be [176]

$$r_H^{(ac)} = 1.18 \ , \quad r_H^{(ii)} = 1.93,$$

respectively.

11.4 High-Magnetic-Field Effects

Several interesting effects exist that occur at high magnetic-field intensities and low temperatures. For reasons of space, we are forced to simply mention them, without any detail.

The previous sections dealt with electron transport in terms of semiclassical dynamics, which can be reasonably applied as long as the collisions interrupt the free electron flights before full cyclotron orbits are completed, i.e., as long as $\omega_c \tau \ll 1$. This product, however, may become comparable with or larger than unity when a sufficiently high magnetic field (typically several Tesla) produces a large ω_c in a very pure material at low temperatures, where a long relaxation time is present. In such conditions, the closed cyclotron orbits involve energy quantization of the electron states in the plane perpendicular to $\boldsymbol{B}$ (see Appendix D). The energy eigenvalues, called *Landau levels*, are separated by the amount $\hbar\omega_c$, so that the energy bands are split into subbands separated by this quantum of energy. Furthermore, the density of states increases linearly with the magnetic field, and the Fermi level oscillates as a function of B. These oscillations are shown, in connection with the quantum Hall effect, in Fig. 21.9. In a degenerate electron gas when, at increasing B, the bottom of a subband crosses the Fermi energy the density of states available for electron scattering has a large discontinuity. Since many physical phenomena depend on the density of states available to the electrons at the Fermi level, many effects show oscillations as functions of the magnetic-field intensity. These effects are particularly evident in metals, but are present also in degenerate semiconductors. In particular, the *Shubnikov-de Haas effect* consists of oscillations of the magnetoresistance, while the *de Haas-van Alphen effect* consists of oscillations of the magnetic susceptibility and other quantities, such as specific heat and sound attenuation. All these effects show a periodicity as a function of $1/B$, and the period of the oscillation yields information on the Fermi surface. In particular, extremal cross-sectional areas of the Fermi surface, orthogonal to the magnetic field, are involved in these effects. When more than one of such extremal cross sections exist, multiple periodicities appear. Such measurements are currently used for the experimental determination of the Fermi surfaces of conducting materials [18].

Another interesting phenomenon related to Landau quantization at high magnetic fields is the *magneto-phonon resonance*. When the energy separation $\hbar\omega_c$ between two Landau levels equals the energy of optical phonons, resonant interaction occurs with an enhancement of electron scattering and, therefore, a depression of the conductivity. Other types of magneto-phonon resonances have been theoretically predicted and observed experimentally, related to different types of electron states and/or phonon branches. A brief review of such phenomena can be found in [176].

11.5 Evaluation of the Momentum Relaxation Times

In the previous sections we have developed the theory of linear transport under the hypothesis that the collision integral in the BE could be conveniently represented by a relaxation time, function of the electron energy, as indicated in (11.4). Now it is time to find what properties the transition rates must have for this hypothesis to be verified, and, in such cases, how it is possible to evaluate the relaxation time starting from the known transition rates.

Let us consider the collision integral in the linearized BE, as given in (11.2). If the detailed balance (10.9) is used, the collision integral becomes

$$\frac{V}{(2\pi)^3}\int \Big\{ f_1(\boldsymbol{k}')P(\boldsymbol{k},\boldsymbol{k}')\left[\mathrm{e}^{\beta(\epsilon'-\epsilon)}[1-f_F(\boldsymbol{k})]+f_F(\boldsymbol{k})\right] - f_1(\boldsymbol{k})P(\boldsymbol{k},\boldsymbol{k}')\left[[1-f_F(\boldsymbol{k}')]+\mathrm{e}^{\beta(\epsilon'-\epsilon)}f_F(\boldsymbol{k}')\right]\Big\}\,\mathrm{d}\boldsymbol{k}'. \tag{11.39}$$

We shall now use this expression to show that a relaxation time can be defined for velocity-randomizing collisions and for elastic collisions.

11.5.1 Relaxation Time for Velocity-Randomizing Collisions

Collisions are said to be *velocity randomizing* when a final state of the collision is equally probable with respect to the final state with opposite velocity. In Sect. 8.2, we have seen that opposite velocities correspond, for time reversal symmetry, to opposite wavevectors, so that the above condition corresponds to

$$P(\boldsymbol{k},\boldsymbol{k}') = P(\boldsymbol{k},-\boldsymbol{k}').$$

In particular, this is true for isotropic collisions, with many examples seen in Chap. 9, where the scattering rate depends upon the energy of the final state and not upon its direction. The two concepts of velocity-randomizing collisions and of isotropic collisions, however, should not be confused.

Let us now remember that $f_1(\boldsymbol{k})$ in (11.39) is proportional to the electric field, and changes sign if we change $\boldsymbol{E}$ into $-\boldsymbol{E}$. This operation must be equivalent to change sign to $\boldsymbol{k}$. The presence of a magnetic field does not interfere, since it remains unchanged with an inversion operation. Thus,

$$f_1(\boldsymbol{k}) = -f_1(-\boldsymbol{k}).$$

From the above considerations, it results that the first integral in (11.39) vanishes. In fact changing sign to $\boldsymbol{k}'$, $f_1(\boldsymbol{k}')$ changes sign while everything else remains unaltered; the integrand is odd, and the integral is zero. The second integral has the form required for the definition of a relaxation time:

$$\left.\frac{\partial f}{\partial t}\right|_{\text{coll}} = -\frac{f_1}{\tau},$$

with

$$\frac{1}{\tau} = \frac{1}{\tau(\boldsymbol{k})} = \frac{V}{(2\pi)^3}\int P(\boldsymbol{k},\boldsymbol{k}')\left[1 - f_F(\boldsymbol{k}')\left(\mathrm{e}^{\beta(\epsilon'-\epsilon)} - 1\right)\right]\mathrm{d}\boldsymbol{k}'. \tag{11.40}$$

In case of nondegenerate statistics, when $f_F \ll 1$, the above reduces to

$$\boxed{\frac{1}{\tau(\boldsymbol{k})} = \frac{V}{(2\pi)^3}\int P(\boldsymbol{k},\boldsymbol{k}')\mathrm{d}\boldsymbol{k}'} \tag{11.41}$$

This expression has a very simple physical interpretation: when the scattering probability is the same for opposite velocities of the state after scattering, at each collision the electron momentum is completely dissipated, in average, and the momentum relaxation time coincides with the inverse scattering rate. We have seen that most of the times the integrated scattering rates depend upon the wavevector only through the energy. Thus, the relaxation time in (11.41) is actually a function of $\epsilon(\boldsymbol{k})$.

11.5.2 Relaxation Time for Elastic Collisions

If the collisions are elastic, $\epsilon(\boldsymbol{k}') = \epsilon(\boldsymbol{k})$, and (11.39) becomes

$$\frac{V}{(2\pi)^3}\int P(\boldsymbol{k},\boldsymbol{k}')\left\{f_1(\boldsymbol{k}') - f_1(\boldsymbol{k})\right\}\mathrm{d}\boldsymbol{k}'.$$

This collision integral moves electrons only within a surface of constant energy. For spherical bands,[5] this means that only electrons within a spherical surface of radius k are involved. The transition rates may be written as

$$P(\boldsymbol{k},\boldsymbol{k}') = P(k,\theta),$$

where θ is the angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$. Furthermore, in absence of magnetic field,[6] the distribution function can depend only upon k and the angle χ

[5] The argument has been extended in [185] to ellipsoidal valleys, where a tensor relaxation time is introduced.

[6] If a magnetic field is present, this argument fails. The calculations of the previous sections indicate that for each energy surface, the distribution is the same as that obtained by an effective field E'. This result, however, is obtained within the relaxation time approximation, so that the argument would be somewhat circular. We can therefore quote Ziman [489], who says *Almost all work on magnetic effects has been based upon the hypothesis, whether justifiable or not, that* $\partial f/\partial t|_{\text{scat}} = -f_1/\tau$.

between $\boldsymbol{k}$ and the electric field (see Fig. 11.5). The linear term of the BE can be written as

$$\dot{\boldsymbol{k}} \cdot \nabla_{\boldsymbol{k}} f_F(\boldsymbol{k}) = (-\mathrm{e}) \frac{\partial f_F}{\partial \epsilon} \boldsymbol{E} \cdot \boldsymbol{v}(\boldsymbol{k})$$

$$= (-\mathrm{e}) \frac{\partial f_F}{\partial \epsilon} E v(k) \cos\chi = \frac{V}{(2\pi)^3} \int P(k,\theta) \left\{ f_1(k,\chi') - f_1(k,\chi) \right\} \mathrm{d}k \mathrm{d}\Omega. \tag{11.42}$$

In this equation, the independent variable is only the direction of $\boldsymbol{k}$ while its modulus, or the energy, plays simply the role of a parameter. We shall now prove that the equation is solved by a function $f_1(k,\chi)$ proportional to $\cos\chi$:

$$f_1(k,\chi) = g(k) \cos\chi.$$

In fact, considering only the angle variables, the r.h.s. of (11.42) becomes proportional to

$$\int P(k,\theta) g(k) [\cos\chi' - \cos\chi] \mathrm{d}\Omega' = \cos\chi \int P(k,\theta) g(k) \left[\frac{\cos\chi'}{\cos\chi} - 1 \right] \mathrm{d}\Omega'. \tag{11.43}$$

Following [291], let us consider a frame of reference with $\boldsymbol{k}$ along z, the polar axis for the angular integration; y is taken in the plane of $\boldsymbol{k}$ and $\boldsymbol{E}$. Then, with reference to Fig. 11.5, we have

$$\boldsymbol{k}' \cdot \boldsymbol{E} = k' E \cos\chi' = k'_x E_x + k'_y E_y + k'_z E_z$$

$$= 0 + k' \sin\theta \sin\phi E \sin\chi + k' \cos\theta E \cos\chi,$$

or

$$\frac{\cos\chi'}{\cos\chi} = \sin\theta \sin\phi \tan\chi + \cos\theta.$$

When inserted in the r.h.s of (11.43), the first term integrates to zero because of the factor $\sin\phi$. The collision integral becomes

$$\frac{V}{(2\pi)^3} g(k) \cos\chi \int P(k,\theta) \left\{ \cos\theta - 1 \right\} \mathrm{d}\boldsymbol{k}'.$$

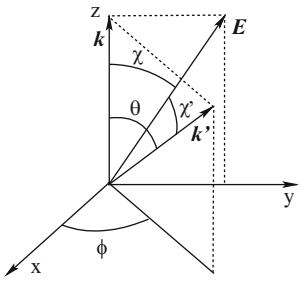


Fig. 11.5. For the evaluation of the relaxation time in case of elastic collisions, see text

The BE (11.42) is thus satisfied, since both sides have the same angular dependence. Furthermore, the collision integral has the form requested by the definition of the momentum relaxation time, which results to be given by

$$\boxed{\frac{1}{\tau(k)} = \frac{V}{(2\pi)^3}\int P(\boldsymbol{k},\boldsymbol{k}')(1-\cos\theta)\mathrm{d}\boldsymbol{k}'} \tag{11.44}$$

Also this expression, as in the case of (11.41), has a direct physical meaning: if the scattering rate is not velocity randomizing, and the modulus of the momentum remains unchanged, only a fraction $(1-\cos\theta)$ of the momentum is lost in the collision, being θ the scattering angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$, and the momentum relaxation time is the average of the collision rate weighted by this fraction. As in the previous case, most of the times the relaxation time results to be a function of the energy $\epsilon(\boldsymbol{k})$.

11.6 Mobilities

At this point, we are in a condition to evaluate the mobilities due to the various scattering mechanisms. We have seen in Chap. 9 that acoustic scattering can be treated, at least at room temperature, as an elastic process; optical and intervalley scattering mechanisms with deformation potential interactions are isotropic; impurity scattering is elastic. Thus all these scattering processes, i.e., all the main ones except polar optical phonons, admit a momentum relaxation time, and simple theoretical evaluations of the corresponding mobilities are at hand. In the following, we shall accomplish this project.

11.6.1 Acoustic–Phonon Scattering, Deformation Potential, Elastic

The scattering rate to use is given in (9.20), which includes both absorption and emission in the equipartition approximation. It results to be not only elastic, but also isotropic, so that the inverse momentum relaxation time coincides with the integrated scattering rate given in (9.21)

$$\frac{1}{\tau_{ae}^{(d)}(\epsilon)} = \frac{\sqrt{2}m^{3/2}K_\mathrm{B}TE_1^2}{\pi\hbar^4\rho v_l^2}\sqrt{\epsilon}. \tag{11.45}$$

Note that since this inverse relaxation time is proportional to the velocity of the electron, its mean free path is constant and is given by

$$l_{ae}^{(d)} = v\tau_{ae}^{(d)} = \frac{\pi\hbar^4\rho v_l^2}{m^2K_\mathrm{B}TE_1^2}.$$

The corresponding mobility is given by (11.10), and the integration yields

$$\mu_{ae}^{(d)} = (-\mathrm{e})\frac{4\sqrt{\pi}\hbar^4\rho v_l^2}{3\sqrt{2}m^{5/2}E_1^2}(K_\mathrm{B}T)^{-3/2}, \tag{11.46}$$

proportional to $T^{-3/2}$, as shown in Fig. 11.6.

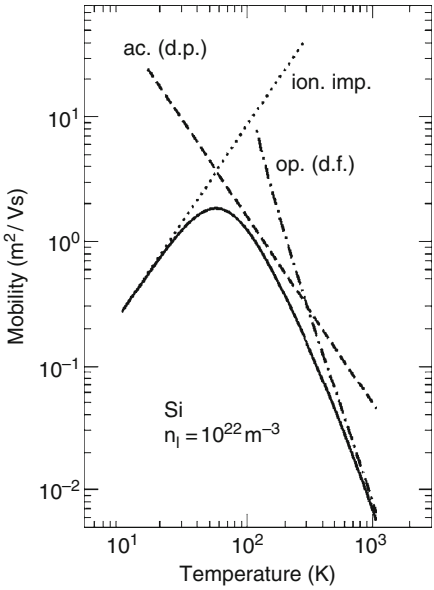


Fig. 11.6. Electron mobilities, as given by (11.46), (11.47), and (11.49), in a material with silicon-like physical constants, as given in Table 11.1, combined with Matthiessen rule. At room temperature, optical and acoustic phonons contribute with similar weight. At lower temperatures, acoustic phonons are more effective than optical phonons, and the opposite is true at higher temperatures. Ionized impurities may become important at low temperatures, depending on their concentration, here assumed to be $10^{22}\,\mathrm{m}^{-3}$. The same concentration has been assumed for screening

11.6.2 Optical–Phonon Scattering, Deformation Potential

Also in this case the scattering is isotropic, and the inverse momentum relaxation time coincides with the integrated scattering rate, given, for the simple spherical and parabolic band, by (9.31). Summing up absorption and emission processes, we have

$$\frac{1}{\tau_{op}^{(d)}(\epsilon)} = \frac{m^{3/2}(D_t K)^2}{\sqrt{2}\pi\hbar^3\rho\omega_{op}} \left[N_{op}\sqrt{\epsilon+\hbar\omega_{op}} + (N_{op}+1)\Re\sqrt{\epsilon-\hbar\omega_{op}}\right],$$

where the real-part operator has been introduced to take into account that emission processes are to be considered only for electron energies higher than the phonon energy. To introduce this result into the expression (11.10) for the mobility, we use the variable $x = \epsilon/K_\mathrm{B}T$, $x_\circ = \hbar\omega_{op}/K_\mathrm{B}T$:

$$\tau_{op}^{(d)}(x) = \frac{\sqrt{2}\pi\hbar^2\rho x_\circ\sqrt{K_\mathrm{B}T}}{m^{3/2}(D_t K)^2} \frac{1}{[N_{op}\sqrt{x+x_\circ} + (N_{op}+1)\Re\sqrt{x-x_\circ}]}.$$

Substitution into (11.10), after simple manipulation, yields

$$\mu_{op}^{(d)} = (-\mathrm{e})\frac{4\sqrt{2\pi}\hbar^2\rho\sqrt{\hbar\omega_{op}}}{3m^{5/2}(D_t K)^2} f(x_\circ), \tag{11.47}$$

where

$$f(x_\circ) = (\mathrm{e}^{x_\circ} - 1)\int_0^\infty x^{\frac{3}{2}}\mathrm{e}^{-x}\left[\sqrt{x/x_\circ + 1} + \mathrm{e}^{x_\circ}\Re\sqrt{x/x_\circ - 1}\right]^{-1}\mathrm{d}x.$$

The integration does not yield a simple formula. A much stronger dependence upon temperature is found by numerical integration, as shown in Fig. 11.6.

11.6.3 Ionized–Impurity Scattering

Ionized-impurity scattering is elastic and therefore allows the definition of a relaxation time. Since it is not velocity randomizing, but peaked in the forward direction, for the evaluation of the relaxation time we must use the form in (11.44). If we insert the scattering rate in (9.52), with unit overlap factor $\mathcal{G}$ and unit charge Z, into (11.44), we obtain

$$\frac{1}{\tau_i(\epsilon)} = \frac{V}{(2\pi)^3}\int\frac{2\pi}{\hbar}n_I\frac{\mathrm{e}^4}{\varepsilon^2 V}\frac{1}{[q_\circ^2 + q^2]^2}\delta(\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k})(1 - \cos\theta)\mathrm{d}\boldsymbol{k}'.$$

In what follows we shall adopt BH approach. If CW approach is used, the result is very similar. A comparison between the two resulting momentum relaxation times can be found, for example, in [402]. The integration can be performed again in polar coordinates with $\boldsymbol{k}$ as polar axis, and the δ function is used for the integration over k'. With the substitution $x = q_\circ^2 + 2k^2(1 - \cos\theta)$, we obtain, after straightforward integration,

$$\frac{1}{\tau_i(\epsilon)} = \frac{n_I\mathrm{e}^4}{16\sqrt{2}\pi\varepsilon^2 m^{1/2}}\epsilon^{-\frac{3}{2}}\left[\log(1 + 4\epsilon/\epsilon_\circ) - \frac{4\epsilon/\epsilon_\circ}{1 + 4\epsilon/\epsilon_\circ}\right], \tag{11.48}$$

where $\epsilon_\circ = \hbar^2 q_\circ^2/2m$, as in (9.55). As before, the mobility controlled by ionized-impurity scattering is obtained inserting this relaxation time into (11.10). With dependence $\epsilon^{-\frac{3}{2}}$ in the prefactor of (11.48) and the $x^{\frac{3}{2}}$ in (11.10), the integrand contains the function $x^3\mathrm{e}^{-x}$, strongly peaked at $x = 3$. Thus, the factor in square bracket in (11.48) can be approximated by the constant

$$F_{\mathrm{BH}} = \log(1 + 12K_{\mathrm{B}}T/\epsilon_\circ) - \frac{12K_{\mathrm{B}}T/\epsilon_\circ}{1 + 12K_{\mathrm{B}}T/\epsilon_\circ},$$

and the BH mobility becomes, after integration,

$$\mu_i^{(\mathrm{BH})} = -\frac{64}{\sqrt{m}}\frac{\sqrt{\pi}\varepsilon^2}{n_I\mathrm{e}^3 F_{\mathrm{BH}}}(2K_{\mathrm{B}}T)^{3/2}, \tag{11.49}$$

Table 11.1. Parameters for "simple semiconductor models" of electrons in silicon and gallium arsenide. Data have been taken from [20, 363]

Quantity	Si	GaAs
Density ρ	$2330\,\mathrm{Kg/m^3}$	$5310\,\mathrm{Kg/m^3}$
Sound velocity v_s	9.0×10^3 m/s	5.2×10^3 m/s
Diel. const. $\varepsilon(0)$	11.9	13.5
High freq. diel. const. $\varepsilon(\infty)$	–	11.6
Effective mass $m/m_\circ$	0.295	0.067
Opt. Phonon eq. temp. T_{op}	450 K	418 K
Ac. def. pot. E_1	9.0 eV	7.0 eV
Opt. def. pot $D_t K$	8.0×10^{10} eV/m	3.0×10^{10} eV/m
Piezoel. constant p	–	$0.16\,\mathrm{C/m^2}$

increasing with temperature as $T^{3/2}$, as shown in Fig. 11.6, as a consequence of the decreasing effectiveness of Coulomb scattering as the electron energy increases. Remember that the minus sign comes from the fact that the mobility is calculated for a carrier with charge (−e).

As an example, Fig. 11.6 shows the mobilities due to acoustic phonons, as in (11.46), optical phonons as in (11.47), and impurity scattering, as in (11.49). Furthermore, the total mobility is shown, as given by Matthiessen rule (11.13). The physical parameters used in this figure are those of a "simple semiconductor" modeled on silicon, as given in Table 11.1. The resulting mobility is not very different from the experimental one, and from that evaluated with a much more complete model, shown in Chap. 15 (see Fig. 15.1), since for the low-field mobility the complete band structure does not play a crucial role in a cubic semiconductor. It may be seen that at room temperature optical and acoustic phonons contribute with similar weight. At lower temperatures, acoustic phonons are more effective than optical phonons, and the opposite is true at higher temperatures. Ionized impurities may become important at low temperatures, depending on their concentration.

11.6.4 Acoustic–Phonon Scattering, Piezoelectric, Elastic

Assuming the elastic approximation with a scattering rate given by (9.43), piezoelectric scattering yields a relaxation time given by

$$\frac{1}{\tau_{ae}^{(p)}(\epsilon)} = \frac{V}{(2\pi)^3}\int \frac{2\pi p^2 \mathrm{e}^2 K_\mathrm{B} T}{\varepsilon^2 \hbar \rho V v_s^2}\left(\frac{q}{q_\circ^2+q^2}\right)^2 \delta[\epsilon(\boldsymbol{k}')-\epsilon(\boldsymbol{k})](1-\cos\theta)\mathrm{d}\boldsymbol{k}'.$$

Now we proceed as for the previous case, and after straightforward integration, the result is

$$\frac{1}{\tau_{ae}^{(p)}(\epsilon)} = \frac{p^2 \mathrm{e}^2 \sqrt{m} K_\mathrm{B} T}{2\sqrt{2}\pi\varepsilon^2\hbar^2\rho v_s^2}\epsilon^{-1/2}\left[1-\frac{\epsilon_\circ}{2\epsilon}\log\left(1+4\frac{\epsilon}{\epsilon_\circ}\right)+\frac{1}{1+4\epsilon/\epsilon_\circ}\right]. \quad (11.50)$$

With this expression for the relaxation time we may now evaluate, as before, the mobility controlled by piezoelectric scattering. The term inside the squared bracket is due to screening. Often it is neglected all together. If we want to take it into account, we may repeat the consideration made in the case of the ionized impurities: with dependence $\epsilon^{-1/2}$ in (11.50) and the $x^{\frac{3}{2}}$ in (11.10), the integrand for the mobility contains the function $x^2 \mathrm{e}^{-x}$, peaked at $x = 2$. Thus, the factor in square bracket in (11.50) can be approximated by the constant

$$F_{PE} = \left[1 - \frac{\epsilon_\circ}{4K_\mathrm{B}T} \log\left(1 + \frac{8K_\mathrm{B}T}{\epsilon_\circ}\right) + \frac{1}{1 + 8K_\mathrm{B}T/\epsilon_\circ}\right],$$

and the mobility due to piezoelectric scattering, after integration, results to be

$$\mu_{ae}^{(p)} = -\frac{16\sqrt{2\pi}}{3} \frac{\varepsilon^2 \hbar^2 \rho v_s^2}{p^2 \mathrm{e} m^{3/2} \sqrt{K_\mathrm{B}T} F_{PE}}, \tag{11.51}$$

shown in Fig. 11.7.

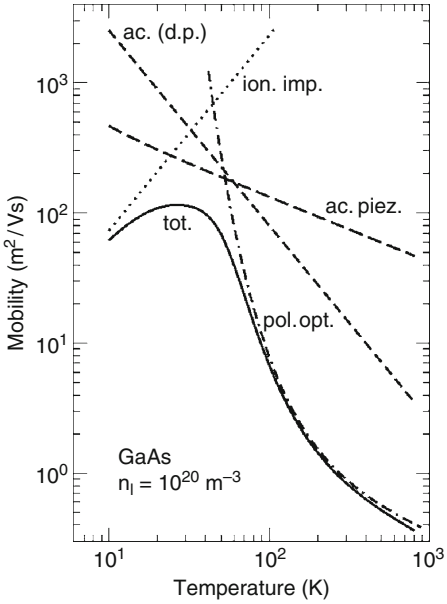


Fig. 11.7. Electron mobilities, as given by (11.46), (11.49), (11.51), and (11.52), in a material with GaAs-like physical constants, as given in Table 11.1, combined with Matthiessen rule. At room temperature, the mobility is controlled by polar optical phonons. At low temperatures, piezoelectric acoustic phonons are more effective. Ionized impurities may become important at low temperatures, depending on their concentration, here assumed to be $10^{20}\,\mathrm{m}^{-3}$. The same concentration has been assumed for screening

11.6.5 Optical–Phonon Scattering, Polar Interaction

The transition rates for electron scattering by optical phonons are given in (9.47). In this case, the scattering is neither elastic nor velocity randomizing. In fact, the electrostatic nature of the interaction is such that forward scattering is favored. Strictly speaking, therefore, a momentum relaxation time does not exist. Various approaches have been used to overcome this difficulty, using the variational principle mentioned at the end of Sect. 11.1 [129,193], or approximate solutions of the BE [152,375], or approximate relaxation times [183].

A manageable expression for a mobility due to polar optical scattering has been given by Stratton [427] and rederived by Conwell [107] using the polar optical–phonon scattering rates in a drifted Maxwellian distribution as approximate solution of the BE (see also [176]):

$$\mu_{\text{pop}} = \frac{3(2\pi\hbar\omega_{LO})^{1/2}}{4m^{1/2}E_\circ n(\omega_{LO})x_\circ^{3/2}\mathrm{e}^{x_\circ/2}K_1(x_\circ/2)}, \tag{11.52}$$

where

$$E_\circ = \frac{me\hbar\omega_{LO}}{4\pi\hbar^2}\left[\frac{1}{\varepsilon(\infty)} - \frac{1}{\varepsilon(0)}\right], \quad n(\omega_{op}) = \frac{1}{\mathrm{e}^{x_\circ}-1}, \quad x_\circ = \frac{\hbar\omega_{op}}{K_{\text{B}}T},$$

and $K_1(t)$ is the Bessel function

$$K_1(t) = t\int_1^\infty \sqrt{z^2-1}\mathrm{e}^{-tz}\mathrm{d}z.$$

Figure 11.7 shows the mobilities due to acoustic phonons with deformation potential interaction as in (11.46), piezoelectric phonons, as in (11.51), polar optical phonons as in (11.52), and impurity scattering, as in (11.49). Furthermore, the total mobility is shown, as given by Matthiessen rule (11.13). The physical parameters used in this figure are those of a "simple semiconductor", modeled on gallium arsenide, as given in Table 11.1. The resulting mobility is not very different from the experimental one, and from that evaluated with a more complete model, shown in Chap. 15 (see Fig. 15.14). It may be seen that at room temperature the mobility is essentially controlled by polar optical phonons.

12

Diffusion, Fluctuations, and Noise

12.1 Fick Laws

In Sect. 10.4, we have seen that the term of the BE containing the space gradient of the distribution function produces a diffusion current that tends to eliminate space inhomogeneities of the particle concentration. Here, we will analyze this important phenomenon in more detail. We shall consider particles ignoring their charges, since diffusion is not specifically related to their electric charges.

Let us first introduce Fick law, the fundamental equation of diffusion, in a phenomenological way: if the concentration is constant, no diffusion can occur; if the concentration is not constant, and a small gradient is present, we may assume a current that in the linear-response regime will be proportional to the gradient:

$$\boxed{\boldsymbol{j}_D = -D\nabla n} \tag{12.1}$$

where D is the *diffusion coefficient*. This is the *first Fick law*, obtained also in the analysis of the first moment of BE (see Sect. 10.4.3). As a constant of proportionality between two vectors, D, in general, is a second-rank tensor, but for isotropic or cubic materials it reduces to a scalar quantity.

If we couple the continuity equation (10.16) to the above (12.1), we obtain

$$\frac{\partial n}{\partial t} = \nabla \cdot [D\nabla n],$$

and, if we assume D independent of position, we get the *second Fick law*:

$$\boxed{\frac{\partial n}{\partial t} = D\nabla^2 n} \tag{12.2}$$

We shall now give an elementary derivation of the first Fick law in a relaxation-time approximation. Let us analyze a situation in which the concentration varies along the z direction and consider the plane orthogonal to

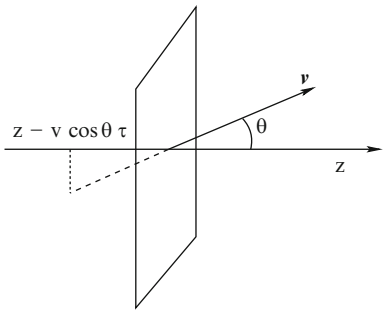


Fig. 12.1. For the derivation of Fick law, see text. The current at a given position is due to particles coming from different positions, where the concentration is different

the z axis crossing this axis at point z, as shown in Fig. 12.1. A particle that at a given time crosses this plane comes from a different position, with z-coordinate given by $z - v\cos\theta\tau$, if v is the velocity of the particle, θ the angle between the particle velocity and the z axis, and τ the relaxation time. In this position the concentration n is different than in the plane under consideration, and the net current can be written as the following average:

$$j = \langle v\cos\theta n(z - v\cos\theta\tau)\rangle = \left\langle v\cos\theta\left[n(z) - \frac{\partial n}{\partial z}v\cos\theta\tau\right]\right\rangle.$$

The average of the first term vanishes, as $\langle\cos\theta\rangle = 0$, and the second term yields

$$j = -\frac{\partial n}{\partial z}\langle v^2\rangle\langle\cos^2\theta\rangle\tau. \tag{12.3}$$

If we define

$$\overline{v} = \sqrt{\langle v^2\rangle}$$

and observe that

$$\langle\cos^2\theta\rangle = \frac{\int_0^\pi \cos^2\theta\sin\theta\mathrm{d}\theta}{\int_0^\pi \sin\theta\mathrm{d}\theta} = \frac{1}{3},$$

we see that (12.3) yields the first Fick law, with

$$D = \frac{1}{3}\overline{v}^2\tau = \frac{1}{3}\overline{v}l, \tag{12.4}$$

where $l = \overline{v}\tau$ is the particle mean free path. This expression for the diffusion coefficient can be compared with the more rigorous expression in (10.29). The latter reduces to the present one in the constant relaxation-time approximation with Maxwellian velocity fluctuations.

12.2 Einstein Relation

An important relation, known as *Einstein relation*, which connects diffusion and mobility, follows from the expression just found for the diffusion coefficient. Taking into account that

$$\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}K_{\mathrm{B}}T,$$

we obtain

$$D = \frac{K_{\mathrm{B}}T}{m}\tau,$$

and remembering the simple expression (11.11) for the mobility,

$$\boxed{D = \frac{\mu K_{\mathrm{B}}T}{q}} \tag{12.5}$$

This is the Einstein relation. It has a much more general validity than that of the present derivation, and we shall shortly see a somewhat more general derivation. Furthermore, this relation is a particular case of a general theorem of linear-response theory, call *fluctuation-dissipation theorem*, which connects fluctuation phenomena, in our case diffusion, to dissipation, in our case conductivity. The quantum version of this theorem is known as Kubo formula, and it will be discussed in Chap. 16.

In simple physical terms, we may understand fluctuation-dissipation theorem by observing that fluctuations are generated by thermal excitations and are damped by friction, and friction determines also the dissipation activated by the conductivity. Thus friction, represented by the inverse mean free path, is the key phenomenon at the basis of both damping of fluctuations and linear response, which dissipates the input power necessary to generate and maintain such response.

The standard macroscopic derivation of the Einstein relation considers a constant, homogeneous, force along, say, the x direction, acting on particles inside an isolated sample, without contacts that can close the circuit. Particles concentrate on one edge of the sample, pushed by the force, and tend to diffuse toward the interior of the sample, as shown in Fig. 12.2. In steady-state conditions, the diffusion current $-D\partial n/\partial x$, due to the concentration gradient, compensates the current $n\mu E$ due to the applied field:

$$n\mu E = D\frac{\partial n}{\partial x}. \tag{12.6}$$

The potential energy of the particles due to the force $F = qE$ is given by

$$V(x) = -Fx = -qEx.$$

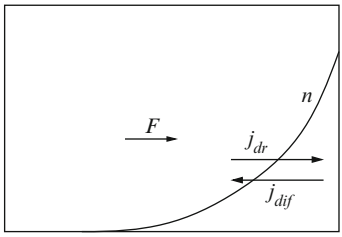


Fig. 12.2. For the derivation of Einstein relation. In steady state, in an open circuit with an applied force, the diffusion current and the drift current cancel each other

At equilibrium, the concentration is proportional to the Boltzmann factor

$$n \propto \mathrm{e}^{qEx/K_\mathrm{B}T},$$

where it has been assumed that the particles do not interact and obey classical statistics. Equation (12.6) becomes

$$\mathrm{e}^{qEx/K_\mathrm{B}T}\mu E = D\mathrm{e}^{qEx/K_\mathrm{B}T}\frac{qE}{K_\mathrm{B}T},$$

from which the Einstein relation (12.5) follows immediately.

12.3 Drift-Diffusion Equation and the Gaussian Solution

Combining drift and diffusion in a one-dimensional system, the current is given by the drift-diffusion equation (cf. 10.30)

$$j = n\mu E - D\frac{\partial n}{\partial x}. \tag{12.7}$$

Taking the derivative with respect to x and using the continuity equation, we obtain the differential equation for the concentration $n(x,t)$:

$$\frac{\partial n}{\partial t} = -\mu E\frac{\partial n}{\partial x} + D\frac{\partial^2 n}{\partial x^2}. \tag{12.8}$$

A solution of particular interest of such equation is provided by a Gaussian distribution which drifts with mean velocity $v_\mathrm{d} = \mu E$ and, at the same time, spreads by diffusion:

$$n(x,t) = \frac{N}{\sqrt{4\pi Dt}}\,\mathrm{e}^{-(x-v_\mathrm{d}t)^2/4Dt}. \tag{12.9}$$

We leave to the reader to verify, with simple straightforward calculations, that this is actually a solution of (12.8). The solution in (12.9) is normalized to the

total number of particles N, as can be easily verified through the evaluation of the normalization integral, with the substitution $(x - v_\mathrm{d}t)/(\sqrt{4Dt}) = \xi$:

$$\mathcal{I} = \int_{-\infty}^{\infty} \frac{N}{\sqrt{4\pi Dt}}\, \mathrm{e}^{-(x-v_\mathrm{d}t)^2/4Dt} \mathrm{d}x = \frac{N}{\sqrt{\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\xi^2}\, \mathrm{d}\xi = N.$$

Furthermore, for $t \to 0$ n becomes a δ function in $x = 0$.

The mean position is

$$\langle x \rangle = \frac{1}{N} \int_{-\infty}^{\infty} x\; n(x,t)\; \mathrm{d}x = \int_{-\infty}^{\infty} (\sqrt{4Dt}\xi + v_\mathrm{d}t) \frac{1}{\sqrt{\pi}}\, \mathrm{e}^{-\xi^2}\, \mathrm{d}\xi.$$

The first integral is zero since the integrand is odd. The second term yields

$$\langle x \rangle = \int_{-\infty}^{\infty} v_\mathrm{d}t \frac{1}{\sqrt{\pi}}\, \mathrm{e}^{-\xi^2}\, \mathrm{d}\xi = v_\mathrm{d}t, \tag{12.10}$$

as expected.

With similar calculations, the variance of the Gaussian distribution is found to be

$$\langle (x - \langle x \rangle)^2 \rangle = \frac{1}{N} \int_{-\infty}^{\infty} (x - v_\mathrm{d}t)^2 \frac{N}{\sqrt{4\pi Dt}}\, \mathrm{e}^{-(x-v_\mathrm{d}t)^2/4Dt} \mathrm{d}x = 2Dt. \tag{12.11}$$

Thus, the distribution (12.9) is the solution of the drift-diffusion equation (12.8), with the initial condition of a δ distribution in the origin. However, if we perform a simulation putting all particles in a given point (initial δ) with random velocities and letting the swarm of particles evolve with some physical scattering mechanism, we shall not obtain the result in (12.11), but rather the result shown in Fig. 12.3. The reason is to be found in the fact that (12.8) describes the diffusion phenomenon only after the time necessary to establish the "correct" correlations between positions and velocities of the particles, as it results also from the above kinetic derivation of Fick law. The result (12.11), on the contrary, comes from the distribution in (12.9), which is

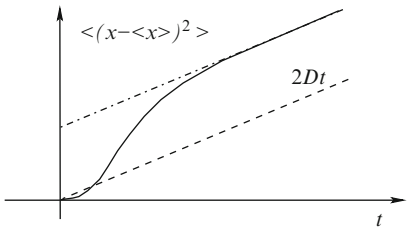


Fig. 12.3. Transient variance of positions of diffusing particles. The continuous line is the result of a physical simulation, the dashed line indicates the linear behavior in (12.11), and the dot-dashed line represents the slope of the simulated curve at long times

a solution of the diffusion equation with the δ as initial condition, only if the equation itself is valid from the initial time $t = 0$. A correct result, in place of (12.11), will be found in next section.

12.4 Moments and Correlations

The first moment of the distribution $n(x, t)$ is the mean value of the position x:

$$M_1(t) = \frac{1}{N} \int_{-\infty}^{\infty} x \; n(x, t) \; \mathrm{d}x.$$

From the diffusion equation, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} M_1(t) = \frac{1}{N} \int_{-\infty}^{\infty} x \; \frac{\partial n}{\partial t} \; \mathrm{d}x = \frac{1}{N} \int_{-\infty}^{\infty} x \; \left(-v_\mathrm{d} \frac{\partial n}{\partial x} + D \frac{\partial^2 n}{\partial x^2} \right) \; \mathrm{d}x.$$

With successive integrations by parts, taking into account that at infinity the distribution and its derivative vanish, we obtain the result

$$\frac{\mathrm{d}}{\mathrm{d}t} M_1(t) = v_\mathrm{d}, \tag{12.12}$$

to be compared with the result (12.10) obtained from the Gaussian solution.

In the same way, for the second moment we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} M_2(t) = 2 v_\mathrm{d} M_1 + 2D.$$

Collecting the above results,

$$2D = \frac{\mathrm{d}M_2}{\mathrm{d}t} - 2 v_\mathrm{d} M_1 = \frac{\mathrm{d}M_2}{\mathrm{d}t} - 2 \frac{\mathrm{d}M_1}{\mathrm{d}t} M_1 = \frac{\mathrm{d}}{\mathrm{d}t} \left[M_2 - M_1^2 \right],$$

or

$$D = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \left\langle (x - \langle x \rangle)^2 \right\rangle, \tag{12.13}$$

to be compared with the result (12.11) obtained from the Gaussian solution. The two results in (12.12) and (12.13) hold whenever the diffusion equation holds and do not depend on the validity of the equation at previous times.

The last equation (12.13) allows us to better focus that diffusion is related to the presence of correlations between particle positions and velocities. Taking the time derivative of (12.13), in fact, we obtain

$$D = \langle (x - \langle x \rangle)(v - \langle v \rangle) \rangle, \tag{12.14}$$

that indeed yields the diffusion coefficient in terms of the correlation between positions and velocities.

We may also give a different form, of significant physical content, to the relation just written. The position of a particle depends on its previous velocity, according to

$$x(t) = x(0) + \int_0^t v(t')\ \mathrm{d}t'.$$

From this, we obtain the deviation from the mean value:

$$x - \langle x\rangle = x(0) + \int_0^t v(t')\ \mathrm{d}t' - \langle x(0)\rangle - \int_0^t \langle v(t')\rangle\ \mathrm{d}t' = \delta x(0) + \int_0^t \delta v(t')\ \mathrm{d}t',$$

where

$$\delta x = x - \langle x\rangle \quad , \qquad \delta v = v - \langle v\rangle.$$

Substituting this result into (12.14), we obtain

$$D = \left\langle \left(\delta x(0) + \int_0^t \delta v(t')\,\mathrm{d}t'\right)\delta v(t)\right\rangle = \langle \delta x(0)\delta v(t)\rangle + \left\langle \int_0^t \delta v(t')\,\mathrm{d}t'\delta v(t)\right\rangle.$$

For times sufficiently far from the initial conditions, the particle velocities are independent of their initial positions, and the first term vanishes. The second term becomes, putting $t - t' = \tau$,

$$D = \int_0^t \langle \delta v(t)\,\delta v(t-\tau)\rangle\ \mathrm{d}\tau.$$

The integrand is the mean value of the product of the velocity fluctuations at a time separation τ. For stationary processes, this mean value has the following property

$$\langle \delta v(t)\,\delta v(t-\tau)\rangle = \langle \delta v(t)\,\delta v(t+\tau)\rangle.$$

Furthermore, for sufficiently distant times the velocity correlation vanishes, and the integral can be extended to infinity:

$$\boxed{D = \textstyle\int_0^\infty C(\tau)\ \mathrm{d}\tau} \tag{12.15}$$

where we have defined the velocity autocorrelation function $C(\tau)$ as

$$C(\tau) = \langle \delta v(t)\,\delta v(t+\tau)\rangle. \tag{12.16}$$

This result is the mathematical expression of the concept introduced above in connection with the fluctuation-dissipation theorem: diffusion is related to the persistence of velocity fluctuations: if the velocity fluctuations are very rapidly damped, the autocorrelation function decreases rapidly to zero; its integral is small, and the diffusion constant in (12.15) is small.

Finally, if we use Einstein relation (12.5), (12.15) above becomes

$$\mu = \frac{q}{K_\mathrm{B}T}\int_0^\infty \langle \delta v(t)\delta v(t+\tau)\rangle\ \mathrm{d}\tau, \tag{12.17}$$

which is strictly analog to Kubo quantum formula (see Sect. 16.4) of fluctuation-dissipation theorem.

12.5 Spectral Density and Wiener–Kintchine Theorem

The autocorrelation function of a stochastic variable, as the velocity fluctuation of a particle around its mean value seen in the previous section, is related to its power spectrum by the Wiener–Kintchine theorem, as we shall now discuss.

Let us consider a *noise* function, i.e., a stochastic, or random, real function of time $y(t)$ with zero mean (as our δv), defined in a finite, but arbitrarily long, time interval T (Fig. 12.4).

The Fourier series of $y(t)$ is obtained (see Appendix A) with the orthonormal basis functions

$$\phi_n(t) = \frac{1}{\sqrt{T}} \mathrm{e}^{\mathrm{i}\omega_n t}, \qquad \omega_n = \frac{2\pi}{T} n, \qquad n = 0,\ \pm 1,\ \pm 2,\ \ldots .$$

Thus, the Fourier series is

$$y(t) = \sum_n A(\omega_n) \frac{1}{\sqrt{T}} \mathrm{e}^{\mathrm{i}\omega_n t} \quad \text{with} \quad A(\omega_n) = \frac{1}{\sqrt{T}} \int_0^T y(t) \mathrm{e}^{-\mathrm{i}\omega_n t} \mathrm{d}t. \tag{12.18}$$

Let us now calculate the variance of $y(t)$. Since $\langle y \rangle$ is zero, and $y(t)$ is a real function, the variance can be evaluated as

$$\langle y^2 \rangle = \langle |y|^2 \rangle = \frac{1}{T} \int_0^T \sum_n A(\omega_n) \phi_n(t) \sum_m A^*(\omega_m) \phi_m^*(t) \mathrm{d}t = \frac{1}{T} \sum_n |A(\omega_n)|^2.$$

Since T is arbitrarily large, we may transform the sum into an integral with a density of states $g(\omega) = 1/\delta\omega = T/2\pi$. We obtain

$$\langle y^2 \rangle = \frac{1}{2\pi} \int_{-\infty}^{\infty} |A(\omega)|^2 \, \mathrm{d}\omega = \frac{1}{\pi} \int_0^{\infty} |A(\omega)|^2 \, \mathrm{d}\omega,$$

where we have taken into account that, for $y(t)$ real, $A(-\omega) = A^*(\omega)$, so that $|A|^2$ is an even function of ω. The *spectral density* of $y(t)$ is defined as the function

$$G(\omega) = \frac{1}{\pi} |A(\omega)|^2. \tag{12.19}$$

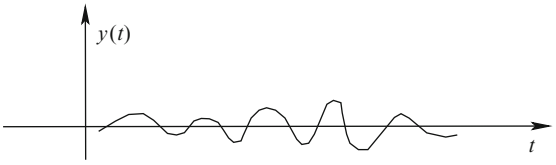


Fig. 12.4. A stochastic variable with zero mean

The above result may then be written as

$$\langle y^2 \rangle = \int_0^\infty G(\omega)\, \mathrm{d}\omega. \tag{12.20}$$

In this expression, the "power" of the noise $y(t)$ is separated into the sum (integral) of its spectral components, i.e., as the sum of its components at the various frequencies.

Consider now the autocorrelation function of $y(t)$:

$$C_y(\tau) = \langle y(t)\, y(t+\tau) \rangle = \frac{1}{T} \int_0^T (y(t)\, y(t+\tau))\, \mathrm{d}t. \tag{12.21}$$

Two comments can be made at this point. First, with reference to the previous section, we now talk of time averages of a single function $y(t)$. Nothing bad if we add an ensemble average. If, furthermore, the system is ergodic, the two averages coincide. Second, to evaluate the integral above the variable t must exceed the interval T of a small amount of time, the time needed to the autocorrelation to vanish. But T is arbitrarily large and we may "sacrifice" a little piece of it in the integration interval without any effect on the average.

If we now insert the Fourier series (12.18) into the autocorrelation function (12.21), we obtain

$$\begin{aligned} C_y(\tau) &= \frac{1}{T} \int_0^T \sum_n \frac{1}{\sqrt{T}} A(\omega_n) \mathrm{e}^{\mathrm{i}\omega_n t} \sum_m \frac{1}{\sqrt{T}} A^*(\omega_m) \mathrm{e}^{-\mathrm{i}\omega_m (t+\tau)}\, \mathrm{d}t \\ &= \frac{1}{T} \sum_n |A(\omega_n)|^2 \mathrm{e}^{-\mathrm{i}\omega_n \tau} = \frac{1}{2\pi} \int_{-\infty}^\infty |A(\omega)|^2 \cos(\omega\tau)\, \mathrm{d}\omega, \end{aligned} \tag{12.22}$$

where we have taken into account that $C(\tau)$ is a real function and the orthogonality of the Fourier functions. In conclusion, using the definition (12.19) and the parity of the integrand:

$$C_y(\tau) = \int_0^\infty G(\omega) \cos(\omega\tau)\, \mathrm{d}\omega. \tag{12.23}$$

If we now invert the Fourier transform in (12.22), we obtain

$$\frac{1}{\sqrt{2\pi}} |A(\omega)|^2 = \int_{-\infty}^\infty C(\tau) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}\omega\tau}\, \mathrm{d}\tau,$$

or, taking into account that $G(\omega)$ is real and $C(\tau)$ is even,

$$\boxed{G(\omega) = \tfrac{1}{\pi} \textstyle\int_{-\infty}^\infty C(\tau) \mathrm{e}^{\mathrm{i}\omega\tau}\, \mathrm{d}\tau = \tfrac{2}{\pi} \int_0^\infty C(\tau) \cos(\omega\tau)\, \mathrm{d}\tau} \tag{12.24}$$

Often the spectral density is considered as a function of frequency $\nu = \omega/2\pi$, such that

$$G'(\nu)\mathrm{d}\nu = G(\omega)\mathrm{d}\omega.$$

Thus,

$$G'(\nu) = 2\pi G(\omega) = 4\int_0^\infty C(\tau)\cos(2\pi\nu\tau)\ \mathrm{d}\tau.$$

The above relations are known as the *Wiener–Kintchine theorem* and show the amount of information in the autocorrelation function of fluctuations: the spectral density is entirely contained in it.

12.6 Nyquist Theorem

In this last section of the chapter devoted to the analysis of diffusion and fluctuations, we shall consider an important application to noise in electrical circuits. Let us consider a sample of material with electrical resistance R, as shown in Fig. 12.5, with no applied voltage, and short-circuited by a perfect conductor.

Charge carriers are subject to thermal excitation. Thus, while the average current is zero, the instantaneous current fluctuates around this value. From Shockley–Ramo theorem [358,415], the current induced in the external circuit by the motion of the internal charges is given by

$$I = \sum_i \frac{1}{L} q v_i,$$

where v_i is the velocity component of the i-th charge orthogonal to the contacts, and L is the length of the sample. Since the mean quantities are zero, the above equation can be written as

$$\delta I = \sum_i \frac{1}{L} q \delta v_i.$$

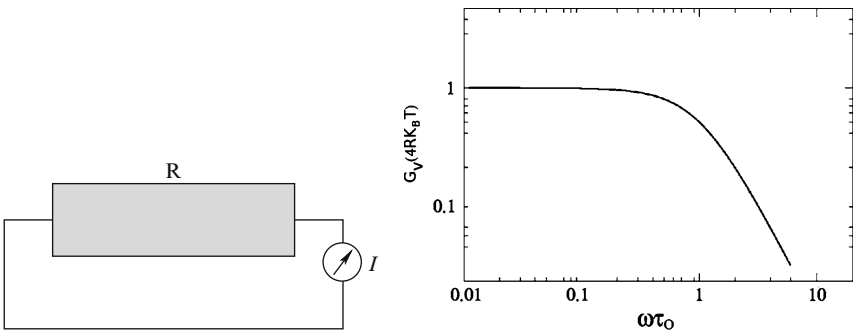


Fig. 12.5. Spectral density of Johnson–Nyquist noise (*right*), in a resistance R (*left*)

It follows:

$$\langle (\delta I)^2 \rangle = \sum_{ij} \frac{q^2}{L^2} \langle \delta v_i \delta v_j \rangle.$$

If the velocities of the different particles are not correlated, the average products with $i \neq j$ vanish:

$$\langle (\delta I)^2 \rangle = \sum_{i} \frac{q^2}{L^2} \langle \delta {v_i}^2 \rangle = N \frac{q^2}{L^2} \langle \delta v^2 \rangle,$$

where N is the total number of charge carriers in the sample. In terms of spectral densities, using (12.20), this yields

$$G_I(\omega) = N \frac{q^2}{L^2} G_v(\omega).$$

Let us now apply Wiener–Kintchine theorem (12.24):

$$G_v(\omega) = \frac{2}{\pi} \int_0^\infty C_v(\tau) \cos(\omega\tau) \mathrm{d}\tau = \Re \frac{2}{\pi} \int_0^\infty C_v(\tau) \mathrm{e}^{\mathrm{i}\omega\tau} \mathrm{d}\tau.$$

If we assume the existence of a constant velocity relaxation time $\tau_\circ$ (the same considered in Sect. 11.3),

$$C_v(\tau) = \langle v^2 \rangle \mathrm{e}^{-\tau/\tau_\circ},$$

and it results (only one component of v is considered)

$$G_v(\omega) = \Re \frac{2}{\pi} \int_0^\infty \langle v^2 \rangle \mathrm{e}^{-\tau/\tau_\circ} \mathrm{e}^{\mathrm{i}\omega\tau} \mathrm{d}\tau = \Re \frac{2}{\pi} \frac{K_\mathrm{B} T}{m} \frac{1}{1/\tau_\circ - i\omega} = \frac{2}{\pi} \frac{K_\mathrm{B} T}{m} \frac{\tau_\circ}{1+(\omega\tau_\circ)^2}.$$

If we now consider that

$$\mu = \frac{q\tau_\circ}{m},$$

we have

$$G_I(\omega) = N \frac{q^2}{L^2} G_v(\omega) = N \frac{q}{L^2} \mu \frac{2}{\pi} K_\mathrm{B} T \frac{1}{1 + (\omega\tau_\circ)^2}.$$

If the sample has area A, its resistance is given by

$$R = \rho \frac{L}{A} = \frac{1}{\sigma} \frac{L}{A} = \frac{1}{nq\mu} \frac{L}{A} = \frac{L^2}{Nq\mu},$$

and the current spectral density becomes

$$G_I(\omega) = \frac{1}{R} \frac{2}{\pi} K_\mathrm{B} T \frac{1}{1 + (\omega\tau_\circ)^2}.$$

If the fluctuations are observed by a low-noise external amplifier with a high input impedance, they are measured as voltage fluctuations applied to the

resistance given by $\delta V = R\delta I$. Thus, we have a voltage spectral density given by

$$G_V(\omega) = R^2 G_I(\omega) = R\frac{2}{\pi}K_{\mathrm{B}}T\frac{1}{1+(\omega\tau_\circ)^2},$$

or, in terms of frequency ν,

$$G_V(\nu) = 4RK_{\mathrm{B}}T\frac{1}{1+(2\pi\nu\tau_\circ)^2}.$$

This is the spectral density of the voltage noise due to velocity fluctuations, called *Johnson–Nyquist noise* and it is shown in Fig. 12.5. For frequencies well below the cutoff frequency $1/2\pi\tau_\circ$, it yields the *white-noise* spectrum, independent of frequency, whose value is given by *Nyquist theorem*:

$$\boxed{G_V(0) = 4RK_{\mathrm{B}}T} \tag{12.25}$$

Nyquist theorem can also be proved with very general macroscopic, thermodynamic arguments [370].

13

Nonlinear Transport

The year 1951 marked an important milestone in the physics of semiconductors. The invention of the transistor by J. Bardeen, W. Shockley, and W. Brattain opened the era of solid-state electronics: through the successive steps of integrated circuits, large-scale integration (LSI), very large-scale integration (VLSI), and finally microprocessors, one of the most impressive and unpredictable technological revolutions of human history was realized. At present, systems are fabricated in single chips that contain almost a billion transistors, and computers can perform almost a billion elementary instructions per second. This fantastic process did not come to an end yet,[1] and it is not easy to predict what kind of new surprises it is going to present us. One of the most important features of such technological evolution is the continuous feedback between science and technology, since the increased knowledge has produced new technological applications and new technologies have provided new instruments, both experimental apparatuses and computational tools, for the advancement of science.

To realize faster and cheaper systems, the dimensions of solid-state devices has steadily decreased, reaching the nanometer scale, comparable with the size of microscopic systems, where quantum mechanics is the only reliable theoretical framework. Furthermore, the voltages applied to such "nano-devices" could not be proportionally decreased, since it is necessary that they dominate over thermal excitations to maintain electric control of the device performances. As a result, the applied electric fields inside the active regions of the devices had to increase steadily.

It is not a coincidence, therefore, that in the same year of the invention of the transistor, papers were published by Ryder and Shockley [387,388,418] reporting experimental evidence of deviations from linearity in the drift

[1] According to the well-known Moore's law, the number of transistors in an integrated circuit has doubled approximately every two years. This trend is still active in modern chips, and Moore's law actually influences the long-term planning of semiconductor industry.

velocity of charge carriers in high electric fields. These papers are often considered the official birth of the *hot-electron problem*, i.e., of the study of the deviations from the linear response regime due to heating of charge carriers above thermal equilibrium, even though a number of theoretical studies had been carried out on this subject in the 1930s, mainly by Russian physicists, such as Landau and Davidov. The concept of an electron temperature T_e which, in the presence of an external high electric field, is higher than the lattice temperature T has been introduced in [156, 157]. References to the early works on this subject can be found in the classical review by Conwell [107]. Several monographs have been successively written where the interested reader may follow the development of this field [23, 64, 364, 406].

13.1 Hot Electrons

As mentioned above, nonlinear transport deals with problems that arise when a sufficiently strong electric field is applied to a semiconductor sample, so that the current deviates from the linear response. This effect is typical of semiconductors and cannot be seen in metals since, owing to their large conductivities, Joule heating would destroy the material before deviations from linearity could be observed.

Carrier heating and its relation to nonlinearity can be easily understood in a "simple model" semiconductor (i.e. with spherical and parabolic bands and the effect of scattering approximated with a relaxation time), by considering the argument that led to the simple explanation for the Ohmic mobility at the end of Sect. 11.3.1. With reference to Fig. 11.2, let us consider again the equation (11.12):

$$\boldsymbol{v}^{(i)}(t) = \boldsymbol{v}_\circ^{(i)} + \boldsymbol{a}\Delta t^{(i)}, \tag{13.1}$$

where $\boldsymbol{v}^{(i)}(t)$ is the instantaneous velocity of the i-th electron at time t, $\Delta t^{(i)}$ is the time elapsed after its last scattering event, and $\boldsymbol{a} = q\boldsymbol{E}/m$ is the electron acceleration, due to a constant and uniform applied electric field $\boldsymbol{E}$. In Sect. 11.3.1 we have seen that, by averaging the above expression, we obtain a simple formula for the mobility at sufficiently low values of the electric field.

If, on the contrary, the field is sufficiently high, the energy gained by the electrons during each single flight is not negligible; the electrons cannot fully dissipate, initially, this energy at each scattering event, and the effect of the field accumulates on the velocities $\boldsymbol{v}_\circ^{(i)}$. The distribution of the flight durations $\Delta t^{(i)}$ also depends upon the field, since the scattering rates are energy dependent, and the linearity between mean velocity and electric field is lost in averaging the expression (13.1). The mean energy of the carriers then increases, and it would continue to increase up to breakdown of the material were it not for scattering mechanisms whose dissipation capabilities increase at increasing electron energy. These scattering mechanisms establish a new stationary state in which the average electron energy is higher

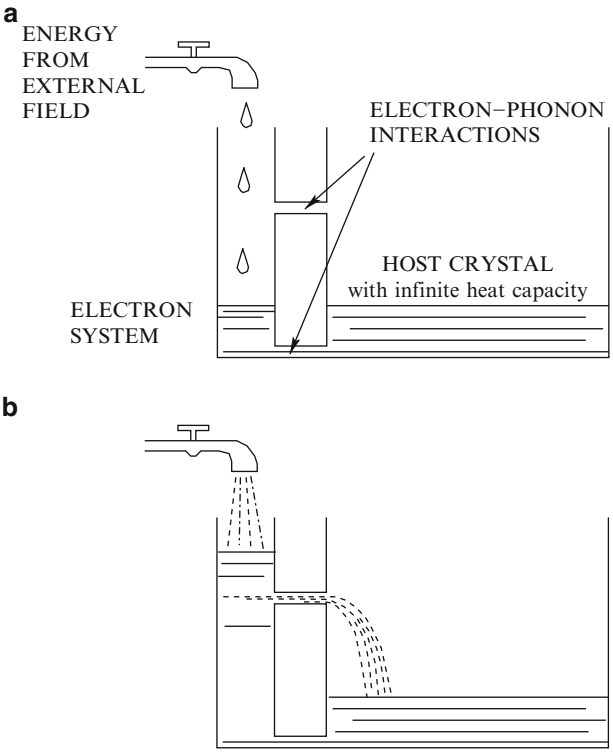


Fig. 13.1. Hydraulic analog of the concept of hot electrons. At increasing input power, both the dissipation mechanisms previously present become more effective, and new mechanisms are activated

than at equilibrium. This is the origin of the expression "hot-electron effects" applied to nonlinear transport phenomena in semiconductors. The region of field strengths where transport starts to deviate from linearity has been called *warm-electron region*.

Figure 13.1 shows a hydraulic analog of the concept of hot electrons. Energy is transferred from the external electric field to the electron system and, through them, to the host crystal. In the low-field case, represented by part (a) of the figure, some interaction mechanisms are capable of maintaining the temperature of the electron system equal to that of the thermal bath. At high fields, part (b) of the figure, the power supplied to the electrons is higher, and a new stationary state is attained by increased efficiencies of the scattering mechanisms already active at low fields and, sometimes, by the onset of new mechanisms.

In what follows, a general picture of hot-electron effects in pure semiconductors is given (see Fig. 13.2), taking into account the general features of the band structures and of the scattering rates [214]. For the sake of simplicity

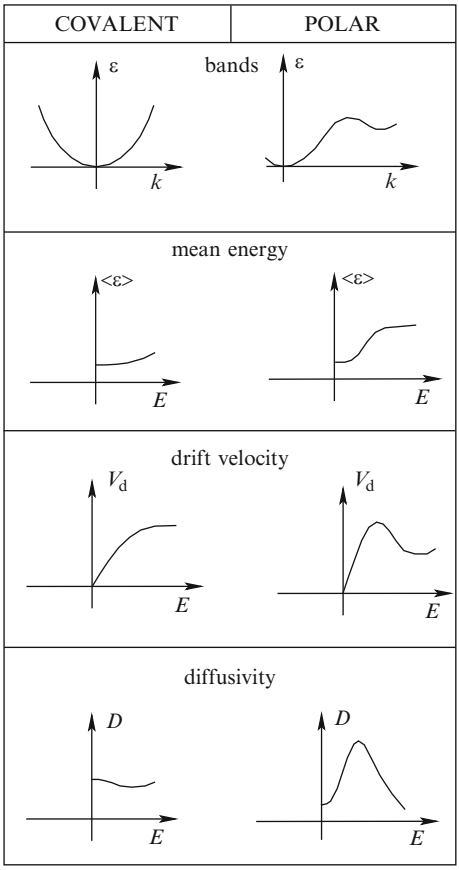


Fig. 13.2. Simplified general physical picture of hot-electron effects in semiconductors (see text)

we shall limit ourselves, in the present discussion, to transport dominated by phonon scattering.

Let us separate all semiconductors into two categories depending on whether the most effective phonon scattering mechanisms are due to electrostatic fields (polar materials) or to deformation potential (covalent materials). Then we take into account that as the external field increases above the linear region, the mean energy of the electrons in steady-state conditions increases.

For covalent materials, the scattering mechanisms become more effective as the electron energy increases, so that (a) the mobility decreases as the field increases and (b) steady state is attained with a small energy rise above thermal equilibrium. The electron mean energy increases slowly with field at the beginning of the deviation from linear-response regime, as shown in Fig. 13.2.

For polar materials, the efficiencies of the scattering mechanisms do not effectively increase (and may even decrease) as the electron mean energy increases. Thus, the mean electron energy increases very fast to reach steady-state conditions, as the applied field increases, while the mobility does not decrease. Eventually polar run-away could occur, but in most cases new scattering mechanisms become effective, as shown in Fig. 13.1 (see Sect. 13.4).

At higher fields, in covalent semiconductors the mobility keeps decreasing toward saturation of the drift velocity and the electron mean energy keeps increasing. Eventually upper valleys are reached; this happens in germanium, but not in silicon, since in Si the upper valleys are so high in energy that breakdown occurs first.

In polar semiconductors, the mean energy of the electrons increases much faster, and above a certain threshold field electrons then populate the upper valleys in the band and the population of these upper valleys increases very rapidly as the external field increases. At this point, the new mechanism of non-equivalent intervalley scattering comes into play. Since this is very effective in dissipating both energy and momentum, above threshold field for polar run-away the energy increase with field is slowed down, and the drift velocity decreases giving rise to the well-known phenomenon of *negative differential mobility* (NDM) at the basis of the Gunn effect [174]. See, for example [82, 373, 410].

If we now consider the diffusivity of the electrons, we must remember that at high fields, in the nonlinear regime, the Einstein relation does not hold. However, we may consider qualitatively a very rough generalization of that relation:

$$D \approx \frac{2}{3}\mu(E)\frac{\langle\epsilon\rangle}{(-\mathrm{e})},$$

where $\mu(E)$ is the electric-field dependent mobility. By taking into account the considerations made above on the drift velocity and on the mean energy, we obtain the general behavior shown in the lowest part of Fig. 13.2. The diffusivity is determined, in general, by the competing effects of reducing $\mu(E)$ and increasing $\langle\epsilon\rangle$ as the external field increases. In covalent semiconductors, the two effects are almost balanced, the former being slightly prevalent, so that D decreases slowly with increasing fields. In polar semiconductors, we have seen, to the contrary, that there is first a sharp increase in $\langle\epsilon\rangle$ not accompanied by a decrease in mobility, so that D increases very rapidly with E. Then, the increase in energy is quickly slowed down, while the mobility decreases: a sharp reduction of D follows.

Similar considerations hold for holes except for the absence of upper valleys.

Almost all hot-electron effects can be described as details of the above general picture.

To obtain more precise theoretical results, the Boltzmann equation should be solved without linearization with respect to the external fields. This, however, is a formidable mathematical problem which has resisted many attacks,

in the past, for several decades and is still unsolved today. A full account of the investigations carried out analytically in this area can be found, for example, in [107].

A big step forward in the solution of the Boltzmann equation has been accomplished with the introduction of powerful numerical techniques. The most important of them is the Monte Carlo (MC) technique, introduced by Kurosawa, as it regards its application to electron transport in semiconductors, at the Semiconductor Conference in Kyoto in 1966 [258].[2] Chapter 14 will be devoted to the description of this method.

13.1.1 The Warm Electron Region

In linear-response regime, the drift velocity $\boldsymbol{v}_\mathrm{d}$ is proportional to the applied field $\boldsymbol{E}$. Outside the linear regime, we may still formally write

$$\boldsymbol{v}_\mathrm{d} = \mu(\boldsymbol{E})\boldsymbol{E}.$$

This time, however, the mobility μ is a function of field.

In most cases, the first deviation from linearity of the drift velocity with field may be approximated as

$$\boldsymbol{v}_\mathrm{d} = \mu_\circ(1 + \beta E^2)\boldsymbol{E}, \tag{13.2}$$

where $\mu_\circ$ is the low-field mobility, and β is independent of E. The region of field strengths in which this approximation holds is called the *warm electron region*. It must be noted, however, that once the transport phenomenon deviates from linearity, the shape of the curve $v_\mathrm{d}(E)$ quickly becomes quite complex so that the warm-electron region is often very narrow.

As it happens to most quantities related to nonlinear transport, the value of β is influenced by the dependence upon energy of the efficiency of the scattering mechanisms in relaxing electron momentum and energy. In the warm electron region, the introduction of an *energy relaxation time* τ_ϵ has been found convenient. This quantity should describe in a phenomenological way the tendency of the mean energy, and therefore of the isotropic part of the distribution function, to decay toward their equilibrium values and is in general a function of field and temperature. It is defined by equating the power

[2] At the same Conference, Budd introduced another numerical technique called *iterative technique* [76]. This method yields a solution of the Boltzmann equation by means of an iterative procedure and processes the whole distribution function in each step of the procedure. For this reason, it can be useful when we deal with physical phenomena which depend on details of the distribution function, such as, for example, impact ionization or penetration into the oxide barrier of a MOSFET, which depend on the high-energy tail of the distribution function. This method has been somewhat popular for some time, but is very little used today, owing to the much better stability and ease of use of the MC approach.

input by the field to the dissipated power:

$$ev_{\mathrm{d}}E = \frac{\langle\epsilon\rangle - \langle\epsilon\rangle_{\circ}}{\tau_{\epsilon}}, \tag{13.3}$$

where $\langle\epsilon\rangle_{\circ}$ is the mean energy at equilibrium, $\langle\epsilon\rangle$ the mean energy in presence of the field, and v_{d} the drift velocity.

The definition of the energy relaxation time, given by (13.3) above, is often used also outside the warm-electron region.

13.2 Electron–Electron Collisions and the Heated and Drifted Maxwell Distribution

For simplicity, electron–electron (e–e) collisions have been ignored in the general discussion of Sect. 13.1. At high concentrations, however, they may have relevant effects. They do not dissipate momentum nor energy of the entire electron gas, since such quantities are transferred from one colliding particle to the other one. Such collisions, instead, may influence the shape of the electron distribution function and, therefore, may be important for understanding the effect, for example, of highly energetic electrons in phenomena, such as impact ionization or tunneling through material barriers, determined by the tail of the distribution function. When e–e collisions are dominant, we may assume that the electron distribution is well approximated by a drifted and heated Maxwellian, introduced in [157]:

$$f(\boldsymbol{v}) \propto \mathrm{e}^{\frac{m(\boldsymbol{v}-\boldsymbol{v}_{\mathrm{d}})^2}{2K_{\mathrm{B}}T_{\mathrm{e}}}}, \tag{13.4}$$

where $\boldsymbol{v}_{\mathrm{d}}$ is the drift velocity, and T_{e} the electron temperature. There have been several studies on the critical electron concentration, necessary to ensure a Maxwell distribution. It depends, of course, by the material of interest and in particular on the competing scattering mechanisms. Generally speaking, we may say that a concentration of the order of 10^{16}–$10^{17}\,\mathrm{cm}^{-3}$ is necessary. However, this approximation of the distribution function has lost much of its interest as an analytical tool, after the introduction of numerical techniques for the solution of the BE, while it has maintained its heuristic value to obtain physical insight, so that often the concept of electron temperature, as an indication of the mean electron energy, is still usefully employed today, even when it is clear that the distribution cannot be approximated by a form like (13.4).

13.3 Anisotropy of Transport Coefficients

One of the main differences between transport properties in linear and in nonlinear regimes, not included in the general picture given above, is related

to the anisotropy of the transport coefficients. The conductivity is defined as the proportionality constant between two vectors, the current density $\boldsymbol{j}$ and the electric field $\boldsymbol{E}$. As such it is, in general, a tensor:

$$\boldsymbol{j} = \sigma \boldsymbol{E}, \qquad j_i = \sum_k \sigma_{ik} E_k.$$

However, as long as the conduction process remains within the linear-response regime, the conductivity is a property of the material, independent of the applied field, and shares with it the symmetry properties. A cubic tensor reduces to a scalar quantity.[3] Thus, the current density results parallel to the applied field and independent of its direction.

When, on the contrary, the field intensity is outside of the linear-response regime, the conductivity becomes a function of the applied field, and its symmetry is lowered, being that of the cubic crystal with the applied field. The current density is different for fields applied along different directions, and when the field is applied along a direction of low symmetry, drift velocity and field are not even parallel [396, 411].

The major source of anisotropy is related to the band structure. In the case of current carried by electrons in many-valley semiconductors, the anisotropy effects are well understood on the basis of a *valley repopulation* of the different valleys (see Fig. 13.3). In fact, when the field strength is small, in the linear regime, and the energy distribution is the equilibrium one, all valleys are equally populated by electrons. Owing to the anisotropy of the equienergetic surfaces, the drift velocities in the different valleys are different. The total drift velocity is obtained as an average, with equal weights, of the drift velocities in the different valleys. If the field orientation is changed, the contributions of the different valleys change, but the final average drift velocity is the same for the symmetry considerations made above.

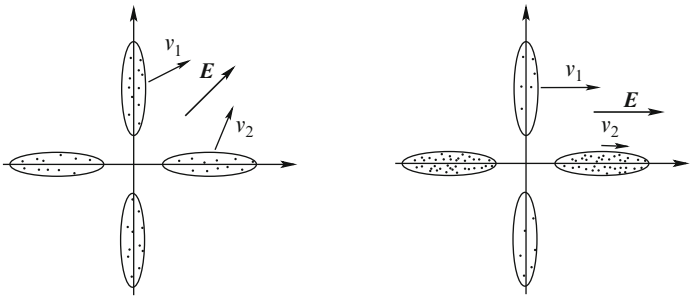


Fig. 13.3. Valley repopulation at high fields is responsible for conduction anisotropy

[3] It is easy to show that if we require to a tensor to remain unaltered for the symmetry operations that bring a cube into itself, we find that the tensor must be diagonal with equal diagonal element, equivalent to a scalar quantity.

When the field strength is increased above the linear regime, valleys with different orientations with respect to the field direction, having different single-valley mobilities, are warmed up to a different extent. The intervalley scattering rates, functions of the electron energy, are different in the different valleys. The probability for an electron in a "hot valley" to make a transition to a "cold valley" is higher than that of an opposite transition for an electron in a cold valley. In steady state, therefore, the valley populations are no longer equal. More precisely, the valleys the present higher effective mass along the direction of the field have lower mobilities, are slower and colder, have a larger population with respect to the equilibrium situation, give larger contributions to the total average drift velocity, and the total drift velocity is reduced with respect to the symmetric case, as shown in Fig. 13.3.

Since intervalley transitions require a phonon of large momentum, and therefore of relative large energy, at low temperatures such transitions are rare, and valley repopulation requires a relative long time. This fact allowed the experimental observation of the time evolution of such phenomenon, with consequent determination of the intervalley repopulation time and, therefore, of the intervalley coupling constant [88].

For the case of holes, the symmetry considerations made above maintain their validity, even though the physical mechanism, which originates anisotropy of the drift velocity is somewhat different. Owing to optical–phonon interaction, at high fields the drift motion assumes a peculiar streaming character [63], which is particularly sensitive to the value of the effective mass along the field direction. Thus, the warped shape of the equi-energetic surfaces of holes, with particular reference to the heavy-hole band, is reflected in the anisotropy of the drift velocity.

13.4 Negative Differential Mobility and Gunn Effect

Negative differential mobility (NDM) is a phenomenon presented in some semiconductor materials, typically compound semiconductors, characterized by the presence of a range of field strengths in which the drift velocity decreases with increasing field strength, and consequently the differential mobility

$$\mu' = \frac{\mathrm{d}\, v_\mathrm{d}}{\mathrm{d}E}$$

is negative.

This phenomenon is due to a transfer mechanism of electrons to upper valleys, as shown in Fig. 13.4. The conduction bands in compound materials contain, besides the main minimum at the center of the BZ, other secondary minima higher, in energy, of a quantity of a few tens of meV. At room or lower temperature and at low field intensities, electrons occupy only the main minimum (part (a) of the figure), and the mechanism that controls the mobility is the scattering with polar optical phonons. At high electric fields, electrons are

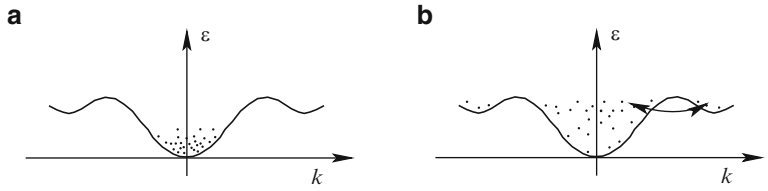


Fig. 13.4. Intervalley scattering from central to upper valleys and electron transfer to upper, slower valleys are responsible of negative differential mobility in GaAs. At low fields (**a**) electrons lie in the bottom of central valley; at higher fields (**b**) they are warmed up and make transitions to upper valleys

heated and at a certain point this heating is such as to allow some electrons to scatter to the upper valleys.

Two phenomena at this point occur: (i) electrons start to populate upper valleys with higher effective masses, and therefore with lower mean velocity, and (ii) a new scattering mechanism, nonequivalent intervalley scattering, becomes active, very effective in dissipating momentum. Both these phenomena contribute to lower the drift velocity, and since they become more important as the field strength increases, the net result is that the drift velocity decreases at increasing fields, or, in other words, a *negative differential mobility* (NDM). The fact that also the increased efficiency of the scattering, and not only the lower drift velocities of the upper valleys as often stated in the literature, is important in determining the NDM, is confirmed by the fact that also the mean velocity of electrons in the central valley alone decreases at increasing field strength, as shown in Fig. 13.5.

The threshold fields for NDM in different materials are consistent with the values predicted by Stratton [427], based on a drifted Maxwellian distribution, for *polar run-away*: in compound semiconductors, the equilibrium of electron energy at low fields is in general maintained by polar scattering with optical phonons. At high energies, the efficiency of such mechanism decreases at increasing electron energy, as shown in Fig. 9.8. Electrons that reach a high energy are no longer able to dissipate the energy gained by the field, and their energy increases indefinitely. This is the phenomenon of polar run-away, shown by Frölich to be the physical origin of dielectric breakdown in polar materials. The above consideration can explain why the energy of the upper valleys is not crucial for the occurrence of NDM, as long as electrons can reach them before impact ionization occurs [59, 212].

The NDM phenomenon is at the basis of the Gunn effect [174]. See, for example [82, 373, 410]. This effect consists in very fast current oscillations, (in the range of 10–30 GHz) obtained with the application of a constant voltage of high enough intensity to samples of materials, such as GaAs and InP, that present NDM. Obviously, such an effect is of great practical importance for the generation of microwaves. The reason for this phenomenon is well understood: NDM implies that a local fluctuation of charge density tends to grow with

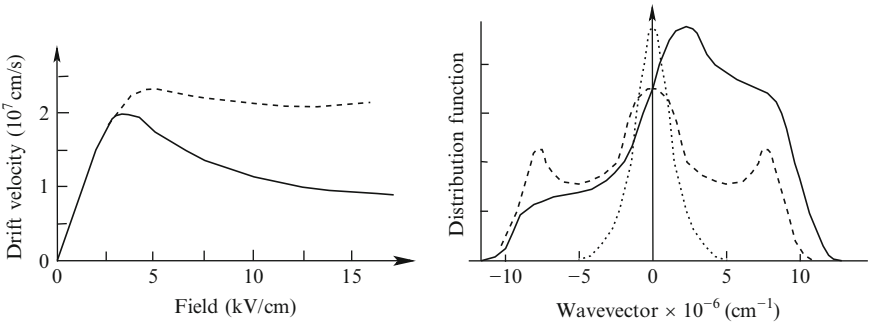


Fig. 13.5. Results of MC simulation [137] of electrons in GaAs. Part (a) shows the drift velocity as a function of the applied field (*full line*) and the average velocity of electrons in the central valley (*dashed line*). Part (b) shows the distribution function in the central valley along lines through $k = 0$ parallel (*full line*) and perpendicular (*dashed line*) to a field of 15 kV/cm. The dotted line represent the equilibrium Maxwellian

time, and this brings about the formation of the high-field domains that lower the current. The domains travel with the drift velocity of the electrons, which is the same inside and outside the domain, owing to the shape of the v_d vs E curve. When they reach the anode they disappear, and the current increases until a new domain is formed. The period of the Gunn oscillations is therefore given by the transit time of the electrons across the sample and can be very short.

13.5 High-Field Diffusivity

Diffusion is heavily influenced by carrier heating in nonlinear regime. In particular, Einstein relation no longer holds, at least in its simple form (12.5) valid at equilibrium. A generalization has been proposed by Price [348], given by

$$D = \frac{\mu' K_\mathrm{B} T_n}{q}, \tag{13.5}$$

where μ' is the *differential mobility*, defined as $\mathrm{d}v_\mathrm{d}/\mathrm{d}E$ (the "standard" mobility $\mu = v_\mathrm{d}/E$ in this context is called *cord mobility*), and T_n is the *noise temperature*.[4] Equation (13.5) is essentially a definition of the noise temperature.[5] However, Price has shown that in suitable conditions it may be approximated by the electron energy T_e, as defined by the mean electron random energy.

[4] The quantities involved in (13.5) have a tensor nature, but for simplicity this fact will be ignored in this presentation.

[5] More precisely, the noise temperature is defined by the spectral density of fluctuations.

A detailed analysis of diffusion, still related to velocity fluctuations, can be obtained [70] by applying a sort of analysis-of-variance theory [477]: the deviation from the mean of the velocity of each electron is separated into several terms, taking into account the valley in which the electron lies, its energy, and finally its momentum:

$$\boldsymbol{v} = \boldsymbol{v}_\mathrm{d} + (\boldsymbol{v}_v - \boldsymbol{v}_\mathrm{d}) + (\boldsymbol{v}_{\epsilon,v} - \boldsymbol{v}_v) + (\boldsymbol{v} - \boldsymbol{v}_{\epsilon,v}) = \boldsymbol{v}_\mathrm{d} + \delta\boldsymbol{v}_v + \delta\boldsymbol{v}_\epsilon + \delta\boldsymbol{v}_k, \quad (13.6)$$

where $\boldsymbol{v}_v$ is the mean velocity of the electrons in valley v so that $\delta\boldsymbol{v}_v(t)$ is the fluctuation associated with the valley in which is the electron at time t, $\boldsymbol{v}_{\epsilon,v}$ is the mean velocity of the electrons in valley v with the energy of the electron under examination so that $\delta\boldsymbol{v}_\epsilon$ is the fluctuation associated with the energy of the electron, and finally $\delta\boldsymbol{v}_k$ is the velocity fluctuation due to the actual velocity of the electron with respect to the mean velocity of the electrons in that valley with that energy. Once the velocity fluctuation has been split into its component, as in (13.6), its autocorrelation may be evaluated for the determination of the diffusion constant, as described in Sect. 12.4.

$$C(\tau) = \langle \delta\boldsymbol{v}(t)\delta\boldsymbol{v}(t+\tau)\rangle = \sum_{ij}\langle \delta\boldsymbol{v}_i(t)\delta\boldsymbol{v}_j(t+\tau)\rangle = \sum_{ij} C_{ij}.$$

Several distinct contributions may thus be associated with noise and, therefore, with diffusion, called *intervalley noise* [347], associated to C_{vv}, *effusion noise* [348], associated to $C_{\epsilon\epsilon}$, and thermal noise associated with C_{kk}. These contributions to noise and diffusion come from the diagonal elements of C_{ij}. Off diagonal terms may also contribute, although to a minor extent and with quantities that tend to cancel each other [70].

13.5.1 Intervalley Diffusion

Intervalley diffusion is of particular interest in many-valley semiconductors. As just indicated, it arises from the term $\delta\boldsymbol{v}_v$ in (13.6). In physical terms, if we associate with each electron exactly the mean velocity of its valley, a velocity fluctuation is present due to the electron transitions between valleys with different drift velocities, as shown in Fig. 13.6. An initial bunch of electrons under the influence of the electric field will spread as an effect of the random character if intervalley transitions, even if any other source of disordered motion is neglected. This effect can be observed along the direction of the external field when the valley drift velocities have different components along $\boldsymbol{E}$, and transverse to the field when valley drift velocities have different components orthogonal to $\boldsymbol{E}$. The upper part of Fig. 13.6 shows the orientations of the valley drift velocities $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ with respect to the drifting force $\boldsymbol{F}$. In the lower part, the time evolutions of the particle distributions in space are shown for bunches of carriers that start a $t = 0$ in $x = 0$. The two lines in the lower part refer to the absence of intervalley transitions (above) and to the presence of only intervalley diffusion (below).

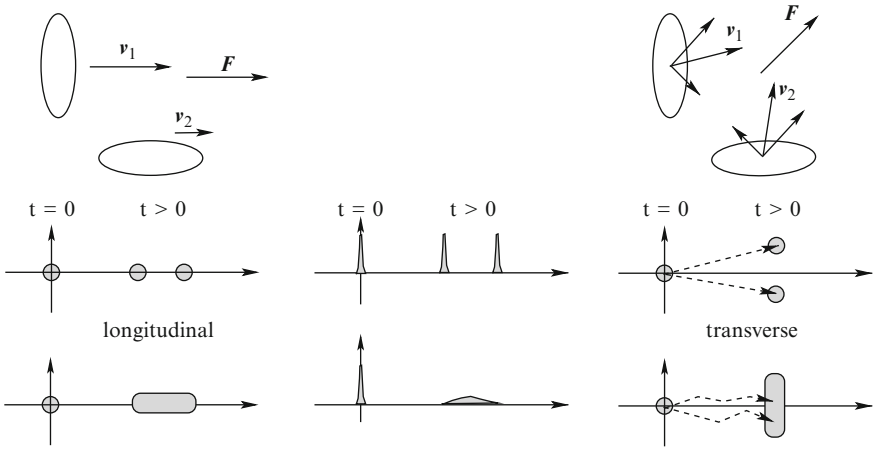


Fig. 13.6. Intervalley diffusion (see text)

An approximate expression for the intervalley diffusion coefficient can be obtained as follows. For simplicity, only the velocity fluctuations along one direction, longitudinal or transverse, will be considered. The intervalley velocity autocorrelation function

$$C_{vv}(\tau) = \langle \delta v_v(t) \delta v_v(t+\tau) \rangle$$

can be expressed as

$$\frac{1}{N} \sum_{p=1}^{N} \delta v_{vp}(t) \delta v_{vp}(t+\tau) = \frac{1}{N} \overline{\sum_{p=1}^{N} \delta v_{vp}(t) \sum_{q=1}^{N} \delta v_{vq}(t+\tau)}, \tag{13.7}$$

where $\delta v_{vp}(t)$ is the intervalley velocity fluctuation of the pth particle at time t and N is the number of particles in the ensemble. The second sum could be inserted by assuming no correlations between different particles. Furthermore, the time average has been added, always possible in mean quantities of stationary systems. In the case we are considering of a two-valley system, if $n_1(t)$ and $n_2(t)$ are the relative valleys populations at time t, and $\overline{n}_1$ and $\overline{n}_2$ their mean values, we have

$$\sum_{p=1}^{N} \delta v_{vp}(t) = \sum_{p=1}^{N} \left[v_{vp}(t) - v_{\mathrm{d}} \right] = N \left[n_1(t) \left(v_1 - v_{\mathrm{d}} \right) + n_2(t) \left(v_2 - v_{\mathrm{d}} \right) \right].$$

Taking into account that $n_1 + n_2 = 1$ and that $v_{\mathrm{d}} = \overline{n}_1 v_1 + \overline{n}_2 v_2$, this becomes

$$= N \left[n_1(t) v_1 + n_2(t) v_2 - v_{\mathrm{d}} \right] = N \left[n_1(t) v_1 + n_2(t) v_2 - \overline{n}_1 v_1 - \overline{n}_1 v_1 \right].$$

Finally, since $\delta n_1 = -\delta n_2$,

$$\sum_{p=1}^{N} \delta v_{vp}(t) = N\left[\delta n_1 v_1 + \delta n_2 v_2\right] = N\delta n_1 (v_1 - v_2).$$

If this result is inserted into (13.7), it yields

$$C_{vv}(\tau) = N(v_1 - v_2)^2 \overline{\delta n_1(t)\delta n_1(t+\tau)}.$$

This autocorrelation function can be used in (12.15) to obtain the intervalley diffusion coefficient:

$$D_{vv} = N(v_1 - v_2)^2 \int_{\circ}^{\infty} \overline{\delta n_1(t)\delta n_1(t+\tau)}\, \mathrm{d}\tau. \tag{13.8}$$

From this equation, it is clear that intervalley diffusion, due to random changes of the valley drift velocity of each electron, when an intervalley transition occurs, corresponds to the noise due to the fluctuations of the valley populations.

To estimate the dependence of the autocorrelation function in (13.8) upon τ let us assume that electrons scatter from valley 1 to valley 2 with constant rate $P_{12} = \tau_{12}^{-1}$ and similarly for the inverse transition: $P_{21} = \tau_{21}^{-1}$. The equation that governs the number N_1 of particles in valley 1 is then

$$\frac{\mathrm{d}}{\mathrm{d}t} N_1(t) = -N_1 P_{12} + N_2 P_{21} = -N_1 P_{12} + (N - N_1) P_{21} = N P_{21} - N_1 (P_{12} + P_{21}).$$

This shows that in stationary conditions, corresponding to zero time derivative, $\overline{N}_1 = NP_{21}/(P_{12} + P_{21})$, as expected, and, more important for us now, that a variation δN_1 decays as

$$\delta N_1(\tau) = \delta N_1(0) \mathrm{e}^{-\tau/\tau_i},$$

where τ_i is $(P_{12} + P_{21})^{-1}$. Since $\delta n_1 = \delta N_1/N$ this equation allows us to write (13.8) as

$$D_{vv} = N(v_1 - v_2)^2 \int_{\circ}^{\infty} \overline{\delta n_1^2} \mathrm{e}^{-\tau/\tau_i} \mathrm{d}\tau = N(v_1 - v_2)^2 \tau_i \overline{\delta n_1^2}. \tag{13.9}$$

It remains to evaluate the variance $\overline{\delta n_1^2}$, and this can be done by observing that if we select electrons at random we catch one in valley 1 with probability $\overline{n_1}$ and from valley 2 with probability $\overline{n_2}$, so that we have the variance of a binomial distribution [370]: $\overline{\delta n_1^2} = \overline{n}_1 \overline{n}_2 / N$, and (13.9) becomes

$$D_{vv} = \overline{n}_1 \overline{n}_2 (v_1 - v_2)^2 \tau_i, \tag{13.10}$$

known as Shockley formula for intervalley diffusion [347, 419].

Note, finally, that intervalley noise is not an equilibrium phenomenon, since a drift velocity must be present, but it is not a "hot-electron" phenomenon, since it does not require that the transport is beyond the linear-response regime.

13.6 Transient Transport

In linear-response regime, the state of the electron gas is close to equilibrium, so that the application of a weak field does not require a profound modification of the system to reach the new stationary state. On the contrary, if the applied field is high, and the response is in the hot-electron regime, the new steady state is far from equilibrium, and the process to reach it may be rather complex. The features of transient transport has received much attention, mainly in connection with the applications to small electronic devices. In fact, if charge carriers with the equilibrium distribution germane to the highly conductive contact with very low field, enter a small device where a high field is present, they start to be accelerated and warmed up by this field. If the device is very small, as those currently used in nanoelectronics, the carriers reach the opposite contact before their velocities have reached the steady distribution. The electronic behavior of the device may thus be dominated, even in steady-state conditions, by the properties of transient transport. In particular, since one of the typical properties of transient transport, as we shall shortly see, is a higher average velocity with respect to steady state in a bulk material, some devices try to exploit this *velocity overshoot* to realize a smaller transit time and therefore a faster operation of the device (faster switching or higher operation frequency).

The duration of the transient response is not known *a priori* and, in general, will be of the order of the longest of the characteristic times of the carrier system. This time may be called *transient-transport time* and depends upon the values of the applied field and temperature. It may roughly correspond to the energy relaxation time or to the time necessary for the repopulation of the different valleys.

A situation similar to that of transient transport may be experienced by charge carriers when the system is subject to an alternating field with frequency higher than the inverse of some electronic relaxation time ($\omega\tau > 1$) or with wavevector higher than the inverse of the electronic mean free path ($ql > 1$).

For space- or time-dependent problems the analytical solution of the BE is even more difficult than for homogeneous and stationary problems, while for MC simulation programs little work needs to be added to attack such problems, as will be discussed in next chapter.

As example of the behavior of transport in transient conditions, the mean velocity and mean energy of an ensemble of carriers in Si and GaAs are, obtained with MC simulations, shown in Figs. 13.7 and 13.8 as functions of the time elapsed after the instantaneous application of a static field [205]. For GaAs also the populations of the different valleys are shown.

In both cases, electrons have initially the equilibrium Maxwell distribution. For the case of silicon, they have been randomly distributed among the six equivalent valleys, while for GaAs they are all in the central valley.

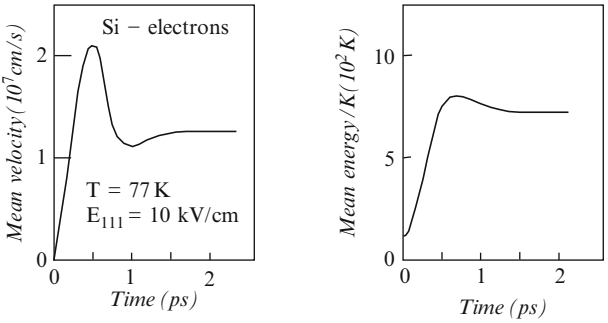


Fig. 13.7. MC results for the mean velocity and mean energy of electrons in Si at 77 K as functions of the time elapsed after the application of a field $E = 10\,\mathrm{kV/cm}$ along a $\langle 111 \rangle$ direction [205]

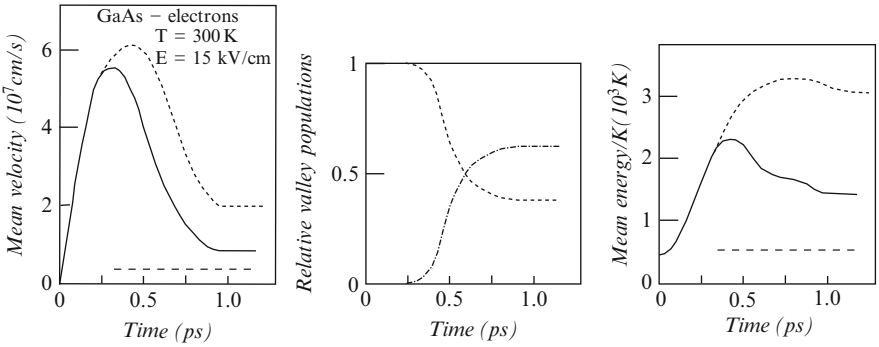


Fig. 13.8. MC results for the mean velocity, the relative valley populations, and the mean energy for electrons in GaAs at 300 K as functions of the time elapsed after the application of a field $E = 15\,\mathrm{kV/cm}$. Continuous curves refer to quantities averaged over all particles. Dashed and dot-dashed curves refer to quantities averaged over electrons in the central and upper valleys, respectively [205]

As can be seen in these figures, a general feature of high-field transient transport, starting from equilibrium conditions, is an overshoot of both the mean velocity $\langle v \rangle$ and the mean energy $\langle \epsilon \rangle$, before reaching their steady-state values. The overshoot of $\langle v \rangle$ is associated with the larger momentum relaxation times of "cold" electrons with respect to electrons heated up by the field. The overshoot of the mean energy is associated with an initial excess of absorbed power as effect of the higher mean velocity. In the case of GaAs, the transient transport is also influenced by the electron transfer to upper valleys, which further reduces average velocity and kinetic energy. In fact, as shown in Fig. 13.8, in correspondence with the population of the upper valleys, the mean energy and velocity in the central valley are higher than the corresponding total quantities because in the upper valleys $\langle v \rangle$ and $\langle \epsilon \rangle$

have values approximately constant and much lower than those of the central valley.

13.7 Hot Phonons

When charge carriers are driven far from their equilibrium conditions, the rates of phonon emission and absorption may be quite different from the detailed balance of thermal equilibrium at the lattice temperature. The phonon lifetime may not be short enough to maintain the phonon population at equilibrium, and in this case the distribution of phonons becomes an unknown of the problem, in the same way as the electron distribution. Since the emission and absorption of phonons by the electrons depend upon the phonons present in the sample, it is necessary to obtain electron and phonon distributions self consistently. This problem was present from the very beginning of the study of hot electrons. The phenomenon was named *hot-phonon problem* and observed experimentally with several techniques [107]. For the solution of such a problem it is necessary, first of all, to have a model for the phonon relaxation. This is due, finally, to the absorption at the thermal contact, but goes in general through intermediate steps, such as phonon–phonon interaction and reabsorption of phonons by the electrons. From the analytical point of view, already the BE with phonons at equilibrium is almost intractable. Thus, attempts to solve the coupled transport equations for electrons and phonons have to resort to even more severe approximations. The Monte Carlo approach has been applied also to the hot-phonon problem [53], mainly to account for phonon reabsorption in the electron relaxation. Even in this case, approximations have to be made, since it is very difficult to follow all the relaxation processes that phonons undergo after their emission.

13.8 Ultrafast Spectroscopy

Spectroscopy is a formidable tool to investigate electronic structures of condensed systems. From the late 1960s, optical measurements are also used for the investigation of the dynamics of free electrons out of equilibrium, and spectroscopy has become part of the experimental approaches to nonlinear transport [381,407]. In continuous-wave laser operation [408,453], radiation is sent on a sample, and the resulting photoluminescence is analyzed to investigate the distribution of electrons and holes generated by the laser beam. With the more powerful technique of sub-picosecond laser pulses [409], it is possible to perform a time-resolved analysis of the hot electron dynamics.

In optical measurements, hot electrons are produced by the radiation, as shown in Fig. 13.9, and in this way their initial distribution can be specifically selected. In the classical *pump-and-probe* experiments, a very short laser pulse of frequency ω, the *pump pulse*, whose duration is of the order of tens of fs,

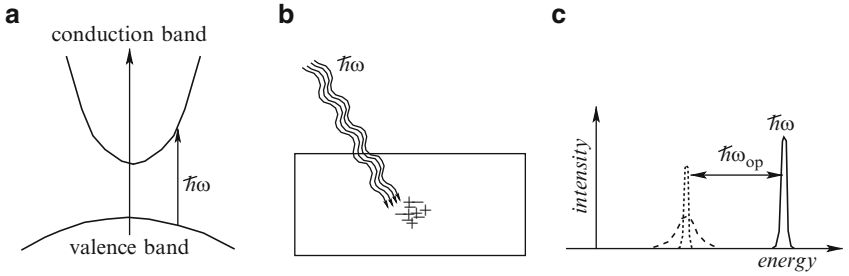


Fig. 13.9. The principles of pump and probe experiments for the analysis of the dynamics of electron energy relaxation. (**a,b**): a short laser pulse is absorbed by electrons in the valence band, thus generating an ensemble of electrons and holes; a fraction of a ps later a second pulse probes the electron distribution. In (**c**), the initial carrier distribution is shown with a continuous line; the distribution that would be obtained after the emission of an optical phonon by the electrons is shown as a dotted line, while the dashed line illustrates the additional effects of electron–electron interaction and short-time measurement

is sent on a semiconductor sample with appropriate wavelength. Absorbed by the electrons in the valence bands, the photons generate electron-hole pairs with reasonably well-defined energy.[6] Then, scattering processes relax the distribution functions toward their equilibrium values. These processes occur in a time scale of the order of several hundreds fs. A second *probe pulse* is sent to the sample with a delay, with respect to the pump sample, of the same order of magnitude. The reflection, or the absorption, of the probe pulse, measured as a function of the time delay of the probe pulse, yields information on the temporal evolution of the ensemble of nonequilibrium carriers generated by the pump pulse. In addition, it must be noted that coherent superpositions of states are generated by the laser radiation, and the dynamics of phase-related quantities can be analyzed [256, 381].

Figure 13.9 illustrates the basic principles of such experiments. Optical–phonon scattering is the main responsible of the electron relaxation, so that *phonon replicas* of the initial electron distribution are detected, as shown in part (c) of the figure. The measured distributions are spread by the quantum effect of short-time measurements [73, 206] and by carrier–carrier interaction. Since optical phonons have very small group velocities, phonons remain around for some time, and re-absorption slows down the energy relaxation of the electrons as a relevant hot-phonon effect [288].

[6] It must be remembered, however, that for the uncertainty relations the shorter is the pulse, the less defined is the energy of the photons. Thus the electrons generated in the conduction band are spread over an energy interval wider, when the pulse is shorter.

14

Monte Carlo Simulation of Bulk Electron Transport

14.1 The Monte Carlo Method

Monte Carlo (MC) methods may be defined as *that branch of experimental mathematics, which is concerned with experiments on random numbers* [178]. Even though it may seem strange, at first, that a well-defined mathematical result can be obtained by means of random numbers, the following classical example may immediately convince us that this is actually the case. Let us consider a square of side a with a circle inscribed, as shown in Fig. 14.1. Then generate pairs of random numbers (x, y), evenly distributed between 0 and a, to be used as coordinates of points inside the square. The expectation value of the fraction f of points falling inside the circle is given by the ratio between the area of the circle $\pi(a/2)^2$ and that of the square a^2, i.e., by $\pi/4$. The number π is then given by $4\langle f\rangle$. In the case of the figure, using ten points we obtain an estimate of π given by 2.8. It is clear that this estimate will be more and more precise as the number of samples increases, a typical property of MC calculations.

According to the definition above, MC approach was already used in remote times. In 1777, the French Encyclopedist Georges-Louis Leclerc, Compte de Buffon, considered the probability that a needle of a given length thrown on a floor with a grid of parallel lines will cut one of such lines. A few years later, Laplace proposed this method for an experimental determination of the number π. Statistical sampling has then been used in several occasions both for research and pedagogical purposes. The official birth of the Monte Carlo method, however, is considered the activity of the Los Alamos group for the development of nuclear weapons, and in particular the paper "The Monte Carlo method" by Metropolis and Ulam of 1949 [307]. The name first appeared in that paper and derives from the Monte Carlo casino, where roulettes generate random numbers. Today, it is applied regularly to practically all fields of natural, economical, and social sciences. Many books are available on this subject, as, for example, [40, 178, 230, 260].

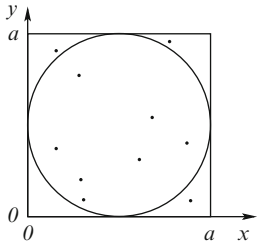


Fig. 14.1. Monte Carlo determination of the number π, see text

Even though the MC method is based on statistical sampling, it can be formulated and used for the solution of deterministic problems, such as the evaluation of integrals or the solution of algebraic or integral equations. When the problem at hand is statistical in nature, the MC method may consist in a direct simulation of the phenomenon under study. At times, the statistical problem to be solved may be formulated in terms of a deterministic equation for average quantities, as is the case of our BE. In such cases, we may apply the method either as a direct simulation of the phenomenon or by a formal solution of the equation. It may even be possible to solve the equation by inventing and simulating a new statistical phenomenon that has the same solution as the one of interest. This last possibility is often chosen to decrease the uncertainty, or variance, of the MC results.

The MC method was introduced to nonlinear electron transport by Kurosawa in 1966 [258] and was received by the hot-electron community with great enthusiasm. It was in fact clear that, with the aid of modern fast computers, it would become possible to obtain exact numerical solutions of the BE for microscopic physical models of considerable complexity. The technique was soon developed to a high degree of refinement by several authors, and in particular by Price [349] and by the Malvern group [49, 137, 361]. Since then the method has been applied to a great variety of materials and structures [213, 351] and it is now the most reliable approach to the simulation of electronic devices [187, 209, 227].

In the case of charge transport, the typical MC approach to the solution of the Boltzmann equation proves to be a direct simulation of the dynamics of charge carriers inside the crystal, so that, while the solution of the equation is being built up, any physical information required can be easily extracted from the simulation. In this respect it should be noted that, once a numerical solution of a given problem is obtained, its subsequent physical interpretation is still very important in gaining an understanding of the phenomenon under investigation. The MC method is a very useful tool toward this end, since it enables the simulation of particular physical situations unattainable in experiments, or even the investigation of nonexisting materials to emphasize special features of the phenomenon under study. This use of the MC method makes

it similar to an experimental technique; the simulated experiment can in fact be compared with analytically formulated theories.

However, as indicated above, the MC method may be also seen as a formal tool for the solution of mathematical problems, such as the evaluation of integrals or the solution of differential or integral (or integro-differential, as in our case) equations.

In the following sections, we shall first describe the MC approach in its standard formulation as direct simulation. Then we will show how the formal MC method may be applied to the solution of the Boltzmann equation. We shall find that this formulation provides a much larger flexibility to the method.

14.2 Direct Monte Carlo Simulation

The method consists of a simulation of the motion of one or more electrons inside the crystal, subject to the action of external forces due to applied electric and magnetic fields and of given scattering mechanisms. The duration of the carrier free flight, i.e., the time between two successive collisions, and the scattering events involved in the simulation are determined stochastically in accordance with given probabilities describing the microscopic processes. As a consequence, any MC method relies on the generation of a sequence of random numbers with given probability distributions. Such a technique takes advantage of the fact that any computer generates sequences of random numbers evenly distributed between 0 and 1 at fast rate. For reasons of space, we shall not discuss here the purely technical problem of the generation of random numbers with given probability distributions starting from random numbers evenly distributed in the interval between 0 and 1. We refer the reader to the specialized literature, as, for example, to [213].

When the purpose of the analysis is the investigation of a steady-state, homogeneous phenomenon without electron–electron interaction, it is in general sufficient to simulate the motion of one single carrier. From ergodicity, we may assume that a sufficiently long path of this sample carrier will give information on the behavior of the entire gas of particles. When, on the contrary, the transport problem under investigation is time or space dependent, then it is necessary to simulate a large number of carriers and follow them in their dynamical histories in order to obtain the desired information on the system of interest. When used in this version, the method is in general called *ensemble Monte Carlo* (EMC) [272, 273], and in space-dependent problems, as for device simulation, it must be coupled to Poisson equation.

Remember that in semiclassical approximation both momentum and position of a particle may be defined, even though for the simulation of homogeneous systems it is not necessary to keep track of the carrier positions.

14.2.1 A Typical Monte Carlo Program for Homogeneous, Stationary Transport

Let us summarize here the structure of a typical Monte Carlo program suited to the simulation of a stationary and homogeneous transport process, in presence of an external electric field $\boldsymbol{E}$. A flowchart is shown in Fig. 14.2. The details of each step of the procedure will be given in the following. The simulation starts with one electron in given initial conditions with wavevector $\boldsymbol{k}_\circ$. Then the duration of the first free flight is chosen with a probability distribution determined by the scattering probabilities.

During the free flight, the external forces are made to act according to the semiclassical dynamics $\hbar\dot{\boldsymbol{k}} = \mathrm{e}\boldsymbol{E}$. The force may of course include also the effect of a magnetic field. For reasons of space, however, this case will not be treated here. In this part of the simulation all quantities of interest, namely velocity, energy, etc., are recorded. Then a scattering mechanism is chosen as responsible for the end of the free flight, according to the relative probabilities of all possible scattering mechanisms. From the differential cross section of the selected mechanism a new $\boldsymbol{k}$ state after scattering is randomly chosen as initial state of the new free flight, and the entire process is iteratively repeated. The results of the simulation become more and more precise as the simulation goes on, and the simulation ends when the quantities of interest are obtained with the desired precision.

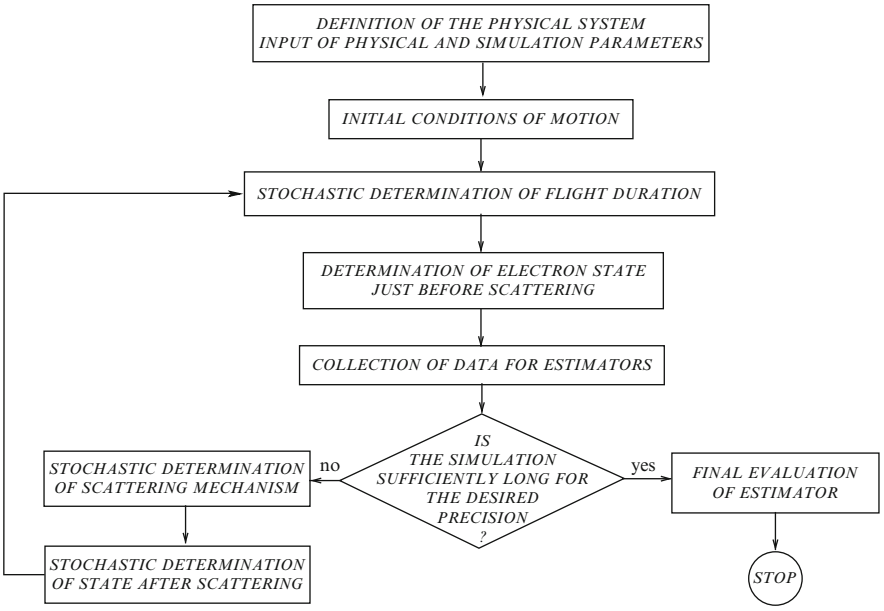


Fig. 14.2. Flowchart of the simplest one-particle Monte Carlo program for homogeneous, steady-state systems

A simple way to determine the precision, i.e., the statistical uncertainty, of transport quantities consists of dividing the entire history into a number of successive subhistories of equal time durations, and making a determination of the quantity of interest for each of them. The average value of this quantity is then evaluated, and its standard deviation is an estimate of its statistical uncertainty.

Figure 14.3 illustrates the principles of the method by showing the simulation in $\boldsymbol{k}$ space and real space and the effect of collecting statistics in the determination of the drift velocity.

Definition of the Physical System

The starting point of the program is the definition of the physical system of interest, including the parameters of the material and the values of physical quantities, such as lattice temperature and applied field. At this level, we also define the parameters that control the simulation: the duration of each sub-history, the desired precision of the results, and so on.

The next step in the program is a preliminary calculation of each scattering rate as a function of electron energy. This step will provide information on the the maximum values of these functions, which will be useful for optimizing the efficiency of the simulation (see below).

Initial Conditions of Motion

In the case under consideration, in which a steady-state situation is simulated, the length of simulation must be large enough for the evaluation of average quantities based on ergodicity. Thus, the initial conditions of the electron motion do not influence the final results. When the simulation is divided into many subhistories, the initial state of each new subhistory is conveniently taken equal to the final state of the previous one.

When the simulation is made to study a transient phenomenon and/or a transport process in a nonhomogeneous system (for example, when the electron transport in a device is analyzed), it is necessary to simulate many electrons at the same time; in this case, the distribution of the initial electron states for the particular physical situation under investigation must be taken into account, and the initial transient may become an essential part of the results aimed at.

Free-Flight Duration – Self-Scattering

The electron wavevector $\boldsymbol{k}$ changes continuously during a free flight because of the applied field. Thus, if $P(\boldsymbol{k}(t))\,\mathrm{d}t$ is the probability that an electron in state $\boldsymbol{k}$ suffers a collision during $\mathrm{d}t$ around t, the probability that an electron, which suffered a collision at time $t = 0$ has not suffered another collision after

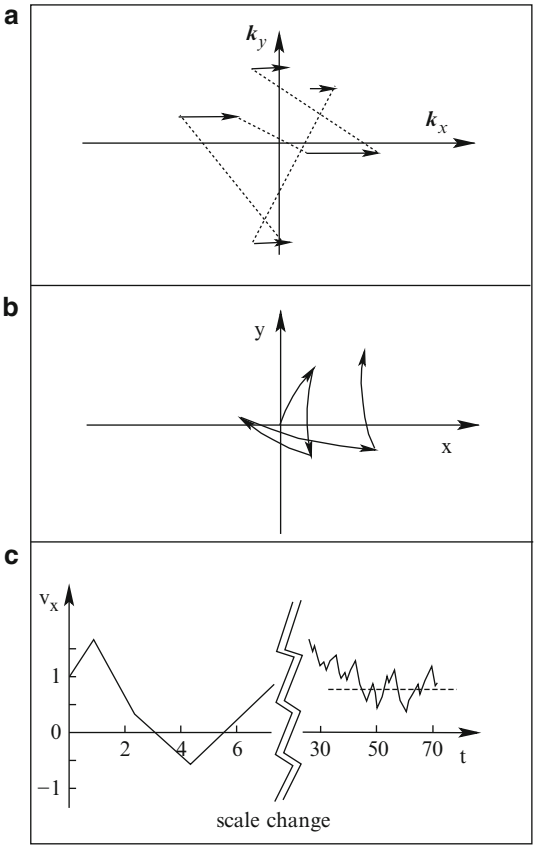


Fig. 14.3. The principles of the MC method. For simplicity a two-dimensional model is considered. (**a**) Trajectory of a simulated particle in $\boldsymbol{k}$ space, subject to an accelerating force oriented along the positive x direction. The heavy segments are due to the effect of the force during free flights; dotted lines represent the discontinuous variations of $\boldsymbol{k}$ due to scattering processes. (**b**) Path of the particle in real space. It is composed of fragments of parabolas corresponding to the free flights in (**a**). (**c**) Average velocity of the particle as a function of simulated time. The left part of this line corresponds to the above parts of the figure; the horizontal dashed line represents the drift velocity obtained with a very long simulation. All units are arbitrary [213]

a time t is

$$\exp\left[-\int_0^t P(\boldsymbol{k}(t'))\,\mathrm{d}t'\right].$$

Consequently, the probability that the electron will suffer its next collision during $\mathrm{d}t$ around t is given by

$$\mathcal{P}(t)\,\mathrm{d}t = \mathrm{e}^{-\int_0^t P(\boldsymbol{k}(t'))\,\mathrm{d}t'} P(\boldsymbol{k}(t))\,\mathrm{d}t. \tag{14.1}$$

The analytical forms of the scattering rates $P(\boldsymbol{k})$ are not in general very simple so that it is impractical to generate stochastic free-flight durations with the distribution (14.1): an integral equation would need to be solved for each scattering event. Rees [360, 361] has devised a very simple method to overcome this difficulty. If $\Gamma = 1/\tau_\circ$ is the maximum value of $P(\boldsymbol{k})$ in the region of $\boldsymbol{k}$ space of interest, a new fictitious *self-scattering* is introduced such that the total scattering rate, including this self-scattering, is constant and equal to Γ. If the carrier undergoes such a self-scattering, its state $\boldsymbol{k}'$ after the collision is assumed to be equal to its state $\boldsymbol{k}$ before the collision, so that in practice the electron path continues unperturbed as if no scattering at all had occurred. Now, with a constant $P(\boldsymbol{k}) = \tau_\circ^{-1}$, (14.1) reduces to

$$\mathcal{P}(t) = \frac{1}{\tau_\circ} \mathrm{e}^{-t/\tau_\circ}, \tag{14.2}$$

and random numbers r, evenly distributed between 0 and 1, can be used very simply to generate stochastic free-flight durations t_r [213]. They will be given by

$$t_r = \tau_\circ \ln(r).$$

If the scattering rate $P(\boldsymbol{k})$ has a large variation in the region of energies of interest, the value of Γ may be taken a stepwise function of energy [213].

Choice of the Scattering Mechanism

During the free flight, the electron dynamics is governed by semiclassical dynamics, and at its end the electron wavevector and energy are known. All scattering rates $P_i(\epsilon)$ can be evaluated, where i indicates the i-th scattering mechanism. The probability of self-scattering is the complement to Γ of the sum of the P_i's. A mechanism must then be chosen among all the possible ones: given a random number r, evenly distributed between 0 and 1, the product $r\Gamma$ is compared with the successive sums of the P_i's, and a mechanism is selected as indicated in Fig. 14.4.

Choice of the State After Scattering

Once the scattering mechanism that caused the end of the free flight has been determined, the new state after scattering, $\boldsymbol{k}_a$, must be chosen as final state of the scattering event. If the free flight ended with a self-scattering, $\boldsymbol{k}_a$ must be taken as equal to $\boldsymbol{k}_b$, the state before scattering. When, instead, a true scattering occurred, $\boldsymbol{k}_a$ must be generated stochastically, according to the differential cross section $P(\boldsymbol{k}, \boldsymbol{k}')$ of that particular mechanism, given in Chap. 9. How to practically realize this step is a technicality outside of the scope of this book, and the interested reader is referred to the specialized literature [213].

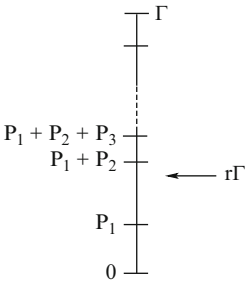


Fig. 14.4. Selection of the scattering mechanism. $r\Gamma$ is a random number evenly distributed between 0 and Γ and it is compared with the successive sums of the scattering rates. If the first sum larger than $r\Gamma$ is $P_1 + P_2 + \ldots + P_j$, the j-th mechanism is selected. In the case of the figure, the second scattering mechanism is chosen to act

Time Averages

In the simulation of a stationary and homogeneous phenomenon, we may obtain the average value of a quantity $\mathcal{A}(\boldsymbol{k}(t))$ (e.g., drift velocity, mean energy, etc.) during a single history of duration T as

$$\langle \mathcal{A} \rangle_T = \frac{1}{T} \int_0^T \mathcal{A}(\boldsymbol{k}(t))\, \mathrm{d}t = \frac{1}{T} \sum_i \int_0^{t_i} \mathcal{A}(\boldsymbol{k}(t'))\, \mathrm{d}t', \tag{14.3}$$

where the integral over the whole simulation time T has been separated into the sum of integrals over all free flights of duration t_i. T should be taken sufficiently long that $\langle \mathcal{A} \rangle_T$ in (14.3) represents an unbiased estimator of the average of the quantity $\mathcal{A}$ over the electron gas.

In a similar way, we may obtain the electron distribution function: a mesh of $\boldsymbol{k}$ space (or of energy) is set up at the beginning of the computer run; during the simulation, the time spent by the sample electron in each cell of the mesh is recorded, and, for large T, this time conveniently normalized will represent the electron distribution function, that is, the solution of the Boltzmann equation [137].

Synchronous Ensemble

Another method of obtaining an average quantity $\langle \mathcal{A} \rangle$ is the so-called *synchronous-ensemble method*, introduced by Price [349, 350]. The method is illustrated in Fig. 14.5. With a constant scattering rate (including self-scattering), the distribution of electron states before each scattering is equal to the distribution of electron states at a given time t. Thus, the mean quantity $\langle \mathcal{A} \rangle$ can be evaluated as the average of the values assumed by $\mathcal{A}$ at the end of each free flight:

$$\langle \mathcal{A} \rangle = \frac{1}{N} \sum_i \mathcal{A}_{bi}, \tag{14.4}$$

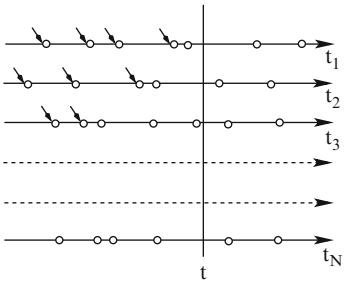


Fig. 14.5. Illustration of the synchronous ensemble. The horizontal lines represent the time axes of the different particles, and circles represent the scattering times. If the flights are generated with a constant Γ, the distribution of electron states before scattering (some of which are indicated by arrows) is equal to the distribution of states at a given time t. The average time between collisions is τ; the vertical line at the observation time t cuts flights of average duration 2τ

where the sum covers all N free flights, and $\mathcal{A}_{bi}$ indicates the value of $\mathcal{A}$ at the end of the i-th free flight, i.e., immediately before the i-th scattering event.

This result may seem, at first, strange. In fact, the states just before the scattering events seem to be influenced more than average by the applied field, since the latter had the whole flight to act. However, one should considered that, while the mean in (14.4) weights equally all free flights, short and long ones, with average duration τ, when an instantaneous picture of the electron gas is taken at a time t, longer free flights are more likely to be caught. In other words, in the latter case the vertical line in Fig. 14.5 crosses free flights whose mean duration is longer than the average over all free flights; in fact, the distribution of the hemi-flights on the right and on the left of the line t reproduces the distribution of flight durations, so that the average length of the flights crossed by t is 2τ. The distribution of durations between the last scattering event and t is the same as that of all flights, and this is the reason for the validity of (14.4). The above argument shows also that the synchronous ensemble method, in the simple form just described, is applicable only when a constant Γ is used, with the inclusion of self-scattering. If a variable Γ is used, a variation of the synchronous ensemble method must be employed [213,351].

14.2.2 Time- and Space-Dependent Phenomena – Ensemble MC

Little work needs to be added to MC programs to attack time- and/or space-dependent problems. This use of the MC method is particularly important for the analysis of small devices (see Sect. 18.7.3), where it is often necessary to consider both the transient dynamic response to voltage changes and the electronic behavior at different points of the device.

Transients

Let us first consider the case of a homogeneous electron gas with time-dependent behavior. In particular, it is of interest to study the transient dynamic response to a sudden change in the value of an applied field. In this situation, we cannot rely on ergodicity and many particles must be independently simulated with the appropriate distribution of initial conditions. Provided the number of simulated particles is sufficiently large, the average value of a quantity of interest, obtained on this sample ensemble as a function of time, will be representative of the average over the entire gas. Instead of the time average in (14.3) or the synchronous ensemble average in (14.4), now the ensemble average

$$\langle \mathcal{A}(t) \rangle = \frac{1}{N} \sum_i \mathcal{A}_i(t)$$

must be used, where i now runs over the N simulated particles.

The "transient-transport time", as discussed in Sect. 13.6, is not known *a priori* and is of the order of the largest of the characteristic times of the electron system. In general, it depends upon the values of the applied field and temperature.

The transient dynamic response obtained by means of the simulation depends obviously upon the initial conditions of the carriers, which must be assumed according to the situation to be explored.

To determine the precision of the results, subensembles of electrons can be considered. The quantity of interest $\mathcal{A}$ is evaluated in each subensemble. Their average value and standard deviation can then be taken as the most probable value and the statistical uncertainty of $\mathcal{A}$, respectively.

Periodic Fields

Even though the application of a periodic field is associated with a time-dependent phenomenon, its analysis with an MC procedure may be performed without resorting to an ensemble of particles. [271, 490]. If a field

$$\boldsymbol{E} = \boldsymbol{E}_\circ + \boldsymbol{E}_1 \sin(\omega t) \tag{14.5}$$

is applied, and the ac term is small enough to be in the linear-response regime, the average electron velocity will be of the form

$$\langle \boldsymbol{v}(t) \rangle = \boldsymbol{v}_\circ + \boldsymbol{v}_1 \sin(\omega t) + \boldsymbol{v}_2 \cos(\omega t).$$

The coefficients $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ can be obtained as sine and cosine Fourier transforms of the velocity of the simulated electron over its history. Since the equation of motion of a particle subject to a field given by (14.5) is known in explicit terms [271], the free flight between scattering events is easily obtained, and the simulation may be realized as for a constant field.

For large periodic fields, outside the linear response regime, the periodic part of the current will contain higher harmonics, which can also be obtained by Fourier analysis of the simulated velocity.

It is also possible to obtain the total response of the electron gas as follows [490]: let us divide the period $2\pi/\omega$ in N parts of duration Δt and "read" from the simulation the electron velocity at times given by $0,\ \Delta t,\ 2\Delta t,\ \ldots\ ; \Delta t = (1/N)(2\pi/\omega)$. The average values of $\boldsymbol{v}$ obtained at times

$$n\Delta t,\ (n+N)\Delta t,\ (n+2N)\Delta t,\ \ldots$$

is an estimator of the average electron velocity, which is a periodic function of t with the same period $2\pi/\omega$, at the times indicated above. A successive Fourier analysis may yield the amplitudes of the different harmonics.

Space-Dependent Phenomena

The simulation of a steady-state phenomenon in a physical system where electron transport depends upon the position in space is of particular interest for the analysis and modeling of devices. This subject will be treated in Sect. 18.7.3 and it is not really pertinent to this chapter, devoted to bulk transport. We simply mention that also in this case an ensemble of particles must be used, and averages must be taken over particles at given positions.

Space- and time-dependent phenomena may present similar features. For instance, if a field is suddenly switched on from zero to a large value, electrons experience a situation similar to that of electrons entering from a metallic contact into a device where a large field is present. However, different averaging procedures must be appropriately considered in the two cases.

14.2.3 Diffusion

Diffusion may be considered a special, important case of a space-dependent phenomenon which is, in general, also time dependent. In the linear-response regime, diffusivity D and mobility μ are related by the Einstein relation (12.5) of Chap. 12, where we have discussed the general problem of diffusion, fluctuations, and noise. It was also noted, there, that at high fields the Einstein relation fails, and the study of hot-electron diffusion yields independent information.

The diffusion coefficient may be determined in an MC simulation by means of (12.13), which describe the spreading of a bunch of particles due to diffusion: a number of particles is independently simulated and their position are recorded at fixed times. For large enough simulation times, the second central moment shows the linearity predicted by (12.13), and from its slope D is obtained. Particular care must be put into the initial spreading of the particles, which does not follow the diffusion equation, as discussed at the end of Sect. 12.3.

In hot-electron conditions, the diffusion coefficient shows its tensor nature, and the various components can be obtained from the simulation as

$$D_{xy} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle (x - \langle x\rangle)(y - \langle y\rangle)\rangle.$$

The diffusion coefficient can also be determined from an MC simulation through the evaluation of the autocorrelation function of velocity fluctuation, as given by (12.16) and (12.17). Finally, MC can be used to evaluate a diffusion coefficient $D(q, \omega)$ to be used when the mean free time and mean free path are shorter than the time or space variations of the concentration, respectively [203, 215].

14.2.4 Ohmic Mobility

When the MC method is used to obtain the drift velocity of charge carriers at low applied fields, the statistical uncertainty originating from thermal motion may become particularly bothersome. To simulate the linear-response mobility, however, it is possible to evaluate the diffusion coefficient at zero field with one of the method presented above and then obtain the Ohmic mobility by means of the Einstein relation. It is worth noting that when no external field is applied, the energy, and therefore the scattering probability, of the particle is constant during a free flight, so that it is not necessary to introduce self-scattering.

As it regards MC simulation at low fields, a note is in order on acoustic-phonon scattering and its elastic approximation. When Ohmic transport is investigated by analytical means, the energy distribution function is the equilibrium distribution (Maxwellian, in case of nondegenerate statistics), and no energy exchange of the electrons with the heat bath is explicitly required. On the contrary, when an MC simulation is undertaken at low fields and temperatures, to obtain a correct energy balance in steady-state conditions, we need a mechanism that can exchange an arbitrary small amount of energy between electrons and the heat bath (the host crystal). Physically, this role is played by the interaction with acoustic phonons. Thus, considering elastic this mechanism is, in general, illegitimate, because in this case acoustic-phonon interaction would never produce a steady-state condition: the power transferred by the applied field would increase the carrier mean energy indefinitely. Furthermore, for a precise energy balance the exact phonon population N_q or, at least, a good approximation of it must be used. When, in contrast, high fields and/or high temperatures are considered, acoustic scattering can be treated as an elastic process, since the average electron energy is of the order of the optical–phonon energy, and optical phonons can assume the task of exchanging energy between the electrons and the crystal. In this case, if, as usually done, the optical–phonon energy is assumed constant, the presence of the external field is essential for smearing out the energy of each single electron through the acceleration process. In fact, in the absence of an external

field, the electron energy would take only its initial value plus or minus an integer number of optical–phonon energy quanta.

14.2.5 Electron–Electron Interaction and Degenerate Statistics

Coulomb interaction between charged particles may be split into two terms: a collective long-range interaction dealt with by Poisson equation, and electron–electron (e–e) short range collisions, usually screened by all other particles. Here we are interested in the latter form of interaction. As it regards Poisson equation, it will be considered in the chapter devoted to devices.

We have already noted in Chap. 9 that in interparticle collisions the total momentum and the total energy of the two colliding particles is conserved, and no dissipation occurs. Thus, this type of interaction does not usually affect transport properties in semiconductors to a large extent. Momentum and energy are, however, redistributed among the particles so that the shape of the distribution function $f(\boldsymbol{k})$ is influenced by e–e interaction. A typical result is shown in Fig. 14.6a. We saw already in Sect. 13.2 that this fact has been used for stating that at high electron densities $f(\boldsymbol{k})$ assumes a Maxwellian shape also far from equilibrium, characterized by a mean drift velocity $\boldsymbol{v}_{\mathrm{d}}$ different from zero and an electron temperature T_{e} higher than the lattice temperature $T_{\circ}$.

Figure 14.6 shows the typical effect of e–e interaction on the distribution function. It is interesting to note that the distribution obtained when the simulation contains only phonon scattering shows a kink at the energy of optical phonons. Above it optical–phonon emission makes the dissipation much more effective, and the slope of the distribution function is essentially the same as that of equilibrium, while below that energy the distribution is much "hotter". This phenomenon has sometimes suggested to use a "two-temperature" model. When e–e is included, the distribution function is considerably smoothed.

From the above considerations, it is clear that e–e interaction affects primarily the transport quantities which are more sensitive to the particular shape of the distribution function, such as valley repopulation, impact ionization, oxide penetration and tunneling.

The inclusion of e–e scattering in MC simulation is not simple, since the scattering probability itself depends on the distribution function, through both the screening of the Coulomb scattering potential and the probability of the sampling electron to collide with another electron of a given momentum. A self-consistent calculation must therefore be performed in which the distribution $f(\boldsymbol{k})$ used to evaluate the scattering probability is the same which results as solution. This is often done by collecting the distribution function during the simulation. Then, the distribution obtained in the previous simulation is used to evaluate the screening, and one of the previous electron states is chosen at random as the scattering electron. The distribution function is periodically updated until convergence is obtained.

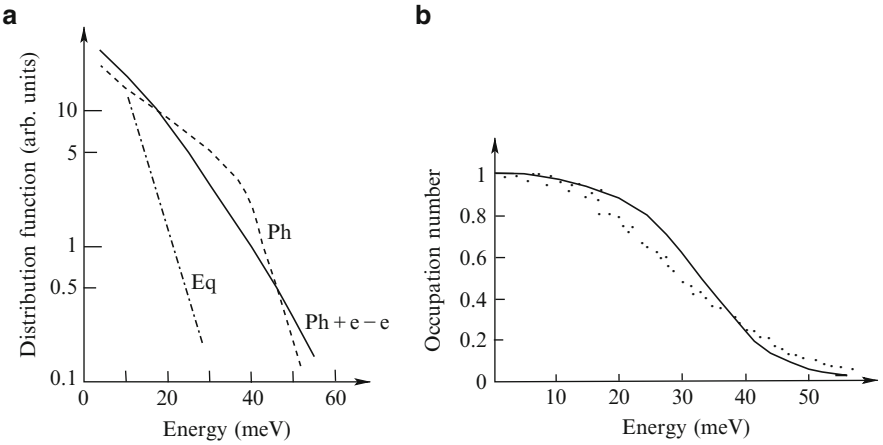


Fig. 14.6. Effects of electron–electron interaction and of Pauli exclusion principle. In (**a**) the distribution functions are shown obtained with MC simulations of a simple silicon model with only phonon scattering (*dashed line*) and with the addition of e–e collisions (*continuous line*) with a concentration of $10^{17}\,\mathrm{cm}^{-3}$ ($T = 45\,\mathrm{K}$, $E = 300\,\mathrm{V/cm}$. The dash-dotted line indicates the equilibrium distribution [204]. In (**b**), the occupation number as a function of energy obtained with an MC simulation (*dots*) for GaAs at 77 K in degenerate conditions, with a concentration $5 \times 10^{17}\,\mathrm{cm}^{-3}$ and a field $E = 900\,\mathrm{V/cm}$. The continuous line indicates, for comparison, the equilibrium Fermi distribution [60]

Another approach sometimes used to include e–e interaction in MC simulation is a combination of the MC technique with the molecular dynamics approach [204]. Many electrons are simulated at the same time, as in the standard ensemble MC method, but the Coulomb interaction among them is included in the determination of the orbits during the free flights. The results shown in Fig. 14.6a above have been obtained with this technique. Difficulties arise, however, due to the long-range nature of Coulomb interaction.

The Pauli exclusion principle, too, is a sort of e–e interaction which brings about nonlinear terms in the Boltzmann equation and requires some self consistent procedure. To account for Pauli principle in the MC approach the distribution function $f(\boldsymbol{k})$ obtained at the current time of the simulation is used to correct all the scattering probabilities by a factor $(1-f(\boldsymbol{k}))$. To include this factor, the rejection technique is employed after the final state $\boldsymbol{k}_a$ has been selected [60]. Again the simulation must continue until convergence of $f(\boldsymbol{k})$ is attained. The results shown in Fig. 14.6b above have been obtained with this technique. This method has been applied to device simulations when the carrier concentration implies degenerate statistics [290].

14.2.6 Impact Ionization

If a carrier gains enough energy from the applied field, it may extract an electron from the valence band and promote it into the conduction band, creating in this way an extra electron-hole pair. This process, called *impact ionization* [402, 431], is of particular importance in devices since, in presence of a high field, it can initiate an avalanche, a mechanism that may lead to breakdown and that is used in avalanche diodes.

Impact ionization has soon been accounted for in ensemble MC simulation [272] by introducing the probability of such event as an independent scattering mechanism and it has often been used in MC approaches to both material and devices analyses (see for example [77, 393, 394, 442]). In [272], after each ionization process, the minority carrier is neglected and one of the resulting $(N + 1)$ particles, chosen at random, is eliminated to maintain the sample size fixed. The pair generation rate g_I per particle per unit time is obtained by counting ionization events and the impact-ionization rate is obtained as $\alpha_I = g_I / v_\mathrm{d}$.

14.2.7 Variance-Reducing Techniques

Since MC is a statistical procedure, the results obtained by means of such a method are always affected by some statistical uncertainty. In previous sections, it was indicated how to evaluate this uncertainty. As a rule, the statistical precision of the results increases as the square root of the number of trials, i.e., of simulated events. Therefore, the amount of computer time necessary to improve appreciably the quality of the results becomes quickly very long. It is thus of particular interest to find *variance-reducing techniques*, which can be defined as procedures which *change or at least distort the original problem in such a way that uncertainty in the answer is reduced* [182] without affecting the correctness of the results. Some variance-reducing techniques specifically applied to electron transport simulation are briefly presented in what follows. Others can be found in [213].

The first variance-reducing technique, introduced to reduce the variance of the drift velocity due to thermal fluctuations [177, 259], is a sort of *antithetic variate* [182]. When the selected scattering mechanism is velocity randomizing, both states after scattering $\boldsymbol{k}_a$ and $-\boldsymbol{k}_a$ are considered. With the next random number both electrons perform the free flight; the average of the quantities evaluated in the two flights is recorded and, finally, one of the two electrons is chosen at random to continue the simulation. In this way, the fluctuation due to the two different actual velocities is canceled, and only the drift term due to the field is retained.

When a physical phenomenon of interest is due to the occurrence of improbable electron states, the standard MC simulation may lead to a large variance of the desired quantity. In this case, the variance can be reduced by the following *splitting procedure* [338]. During a simulation, when the sample

electron enters a given "rare region", its entering state is recorded and used for generating a number N of different parallel simulations; each of these uses different random numbers and ends when the electron exits from the rare region. Then, starting from one of the N exit states chosen at random, the simulation proceeds in the usual way until the rare region is reached again. In the averaging procedure, a weight $1/N$ must be given to each parallel simulation.

The variance-reducing techniques above may be very useful in some cases. It must be remembered, however, that they are realized by means of a distortion of the original phenomenon, which causes the program to deviate from the strict simulation of possible electron histories. In particular, noise and diffusion, due to velocity fluctuations, are heavily distorted. A strict simulation, on the contrary, has the advantage of yielding a simple and straightforward physical interpretation of the phenomenon under investigation. Consequently, as a rule, a correct balance between computing time, transparency of the simulative procedure, and complexity of the computer program must be found in connection with the particular needs of each single case.

14.2.8 Full-Band Monte Carlo

During the first decades of its application, the MC approach has enormously widened the horizon of problems that could be solved within the semiclassical approximation of electron transport in both semiconductor materials and devices. This progress quickly made people desire to extend the physical models beyond what was used until then. In particular, the ever smaller dimensions of the physical systems investigated required on one side to account for space quantization of the electronic states, and on the other side to analyze the effects of very high energies reached by the electrons in regions where very intense electric fields are present. This latter effect arose the problem of extending transport theories in the direction of a more rigorous quantum framework. But it also indicated the need to account for the band structure much better that previously done with parabolic bands or with the simple analytical models of nonparabolicity (see Sect. 8.7.3) and of many-valley models.

The density of states resulting from full-band calculations, as shown in Figs. 6.4 and 6.5, is very different from that of a many-valley model, even accounting for nonparabolicity. For electrons in silicon, the difference becomes increasingly important above about 2 eV. Since transition rates are dominated by the density of states, the simulations performed with many-valley models become totally unreliable when the resulting electron energies approach this value. This happens for fields of the order of 10^5–10^6 V/cm (see, for example [148, 257]).

After some investigations on the effect of the inclusion of the full-band structure on impact ionization processes in GaAs [413] and in Si [434], a paper was published by Fischetti and Laux [148] that became the standard reference for the so-called *full-band Monte Carlo simulation*.

In full-band MC, the band structure of the material of interest is evaluated, usually with a pseudo-potential method (see Sect. 6.4.3), in a grid of points that covers the whole BZ, and stored in a look-up table. During the simulation, when the band and its gradient are required at a precise value of the electron momentum, appropriate interpolation schemes are adopted. For low values of the electron energy, a parabolic approximation is conveniently used.

As it regards the free flights, they are still governed by the semiclassical equations

$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \frac{1}{\hbar}\nabla_k \epsilon(\boldsymbol{k}) \quad , \qquad \frac{\mathrm{d}(\hbar\boldsymbol{k})}{\mathrm{d}t} = (-\mathrm{e})\boldsymbol{E}.$$

However, since now the band is known only numerically, the above equations cannot be integrated analytically, and the dynamics must be solved by finite differences. This is not a serious problem since in ensemble MC for device simulations very small intervals of time, of the order of 10^{-16} s, are used to resolve plasma oscillations (see Sect. 18.7.3).

In the evaluation of the scattering rates to be used in full-band MC, look-up tables are again used where scattering rates are memorized as functions of initial and final states of the transitions. In fact the scattering rates, treated in Chap. 9, contain a δ-function of energy conservation. When the analytical form of the band is known, as in standard simple MC simulations, this δ function is used to obtain the total scattering probability for each scattering mechanism, and the integration contains the information of the density of states in energy, brought about by the integration through the delta function[1]. When, on the contrary, the band is known only numerically, the δ-function cannot be used for analytical integration. A numerical algorithm is then used, suggested in [164]. For each cell of the grid in the BZ, it is first analyzed if an intersection exists of the energy surface of energy conservation internal to that cell. The density of states available as final states for a scattering process is then given by the area of this interception, divided by the group velocity pertaining to that cell and multiplied by the density of states in $\boldsymbol{k}$-space $V/(2\pi)^3$ (see (8.24)). This evaluation is repeated for all possible final states in the grid and for all scattering mechanisms. The results are again stored in look-up tables.

When the electron energies are such that a full-band MC is appropriate, acoustic phonons involved in the transitions may have very large wavevectors, corresponding to a significant fraction of the BZ. Thus, the dispersion for the phonon branches cannot be approximated by the linear relation $\omega = qv_s$ and better approximations of the real acoustic dispersion must be used.

Figure 14.7 shows schematically the electron distribution inside the BZ of silicon for several values of the applied field. The low-field case, at room temperature, may be valid up to about 10^3 V/cm; the intermediate case up to about 10^5 V/cm; above this field a full-band MC should be considered.

[1] When $\delta(f(x))$ is integrated with respect to x, the absolute value of the derivative of f appears in the denominator (see (A.22) in Appendix A). If the function f is the band, this yields exactly the density of states (see (8.24)).

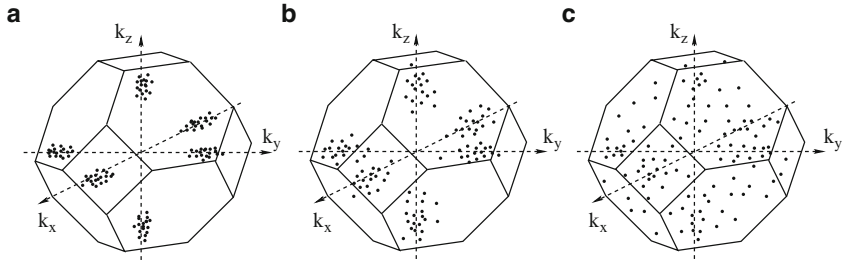


Fig. 14.7. Schematic representation of the distribution of electrons in the BZ of silicon. At low fields (**a**) electrons occupy the bottom of the six valleys and the intervalley model is representative of the real situation; at intermediate fields (**b**) electrons reach higher energies, but it is still possible to identify the valleys they are in; at very high fields (**c**) they are distributed in the entire BZ, and a full-band model is necessary

In the above pages, we could treat only the main principles and features of MC simulation. The interest reader is of course referred to the specialize literature. Let us simply add that, since this approach is very computer-time consuming, several methods have been developed (see, for example [226, 250, 395]) to increase the efficiency of the approach, without losing too much in the correctness of the results.

14.3 Formal Monte Carlo Solution of the BE – Weighted Monte Carlo

As already mentioned, the direct simulation of electron transport, is only one possible application of the MC technique. A much more general approach consists in a formal theory of MC solution of integral equations [211, 322, 323, 380, 382]. In the following, this formal theory is presented, and it is shown that the direct simulation considered in the previous pages may be generalized to arbitrary choices of the probabilities of all possible events, with possible variance-reduction applications.

14.3.1 Monte Carlo Evaluation of Sums and Integrals

In this section, a few mathematical techniques for MC evaluations of sums and integrals is reviewed. They will be used later for the development of the MC approach to electron transport, in both semiclassical and quantum theories.

MC Evaluation of a Sum

Given the sum

$$S = \sum_i a_i,$$

that may also contain an infinite number of terms, a possible MC algorithm for its evaluation is the following: a set of arbitrary probabilities p_i are defined, subject to the conditions

$$p_i \geq 0 \quad (>0 \ \text{ if } \ a_i \neq 0) \ , \quad \sum_i p_i = 1.$$

Then, terms a_i are selected at random with probabilities p_i, and the estimator

$$s = \frac{a_i}{p_i}$$

is evaluated. This is a correct estimator of the sum S. In fact, its expectation value is

$$\langle s \rangle = \sum_i p_i \frac{a_i}{p_i} = S.$$

Generalization to a Number of Sums

If, instead of a single sum S, we have to evaluate a set of sums

$$S_k = \sum_i a_{ki}, \tag{14.6}$$

a very similar procedure can be followed. A a set of arbitrary probabilities p_{ki} are defined, subject to the conditions

$$p_{ki} \geq 0 \quad (>0 \ \text{ if } \ a_{ki} \neq 0) \ , \quad \sum_{ki} p_{ki} = 1.$$

Terms a_{ki} are then selected at random with probabilities p_{ki} and the estimators

$$s_j = \frac{a_{ki}}{p_{ki}} \delta_{kj}$$

are evaluated, where δ_{kj} is the Kronecker symbol. These are correct estimators of the sums S_j since their expectation values are

$$\langle s_j \rangle = \sum_{ki} p_{ki} \frac{a_{ki}}{p_{ki}} \delta_{kj} = \sum_i a_{ji} = S_j.$$

Let us point out that the selection of a single term of the matrix a_{ki} yields an estimate of all the sums in (14.6): This estimate is a_{ki}/p_{ki} for the k-th sum and zero for all the other sums.[2]

[2] An example may clarify the procedure. Let us evaluate the three sums

$$S_1 = 1 + 2 + 3 + 4$$
$$S_2 = 8 - 7 + 6 - 5$$
$$S_3 = 1 - 2 + 3 - 4$$

If we assign to all terms equal probabilities $p_{ij} = 1/12$, and the term $i, j = 3, 2$ has been extracted, then the estimates are $S_1 = 0,\ S_2 = 0,\ S_3 = -2/(1/12) = -24$.

Generalization to Integrals

Another generalization of the above algorithm is obtained by the substitution of the discrete sum by a continuous integral. Let us assume that we have to evaluate the integral

$$F = \int_{y_1}^{y_2} g(y)\,\mathrm{d}y.$$

We then define an arbitrary probability density $p(y)$ subject to the conditions

$$p(y) \geq 0 \quad (> 0 \ \text{ if } \ g(y) \neq 0) \ , \quad \int_{y_1}^{y_2} p(y)\ \mathrm{d}y = 1.$$

With such a probability density, we generate a value y' and evaluate the estimator

$$f = \frac{g(y')}{p(y')}.$$

This is a correct estimator of the integral F since its expectation value is

$$\left\langle \frac{g(y')}{p(y')} \right\rangle = \int_{y_1}^{y_2} p(y')\ \frac{g(y')}{p(y')}\ \mathrm{d}y' = \int_{y_1}^{y_2} g(y)\ \mathrm{d}y = F.$$

The extension to an integral function $F(x)$ is again straightforward. If we have to evaluate the function

$$F(x) = \int_{y_1}^{y_2} g(x,y)\,\mathrm{d}y,$$

we define an arbitrary probability density $p(x,y)$ subject to the conditions

$$p(x,y) \geq 0 \quad (> 0 \ \text{ if } \ g(x,y) \neq 0) \ , \quad \int \mathrm{d}x \int_{y_1}^{y_2} p(x,y)\ \mathrm{d}y = 1.$$

With such a probability density, we generate a pair of values (x', y') and evaluate the estimator

$$f(x) = \frac{g(x',y')}{p(x',y')}\ \delta(x - x'),$$

where now $\delta(x - x')$ is the Dirac δ. This is a correct estimator of the function $F(x)$ since its expectation value is

$$\left\langle \frac{g(x',y')}{p(x',y')}\delta(x - x') \right\rangle = \int \mathrm{d}x' \int_{y_1}^{y_2} \mathrm{d}y'\ p(x',y')\ \frac{g(x',y')}{p(x',y')}\ \delta(x - x') = F(x).$$

Again, the selection of a single value of the integrand function $g(x', y')$ yields an estimate of all the function F. This estimate is $\frac{g(x',y')}{p(x',y')}$ for the $x = x'$ and zero for all other values of x.

Finally, a Function Defined as an Infinite Sum of Multiple Integrals

Let us consider, for example, the series of integrals

$$F(x) = g_\circ(x) + \int_{y_1}^{y_2} g_1(x, y')\,\mathrm{d}y' + \int_{y_1}^{y_2} \mathrm{d}y' \int_{y_1}^{y'} \mathrm{d}y'' g_2(x, y', y'')\,\mathrm{d}y'' + \cdots + \int_{y_1}^{y_2} \mathrm{d}y' \int_{y_1}^{y'} \mathrm{d}y'' \ldots \int_{y_1}^{y^{(n-1)}} g_n(x, y', y'', \ldots, y^{(n)})\,\mathrm{d}y^{(n)} + \ldots . \tag{14.7}$$

We can evaluate this series with the following MC procedure. First we select the $\overline{n}$-th term of the series with probabilities $P(n), n = 0, 1, \ldots$; then, with probabilities $p_x(x), p_1(y'), p_2(y''), \ldots, p_{\overline{n}}(y^{(\overline{n})})$, we select the value of the arguments of the integrand function $\overline{x}, \overline{y'}, \overline{y''}, \ldots, \overline{y^{(\overline{n})}}$. According to the foregoing, the estimator of the series is

$$f(x) = \frac{g_n(\overline{x}, \overline{y'}, \overline{y''}, \ldots, \overline{y^{(\overline{n})}})}{P(\overline{n}) p_x(\overline{x}) p_1(\overline{x'}) \ldots p_n(\overline{y^{(\overline{n})}})} \delta(x - \overline{x}). \tag{14.8}$$

14.3.2 The Integral Boltzmann Equation with Approximate Total Scattering Rate

We are now ready to apply the above arguments to the MC solution of BE. Let us consider the derivation of Chambers integral equation in Sect. 10.5.2. There, after moving to path variables, we reached the equation (cf. (10.36))

$$\frac{\partial}{\partial t^*} f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = \int f^*(\boldsymbol{r}^*, \boldsymbol{p}', t^*) P(\boldsymbol{p}', \boldsymbol{p}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*)) \mathrm{d}\boldsymbol{p}' - \lambda(\boldsymbol{p}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*)) f^*. \tag{14.9}$$

Now we consider the function

$$\tilde{f}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = \mathrm{e}^{\Gamma t^*} f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*). \tag{14.10}$$

This function differs from the one introduced in (10.37) since the exponent is now a constant times t^* instead of the integral of the scattering rate. This substitution will bring about the possibility to include self-scattering, with much more freedom of choice. The time derivative of (14.10) is

$$\frac{\partial}{\partial t^*} \tilde{f}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = \Gamma \tilde{f} + \mathrm{e}^{\Gamma t^*} \frac{\partial}{\partial t^*} f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*),$$

or, using (14.9),

$$\frac{\partial}{\partial t^*} \tilde{f}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = \mathrm{e}^{\Gamma t^*} \int f^*(\boldsymbol{r}^*, \boldsymbol{p}', t^*) P(\boldsymbol{p}', \boldsymbol{p}(t^*)) \mathrm{d}\boldsymbol{p}' - [\lambda(\boldsymbol{p}(t^*)) - \Gamma] \tilde{f}.$$

The "out" term has not completely absorbed in the exponential. A new correction term is present, proportional to the difference between the exact rate λ, variable with time, and its constant approximation Γ. Integration with respect to t^* yields

$$\tilde{f}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = \tilde{f}(\boldsymbol{r}^*, \boldsymbol{p}^*, 0) + \int_0^{t^*} \mathrm{d}t^{*\prime} \mathrm{e}^{\Gamma t^{*\prime}} \int f^*(\boldsymbol{r}^*, \boldsymbol{p}', t^{*\prime}) P(\boldsymbol{p}', \boldsymbol{p}(t^{*\prime})) \mathrm{d}\boldsymbol{p}' - \int_0^{t^*} \mathrm{d}t^{*\prime} \tilde{f}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^{*\prime}) [\lambda(\boldsymbol{p}(t^{*\prime})) - \Gamma].$$

Now we go back to the function f^* using (14.10) and noting that the two functions $\tilde{f}$ and f^* coincide for $t^* = 0$:

$$\begin{aligned} f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = {} & f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, 0) \mathrm{e}^{-\Gamma t^*} \\ & + \int_0^{t^*} \mathrm{d}t^{*\prime} \mathrm{e}^{-\Gamma(t^* - t^{*\prime})} \int f^*(\boldsymbol{r}^*, \boldsymbol{p}', t^{*\prime}) P(\boldsymbol{p}', \boldsymbol{p}(t^{*\prime})) \mathrm{d}\boldsymbol{p}' \\ & - \int_0^{t^*} \mathrm{d}t^{*\prime} \mathrm{e}^{-\Gamma(t^* - t^{*\prime})} f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^{*\prime}) [\lambda(\boldsymbol{p}(t^{*\prime})) - \Gamma]. \end{aligned}$$

If we now return to the old variables $\boldsymbol{r}$, $\boldsymbol{p}$, and t, we obtain

$$\begin{aligned} f(\boldsymbol{r}, \boldsymbol{p}, t) = {} & f(\boldsymbol{r}(0), \boldsymbol{p}(0), 0) \mathrm{e}^{-\Gamma t} \\ & + \int_0^{t} \mathrm{d}t' \mathrm{e}^{-\Gamma(t - t')} \int f(\boldsymbol{r}(t'), \boldsymbol{p}', t') P(\boldsymbol{p}', \boldsymbol{p}(t')) \mathrm{d}\boldsymbol{p}' \\ & + \int_0^{t} \mathrm{d}t' \mathrm{e}^{-\Gamma(t - t')} f(\boldsymbol{r}(t'), \boldsymbol{p}(t'), t') [\Gamma - \lambda(\boldsymbol{p}(t'))], \end{aligned} \quad (14.11)$$

where $\boldsymbol{r}(t') = \boldsymbol{r}(\boldsymbol{r}^*, \boldsymbol{p}^*, t')$ is the position of the particle in the trajectory $(\boldsymbol{r}^*, \boldsymbol{p}^*)$ at time t'. The first term in the above equation represents the ballistic contribution diminished by the approximate probability that electrons are not scattered from the initial time to time t; the second term is the contribution of electrons scattered into the right trajectory at the time t' between $t = 0$ and the time t and not scattered out of the trajectory before the observation time t, according to the approximate scattering rate; The third term gives the correction to the out scattering owing to the fact that an approximate scattering rate has been used in the previous terms.

14.3.3 The Neumann Expansion

If we insert (14.11) into itself, we obtain

$$
\begin{aligned}
f(\boldsymbol{r},\boldsymbol{p},t) = f(\boldsymbol{r}(0),\boldsymbol{p}(0),0)\mathrm{e}^{-\Gamma t} + \int_0^t \mathrm{d}t' \mathrm{e}^{-\Gamma(t-t')} \int \Bigg\{ f(\boldsymbol{r}'(0),\boldsymbol{p}'(0),0)\mathrm{e}^{-\Gamma t'} \\
+ \int_0^{t'} \mathrm{d}t'' \mathrm{e}^{-\Gamma(t'-t'')} \int f(\boldsymbol{r}'(t''),\boldsymbol{p}'',t'')P(\boldsymbol{p}'',\boldsymbol{p}'(t''))\mathrm{d}\boldsymbol{p}'' \\
+ \int_0^{t'} \mathrm{d}t'' \mathrm{e}^{-\Gamma(t'-t'')} f(\boldsymbol{r}'(t''),\boldsymbol{p}'(t''),t'')[\Gamma - \lambda(\boldsymbol{p}'(t''))] \Bigg\} \\
\times P(\boldsymbol{p}',\boldsymbol{p}(t'))\mathrm{d}\boldsymbol{p}' + \int_0^t \mathrm{d}t' \mathrm{e}^{-\Gamma(t-t')} \Bigg\{ f(\boldsymbol{r}(0),\boldsymbol{p}(0),0)\mathrm{e}^{-\Gamma t'} \\
+ \int_0^{t'} \mathrm{d}t'' \mathrm{e}^{-\Gamma(t'-t'')} \int f(\boldsymbol{r}(t''),\boldsymbol{p}',t'')P(\boldsymbol{p}',\boldsymbol{p}(t''))\mathrm{d}\boldsymbol{p}' \\
+ \int_0^{t'} \mathrm{d}t'' \mathrm{e}^{-\Gamma(t'-t'')} f(\boldsymbol{r}(t''),\boldsymbol{p}(t''),t'')[\Gamma - \lambda(\boldsymbol{p}(t''))] \Bigg\} [\Gamma - \lambda(\boldsymbol{p}(t'))].
\end{aligned}
\tag{14.12}
$$

This expression is rather cumbersome, but it has a simple physical interpretation. It contains a term of order zero in the scattering rate P:

$$
f^{(0)}(\boldsymbol{r},\boldsymbol{p},t) = f(\boldsymbol{r}(0),\boldsymbol{p}(0),0)\,\mathrm{e}^{-\Gamma t}. \tag{14.13}
$$

This term represents the contribution to the distribution function of the electrons that reach the point $(\boldsymbol{r},\boldsymbol{p})$ at time t from the initial point of the appropriate trajectory. This contribution is weighted with the factor $\mathrm{e}^{-\Gamma t}$ that indicates the (approximate) probability that such electrons did not suffer any collision during the interval of time from the initial $t = 0$ to the final t, if the electron scattering rate were Γ.

Then two terms of first order are present:

$$
\begin{aligned}
f^{(1)}(\boldsymbol{r},\boldsymbol{p},t) = \int_0^t \mathrm{d}t' \mathrm{e}^{-\Gamma(t-t')} \int f(\boldsymbol{r}'(0),\boldsymbol{p}'(0),0)\mathrm{e}^{-\Gamma t'} P(\boldsymbol{p}',\boldsymbol{p}(t'))\mathrm{d}\boldsymbol{p}' \\
+ \int_0^t \mathrm{d}t' \mathrm{e}^{-\Gamma(t-t')} f(\boldsymbol{r}(0),\boldsymbol{p}(0),0)\mathrm{e}^{-\Gamma t'}[\Gamma - \lambda(\boldsymbol{p}(t'))].
\end{aligned}
\tag{14.14}
$$

"Reading" the equation from the left, the first term represents the contribution of electrons that at time $t = 0$ leave $\boldsymbol{r}'(0)$ with $\boldsymbol{p}'(0)$; do not suffer any scattering until the time t' (probability evaluated again with the approximate lifetime $1/\Gamma$); at t' undergo a scattering process that changes their momenta from $\boldsymbol{p}'(t')$ to $\boldsymbol{p}(t')$, i.e., are put in the correct trajectory that will lead them to the right position $(\boldsymbol{r},\boldsymbol{p})$ at the right time t, multiplied by the factor that considers the approximate probability that they are not scattered again during this second time interval.

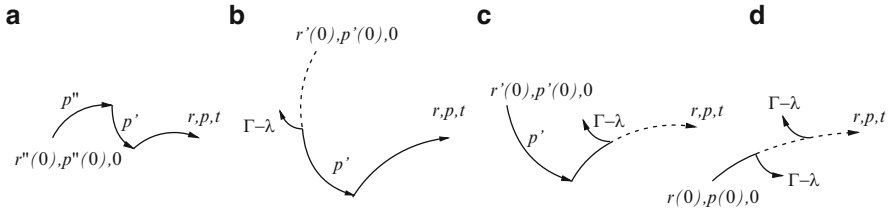


Fig. 14.8. Four terms of second order. (**a**) represents the contribution of paths with two scattering events, evaluated with the approximate lifetime; (**b**) and (**c**) represent the first-order correction to the one-scattering paths, and (**d**) represents the second order correction to the zero-scattering path

The second term in (14.14) is the first-order correction to the probability of no scattering during the free flight from $t = 0$ to t, previously approximated with $\exp(-\Gamma t)$ in (14.13). It corresponds to electrons which leave $\boldsymbol{r}(0)$ with $\boldsymbol{p}(0)$ at $t = 0$ and are not scattered (according to the approximate lifetime) until t', when they suffer a scattering process that takes them away from the path that would lead them to $\boldsymbol{r}$ and $\boldsymbol{p}$ at time t. This process, however, is supposed to happen with a probability given by only the difference between the correct probability and the approximate one, since only the correction must be accounted for.

The remaining terms in the integral equation (14.12) still contain the unknown function f. If the iteration is continued, inserting the expression (14.11) for f in (14.12), we obtain four second-order terms, by now easily interpretable, represented in Fig. 14.8.

If we continue to insert (14.11) in place of f, we obtain the Neumann series of the original integral equation. This series yields the distribution function in $(\boldsymbol{r}, \boldsymbol{p})$ at time t as the sum of contributions corresponding to electron paths with increasing numbers of scattering events and corrections to paths previously evaluated with approximate lifetimes. The various terms of the series can be evaluated with MC sampling as discussed in the previous section. We shall shortly see that appropriate choices of the sampling probabilities provide the same algorithm as the direct simulation described in Sect. 14.2. However, other choices of the sampling probabilities can provide correct algorithms as well. Thus, the formal MC solution of the BE provides a very flexible tool; the direct simulation is only one of its possibilities.

14.3.4 Sampling

The series expansion described in the previous section can now be sampled with the MC technique described in Sect. 14.3.1. From the graphic representation, we understand that the selection of a value of the integrand function of a given term of the series corresponds to the selection of a given electron trajectory. The estimator (cf. (14.8)) will be of the form

$$\frac{g(traj)}{P(traj)}\delta(\boldsymbol{r}-\overline{\boldsymbol{r}})\delta(\boldsymbol{p}-\overline{\boldsymbol{p}}), \tag{14.15}$$

where $g(traj)$ is value of the multiple integrand function corresponding to the selected trajectory, $P(traj)$ is the probability of selecting that trajectory, $\overline{\boldsymbol{r}}$ and $\overline{\boldsymbol{p}}$ are the coordinates where the selected trajectory ends.

In what follows we shall show that particular choices of the sampling probabilities lead to the "standard" direct MC simulation, described in Sect. 14.2.

First, an initial state $(\boldsymbol{r}(0), \boldsymbol{p}(0))$ is randomly selected. If the known distribution at the initial time is used, the probability at the denominator of the estimator cancels the initial distribution that appears in all terms of the Neumann expansion.

Then, a time t_1 is generated with probability distribution

$$P(t_1)\mathrm{d}t_1 = \mathrm{e}^{-\Gamma t_1}\Gamma \mathrm{d}t_1.$$

If t_1 is larger than the time t at which the distribution function is to be evaluated, the zero-order term is selected. The probability that such an event occurs is the probability that a particle with scattering rate Γ did not scatter before t, given by $\mathrm{e}^{-\Gamma t}$.

If $t_1 < t$, t_1 is chosen as the end of the first free flight, and a second flight duration Δt_2 is selected with the same probability distribution:

$$P(\Delta t)\ \mathrm{d}t_2 = P(t_2 - t_1)\ \mathrm{d}t_2 = \mathrm{e}^{-\Gamma \Delta t}\Gamma \mathrm{d}t_2.$$

If $t_2 > t$, the first-order term is selected, otherwise t_2 is chosen as the end of the second free flight, and a third flight duration is generated in the same way. The process continues until a free flight terminates beyond the final time t. If the sequence of times

$$t_1, \quad t_2, \quad \ldots \quad t_n \tag{14.16}$$

is generated, then the term of order n is selected and the above t_i values are used as values of t', t'', ... (in reverse order) to sample the time integrals of the Neumann expansion. The probability that a given sequence of times (14.16) is selected is given by

$$\begin{aligned} P(t_1, t_2, \ldots, t_n)\mathrm{d}t_1\mathrm{d}t_2 \ldots \mathrm{d}t_n &= \mathrm{e}^{-\Gamma t_1}\Gamma \mathrm{d}t_1 \mathrm{e}^{-\Gamma(t_2-t_1)}\Gamma \mathrm{d}t_2 \ldots \mathrm{e}^{-\Gamma(t-t_n)} \\ &= \mathrm{e}^{-\Gamma t}\Gamma^n \mathrm{d}t_1\mathrm{d}t_2 \ldots \mathrm{d}t_n. \end{aligned} \tag{14.17}$$

When this probability is used in the evaluation of the estimator in (14.15), the exponential factor cancels the same factor that appears in the integrand function of all integrals of the series expansion.

At each time t_i, we must now choose whether an *in* scattering event or a "self-corrected" *out* scattering event occurs, corresponding, in the first order

case, to the two terms in (14.14). Let us choose between them with the probabilities

$$P_i = \frac{\lambda(\boldsymbol{p}(t_i))}{\Gamma} \quad , \qquad P_s = \frac{\Gamma - \lambda(\boldsymbol{p}(t_i))}{\Gamma}. \tag{14.18}$$

Here, we have assumed that Γ is always larger than $\lambda(\boldsymbol{p}(t_i))$, so that both probabilities are positive, as done in the standard simulation where a self-scattering is added to make the total scattering rate Γ constant (such condition is not required in general in the formal MC solution). Each time the choice in (14.18) is made, a Γ appears in the denominator of the probability. Thus, in a term of order n, a factor Γ^n is present in the denominator of the probability that cancels the identical factor in (14.17) for the probability of the sequence of times.

If a self-scattering is chosen with the probability P_s in (14.18), the factor $(\Gamma - \lambda)$ that appears in P_s cancel the identical factor in the second term of (14.14) and in each similar corrective term. In such a case the orbit continues unperturbed.

If a real scattering is chosen, a factor $\lambda(\boldsymbol{p}(t_i))$ remains in the denominator of the estimator. However, the function $P(\boldsymbol{p}', \boldsymbol{p}(t'))$ in the first term of (14.14) must be sampled. If, for the moment, we assume that we are dealing with the event at time t_1, the value of $\boldsymbol{p}'$ is already determined by the previous choices of the initial state of the trajectory and of the time t_1. When the total scattering rate is the sum of different contributions due to different scattering mechanisms, we must first select which mechanism has been active. To this purpose, we select one of them on the basis of the relative integrated scattering rates, i.e., with probabilities $P_i(\boldsymbol{p})/\lambda$. This λ, in the denominator of the probability, cancels the identical factor left over above. When finally the state after scattering is selected with probability $P_i(\boldsymbol{p}', \boldsymbol{p}(t'))/P_i(\boldsymbol{p})$, the denominator of this probability cancels the numerator of the last probability used, and the numerator cancels the integrand function to be sampled. If the event is not the first one, the same argument holds for each single scattering event.

In conclusion, with the above procedure for sampling the Neumann series of the integral transport equation with the "natural" probabilities, all factors mutually cancel, and a unit counting must be attributed to the state reached by the simulated trajectory, exactly as it is done in the direct simulation.

As already mentioned, the above formal MC approach indicates that arbitrary probabilities can be used in the choice of the simulated trajectory, as long as the estimator is weighted with a factor given by the ratio between the natural probability of the selected event and the probability used for its selection. This freedom may be used, for example, to simulate trajectories backward in time, to devote all the computer time to the calculation of the distribution function in a given point of phase space [211], or to increment the statistics of rare events, when these must be known with great accuracy [344, 462].

When probabilities are used different from the "natural" ones, the method is often called *weighted Monte Carlo* [344].

The present formal development of the MC method may be applied, in principle, to transport equations in a quantum framework, such as the von Neumann equation for the density matrix or the equation for the Wigner function. In such cases, however, the paths to be sampled are essentially the Feynman interferencing paths, and a huge number of them (almost canceling each other) must be added to obtain a correct result. This results in serious computational difficulties.

15

Bulk Transport Properties of Main Semiconductors

The number of semiconductor materials that are today studied and employed by the electronic industry is very large and continuously increasing, in particular after the introduction of semiconductor heterostructures. For space reasons, however, in this textbook we will limit ourselves to the two most "popular" materials, namely silicon and gallium arsenide. Hopefully, the analysis of the transport properties of these two examples will enable the reader to understand the main electronic transport properties of most materials. Silicon is by far the most used material in semiconductor industry, both because of its large availability and because of the existence of a "natural oxide", very suitable for the realization of electronic devices. Gallium arsenide, on the contrary, is much more convenient for optoelectronic applications, owing to its direct energy gap, appropriate for a transformation between electronic and optical energies. Furthermore, its small electron effective mass (0.067 compared to an average 0.295 in Si) provides a higher electron mobility.

The need to further improve the performances of electronic devices, and the search for physical features opening the way to new applications, have pushed the scientific community into a continuous search for new semiconductor materials and improvements in the properties of materials already known, but not yet fully exploited. There are many books on special materials, and new ones are published every year, in which the interested reader may find important examples and updated references to the literature.

15.1 Electrons in Silicon

Among the transport properties of a semiconductor, a chief role is played by the carrier mobility. It indicates, in fact, the potentiality of the charge to move in the material and to respond to field changes. At higher mobilities, transit times and switching times are shorter, and frequency cutoffs are higher.

The mobility of electrons in Si is presented in Fig. 15.1 as a function of temperature and of impurity content. The temperature dependence of the

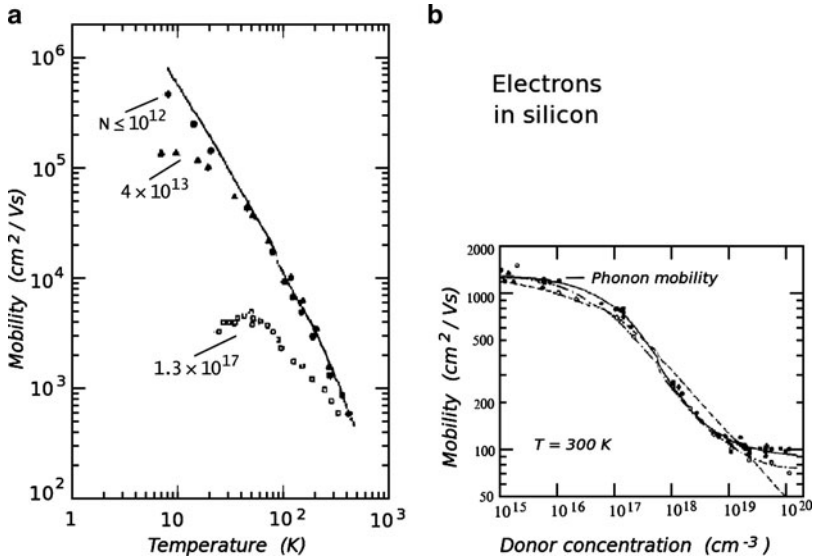


Fig. 15.1. Mobility of electrons in Si as a function of temperature (**a**) and of impurity content at room temperature (**b**) [208]. In (**a**), symbols indicate experimental results obtained by several authors [88, 329] with several techniques in materials with different impurity concentrations; the continuous line indicates the pure lattice mobility obtained with MC simulation. In (**b**), open and closed circles indicate experimental results [200,315]; the continuous, dashed, and dot-dashed line represent best fit obtained with phenomenological analytical expressions given by [21,190,397], respectively

mobility shown in part (a) of the figure is not very far from that obtained with a simple-model semiconductor, shown in Fig. 11.6. Below about 50 K, the lattice mobility is dominated by acoustic scattering, while above this temperature several intervalley scattering mechanisms become important. At high impurity concentrations and low temperatures, the mobility is dominated by impurity scattering, and the deviation from the pure lattice mobility occurs at higher temperatures in less pure materials. The influence of the impurity concentration N on the electron mobility at room temperature is shown in part (b) of the figure. The effect becomes appreciable around $N = 10^{16}\,\mathrm{cm}^{-3}$, and tends to saturate above about $10^{19}\,\mathrm{cm}^{-3}$. This saturation is interpreted as being due to the merging of bound states into the conduction band [151].

We know from the previous chapters that the electrons in Si occupy several valleys having different orientations, and that the Ohmic mobility is an average of the mobilities of the electrons in the different valleys. By symmetry, this average is independent of the field orientation. However, if a mechanism is activated that shifts in energy the valleys oriented along given directions, this symmetry is lost. Thus, a method to increase the electron mobility in Si consists in the application of a uniaxial stress along a direction such that

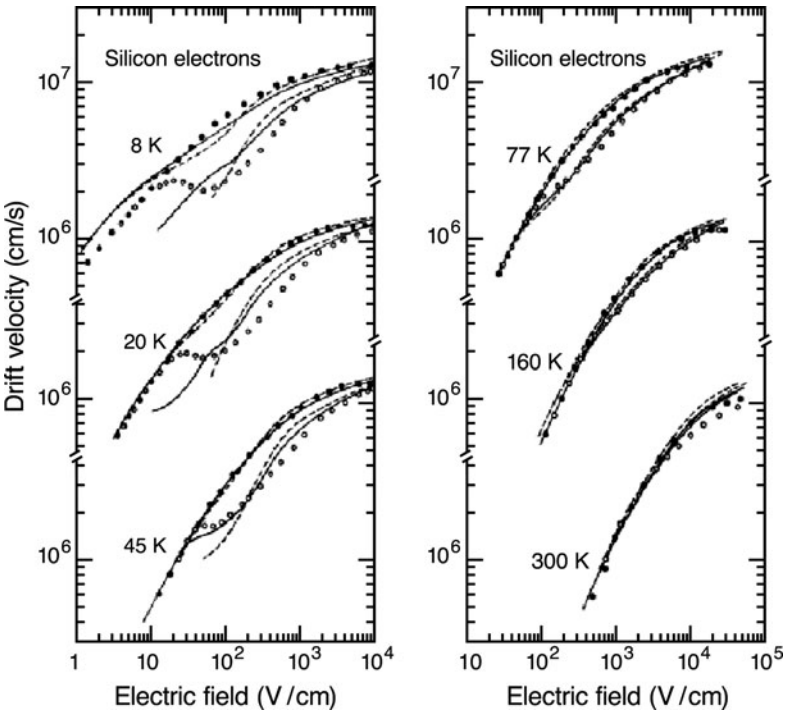


Fig. 15.2. Electron drift velocity in Si as a function of the electric field at different temperatures [88]. Closed and open circles indicate experimental data obtained with the field parallel to ⟨111⟩ and ⟨100⟩ directions, respectively. Continuous and broken lines indicate theoretical MC results obtained with different electron–phonon couplings and neglecting impurity scattering

the valleys presenting the smaller effective mass in the direction of the field are lower in energy than the other ones [110]. Electrons at equilibrium will populate more these "faster valleys", and the overall mobility will increase. A similar effect is also present in holes, where heavy- and light-hole bands are split by uniaxial pressure. Higher drift velocities are obtained in thin layers of Si epitaxially grown next to a Si–Ge alloy where the stress necessary to shift the valley energies is produced by the strain induced by the different lattice constants of the two materials (see, for example [149, 299, 366, 443, 454]).

Figures 15.2–15.4 show the drift velocity of electrons in Si as a function of the applied field at different temperatures. Experimental results obtained with the field oriented along different directions are compared, in Fig. 15.2, with MC simulations obtained with different electron–phonon couplings and neglecting impurity scattering. Nonparabolicity of the conduction band is also neglected in these simulations.

The anisotropy effect, due to valley repopulation as described in Sect. 13.3, increases at decreasing temperatures. It tends to disappear at the highest field

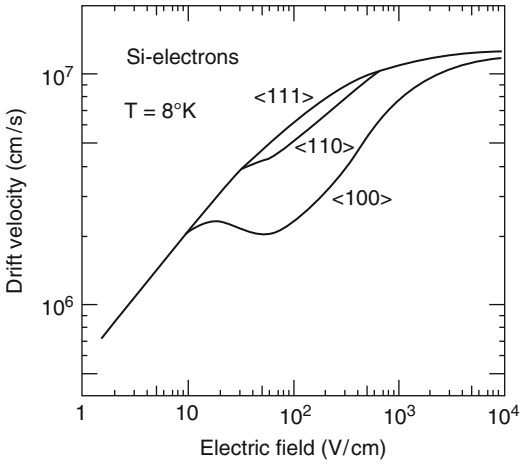


Fig. 15.3. Experimental (average data) electron drift velocity in Si as a function of electric field at $T = 8\,\mathrm{K}$ with the field applied along three high-symmetry directions, as indicated [88]

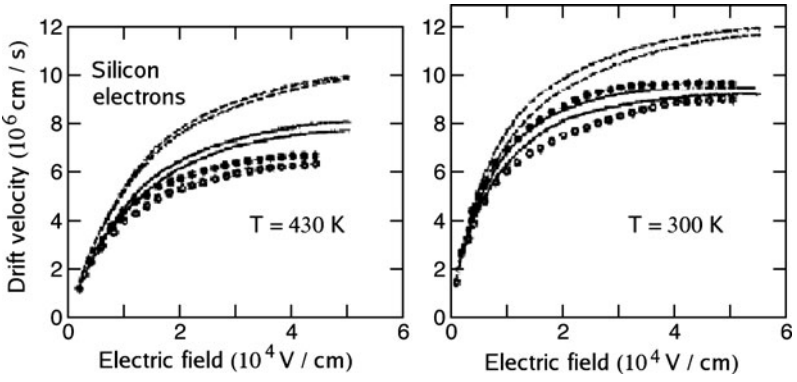


Fig. 15.4. Drift velocity of electron in Si as a function of field at the indicated temperatures. Closed and open circles refer to representative experimental data obtained with field parallel to ⟨111⟩ and ⟨100⟩ directions, respectively. The continuous (*broken*) lines indicate MC results obtained with nonparabolic (*parabolic*) model [210]. The nonparabolicity parameter is $\alpha = 0.5\,\mathrm{eV}^{-1}$

strengths, when the electron energy becomes very high, and the intervalley scattering is very efficient in equalizing electron energies and populations in the different valleys. This picture is validated by Figs. 15.5 and 15.6, where electron mean energies and valley repopulations are presented at various lattice temperatures as a function of the applied field. In Fig. 15.2, the MC curves along the ⟨100⟩ direction are interrupted at the lowest fields and temperatures since the repopulation times are too long, as discussed in Sect. 13.3.

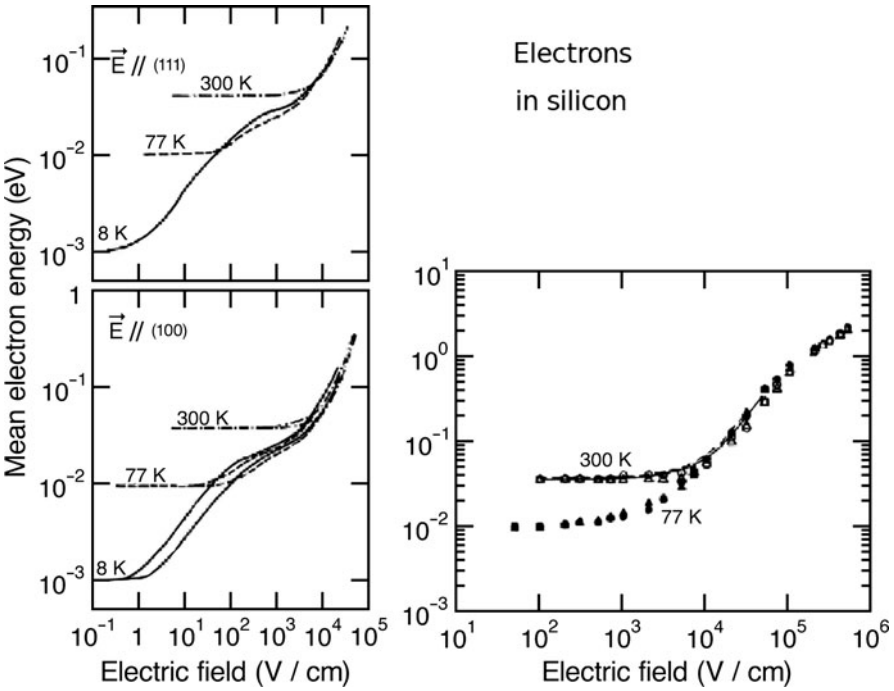


Fig. 15.5. Mean energy of electrons in Si as a function of field strength at different temperatures and field orientations. The results shown in the left part of the figure have been obtained with MC simulations with a many-valley model [88]; in the right part circles and triangles indicate full-band MC results [148] along the ⟨111⟩ and ⟨100⟩ directions, respectively, and compared, at 300 K, with the results in [88]

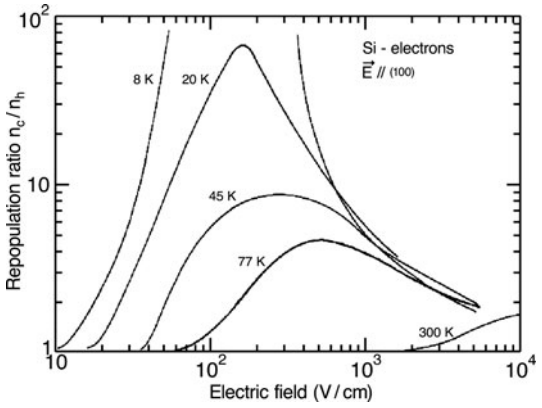


Fig. 15.6. Repopulation ratio of cold-to-hot valleys for electrons in Si as a function of field strength applied along a ⟨100⟩ direction at various temperatures [88]

The geometry of the valleys of the conduction band of Si is such that three different high-field drift velocities are obtained along the three high-symmetry directions ⟨111⟩, ⟨110⟩, and ⟨100⟩. The difference between the first two directions, however, is very small and easily observable only at the lowest temperatures. Figure 15.3 shows experimental results obtained at 8 K [88].

Another important property of the electron drift velocity in Si is the tendency to saturate at the highest fields. The saturation of the drift velocity of electrons at high fields is one of the most characteristic features used in semiconductor electronics. Figure 15.4 shows that at room temperature, nonparabolicity of the conduction band must be included in the model to obtain saturation at fields around 10^5 V/cm, as shown by experimental results [210].

Figure 15.7 shows theoretical results for the drift velocity of electrons in Si, extended to much higher fields using a full-band MC [148].

Analytical expressions of mobility and drift velocity vs field of electrons in Si for numerical applications are reported in [208].

The mean energy of electrons in Si as a function of applied field at various temperatures is shown in Fig. 15.5, obtained with MC simulations [88, 148]. They show that at the highest field strengths the mean energy becomes independent of both lattice temperature and field direction.

Figure 15.6 shows the repopulation of the valleys when the field is oriented along a ⟨100⟩ direction at various lattice temperatures, obtained with MC simulations [88]. It may be seen that the repopulation first increases with field strength because of the different heating in longitudinal and transverse valleys. At higher fields, it decreases again because at very high mean energies f-scattering between perpendicular valleys is very effective in equalizing the

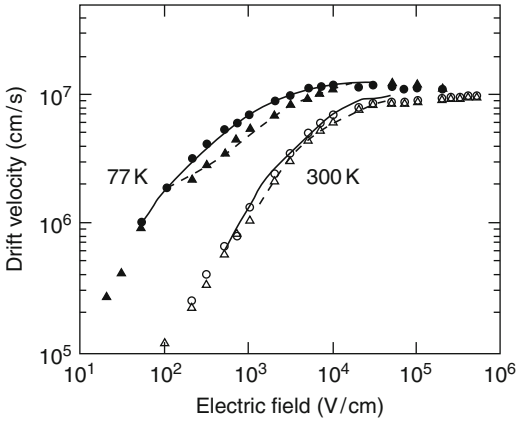


Fig. 15.7. Electron drift velocity in Si as a function of the electric field at different temperatures. Continuous (*broken*) lines indicate results of [88] along ⟨111⟩ (⟨100⟩) directions; circles and triangles indicate full-band MC results [148] along the same directions

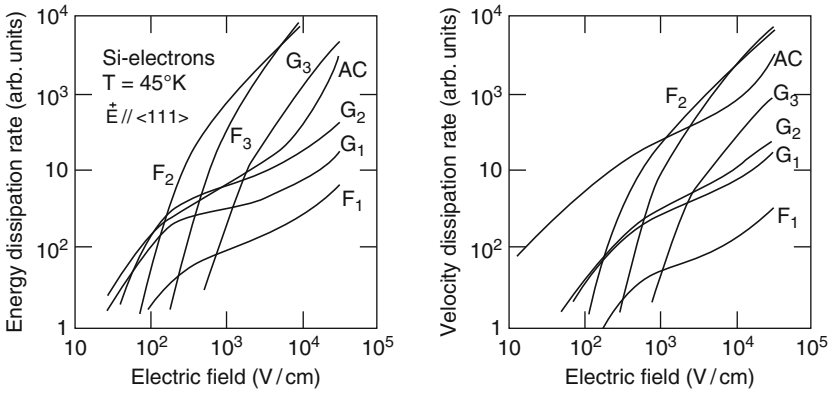


Fig. 15.8. Energy and velocity dissipation rates obtained with an MC many-valley simulation [88] at $T = 45\,\mathrm{K}$, for the various scattering mechanisms. The symbols refer to different scattering mechanisms: acoustic (AC), three f-scattering phonons (F1,F2,F3), and three g-scattering phonons (G1,G2,G3)

mean energy among the different valleys, as also confirmed by the results shown in Fig. 15.5.

One of the most effective investigations we may carry out using the MC simulation is about the rate of momentum and energy dissipation due to each scattering mechanism. In Fig. 15.8, this kind of information is shown for electrons in Si at $T = 45\,\mathrm{K}$ as a function of field strength, applied along a $\langle 111 \rangle$ direction. We may see here that at this low temperature, and at low field strengths, the energy dissipation due to acoustic scattering is as important as that due to intervalley scatterings, while it dominates the velocity dissipation (in pure samples). At higher temperatures and/or field strengths, the contributions of the more energetic intervalley phonons become more important. The energy dissipation due to acoustic phonons at high fields is overestimated in the results shown in Fig. 15.8 because of the linear dispersion relation assumed in the MC model in the whole range of acoustic–phonon wavevectors.

The energy distribution function of electrons, shown in Fig. 15.9, reflects the efficiency of the various scattering mechanisms. For energies above those of the most effective phonons, the slope of the distribution function is close to that of thermal equilibrium.

Finally, the longitudinal diffusion coefficient of electrons in Si is shown in Fig. 15.10 as a function of field strengths at different temperatures and field directions. A discussion of longitudinal and transverse diffusivity (D_l and D_t) of electrons in Si at high fields can be found in [72]. Intervalley diffusion (see Sect. 13.5.1) plays an important role in both D_l and D_t.

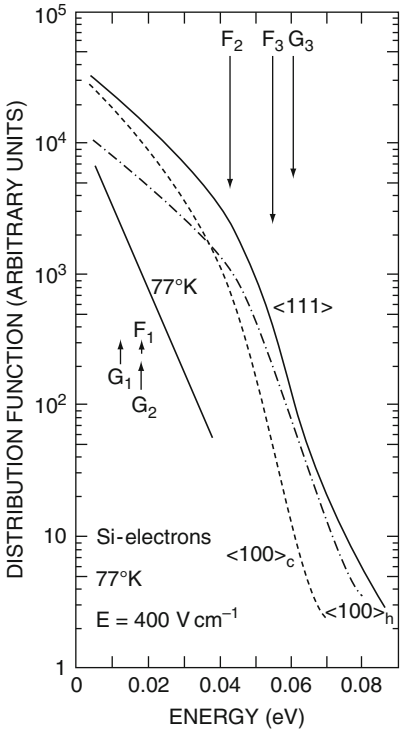


Fig. 15.9. Energy distribution function of electrons in Si obtained at $T = 77\,\mathrm{K}$ with MC simulation [88] for a field strength of $E = 400\,\mathrm{V/cm}$. The continuous line refers to a field oriented along a $\langle 111 \rangle$ direction; the dashed and dot-dashed lines refer to electrons in cold and hot valleys, respectively, with a field oriented along a $\langle 100 \rangle$ direction. The slope of the straight line indicates the lattice temperature. The arrows indicate the energies of intervalley phonons, and their lengths are proportional to the corresponding coupling constants

15.2 Holes in Silicon

Several factors make the calculation of transport properties of holes in cubic semiconductors a difficult task. The first cause of difficulty is the complexity of the valence band, formed by heavy-hole and light-hole bands, degenerate at $k = 0$, and a third split-off band, as described in Sect. 8.7. Interband as well as intraband transitions must therefore be taken into account[1]. Furthermore, anisotropy and nonparabolicity of the bands must be accounted for. Finally, the complexity of the symmetry of the wavefunctions (p-like symmetry around $k = 0$ and a mixture with other symmetries at higher k) has the consequence

[1] In a BE formulation of transport, three distribution functions $f^{(b)}(\boldsymbol{k})$ should be considered, one per band. This implies three integro-differential equations, coupled by interband scattering rates.

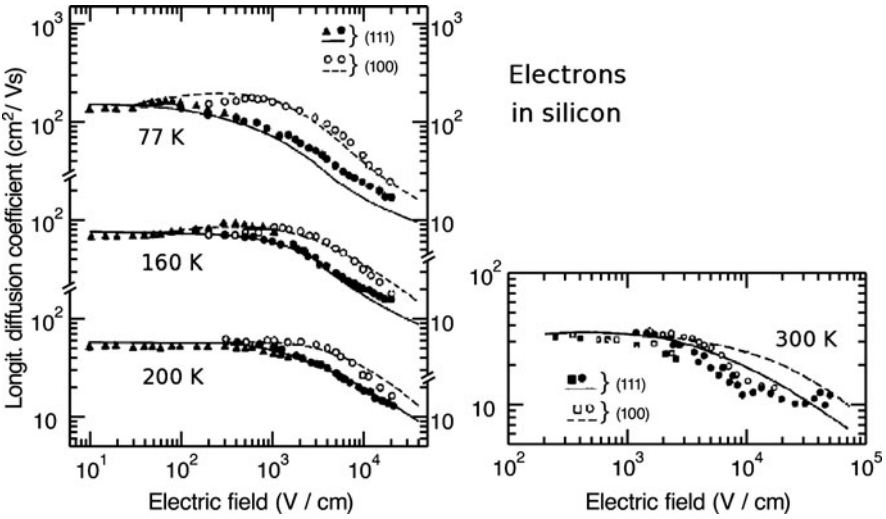


Fig. 15.10. Longitudinal diffusion coefficient of electrons in Si as a function of field at different temperatures. Symbols indicate experimental data, lines the results of MC calculations [72]

of a complicated G-factor in the scattering probabilities, dependent upon the scattering angle (see Sect. 9.2).

All the above complexity must be considered for a rigorous calculation of hole transport. Often, however, simple models are used for practical applications, accounting only for the heavy holes.

As it regards phonons, both acoustic- and optical–phonon scatterings are allowed by symmetry within and between the two bands degenerate at $k = 0$. Furthermore, also transverse acoustic phonons are effective because of the p-like symmetry of the hole wavefunctions [130].

The mobility of holes in Si is presented in Fig. 15.11 as a function of temperature and of impurity content. The temperature dependence of the mobility is shown in part (a) of the figure. The general considerations made for electrons hold also for holes: the lattice mobility is dominated by acoustic modes for temperatures below about 100 K, but does not follow the $T^{-3/2}$ dependence (as in (11.46)) owing to nonparabolicity of the heavy-hole band [333]. Around and above room temperature, the hole mobility is dominated by optical–phonon scattering.

Similar to the case of electrons, the mobility at low temperatures is dominated by ionized-impurity scattering, and the deviation from the pure lattice mobility occurs at higher T in less pure materials. The influence of impurity concentration on the hole mobility at room temperature is shown in part (b) of Fig. 15.11. The effect becomes appreciable around $N = 10^{16}$ cm^{-3}, and tends to saturate above about 10^{19} cm^{-3}.

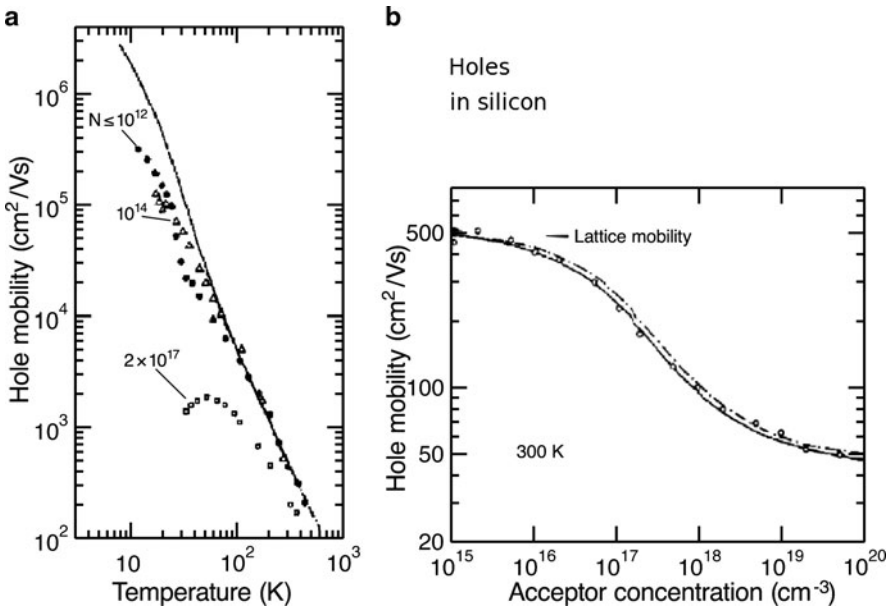


Fig. 15.11. Mobility of holes in Si as a function of temperature (**a**) and of impurity content at room temperature (**b**) [208]. In (**a**), symbols indicate experimental results obtained by several authors [285, 312, 333] with several techniques in samples with different impurity contents; the continuous line indicates MC results for pure lattice mobility [333]. In (**b**), circles indicate experimental results [200]; continuous and dot-dashed lines represent best fitting analytical curves by [98] and [397], respectively

The hole drift velocity is shown in Fig. 15.12 as a function of field, applied along different directions, at different temperatures. Experimental results are again compared with MC simulations. The anisotropy reflects the warped shape of the equienergetic surfaces [333] (see Sect. 8.7).

Experimental and theoretical results for the longitudinal diffusion coefficient of holes in Si are shown in Fig. 15.13 as a function of field strength, for several temperatures and field orientations. The anisotropy of the longitudinal diffusion coefficient again reflects the warped shape of the equienergetic surfaces of heavy holes. The tendency to saturate at the highest fields is ascribed to nonparabolicity of the heavy-hole band.

15.3 Electrons in Gallium Arsenide

The Ohmic mobility of electrons in GaAs is shown in Fig. 15.14 as a function of temperature and of impurity content. The temperature dependence of the mobility shown in part (**a**) of the figure is again coherent with the results obtained with a simple-model semiconductor with parameters adjusted to

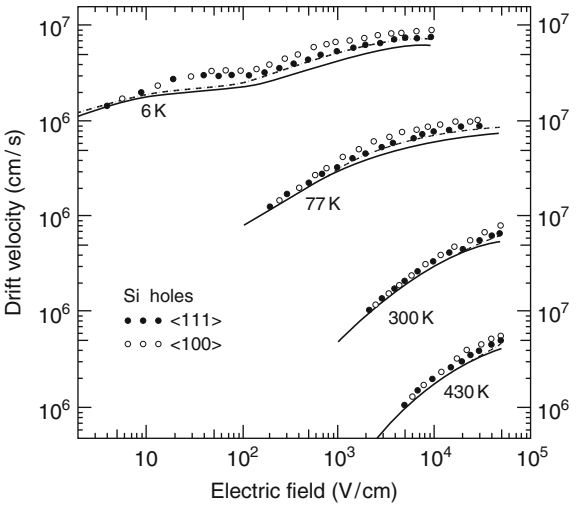


Fig. 15.12. Drift velocity of holes in Si as a function of field strength at various temperatures and field orientations [213, 333]. Circles refer to experimental results and lines to MC simulations

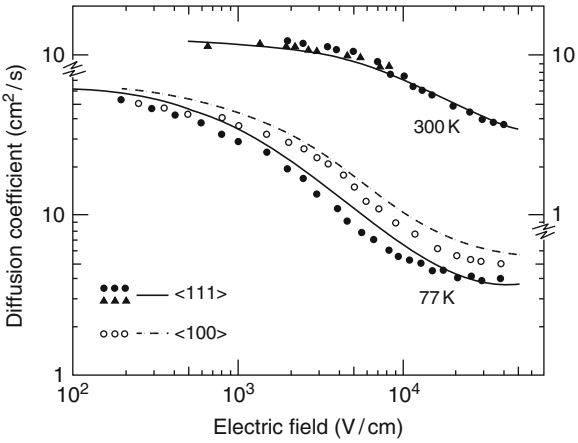


Fig. 15.13. Longitudinal diffusion coefficient of holes in Si as a function of field strength, for different temperatures and field orientations. Circles refer to experimental results and lines to MC simulations [213]

gallium arsenide, shown in Fig. 11.7. The various curves in part (**a**) of the figure refer to different samples with different impurity contents. In particular, the curve with the highest mobility at low temperatures is obtained in a quantum well with spacer layers that keep the donors away from the region of the current in order to minimize impurity scattering [392]. In part (**b**) of Fig. 15.14, the mobility is given as a function of carrier concentration at

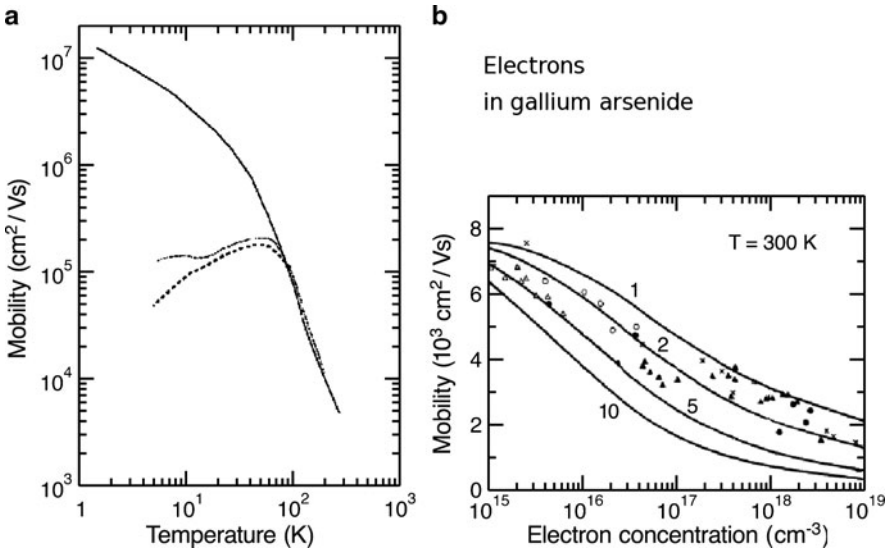


Fig. 15.14. Mobility of electrons in gallium arsenide as a function of temperature (**a**) and of carrier concentration at room temperature (**b**). In (**a**), various curves refer to different samples with different impurity contents [66, 392]. Part (**b**) represents various experimental Hall mobilities as a function of carrier concentration n, and the lines indicate theoretical results obtained assuming different compensation ratios $(N^+ + N^-)/n$ as indicated by the numbers on the lines [376]

room temperature, and the different curves are calculated assuming different compensation ratios, as indicated [376].

The drift velocity of electrons in gallium arsenide is shown in Fig. 15.15 together with the hole drift velocity, as obtained with MC simulations [148, 281]. It presents the well-known phenomenon of negative differential mobility (NDM) at the basis of the Gunn effect, as discussed in Sect. 13.4.

The mean energy of electrons in GaAs is shown in Fig. 15.16 as a function of the applied field at room temperature [169], obtained with an MC simulation using a five-valley model. Full-band MC [148] provided similar results. The rapid increase of the electron mean energy just below the threshold field for NDM is due to the features of optical–phonon scattering, as discussed in Sect. 13.1. After the onset of NDM, when intervalley scattering becomes effective, the mean energy increases much more slowly. The dissipation effect of intervalley scattering is also evident in Fig. 15.17 where the energy distribution of electrons in the central valley, as obtained with MC simulation [137], is shown for different applied fields at room temperature. For energies above the bottom of the upper valleys the distribution is much colder. For lower energies, at high fields a population inversion is also present.

The distribution function of electrons in $\boldsymbol{k}$ space in GaAs at a high field has been shown in Fig. 13.5.

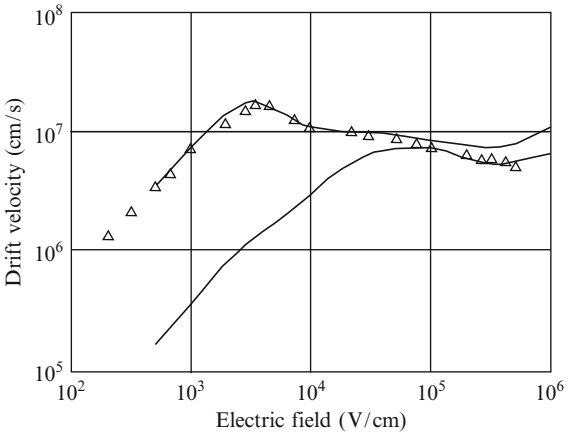


Fig. 15.15. Electron (*upper curve*) and hole (*lower curve*) drift velocity as a function of field strength in gallium arsenide at room temperature obtained with MC simulation [281]. Triangles represent full-band MC results [148]

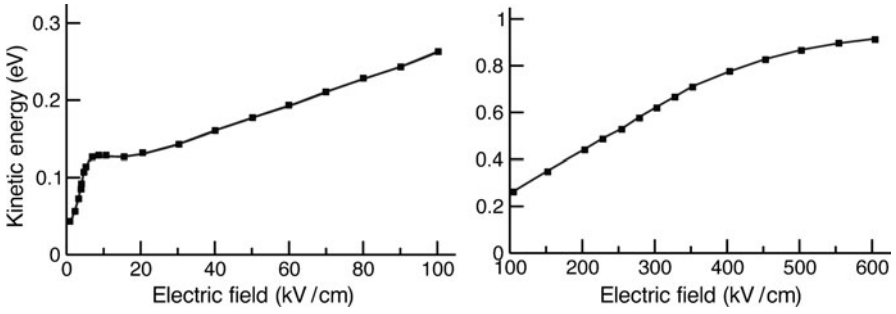


Fig. 15.16. Average kinetic energy of electrons in GaAs as a function of electric field at room temperature obtained with an MC simulation [169]

Figure 15.18 shows the diffusion coefficient of electrons in GaAs as a function of the applied field. The general shape of the curves is that discussed in qualitative terms in Sect. 13.1: the initial increase of the diffusivity is due to the heating of the electrons, and the decrease at higher fields reflects the decreasing mobility associated to transfer of electrons to the upper valleys.

The marked difference between longitudinal and transverse diffusion coefficients, shown in part (**b**) of the figure, has been explained in terms of the microscopic motion of the electrons [138]: an important contribution to the velocity fluctuations comes from particular flights of electrons in the central valley. These are electrons that are scattered into the central valley from the upper valleys with a negative component of the velocity in the direction of the electric force. They are decelerated by the field and, having an energy not sufficient for intervalley scattering, perform long flights until their velocity

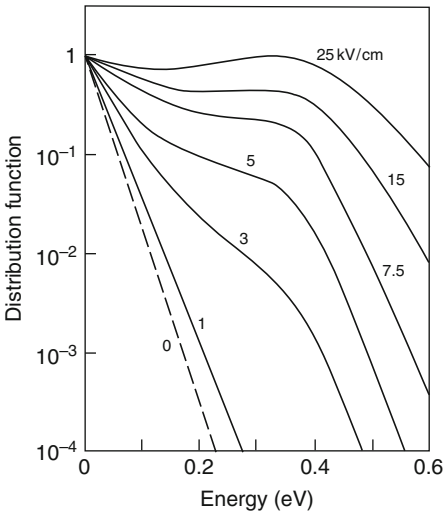


Fig. 15.17. Energy dependence of the spherically symmetric part of the electron distribution function in the central valley of GaAs obtained by means of MC simulation for different applied fields at room temperature [137]. Numbers in the curves indicate the applied electric fields. The broken line is the equilibrium zero-field Maxwellian

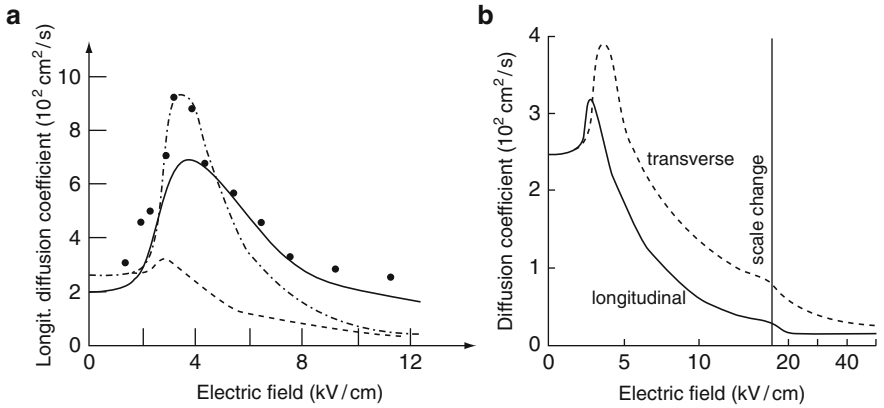


Fig. 15.18. (**a**) Longitudinal diffusion coefficient of electrons in GaAs as a function of electric field at $T = 300\,\mathrm{K}$ [71]. Dots indicate experimental results [383]; lines show different theoretical results in [138] (*dashed*), [332] (*continuous*), and [345] (*dot-dashed*). (**b**) Comparison between longitudinal and transverse diffusivities obtained in [138] with MC simulations

becomes approximately opposite to the initial one and can be scattered again into an upper valley. These flights correspond to a small total space displacement since they consist in oscillations that bring the electrons near their initial positions, and therefore their contribution to the diffusivity is small.

By contrast, the transverse velocity is large and constant during these flights. This difference is reflected in a large transverse diffusion constant compared with the parallel component. The higher values of the longitudinal diffusion constant measured by [383] has been attributed in [138] to possible additional scattering that interrupts the long flights described above.

15.4 Holes in Gallium Arsenide

We have already pointed out, when talking of holes in Si, the difficulties to be faced for determining a rigorous theory of transport properties of holes in our materials. Several calculations have been developed for hole transport in GaAs that differ in details of the band structure and/or of the scattering mechanisms and/or of the method of solution of the BE. See, for example, [67, 111, 251, 326, 432].

Figure 15.19 shows the hole mobility in GaAs as a function of temperature and impurity content.

Experimental data of drift velocity of holes in GaAs as a function of the applied field are rare. Figure 15.20 shows the results of [114].

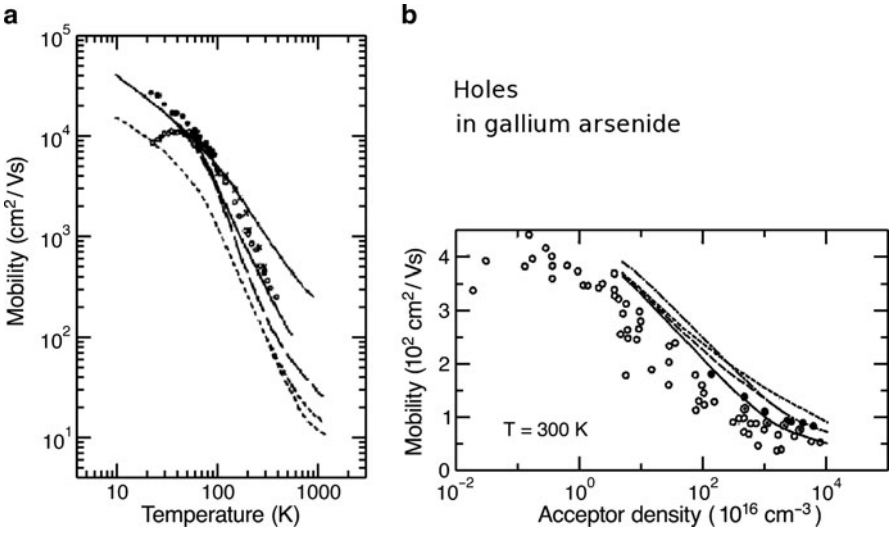


Fig. 15.19. Ohmic mobility of holes in gallium arsenide as a function of temperature (**a**) [432] and as a function of acceptor density at room temperature (**b**) [286]. In part (**a**), symbols indicate experimental results [188, 303, 492] and lines indicate results obtained with different theoretical approaches [432]. In part (**b**), circles represent various experimental results [159, 242, 475]; lines indicate results of different theoretical calculations [286]

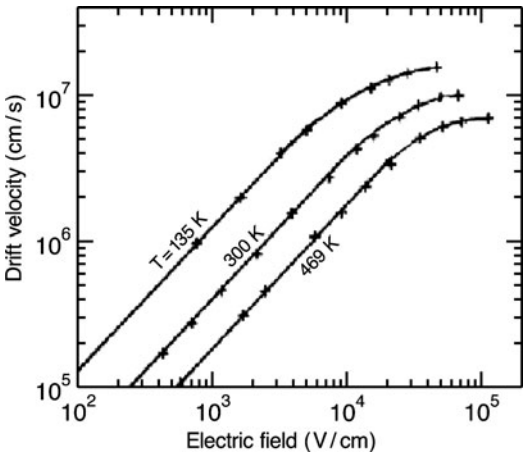


Fig. 15.20. Experimental (average values) hole drift velocity as a function of field strength in GaAs at the indicated temperatures [114]

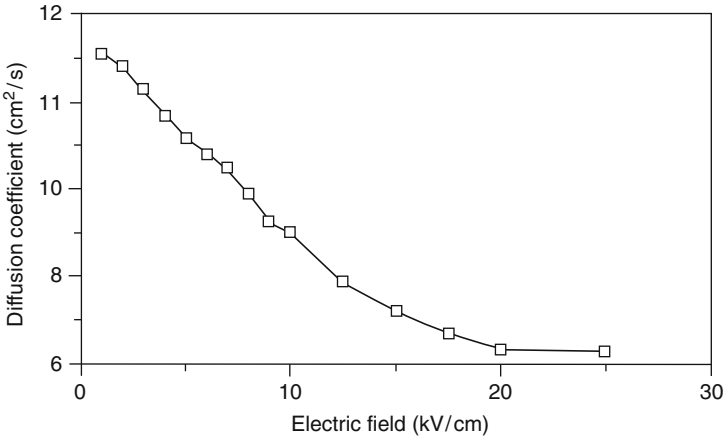


Fig. 15.21. Longitudinal hole diffusivity in GaAs at room temperature as a function of electric field, obtained with MC simulation [225]

Finally, Fig. 15.21 shows the longitudinal diffusion coefficient of holes in GaAs as a function of applied field at room temperature, obtained with an MC simulation accounting for two isotropic hole bands [225].

All the transport properties dealt with in the previous pages refer to stationary states. We have already seen in Sect. 13.6 that when an electric field is suddenly switched on or charge carriers enter, from a low-field region, a region where a high field is present, a *transient-transport* situation occurs where both the mean velocity and the mean energy of the electron gas may be higher than in steady-state conditions. This phenomenon, known as *velocity overshoot* is

of particular interest since it may be used to shorten the transit times across nanodevices and increase their switching speed. Some examples are shown in Sect. 13.6.

15.5 Organic Semiconductors

Before leaving this chapter on the electron transport properties of the main semiconductors, let us briefly mention the organic semiconductors, a class of materials that are receiving increasing interest for their potential applications (see, for example, [75, 135]).

The discovery of semiconductivity in polymers in 1977 [414] opened the way for the development of applications based of organic semiconductors in electronics and optoelectronics. At present, the activity on this field can be divided into two main categories. In one line of work, organic polymers are considered new materials to be used in traditional electronic devices. In the other more challenging activity, single organic molecules are contacted and their transport properties are studied to achieve electronic functionalities. Here we shall briefly consider only the first type of activity, closer to the traditional semiconductor physics analyzed in this book.

Organic semiconductors can be formed by small molecules (e.g., anthracene or pentacene) or by polymers (e.g., polythiophene). π-bonds formed by sp^2 hybridization of the carbon atoms cause delocalization of electrons over certain molecular regions. These electrons are therefore able to move within such regions under the action of an applied electric field. The mobility within such regions is high, but the transfer from one region to the next require passing a barrier, so that the conduction occurs through a mixture of band transport and hopping. For such a reason at the early stage of development, the electron mobility in such materials was very low. However, as time went by, with technological improvements, the mobility has been increased by several orders of magnitude, reaching values comparable to that of amorphous silicon. Nevertheless, the conductivity is lowered by the presence of traps which can immobilize charge carriers.

The role of the valence band in polymer semiconductors is played by the highest occupied molecular orbitals (HOMO), while the conduction band is represented by the lowest unoccupied molecular orbital (LUMO). The band gap is thus the energy difference between the LUMO and HOMO levels, and organic semiconductors can be of p-type or n-type, depending on the position of the Fermi level in such gap. They can be also doped with the addition of dopants. The p-type or n-type character of an organic semiconductor may depend also on the working function of the contacts: if it is closer to the LUMO level, the contact works as an electron injector, whereas if it is closer to the HOMO, injection of holes takes place.

As it regards the electronic applications (see, for example, [75, 287]), the general advantages of organic semiconductors include their low production

cost, flexibility in the substrate choice, possibility of large area deployment, and, more generally, the relative ease of tailoring their properties to specific applications. On the other side, the major limitation is a still inadequate stability.

The main applications are thin-film transistors, solar cells, photodiodes, and light-emitting diodes, called OLED (organic light-emitting diodes), already used in flexible displays. Light emission results from the recombination of electrons, injected into the organic layer from electrodes with lower work function, with holes, injected from electrodes with higher work function.

Part III

Quantum Transport in Bulk Semiconductors

16

Quantum Transport in Homogeneous Systems

This part of the book is devoted to the quantum theory of electron transport. There are situations where it is obvious that we have to resort to quantum dynamics. When the linear dimensions of the system under consideration are comparable to the wavelength of the carriers, when potential barriers exist where tunneling may occur, when potential profiles give rise to resonances, there is no doubt that quantum theory must be used. These situations, however, are all related to the space scale that approaches microscopic dimensions, while this part of the book deals with bulk homogeneous semiconductors. Quantum effects to be considered here must be related to time scales somehow connected to the scattering events. Quantum effects due to the space dependence of the potential profile will be dealt with in later chapters.

16.1 Introduction to Quantum Transport

16.1.1 Semiclassical Transport and Quantum Physics

The analysis of electron transport in semiconductors developed in the previous chapters is heavily based on the semiclassical theory of electron dynamics. As a matter of fact, many elements of that analysis, if not all, are based on quantum theory, from the energy bands of the electron states, to the perturbation theory for the scattering mechanisms. Quantum physics is responsible, paradoxically, even for the possibility to describe electron dynamics in semiclassical terms. It would be very difficult, in fact, to understand how an electron can move almost freely in the dense forest of atoms that fill most of the space in the crystal. Quantum physics indicates that the wave nature of the electron dynamics is such that with continuous reflections and interferences, the electron wavefunction can propagate freely, in a perfect crystal, with the only difference, with respect to an electron in empty space, of the band function $\epsilon(\boldsymbol{k})$ substituting the free dispersion relation $\epsilon_\circ(\boldsymbol{k}) = \hbar^2 k^2/2m_\circ$. This is essentially the content of the effective-mass theorems developed in Chap. 7.

After having recovered the dynamics of a free electron in the new space where the kinetic energy is given by the band function, we have done more work with quantum physics, and have shown that for applied fields slowly varying in space and time, we may make the approximations that in free space lead to the classical dynamics, and in a crystal lead to the semiclassical dynamics, synthesized by (7.44).

In the BE, however, not only semiclassical dynamics is involved: the scattering integral contains the transition rates from the different electronic states which are essential, and contain the most critical assumptions, for the validity of the BE.

16.1.2 From Reversible Dynamics to Irreversible Boltzmann Equation

When we ask ourselves whether the BE is to be substituted with a more rigorous quantum transport equation, we must consider the conditions and the approximations that led to the formulation of the BE. This question is relevant also in classical physics. We all know, in fact, that both classical Hamilton and quantum Schrödinger equations are reversible. On the other hand, we all see that nature behaves in an irreversible way. Naïvely, the problem may seem to be more a practical one than one of principle, related to the impossibility to prepare a system with perfectly reversed initial conditions and to the fact that extremely small variations in the initial conditions soon produce large differences of the state of the system, favoring macroscopic states with larger available phase space.[1] Nevertheless, it has always been recognized the importance of the identification of the mathematical steps that, starting from the fundamental laws of dynamics, reversible, lead to rate equations, like the Boltzmann equation, that contain irreversibility and eventually, through the H theorem, to equilibrium. The importance of these mathematical steps rely mainly on the identification of the physical properties of the system that allow such steps, and therefore that are responsible for the irreversible behavior of nature. Thus, the identification of the hypotheses that justify the use of the semiclassical BE, starting from the fundamental quantum equations, has an importance that goes beyond our present problem and it is shared by classical and quantum statistical physics.

[1] It is interesting to note that in numerical computer simulations of classical many-body systems, the motion of the various "particles" can actually be reversed for an interval of time which is longer when the precision of the calculations is increased. However, no matter how precise is the calculations, owing to the finite representation of the numbers in the computer, sooner or later the state of the system diverges with respect to the reversed motion. It is significant that for perfect reversibility an infinite (nonarbitrarily large, but infinite) precision is necessary, living alone the fact that, in the orthodox Copenhagen interpretation, a quantum measurement process cannot be described by the reversible dynamical equations.

As it regards the approach to equilibrium, starting from a situation far from equilibrium, the representative points of a statistical ensemble should spread in phase space until they occupy uniformly the available space. On the other hand, according to Liouville theorem, discussed in Sect. 3.2, the density of these points remains constant. The solution of such apparent paradox resides in the fact that as time increases, the distribution f of the representative points in phase space may become very filamentous up to the point that we have to substitute its exact value with its mean value inside a cell in which the distribution function is defined. At this point irreversibility is introduced. In fact, the exact knowledge of the state of the system to be reversed is lost. Once again, no matter how small the cells are assumed, sooner or later the substitution of the exact value of f with its *coarse-grained* value introduces irreversibility.

Furthermore, a deep discussion among the founding fathers of statistical physics, and in particular of the kinetic theory of gases, arrived to the identification of the crucial *hypothesis of molecular chaos*, necessary to prove the H theorem of the approach to equilibrium (see, for example, [440]). According to this hypothesis, the velocities of colliding particles must be uncorrelated.[2]

As it regards quantum statistical physics, the problem of obtaining a master equation from the principles of quantum physics has been widely discussed (see, for example [495]). L. Van Hove, in a series of papers [219,457–460], considered the problem of the transition from the reversible Schrödinger equation to a rate equation that leads to irreversibility.[3]

The problem can be stated in the following terms. An isolated physical system is described by the time-independent Hamiltonian

$$\mathcal{H} = \mathcal{H}_\circ + \lambda \mathcal{V},$$

where the unperturbed Hamiltonian $\mathcal{H}_\circ$ describes, say, independent particles, and $\lambda\mathcal{V}$ is the perturbation responsible for the approach to equilibrium. Its dynamics is described by the (reversible) Schrödiger equation, and its state at time t is given by

$$|\Psi(t)\rangle = \exp\left\{-\mathrm{i}\frac{\mathcal{H}}{\hbar}(t - t_\circ)\right\}|\Psi(t_\circ)\rangle. \tag{16.1}$$

For the description of the states of the system, we may take the basis of the eigenstates $|\phi_\epsilon^{(r)}\rangle$ of the "unperturbed" Hamiltonian $\mathcal{H}_\circ$ such that

$$\mathcal{H}_\circ|\phi_\epsilon^{(r)}\rangle = \epsilon|\phi_\epsilon^{(r)}\rangle,$$

[2] Boltzmann recognized the necessity of such hypothesis and justified it on the basis of mean free paths of the molecules much longer than the mean intermolecular distance.

[3] I am grateful to Massimo Fischetti for pointing out to me the importance of these papers.

where r distinguishes between degenerate states. Any state of the system can then be written as the linear combination

$$|\psi\rangle = \sum_{\epsilon,r} C(\epsilon, r)|\phi_\epsilon^{(r)}\rangle. \tag{16.2}$$

The states in (16.2) are not, in general, eigenstates of the total Hamiltonian and therefore are not stationary states. Under what conditions it is possible to obtain, from (16.1), a *master equation* like

$$\mathrm{d}P_\alpha/\mathrm{d}t = \sum_\beta (W_{\alpha\beta}P_\beta - W_{\beta\alpha}P_\alpha), \tag{16.3}$$

describing the irreversible approach to equilibrium? In the above equation, P_α is the probability of finding the system in the eigenstate (or in a group of eigenstates) of $\mathcal{H}_\circ$ labeled by α, and $W_{\alpha\beta}$ are the corresponding transition rates.[4]

The path from (16.1) to (16.3) requires, in general, that the perturbation $\lambda\mathcal{V}$ is small, and that the system has a large number of degrees of freedom. Furthermore, Van Hove [457] showed that the master equation can be obtained in the (not exhaustive) following cases, related to the initial state of the system:

(a) If the coefficients $C(\epsilon, r)$ in (16.2) of the initial state vary slowly over energy intervals of the order of $\delta\epsilon$, a master equation can be written and equilibrium is approached with a relaxation time of the order of $\hbar/\delta\epsilon$. A key point in the derivation by Van Hove is that interference between the waves resulting from two successive transitions is negligible if the transitions occur at a time distance greater than $\hbar/\delta\epsilon$.

(b) If the coefficients $C(\epsilon, r)$ in (16.2) of the initial state are different from zero only inside an interval of energy $\delta\epsilon$, a master equation can be written and equilibrium is approached with a relaxation time of the order of $\hbar/\delta\epsilon$.

The reader will easily identify the connections between points (a) and (b) above with the time-energy uncertainty relations: since the time of approach to equilibrium is a time during which the system varies appreciably, it must be related to the uncertainty on the energy of the system implied in the properties required to the initial states.

(c) In case the coefficients $C(\epsilon, r)$ in (16.2) of the initial states have random phases, a master equation like (16.3) can always be written. Van Hove showed that the random phases are necessary only at the initial times, and not at all times, as previously assumed.

To proceed in our analysis, it is useful to introduce now the concept of coherence time, strictly related to those of dephasing and entanglement.

[4] Note that to obtain the BE for the electrons in a crystal we have to make a further step, because in the above we have considered the whole isolated system (for our case the system of the electrons plus the phonons), while for the BE we have to reduce the problem to the description of the distribution function of only electrons. Van Hove considers this problem in [460].

16.1.3 Coherence, Dephasing, and Entanglement

Let us consider the compound system formed by an electron and the rest of the crystal in which the electron is traveling. Let us also assume that at the initial time $t_\circ$ the compound system is described by a factorized wavefunction

$$\Psi(\boldsymbol{r}, \boldsymbol{R}, t_\circ) = \Psi_e(\boldsymbol{r}, t_\circ)\, \Psi_c(\boldsymbol{R}, t_\circ), \tag{16.4}$$

where $\boldsymbol{r}$ is the electron coordinate, and $\boldsymbol{R}$ represents all the other coordinates describing the state of the crystal as, for example, the positions of all atoms. If the position of the electron is measured, the probability density of finding it in $\boldsymbol{r}$ is given by

$$P(\boldsymbol{r}) = \int |\Psi(\boldsymbol{r}, \boldsymbol{R}, t_\circ)|^2 \, \mathrm{d}\boldsymbol{R} = |\Psi_e(\boldsymbol{r}, t)|^2 \int |\Psi_c(\boldsymbol{R}, t)|^2 \, \mathrm{d}\boldsymbol{R} = |\Psi_e(\boldsymbol{r}, t)|^2 ,$$

where the electronic and crystal wavefunctions have been separately normalized to unity. Thus, the electronic wavefunction $\Psi_e(\boldsymbol{r}, t)$ may yield the well-known interference phenomena if, for example, it is separated into two parts

$$\Psi_e(\boldsymbol{r}, t) = \Psi_e^{(1)}(\boldsymbol{r}, t) + \Psi_e^{(2)}(\boldsymbol{r}, t),$$

and the two parts are brought to converge in the same space region. This interference may occur as long as the total wavefunction maintains the factorization as given by (16.4).

The Hamiltonian of the system contains three terms:

$$\mathcal{H} = \mathcal{H}_e + \mathcal{H}_c + \mathcal{H}_{ec}. \tag{16.5}$$

The first two terms determine, separately, the dynamics of the electron and of the crystal, respectively. Their applications generate the evolutions of the two separate wavefuntions, keeping their factorization as in (16.4), so that the electron interference phenomena are maintained. The third term in (16.5) is the Hamiltonian that describes the interaction between the electron and the rest of the crystal, and its application destroys the factorization. If, to be more specific, we assume that after some time the total wavefunction is the sum of two separate products

$$\Psi(\boldsymbol{r}, \boldsymbol{R}, t) = \Psi_e^{(1)}(\boldsymbol{r}, t)\, \Psi_c^{(1)}(\boldsymbol{R}, t) + \Psi_e^{(2)}(\boldsymbol{r}, t)\, \Psi_c^{(2)}(\boldsymbol{R}, t),$$

the two parts of the electron wavefunction will not generate interference because in the squared modulus they cannot be factorized.

We define *coherence time* the time during which the wavefunction maintains the initial factorization leading to electron interference phenomena.

This concept is generalizable to any compound system formed by two subsystems A and B. If the initial state may be written as a direct product of vector states of the two subsystems,

$$|\Psi\rangle = |\Psi_a(1)\rangle|\Psi_b(2)\rangle,$$

after the coherence time, which is longer when the interaction between the two subsystems is weaker, the interaction Hamiltonian will destroy the factorization. In particular, when the vector state is not the product of separate states of the subsystems, the two subsystems are said to be *entangled*.[5] In the case, of interest for us, of an electron in the crystal, it becomes entangled with the phonon system. As seen from the point of view of the electron subsystem, it loses the description in terms of a single particle wavefunction, and we say that a *dephasing process* occurs. Thus, decoherence and entanglement are two faces of the same physical phenomenon, and the *dephasing time* is defined essentially in the same way as the coherence time.

As it regards the interaction of electrons with impurities, if the states of the impurities are not affected by the scattering processes, their wavefunctions will always be factorisable, so that the electron does not lose its coherence. In fact, the constant potential field due to the impurity may be included in the electronic Hamiltonian, so that the interaction process can be included in the unperturbed single-particle dynamics.

As it regards, finally, the scattering with other electrons, it produces dephasing of each single electron, since the many-electron wavefunction cannot be described as the product of the wavefunction of the single electron under consideration, moreover indistinguishable, times the wavefunction of all other electrons in the crystal.

16.1.4 When is Quantum Transport Necessary?

Characteristic Times

From the point of view of the time scale, for the present discussion on the need of a quantum theory of transport, it is useful to identify a number of time scales characteristic of electron transport.

Free-flight time τ_f: This time is also often called *mean free time* or *scattering time*. In a semiclassical description collisions occur at well-defined times,

[5] This concept is clearly related to the quantum measurement problem. In fact, if the state of the total system, formed by the subsystems "1" and "2", is given by the vector

$$|\Psi\rangle = |\Psi_a(1)\rangle|\Psi_b(2)\rangle + |\Psi_{a'}(1)\rangle|\Psi_{b'}(2)\rangle, \tag{16.6}$$

the two subsystems are entangled and can also be spatially separated. Now assume that a measurement is performed on one of the two subsystems, say on the subsystem "1", able to discriminate between the state $|\Psi_a\rangle$ and $|\Psi_{a'}\rangle$ and assume that as a result of the measurement the subsystem "1" is found in the state $|\Psi_a\rangle$. According to the orthodox Copenhagen interpretation, as a consequence of this measurement, the state of the compound system "collapses" into the first term of the combination in (16.6), and a successive measurement on the subsystems "2" will necessarily yield the result corresponding to the state $|\Psi_b\rangle$.

and τ_f is defined as the average time elapsed between two successive collisions. In quantum terms may be defined as the *lifetime* of the electron state, due to electron collisions.

Momentum relaxation time τ: It is the time required by an electron to lose memory of its initial momentum. It is therefore also the time necessary to the scattering processes to relax a fluctuation of the mean electron momentum. It has been widely discussed in Chap. 11, in connection with the elementary theory of linear conductivity.

Energy relaxation time τ_ϵ: It is the time required by an electron to lose memory of its initial energy. It is therefore also the time necessary to the scattering processes to relax a fluctuation of the mean electron energy. Since in linear transport the energy distribution of the electrons remains equal to the equilibrium one, the energy relaxation time is important only in nonlinear regime. In fact, it has been seen in Chap. 13 in connection with nonlinear transport.

Coherence time τ_ϕ: This is the time introduced in the previous section. It is typical of a quantum description of the electron dynamics and corresponds to the time during which the dynamics of the electron is correctly described by a Schrödinger equation with a single-particle Hamiltonian.

Collision-duration time τ_c: In the semiclassical description of transport, electrons are considered to perform free flights interrupted by instantaneous changes of momentum due to collisions. Nevertheless, even in a purely classical description a collision is not instantaneous. The collision duration, in classical terms, can be defined as the time spent by the colliding particle within the region where the scattering field is present. In quantum terms, the definition is ill-defined, and several interpretations have been proposed. (see, for example, [283] and [56]). Numerical estimates of the different approaches, however, seem to reach similar results, as we shall now see.

Sometimes the collision duration is associated with the time which is necessary, in a collision, to recover the δ-function of energy conservation. This time would depend on the uncertainty we are willing to accept in the electron energy. If the required precision is of the order of 1 meV, this time is of the order of 10^{-13} s. In the meantime, however, a new collision may occur. The collisional broadening of the electron energy does not depend on our tolerance, but on the average time interval between two scattering events. In fact, the collisional broadening is simply related to the electron lifetime by the uncertainty relation.

In the specific case of electron–phonon interaction, a different approach to define a collision duration in quantum terms may consider the time necessary for the electron to feel the periodicity of the phonon. For low-frequency phonons and fast electrons this time is given by the time taken by the electron to travel a wavelength, of the order of $1/qv$. Since the wavevector q of the exchanged phonon is of the order of the electron wavevector, this time is of the order of $\hbar/KT$. At ordinary temperature this time is of the order of 10^{-13}–10^{-14} s. For slow electrons, the time under examination is of the

order of the phonon period $\hbar/KT_{op}$, if KT_{op} is the phonon equivalent energy, quantitatively similar to the previous one.

A third approach to the problem of the collision duration considers the scattering completed when a successive scattering does not interfere with the previous one. According to the discussion above on the analysis of Van Hove in [457], this time is related to the energy broadening of the electron. In [56] the collision duration is also identified as the time required to build up correlation between the initial an the final state, and then to destroy this correlation as the collision is completed, and a value of the order of 10^{-15}–10^{-14} s is obtained.

Electrons are not Classical Particles

At this point, we should be able to find out when we have to abandon Boltzmann equation for a more rigorous quantum transport equation. For this purpose, let us analyze the approximations made for writing the BE and consider when they fail.

First of all, in semiclassical transport we assign to each electron "reasonably well defined" position and momentum during the free flight between two successive collisions. We know that this is possible only within the limits dictated by the uncertainty relations. For our approximation to be acceptable, we must be able to conceive wave packets with momentum uncertainty Δp much less than their average momentum p and, at the same time, a position uncertainty δx much less than the mean free path l:

$$\Delta p \ll p \quad , \qquad \delta x \ll l.$$

From the uncertainty relations, we then have

$$\hbar \sim \Delta p \Delta x \ll pl \sim p\left(\frac{p}{m}\tau_f\right) \sim K_{\mathrm{B}} T \tau_f,$$

or

$$\tau_f \gg \frac{\hbar}{KT} \sim 10^{-13} - 10^{-14}\,\mathrm{s} \tag{16.7}$$

at ordinary temperatures.

As second point, we already noted that in the traditional treatment of the BE collisions are, in general, assumed instantaneous in time. For this approximation to be acceptable, the collision duration should be much shorter than the time between collisions. From the estimates made above of the latter, we find again the condition given in (16.7).

A further problem related to the collision duration is the so-called *intracollisional field effect* [26,105]. In the BE, the transition rates are usually assumed to be independent of applied field, even though this is not an assumption necessary to write down the BE. During the collision time, however, an applied electric field may act on the initial and final electron states, changing their wavevectors with time. As a consequence, the energy difference between the

two states varies with time, and the wave dynamics that generate the transition is modified by the field. Such a process, called intracollisional field effect, modifies the transition rates. To estimate when this effect may be relevant, let us compare the energy $\Delta\epsilon$ transferred to the electron by the field during the collision, with the average electron energy:

$$\frac{\Delta\epsilon}{KT} \sim v\tau_c \frac{eE}{KT},$$

which is of the order of one for 10^5 V/cm. Modern technology provides physical systems where local fields reach values of the order of 10^6 V/cm, and the intracollisional field effect has been found to be actually relevant [378].

As final important point, let us consider that, apart from isolated scattering events, electrons are described by single-particle states with well-defined energy. On the other hand, interactions induce collisional broadening. A rigorous definition of such phenomenon will be seen in the last part of the book, devoted to the Green-function method, but we have already seen that this quantity is related to the mean free time τ_f through the uncertainty relation. If we require that the uncertainty on the electron energy is much smaller than the energy itself, we need

$$\Delta\epsilon \approx \hbar/\tau_f \ll \epsilon \sim K_{\mathrm{B}}T,$$

and we find again the condition in (16.7)

All the above considerations appear to be related to each other and due to the wave nature of electron dynamics, synthesized by the uncertainty relations. The critical parameter for the applicability of the semiclassical transport theory is, as physically plausible, the time τ_f between collisions in the semiclassical theory itself, that should be longer than 10^{-14} s. At the energy reached by hot electrons, of the order or greater than the eV, this condition is not always fulfilled, and a quantum theory of transport must be pursued. The success of the semiclassical BE in describing the behavior of modern electronic devices, when no specific quantum effects are present, such as tunneling or resonances, seems to indicate, however, that the validity of semiclassical theory goes beyond the above limitations. We shall see in Chap. 17, devoted to the Wigner function, a good reason for such a success.

16.2 The Density Matrix

Given a physical system, described by the state vector $|\phi\rangle$, the expectation value of a measurement of a quantity A, according to quantum physics, is given by

$$\langle A\rangle = \langle\phi|\mathcal{A}|\phi\rangle, \tag{16.8}$$

where $\mathcal{A}$ is the operator representing the physical quantity of interest, and the state vector $|\phi\rangle$ is supposed normalized to unity.

If the state of the system is not precisely known, we have to use the methods of statistical physics to deal with our incomplete knowledge, as described in Chap. 3, and the expectation value given by (16.8) is replaced by its ensemble average:

$$\overline{\langle A\rangle} = \overline{\langle\phi|\mathcal{A}|\phi\rangle},$$

where the overbar indicates an average to be performed over a suitable statistical ensemble that accounts for our partial knowledge of the system. If $\{|\varphi_i\rangle\}$ is a complete set of basis vectors, the above equation can be written as (cf. (A.8) of Appendix A)

$$\overline{\langle A\rangle} = \sum_i \overline{\langle\phi|\mathcal{A}|\varphi_i\rangle\langle\varphi_i|\phi\rangle} = \sum_i \langle\varphi_i|\overline{|\phi\rangle\langle\phi|}\mathcal{A}|\varphi_i\rangle = Tr(\overline{|\phi\rangle\langle\phi|}\mathcal{A}). \tag{16.9}$$

From the above result, it is easy to recognize that in quantum statistical physics the mathematical instrument to be used is the *density matrix* operator ρ defined as $\overline{|\phi\rangle\langle\phi|}$, and that the average physical quantities are given by

$$\boxed{\overline{\langle A\rangle} = Tr(\rho\mathcal{A})\ , \qquad \rho \equiv \overline{|\phi\rangle\langle\phi|}} \tag{16.10}$$

It is easy to show that the diagonal element ρ_{ii} of the density matrix gives the probability P_i of finding the system in the state $|\varphi_i\rangle$. In fact, $P^{(j)} = 1/N$ is the probability of selecting at random the j-th system in the ensemble with N elements. Once this system, in the state $|\varphi^{(j)}\rangle$, has been selected, the probability that a proper measurement yields the result corresponding to the state $|\varphi_i\rangle$ is given by

$$P(i|j) = |\langle\varphi_i|\phi^{(j)}\rangle|^2.$$

Thus

$$P_i = \sum_j P^{(j)}P(i|j) = \sum_j \frac{1}{N}|\langle\varphi_i|\phi^{(j)}\rangle|^2 = \sum_j \frac{1}{N}\langle\varphi_i|\phi^{(j)}\rangle\langle\phi^{(j)}|\varphi_i\rangle,$$

or

$$P_i = \langle\varphi_i|\overline{|\phi\rangle\langle\phi|}\varphi_i\rangle = \rho_{ii}.$$

To some extent, therefore, the diagonal elements of the density matrix are the quantum analog of the classical distribution function, and its nondiagonal elements account for the fact that the states can be in a superposition of different states of the chosen representation $|\varphi_i\rangle$.

A thorough discussion on the properties and use of the density matrix is given in [441].

Time Evolution of the Density Matrix

When dealing with transport problems, and in general with nonequilibrium properties of a systems, we are interested in studying the time evolution of the

average quantities in (16.10). Since in the Schrödinger picture the time evolution is carried by the state vectors, in the expression (16.10) the time evolution is assigned to the density matrix. We may then start by the operator $\mathcal{U}(t, t_\circ)$ that generates the time evolution of the state vectors of the ensemble as in (2.2). The time evolution of the density matrix operator in the Schrödinger picture is then immediately obtained from its definition (16.10) and the time evolution of the state vectors:

$$\rho(t) = \mathcal{U}(t, t_\circ)\rho(t_\circ)\mathcal{U}^\dagger(t, t_\circ). \tag{16.11}$$

By differentiating with respect to time, we obtain the *von Neumann equation*, often called Liouville–von Neumann equation, for the density matrix in the Schrödinger picture:

$$\boxed{\mathrm{i}\hbar \tfrac{\partial}{\partial t}\rho(t) = [\mathcal{H}, \rho(t)]} \tag{16.12}$$

In the Heisenberg picture, the density matrix does not depend on time, since the state vectors are constant.

If the total Hamiltonian is split into two parts, $\mathcal{H}_\circ + \mathcal{H}'$, where $\mathcal{H}'$ is considered as a perturbation, the interaction picture can be used (see Sect. 2.2.4); the time evolution of the state vector is given by (2.8), and the von Neumann equation in the interaction picture is

$$\mathrm{i}\hbar \frac{\partial}{\partial t}\rho_I(t) = [\mathcal{H}'_I(t), \rho_I(t)]\,, \tag{16.13}$$

where ρ_I and $\mathcal{H}'_I$ are the density matrix and the interaction Hamiltonian in the interaction picture, respectively. The above equation is convenient in perturbation expansions since only the perturbation Hamiltonian appears explicitly in the commutator.

The time evolution of an average quantity $\overline{\langle A \rangle}$, which is always given by (16.10), is due to the evolution of ρ in the Schrödinger picture and to the evolution of $\mathcal{A}$ in the Heisenberg picture. In the interaction picture, $\mathcal{A}$ carries the time evolution due to the unperturbed Hamiltonian, and ρ the time evolution due to the perturbation.

Density Matrix at Equilibrium

The time evolution for the matrix elements of ρ in the energy representation is

$$\mathrm{i}\hbar \frac{\partial}{\partial t}\rho_{ij}(t) = ([\mathcal{H}, \rho(t)])_{ij} = \langle \varphi_i | \mathcal{H}\rho - \rho\mathcal{H} | \varphi_j \rangle = (\epsilon_i - \epsilon_j)\rho_{ij},$$

where $|\varphi_i\rangle$ are the Hamiltonian eigenstates with eigenvalues ϵ_i. Thus, the condition for the density matrix to be in equilibrium is to be diagonal in the energy representation. Since the diagonal elements yield the probabilities of finding the system with energy ϵ_i and in the canonical ensemble these

probabilities are given by the canonical distribution in (3.14), the equilibrium density-matrix operator in the canonical ensemble is

$$\boxed{\rho_\circ = \frac{\mathrm{e}^{-\beta\mathcal{H}}}{Tr(\mathrm{e}^{-\beta\mathcal{H}})} = \frac{1}{Z}\mathrm{e}^{-\beta\mathcal{H}} \quad , \quad Z \equiv Tr(\mathrm{e}^{-\beta\mathcal{H}})} \tag{16.14}$$

where $\beta = 1/K_{\mathrm{B}}T$, and Z is the partition function. In the grand-canonical ensemble the grand-canonical distribution is given in (3.15), and the equilibrium density-matrix operator is

$$\rho_\circ = \frac{\mathrm{e}^{-\beta(\mathcal{H}-\mu\mathcal{N})}}{Tr(\mathrm{e}^{-\beta(\mathcal{H}-\mu\mathcal{N})})}, \tag{16.15}$$

where μ is the electrochemical potential, and $\mathcal{N}$ the number of particles in the system.

16.3 Reduced Density Matrix

Often we are interested in physical quantities which depend on a subsystem of the entire system under study. In the case of interest for us, we are dealing with electronic quantities in a system formed by N electrons interacting with phonons. In the coordinate representation the state vectors, and therefore the density matrix, will be functions of the electron coordinates $\boldsymbol{x} \equiv (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ and the phonon variables ξ:

$$\langle \boldsymbol{x}, \xi | \, \rho \, | \boldsymbol{x}', \xi' \rangle \equiv \rho(\boldsymbol{x}, \xi, \boldsymbol{x}', \xi').$$

If an observable $\mathcal{A}^{(el)}$ acts only on the electron variables, then

$$A^{(el)}(\boldsymbol{x}, \xi, \boldsymbol{x}', \xi') = \langle \boldsymbol{x}, \xi | A^{(el)} | \boldsymbol{x}', \xi' \rangle = A^{(el)}(\boldsymbol{x}, \boldsymbol{x}')\delta(\xi - \xi'),$$

and its mean value is given by

$$\overline{\langle A^{(el)} \rangle} = Tr(\rho \mathcal{A}^{(el)}) = \int\int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \int\int \mathrm{d}\xi\mathrm{d}\xi' \rho(\boldsymbol{x}, \xi, \boldsymbol{x}', \xi') A^{(el)}(\boldsymbol{x}', \boldsymbol{x})\delta(\xi - \xi'),$$

or

$$\overline{\langle A^{(el)} \rangle} = Tr\left[\rho^{(el)} \mathcal{A}^{(el)}\right], \tag{16.16}$$

where

$$\rho^{(el)}(\mathbf{x}, \mathbf{x}') \equiv \int \rho(\mathbf{x}, \xi, \mathbf{x}', \xi) \, \mathrm{d}\xi \equiv Tr_\xi(\rho) \tag{16.17}$$

is the *reduced electron density matrix*.

In practice, the observable $\mathcal{A}^{(el)}$ is often the average over the particles of a single-particle quantity $\mathcal{A}^{(sp)}$. This is the case, for example, of the drift velocity or the mean kinetic energy of the electron gas. In such a case, the average $\overline{\langle \mathcal{A}^{(el)} \rangle}$ is expressed as

$$\overline{\langle \mathcal{A}^{(el)} \rangle} = \int \mathrm{d}\mathbf{x} \int \mathrm{d}\mathbf{x}' \rho^{(el)}(\mathbf{x}, \mathbf{x}') \frac{1}{N} \sum_{i=1}^{N} \mathcal{A}^{(sp)}(\mathbf{r}_i', \mathbf{r}_i). \tag{16.18}$$

Owing to the symmetry of the wavefunctions of identical particles, and therefore of the corresponding density matrix, the N terms in the sum yield the same integral, and the above becomes

$$\int \mathrm{d}\mathbf{r}_1 \ldots \int \mathrm{d}\mathbf{r}_N \int \mathrm{d}\mathbf{r}_1' \ldots \int \mathrm{d}\mathbf{r}_N' \rho^{(el)}(\mathbf{r}_1, \ldots, \mathbf{r}_N; \mathbf{r}_1', \ldots, \mathbf{r}_N') \mathcal{A}^{(sp)}(\mathbf{r}_1', \mathbf{r}_1).$$

Such a result suggests the introduction of a *single-particle density matrix* $\rho^{(sp)}$ as the integral of the reduced electronic density matrix in (16.17) over all coordinates but one:

$$\rho^{(sp)}(\mathbf{r}, \mathbf{r}') \equiv \int \mathrm{d}\mathbf{r}_2 \ldots \int \mathrm{d}\mathbf{r}_N \int \mathrm{d}\mathbf{r}_2' \ldots \int \mathrm{d}\mathbf{r}_N' \rho^{(el)}(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N; \mathbf{r}', \mathbf{r}_2', \ldots, \mathbf{r}_N').$$

The average value in (16.18) is then given by

$$\overline{\langle A^{(el)} \rangle} = \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \rho^{(sp)}(\mathbf{r}, \mathbf{r}') A^{(sp)}(\mathbf{r}', \mathbf{r}) = Tr \left[\rho^{(sp)} \mathcal{A}^{(sp)} \right]. \tag{16.19}$$

The reduction of the density matrix in (16.17) is obtained by "tracing away" the variables we are not interested in. This, however, by no means imply separation of the dynamics. If we try to take the trace over the phonon variables of the von Neumann equation (16.12), we shall not obtain a close equation for the reduced density matrix since we cannot take the trace inside the commutator: the trace operation does not commute with the Hamiltonian since the latter contains phonon coordinates. The separation between phonon and electron dynamics is a major problem in electron transport theory, more or less explicitly dealt with throughout this book. Specific discussions on this point can be found, for example, in [15, 458, 494, 495].

16.4 Kubo Formula

In Chap. 12, it was shown that the velocity autocorrelation function is strictly related to the diffusivity and, through the Einstein relation, to the conductivity. It was mentioned, there, that this relation is a special case of a general theorem, called *fluctuation-dissipation theorem*, whose content can be physically understood taking into account that both phenomena are due to friction: friction determines both the damping of thermally generated fluctuations and the resistance offered to an external driving force, generating, in this way, dissipation. In this chapter, we shall deal with the quantum version of the Einstein relation, i.e., with the *Kubo formula* [253] for the linear response. It is often called Kubo–Kirkwood formula, with reference to a previous work of Kirkwood [240], a major step in the fluctuation-dissipation theory (see, for

example, [87, 254, 255]). We shall see that the response to an external driving force in the linear regime is strictly related to the fluctuations of the same quantity *in equilibrium conditions*. The importance of this result lies in the fact that, to get the linear response to an external agent, it is not necessary to obtain the out-of-equilibrium state of the system.

Linearization of the Von Neumann Equation

Let us consider a Hamiltonian given by the sum of two terms:

$$\mathcal{H}(t) = \mathcal{H}_\circ + \mathcal{H}'(t),$$

where $\mathcal{H}_\circ$ is the time-independent unperturbed Hamiltonian, which may contain also internal interactions, and $\mathcal{H}'(t)$ is a time-dependent external perturbation, driving the system out of equilibrium. The time dependence of $\mathcal{H}'$ is assumed to be the usual Fourier component:

$$\mathcal{H}'(t) = \mathcal{A}\mathrm{e}^{\mathrm{i}(\omega - \mathrm{i}\epsilon)t}. \tag{16.20}$$

The real part in the exponent is inserted for convergence, and corresponds to the physical assumption of an adiabatic switching of the perturbation from $t = -\infty$, when we assume that the density matrix was in equilibrium with the unperturbed Hamiltonian, given by (16.14). At the end of the calculation, the limit for $\epsilon \to 0$ must be taken.

In the interaction picture, the equation that governs the dynamics of the density matrix is given by the von Neumann equation (16.13) of the previous section where only the perturbation Hamiltonian is explicitly present. The information on the unperturbed system is contained in the picture. By integration, (16.13) yields

$$\rho_I(t) = \rho_I(-\infty) + \frac{1}{\mathrm{i}\hbar}\int_{-\infty}^{t} [\mathcal{H}'_I(t'), \rho_I(t')]\,\mathrm{d}t'.$$

Let us now move to the Schrödinger picture, knowing that at the initial condition the density matrix $\rho_\circ(\mathcal{H}_\circ)$ commutes with $\mathcal{H}_\circ$:

$$\rho(t) = \rho_\circ(\mathcal{H}_\circ) + \frac{1}{\mathrm{i}\hbar}\int_{-\infty}^{t} \mathrm{e}^{-\mathrm{i}(\mathcal{H}_\circ/\hbar)(t-t')}[\mathcal{H}'(t'), \rho(t')]\mathrm{e}^{\mathrm{i}(\mathcal{H}_\circ/\hbar)(t-t')}\mathrm{d}t'. \tag{16.21}$$

Since we are looking for the linear response of ρ to $\mathcal{H}'$, let us put

$$\rho(t) = \rho_\circ(\mathcal{H}_\circ) + \Delta\rho(t).$$

In (16.21) the integral term is the difference $\Delta\rho$, and inside the integral the density matrix is multiplied by $\mathcal{H}'$ so that, to first order, we may substitute it with $\rho_\circ$ as given by (16.14):

$$\Delta\rho(t) = \frac{1}{\mathrm{i}\hbar Z}\int_{-\infty}^{t} \mathrm{e}^{-\mathrm{i}(\mathcal{H}_\circ/\hbar)(t-t')}[\mathcal{H}'(t'), \mathrm{e}^{-\beta\mathcal{H}_\circ}]\mathrm{e}^{\mathrm{i}(\mathcal{H}_\circ/\hbar)(t-t')}\mathrm{d}t'. \tag{16.22}$$

A General Identity

The commutator in (16.22) contains the effect of the "driving perturbation" $\mathcal{H}'$ on the equilibrium density matrix. We shall now transform this commutator deriving a general, important relation, due to Kubo [253]. In doing so, we shall keep the symbols $\mathcal{H}_\circ$, $\mathcal{H}'$, and β, with the purpose of staying in touch with the problem at hand, but the relation has a very general validity.

If $\mathcal{H}'$ and $\mathcal{H}_\circ$ are two operators independent of the number β, let us consider the following derivative:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}(\mathrm{e}^{\beta\mathcal{H}_\circ}\mathcal{H}'\mathrm{e}^{-\beta\mathcal{H}_\circ}) = \mathrm{e}^{\beta\mathcal{H}_\circ}[\mathcal{H}_\circ, \mathcal{H}']\mathrm{e}^{-\beta\mathcal{H}_\circ}.$$

Integrate from 0 to β:

$$\mathrm{e}^{\beta\mathcal{H}_\circ}\mathcal{H}'\mathrm{e}^{-\beta\mathcal{H}_\circ} - \mathcal{H}' = \int_0^\beta \mathrm{e}^{\beta'\mathcal{H}_\circ}[\mathcal{H}_\circ, \mathcal{H}']\mathrm{e}^{-\beta'\mathcal{H}_\circ}\mathrm{d}\beta'.$$

If we now multiply on the left by $\mathrm{e}^{-\beta\mathcal{H}_\circ}$, we obtain the general formula

$$[\mathcal{H}', \mathrm{e}^{-\beta\mathcal{H}_\circ}] = \mathrm{e}^{-\beta\mathcal{H}_\circ}\int_0^\beta \mathrm{e}^{\beta'\mathcal{H}_\circ}[\mathcal{H}_\circ, \mathcal{H}']\mathrm{e}^{-\beta'\mathcal{H}_\circ}\mathrm{d}\beta'. \tag{16.23}$$

This relation is of fundamental importance to understand the physical meaning of the Kubo formula. When we apply it to (16.22), we transfer the effect of the perturbation on the equilibrium density matrix (contained in the commutator in the l.h.s.) to the *equilibrium* dynamics of the perturbation (contained in the commutator in the r.h.s.), and therefore on the physical quantity that produces the response. For example, if the perturbation is due to an electric field, the perturbation Hamiltonian is proportional to the position x; the evolution of x is related to the velocity, so that we expect that the response of the system to an electric field is related to the behavior of the velocities of the system *at equilibrium*, as we shall see below.

Kubo Formula

Substitute (16.23) into (16.22) and obtain, after simple steps

$$\Delta\rho(t) = \frac{\rho_\circ}{\mathrm{i}\hbar}\int_0^\infty \mathrm{e}^{-\mathrm{i}(\mathcal{H}_\circ/\hbar)\tau}\mathrm{e}^{\mathrm{i}(\omega-i\epsilon)(t-\tau)}\int_0^\beta \mathrm{d}\beta'\mathrm{e}^{\beta'\mathcal{H}_\circ}[\mathcal{H}_\circ, \mathcal{A}]\mathrm{e}^{-\beta'\mathcal{H}_\circ}\mathrm{e}^{\mathrm{i}(\mathcal{H}_\circ/\hbar)\tau}\mathrm{d}\tau, \tag{16.24}$$

where we have used (16.14) and (16.20) and put $\tau = t - t'$. To be specific, let us now consider the response to an external, oscillating, electric field. The procedure can be applied identical to any other driving perturbation:

$$\mathcal{A} = \mathcal{P}\cdot\boldsymbol{E},$$

where $\mathcal{P}$ is the polarization operator

$$\mathcal{P} = \sum_i q_i \boldsymbol{r}_i,$$

due to charges q_i at positions $\boldsymbol{r}_i$. Furthermore,

$$\frac{1}{\mathrm{i}\hbar}[\mathcal{H}_\circ, \mathcal{P}\cdot\mathbf{E}] = \mathbf{E}\cdot\dot{\mathcal{P}} = \mathbf{E}\cdot\mathcal{J}\ , \quad \mathcal{J} \equiv \sum_i q_i \mathbf{v}_i. \tag{16.25}$$

Let us now consider the current evolved by the unperturbed Hamiltonian:

$$\mathcal{J}(t) \equiv \mathrm{e}^{\mathrm{i}\mathcal{H}_\circ t/\hbar}\mathcal{J}\mathrm{e}^{-\mathrm{i}\mathcal{H}_\circ t/\hbar}. \tag{16.26}$$

Then (16.24) becomes

$$\Delta\rho(t) = \rho_\circ \int_0^\infty \mathrm{e}^{-\mathrm{i}(\mathcal{H}_\circ/\hbar)\tau}\mathrm{e}^{\mathrm{i}(\omega-\mathrm{i}\epsilon)(t-\tau)} \int_0^\beta \boldsymbol{E}\cdot\mathcal{J}(-\mathrm{i}\beta'\hbar)\mathrm{d}\beta'\mathrm{e}^{\mathrm{i}(\mathcal{H}_\circ/\hbar)\tau}\mathrm{d}\tau,$$

Now we use this expression to evaluate the μ component of the mean current at time t:

$$\langle\mathcal{J}_\mu\rangle_t = Tr\{\mathcal{J}_\mu\rho(t)\} = Tr\{\mathcal{J}_\mu\left[\rho_\circ(t) + \Delta\rho(t)\right]\}. \tag{16.27}$$

The equilibrium density matrix $\rho_\circ$ does not yield any current, and we may substitute, in (16.27), the expression found above for $\Delta\rho$. Furthermore, if we consider for simplicity a homogeneous system of volume V, the current density $j_\mu(t) = \langle\mathcal{J}_\mu(t)\rangle/V$ is given by

$$\frac{1}{V}Tr\left\{\mathcal{J}_\mu\rho_\circ \int_0^\infty \mathrm{d}\tau\mathrm{e}^{-\mathrm{i}(H_\circ/\hbar)\tau}\mathrm{e}^{\mathrm{i}(\omega-\mathrm{i}\epsilon)(t-\tau)} \int_0^\beta \mathrm{d}\beta' \sum_\nu E_\nu\mathcal{J}_\nu(-\mathrm{i}\hbar\beta')\mathrm{e}^{\mathrm{i}(H_\circ/\hbar)\tau}\right\}.$$

Comparing this with the definition of conductivity, $j_\mu(t) = \sum_\nu \sigma_{\mu\nu}(\omega)E_\nu\mathrm{e}^{\mathrm{i}\omega t}$, we obtain

$$\sigma_{\mu\nu}(\omega) = \frac{1}{V}Tr\left\{\mathcal{J}_\mu\rho_\circ \int_0^\infty \mathrm{d}\tau\mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\epsilon)\tau}\mathrm{e}^{-\mathrm{i}(H_\circ/\hbar)\tau} \int_0^\beta \mathrm{d}\beta'\mathcal{J}_\nu(-\mathrm{i}\hbar\beta')\mathrm{e}^{\mathrm{i}(H_\circ/\hbar)\tau}\right\}.$$

Let us now observe that $\rho_\circ$ commutes with $H_\circ$, so that we may move $\rho_\circ$ in front of the integral in β'; then we apply the cyclic property of the trace moving $\mathcal{J}_\mu$ and the exponential with $H_\circ$ to the right; finally we observe that a trace with $\rho_\circ$ means a mean value at equilibrium:

$$\sigma_{\mu\nu}(\omega) = \frac{1}{V}\left\langle \int_0^\infty \mathrm{d}\tau\mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\epsilon)\tau} \int_0^\beta \mathrm{d}\beta'\mathcal{J}_\nu(-\mathrm{i}\hbar\beta')\mathrm{e}^{\mathrm{i}(H_\circ/\hbar)\tau}\mathcal{J}_\mu\mathrm{e}^{-\mathrm{i}(H_\circ/\hbar)\tau}\right\rangle.$$

or, remembering (16.26),

$$\boxed{\sigma_{\mu\nu}(\omega) = \tfrac{1}{V}\int_0^\infty \mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\epsilon)\tau}\int_0^\beta \langle \mathcal{J}_\nu(-\mathrm{i}\hbar\beta')\mathcal{J}_\mu(\tau)\rangle \mathrm{d}\beta'\mathrm{d}\tau} \tag{16.28}$$

This is Kubo formula in its current-correlation version and must be evaluated in the limit of $\epsilon \to 0$. The mean value that appears in this formula is evaluated in the equilibrium ensemble, where the mean current is zero. Thus, the currents which appear in Kubo formula are current fluctuations. The imaginary time as argument of the first current factor is symbolic, related to the expression (16.26), which in this case carries the effect of the temperature on the fluctuations.

As mentioned above, Kubo formula (16.28) is very general, valid for the linear response of a system to any external driving perturbation. On its basis, a number of symmetry relations of the linear-response coefficients can be obtained [255], such as Onsager relations, already encountered in Sect. 11.3.3

The Classical Limit

We may easily obtain the classical limit of Kubo formula assuming $\hbar$ negligibly small. The integrand in the second integral in (16.28) becomes independent of β' and (16.28) reduces to

$$\sigma_{\mu\nu}(\omega) = \frac{1}{V}\frac{1}{KT}\int_0^\infty \mathrm{e}^{-\mathrm{i}\omega\tau}\langle J_\nu(0)J_\mu(\tau)\rangle \mathrm{d}\tau. \tag{16.29}$$

According to the definition of $\mathcal{J}$ in (16.25), the above becomes

$$\sigma_{\mu\nu}(\omega) = \frac{1}{V}\frac{q^2}{KT}\int_0^\infty \mathrm{e}^{-\mathrm{i}\omega\tau}\left\langle \sum_{ij} v_{i\nu}(0)v_{j\mu}(\tau)\right\rangle \mathrm{d}\tau.$$

If there is no correlation between the particles, this becomes

$$\sigma_{\mu\nu}(\omega) = \frac{N}{V}\frac{q^2}{KT}\int_0^\infty \mathrm{e}^{-\mathrm{i}\omega\tau}\,\langle v_\nu(0)v_\mu(\tau)\rangle\,\mathrm{d}\tau,$$

where N is the total number of particles. The above equation yields the frequency-dependent classical conductivity in terms of the Fourier transform of the velocity autocorrelation function and therefore of the diffusivity. Thus, the results of Chap. 12 on Einstein relation and Johnson–Nyquist noise are recovered.

16.5 The Path-Integral Approach

Concluding this introduction to quantum transport, we quickly mention a method, the *path-integral approach*, which is sometimes used in specific calculations. It is not easy to implement from a numerical point of view, as most

quantum approaches to electron transport. In fact, it is very fundamental in its principle, and very often other methods, based on numerical evaluations of the reduced density matrix or of the Wigner function, which will be seen in next chapter, can be recognized as special versions of the path-integral approach.

The Feynman path-integral theory [147] is an alternative approach to quantum mechanics, equivalent to Schrödinger equation. It states that all possible paths of a system from an initial state to a final one are to be considered as simultaneously realized, and their complex amplitudes add up, interfering, to give the probability amplitude of finding the system in the final state. The fundamental equation in this approach is an explicit expression for the evolution operator as an integral over all possible paths of the exponential of the classical action. If q represents a set of classical Lagrangian variables of the system, and $q(\tau)$ one given trajectory, or path, from the initial values $q_i = q(t_i)$ to the final values $q_f = q(t)$, the evolution operator, in q representation, is written as

$$U(q_f, q_i, t, t_i) = \int_{q_i,t_i}^{q_f,t} \mathcal{D}q(\tau) \mathrm{e}^{(\mathrm{i}/\hbar)S[q(\tau)]}. \tag{16.30}$$

Here $\int \mathcal{D}q(\tau)$ indicates the integral over all paths that connect the initial state $\{q_i, t_i\}$ to the final state $\{q_f, t\}$; $S[q(\tau)]$ is the classical action evaluated over each given trajectory in the integral, defined as

$$S[q(\tau)] = \int_{t_i}^{t} L(q(\tau), \dot{q}(\tau), \tau)\ \mathrm{d}\tau,$$

where $L(q, \dot{q}, \tau)$ is the Lagrangian of the system.

By applying the expression of the evolution operator given in (16.30) to the evolution of the density matrix given in (16.11), the following expression for ρ is obtained:

$$\begin{aligned}\rho(q, q', t) = \\ = \int \mathrm{d}q_i \int \mathrm{d}q_i' \int_{q_i,t_i}^{q,t} \mathcal{D}q(\tau) \int_{q_i',t_i}^{q',t} \mathcal{D}q'(\tau) \mathrm{e}^{(\mathrm{i}/\hbar)[S(q(\tau))-S(q'(\tau))]} \rho(q_i, q_i', t_i).\end{aligned} \tag{16.31}$$

The above equation may be elaborated in a useful way by factorizing the effect of "perturbing" agents with respect to the system of interest. In our case, if x indicates the variables of the subsystem formed by an electron, and X indicates the phonon variables, the exponential in (16.31) can be factorized as follows:

$$\begin{aligned}\int \dots \mathrm{e}^{(\mathrm{i}/\hbar)[S(x)-S(x')]} \mathrm{e}^{(\mathrm{i}/\hbar)[S(X)-S(X')+S(x,X)-S(x',X')]} \\ \times \mathcal{D}x(\tau)\mathcal{D}x'(\tau)\mathcal{D}X(\tau)\mathcal{D}X'(\tau),\end{aligned}$$

where $S(x)$ and $S(X)$ are the actions for the electron and the phonons, respectively, and $S(x, X)$ is the action of interaction between the two subsystems. The integral over the paths of the interacting system involves only the second exponential, and an *influence functional* can be defined [146],

$$\mathcal{F}(x(\tau), x'(\tau)) = \int_{X_i, t_i}^{X, t} \mathcal{D}X(\tau) \int_{X_i', t_i}^{X', t} \mathcal{D}X'(\tau) \mathrm{e}^{(\mathrm{i}/\hbar)[S(X) - S(X') + S(x,X) - S(x', X')]},$$

such that the evolved density matrix is written as

$$\rho(q, q', t) = \int \mathrm{d}q_i \int \mathrm{d}q_i' \int \mathcal{D}x(\tau) \int \mathcal{D}x'(\tau)\, \mathcal{F}(x(\tau), x'(\tau)) \times \mathrm{e}^{(\mathrm{i}/\hbar)[S(x) - S(x')]} \rho(q_i, q_i', t_i). \quad (16.32)$$

Here, for any given path of the system of interest, $\mathcal{F}$ carries the information of the influence on that path by all possible paths of the interacting system.

However, the explicit evaluation of the influence functional and the subsequent evaluation of the path integral in (16.32) are in general prohibitively difficult, and approximations must be made, as in other quantum formulations of the transport problem. For systems where the coupling action is a linear function of the coordinates of the interacting parts and for systems weakly coupled, an analytical evaluation of the influence functional is possible [146]. For general systems, however, this is not true.

The main problem with the numerical application of the path-integral approach is that most of the paths in the integral yield contributions that cancel each other almost exactly, apart from the paths very close to the "extremant" one that yields the classical trajectory. In general, however, we are not allowed to cut out paths outside any given region, since in this way the necessary cancelation is interrupted. *This is a special aspect of the general problem always faced in all numerical approaches to quantum electron transport: that of the numerical evaluation of strongly oscillating functions.*

17

The Wigner-Function Approach to Quantum Transport

17.1 Introduction

Wigner introduced the function that carries his name in 1932 [472, 473] as an instrument to study quantum corrections to classical statistical mechanics. Even though the Wigner function (WF) cannot be strictly interpreted as a probability density, as demonstrated by the fact that it can assume negative values (see for example [439]), the very fact that it is defined in a phase space, together with its main properties and dynamical equation, makes it particularly useful to study quantum corrections to classical results and the classical limit to quantum physics. The WF has been widely employed in several fields of quantum statistical physics, such as molecular, atomic, and nuclear physics, quantum optics, quantum chemistry, quantum entanglement and entropy [496]. As it regards the use of the WF in electron transport, it has received great attention since the 1980s, when technological improvements required the development of a full quantum theory of electronic transport.

The approach to quantum transport based on the WF can also be particularly useful to understand why, apart from specific cases in which quantum effects are crucial, the semiclassical theory, through its Boltzmann transport equation, works much better than its limits of validity would justify, as will be shown later in this chapter.

The WF is quite suitable, in particular, for studying mesoscopic systems. Their typical dimensions are such that transport cannot be assumed to be totally coherent since dissipative scattering begins to take place. In such a condition, the Schrödinger equation for isolated electrons cannot be used. On the other hand, dimensions are so small that (1) coherent quantum effects are present, and (2) the system does not present within itself a sufficiently large number of microscopic situations to justify configuration averages: self-averaging cannot be considered a good hypothesis.

Furthermore, the WF is an appropriate tool to treat the problem of the contacts in electronic devices. In fact, being defined in a phase space, both position and momentum can be considered simultaneously, and the connection

with semiclassical systems described by a classical distribution function in the contacts is straightforward.

As we shall see, the dynamical equation that governs the evolution of the WF looks very similar to the transport Boltzmann equation, and yet is much too complicated to be solved exactly, even with numerical techniques. The Monte Carlo (MC) approach that proved to be so successful in the semiclassical case can be extended to the WF equation, and this subject will be developed in the present chapter. Nevertheless, all MC approaches to quantum transport present the serious convergence problems, mentioned at the end of last chapter, typical of the numerical evaluations of Feynman path integrals, and the MC simulation of the WF is not immune from such a drawback.

In this chapter, the definition of the WF, its main properties and its dynamical equation will be treated, together with some applications to homogeneous systems. Applications to devices and the connection of the WF with Green functions will be found in later chapters.

17.2 Definition and Main Properties

17.2.1 Weyl–Wigner Transformation

There are several possible ways to introduce the WF. Here, following [316], we start from the more general problem of finding a numerical function $F(\boldsymbol{r},\boldsymbol{p})$ of the real variables $\boldsymbol{r}$ and $\boldsymbol{p}$ that "corresponds", in a sense specified below, to an operator function $\mathcal{F}$ of the position and momentum operators r and p.

A simple requirement could be that the integral of F over the entire phase space is equal to the sum of all the eigenvalues of $\mathcal{F}$:

$$Tr\{\mathcal{F}\} = \frac{1}{\hbar^3}\int\int F(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p}.$$

The Planck constant h^3 has been inserted to assign to F the same dimensions as $\mathcal{F}$. Normalization will be discussed shortly. The above requirement, however, is not sufficient to determine unambiguously F. We may further require that the same property holds for its Fourier transform:

$$Tr\{\mathcal{F}\mathrm{e}^{\mathrm{i}(\boldsymbol{q}\mathsf{r}+\boldsymbol{s}\mathsf{p})/\hbar}\} = \frac{1}{\hbar^3}\int\int F(\boldsymbol{r},\boldsymbol{p})\mathrm{e}^{\mathrm{i}(\boldsymbol{q}\boldsymbol{r}+\boldsymbol{s}\boldsymbol{p})/\hbar}\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p} \tag{17.1}$$

for all $\boldsymbol{q}$ and $\boldsymbol{s}$. This requirement is sufficient for the definition of F since the Fourier transform in the r.h.s. of (17.1) can be inverted, and the equation solved with respect to F.

The l.h.s. of (17.1) can be evaluated in either r or p representation. In the former case, we first apply the relation

$$\mathrm{e}^{\mathrm{i}(\boldsymbol{q}\mathsf{r}+\boldsymbol{s}\mathsf{p})/\hbar} = \mathrm{e}^{\mathrm{i}\boldsymbol{s}\mathsf{p}/2\hbar}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\mathsf{r}/\hbar}\mathrm{e}^{\mathrm{i}\boldsymbol{s}\mathsf{p}/2\hbar},$$

that can be easily derived from the Campbell–Baker–Hausdorf theorem [306], and obtain

$$Tr\{\mathcal{F}\mathrm{e}^{\mathrm{i}\boldsymbol{s}\mathsf{p}/2\hbar}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\mathsf{r}/\hbar}\mathrm{e}^{\mathrm{i}\boldsymbol{s}\mathsf{p}/2\hbar}\}$$

$$= \int\int\int \langle \boldsymbol{r}|\mathcal{F}|\boldsymbol{r}'\rangle \mathrm{d}\boldsymbol{r}' \langle \boldsymbol{r}'|\mathrm{e}^{\mathrm{i}\boldsymbol{s}\mathsf{p}/2\hbar}|\boldsymbol{r}''\rangle \mathrm{d}\boldsymbol{r}'' \langle \boldsymbol{r}''|\mathrm{e}^{\mathrm{i}\boldsymbol{q}\mathsf{r}/\hbar}|\boldsymbol{r}'''\rangle \mathrm{d}\boldsymbol{r}''' \langle \boldsymbol{r}'''|\mathrm{e}^{\mathrm{i}\boldsymbol{s}\mathsf{p}/2\hbar}|\boldsymbol{r}\rangle.$$

The lateral exponentials are translation operations (cf. (7.1) in Chap. 7), so that the above expression becomes

$$\int\int\int \langle \boldsymbol{r}|\mathcal{F}|\boldsymbol{r}'\rangle \mathrm{d}\boldsymbol{r}' \langle \boldsymbol{r}'|\boldsymbol{r}'' - \boldsymbol{s}/2\rangle \mathrm{d}\boldsymbol{r}'' \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}'''/\hbar} \langle \boldsymbol{r}''|\boldsymbol{r}'''\rangle \mathrm{d}\boldsymbol{r}''' \langle \boldsymbol{r}'''|\boldsymbol{r} - \boldsymbol{s}/2\rangle.$$

The scalar products yield a number of δ functions, and the final result is

$$Tr\{\mathcal{F}\mathrm{e}^{\mathrm{i}(\boldsymbol{q}\mathsf{r}+\boldsymbol{s}\mathsf{p})/\hbar}\} = \int \langle \boldsymbol{r}|\mathcal{F}|\boldsymbol{r} - \boldsymbol{s}\rangle \mathrm{e}^{\mathrm{i}\boldsymbol{q}(\boldsymbol{r}-\boldsymbol{s}/2)/\hbar}\, \mathrm{d}\boldsymbol{r}.$$

After substitution of this result into (17.1), the Fourier transform can be inverted, yielding

$$F(\boldsymbol{r},\boldsymbol{p}) = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \langle \boldsymbol{r} + \boldsymbol{s}/2|\mathcal{F}|\boldsymbol{r} - \boldsymbol{s}/2\rangle \, \mathrm{d}\boldsymbol{s}. \tag{17.2}$$

The above is known as Weyl–Wigner transformation [471] and will be used here to introduce the Wigner function, starting from the density-matrix operator.

An equivalent definition can be obtained from (17.2) by moving to the momentum representation:

$$F(\boldsymbol{r},\boldsymbol{p}) = \int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}/\hbar} \langle \boldsymbol{p} + \boldsymbol{q}/2|\mathcal{F}|\boldsymbol{p} - \boldsymbol{q}/2\rangle \, \mathrm{d}\boldsymbol{q}. \tag{17.3}$$

Note that, when an operator function $\mathcal{F}(\mathsf{r},\mathsf{p})$ depends only upon the operator r, i.e. $\mathcal{F}(\mathsf{r},\mathsf{p}) = \mathcal{G}(\mathsf{r})$, then the Weyl–Wigner transform reduces to the same numerical function G of the numerical variable $\boldsymbol{r}$:

$$F(\boldsymbol{r},\boldsymbol{p}) = G(\boldsymbol{r}).$$

Similarly, if $\mathcal{F}(\mathsf{r},\mathsf{p})$ is a function only of p, i.e. $\mathcal{F}(\mathsf{r},\mathsf{p}) = \mathcal{G}(\mathsf{p})$,

$$F(\boldsymbol{r},\boldsymbol{p}) = G(\boldsymbol{p}).$$

17.2.2 Transformation Between the Matrix Elements of an Operator and Its Weyl–Wigner Transform

Starting from (17.2), it is easy to find in any given basis a matrix of functions that generate the transformation between the matrix elements of an operator and its Weyl–Wigner transform, and vice versa [55, 96, 316]. In fact, if an

orthonormal basis $|\phi_n\rangle$ is inserted in (17.2), we obtain

$$F(\boldsymbol{r},\boldsymbol{p}) = \sum_{nm} \int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar} \langle \boldsymbol{r}+\boldsymbol{s}/2|\phi_n\rangle\langle\phi_n|\mathcal{F}|\phi_m\rangle\langle\phi_m|\boldsymbol{r}-\boldsymbol{s}/2\rangle \,\mathrm{d}\boldsymbol{s},$$

or

$$F(\boldsymbol{r},\boldsymbol{p}) = \sum_{nm} f_{nm}(\boldsymbol{r},\boldsymbol{p}) F_{nm} \ , \qquad F_{nm} \equiv \langle\phi_n|\mathcal{F}|\phi_m\rangle, \tag{17.4}$$

where

$$f_{nm}(\boldsymbol{r},\boldsymbol{p}) \equiv \int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar} \, \langle \boldsymbol{r}+\boldsymbol{s}/2|\phi_n\rangle \, \langle\phi_m|\boldsymbol{r}-\boldsymbol{s}/2\rangle \, \mathrm{d}\boldsymbol{s}. \tag{17.5}$$

The above coefficients verify the symmetry property

$$f_{nm}(\boldsymbol{r},\boldsymbol{p}) = f^*_{mn}(\boldsymbol{r},\boldsymbol{p}),$$

as can be immediately verified. They constitute a unitary transformation, as shown by the following properties,

$$\sum_{nm} f_{nm}(\boldsymbol{r},\boldsymbol{p}) f^*_{nm}(\boldsymbol{r}',\boldsymbol{p}') = h^3 \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(\boldsymbol{p}-\boldsymbol{p}') \tag{17.6}$$

and

$$\frac{1}{h^3} \int\int f_{nm}(\boldsymbol{r},\boldsymbol{p}) f^*_{n'm'}(\boldsymbol{r},\boldsymbol{p}) \,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p} = \delta_{nn'}\delta_{mm'}, \tag{17.7}$$

directly derivable from their definitions. With the aid of relation (17.7), we can easily invert (17.4), obtaining

$$F_{nm} = \frac{1}{h^3} \int\int f^*_{nm}(\boldsymbol{r},\boldsymbol{p}) F(\boldsymbol{r},\boldsymbol{p}) \,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p}. \tag{17.8}$$

17.2.3 Definition of the Wigner Function

We saw in Chap. 16 that the basic instruments in quantum statistical physics is the density-matrix operator ρ, defined as $\rho = \overline{|\Psi\rangle\langle\Psi|}$ where $|\Psi\rangle$ is the state of the system and the overline indicates the ensemble average. If we now apply the Weyl–Wigner transformation (17.2) to the density-matrix operator, we obtain the WF (here, as in the following, the ensemble average is understood):

$$\boxed{f_w(\boldsymbol{r},\boldsymbol{p}) \equiv \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \langle \boldsymbol{r}+\boldsymbol{s}/2|\Psi\rangle\langle\Psi|\boldsymbol{r}-\boldsymbol{s}/2\rangle \,\mathrm{d}\boldsymbol{s}} \tag{17.9}$$

i.e.,

$$f_w(\boldsymbol{r},\boldsymbol{p}) = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \Psi(\boldsymbol{r}+\boldsymbol{s}/2)\Psi^*(\boldsymbol{r}-\boldsymbol{s}/2) \,\mathrm{d}\boldsymbol{s}, \tag{17.10}$$

or, using (17.3),

$$f_w(\boldsymbol{r},\boldsymbol{p}) = \int \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}/\hbar}\Phi(\boldsymbol{p}+\boldsymbol{q}/2)\Phi^*(\boldsymbol{p}-\boldsymbol{q}/2)\,\mathrm{d}\boldsymbol{q}, \tag{17.11}$$

where $\Psi(\boldsymbol{r})$ and $\Phi(\boldsymbol{p})$ are the wavefunctions of the ensemble states in $\boldsymbol{r}$ and $\boldsymbol{p}$ representation, respectively. Note that the WF is a real function.

The above definition (17.10) coincides with that given (for many particles) by Wigner in his original paper, apart from a normalization constant h^3. We shall see that the dynamical equation of the WF is homogeneous, so that its normalization must be chosen with some physical criterion. If we integrate (17.10) over the whole phase space, we obtain

$$\int\int f_w(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p} = h^3\int\langle\boldsymbol{r}|\Psi\rangle\langle\Psi|\boldsymbol{r}\rangle\,\mathrm{d}\boldsymbol{r} = h^3\int|\Psi(\boldsymbol{r})|^2\,\mathrm{d}\boldsymbol{r} = h^3. \tag{17.12}$$

To understand the physical meaning of this normalization, let us separate the phase space into cells of volume h^3 and evaluate the above normalization integral as

$$1 = \frac{1}{h^3}\int\int f_w(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p} = \frac{1}{h^3}\sum_i\int\int_{\mathrm{cell}_i} f_w(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p}, \tag{17.13}$$

and let us assume, for the moment, that f_w does not change appreciably inside a cell. Then the above equation reduces to

$$\frac{1}{h^3}\sum_i \Delta\boldsymbol{r}\Delta\boldsymbol{p}\; f_w(\boldsymbol{r}_i,\boldsymbol{p}_i) = \sum_i f_w(\boldsymbol{r}_i,\boldsymbol{p}_i) = 1, \tag{17.14}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{p}_i$ are values inside the cell i. If N is the total number of particles in the system, then $Nf_w(\boldsymbol{r}_i,\boldsymbol{p}_i)$ gives the number of particles in cell i and therefore it can be compared with unity to evaluate the incidence of the Pauli exclusion principle.

17.2.4 Main Properties

Important properties of the WF may be obtained by integration over $\boldsymbol{p}$ or $\boldsymbol{r}$. From (17.10) and (17.11), we obtain:

$$\frac{1}{h^3}\int f_w(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{p} = |\Psi(\boldsymbol{r})|^2, \qquad \frac{1}{h^3}\int f_w(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r} = |\Phi(\boldsymbol{p})|^2. \tag{17.15}$$

The above results show that by integration over momentum (real) space, we obtain the density in real (momentum) space, as for the classical distribution function (remember that an ensemble average is understood).

Another important property of the WF refers to the scalar product of two pure states. If f_{w1} and f_{w2} are the WF corresponding to the pure quantum

states $|\Psi_1\rangle$ and $|\Psi_2\rangle$, then this property reads

$$\frac{1}{h^3}\int\int f_{w1}(\boldsymbol{r},\boldsymbol{p})f_{w2}(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p} = |\langle\Psi_1|\Psi_2\rangle|^2. \tag{17.16}$$

Two other important properties of the WF should be mentioned at this point, that are not represented by equations. One is the fact that it can be different from zero in regions of space where the wavefunction vanishes and the particle cannot be found. The other property, already mentioned in the introduction, is that the WF may take negative values, so that it cannot be interpreted as a probability density. With regard to this point, a local Gaussian average of the WF has been defined by Husimi [189,195] that assumes always positive values. However, the Husimi function does not posses the properties of the WF that make it so similar to the classical distribution function, in particular the one discussed in the next section.

Average Physical Quantities

The property of the WF which makes its analogy with the classical distribution function particularly strong, is its use for the calculation of average physical quantities. If we work in the $\boldsymbol{r}$ representation, the average value of a general physical quantity represented by the hermitian operator $\mathcal{A}$ is given by

$$\begin{aligned}\langle\mathcal{A}\rangle = \langle\Psi|\mathcal{A}|\Psi\rangle &= \int\int\langle\Psi|\boldsymbol{r}_1\rangle\mathrm{d}\boldsymbol{r}_1\langle\boldsymbol{r}_1|\mathcal{A}|\boldsymbol{r}_2\rangle\mathrm{d}\boldsymbol{r}_2\langle\boldsymbol{r}_2|\Psi\rangle \\ &= \int\int\Psi^*(\boldsymbol{r}_1)A(\boldsymbol{r}_1,\boldsymbol{r}_2)\Psi(\boldsymbol{r}_2)\,\mathrm{d}\boldsymbol{r}_1\,\mathrm{d}\boldsymbol{r}_2,\end{aligned}$$

where, again, the ensemble average is understood. Now let us substitute $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ in the integral with $\boldsymbol{r} = (\boldsymbol{r}_1+\boldsymbol{r}_2)/2$ and $\boldsymbol{r}' = (\boldsymbol{r}_2-\boldsymbol{r}_1)$, obtaining:

$$\langle\mathcal{A}\rangle = \int\int\Psi^*(\boldsymbol{r}-\boldsymbol{r}'/2)A(\boldsymbol{r}-\boldsymbol{r}'/2,\boldsymbol{r}+\boldsymbol{r}'/2)\Psi(\boldsymbol{r}+\boldsymbol{r}'/2)\,\mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{r}.$$

The value of the integral does not change if we multiply the integrand by the delta function $\delta(\boldsymbol{s}-\boldsymbol{r}')$ and integrate over $\boldsymbol{s}$, and if we use the plane-wave representation for the delta function, we obtain

$$\boxed{\langle\mathcal{A}\rangle = \tfrac{1}{h^3}\int\int f_w(\boldsymbol{r},\boldsymbol{p})\;A_w(\boldsymbol{r},\boldsymbol{p})\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p}} \tag{17.17}$$

where

$$A_w(\boldsymbol{r},\boldsymbol{p}) = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar}A(\boldsymbol{r}+\boldsymbol{s}/2,\boldsymbol{r}-\boldsymbol{s}/2)\,\mathrm{d}\boldsymbol{s}$$

is the Weyl–Wigner transform of the operator $\mathcal{A}$ that represents the quantity under consideration. Equation (17.17) is completely analogous to the classical expression of an average quantity in an ensemble.

17.3 Coherent Evolution of the Wigner Function

If we consider a particle of mass m subject to the Hamiltonian

$$\mathcal{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r}) \tag{17.18}$$

and assume that its eigenvalues ϵ_n and eigenstates $|\phi_n\rangle$ are known, then it is easy to find the coherent evolution of the WF. To this purpose, we first obtain the density matrix in the energy representation $\{|\phi_n\rangle\}$ at the initial time $t_\circ$ from the WF $f_w(\boldsymbol{r}, \boldsymbol{p}, t_\circ)$, supposed to be known, using (17.8):

$$\rho_{nm}(t_\circ) = \frac{1}{h^3}\int\int f^*_{nm}(\boldsymbol{r}, \boldsymbol{p}) f_w(\boldsymbol{r}, \boldsymbol{p}, t_\circ)\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p}.$$

Then we evolve the density matrix

$$\rho_{nm}(t) = \mathrm{e}^{-\mathrm{i}(\omega_n-\omega_m)(t-t_\circ)}\frac{1}{h^3}\int\int f^*_{nm}(\boldsymbol{r}, \boldsymbol{p}) f_w(\boldsymbol{r}, \boldsymbol{p}, t_\circ)\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{p}$$

and move back from the density matrix to the WF at time t using (17.4):

$$f_w(\boldsymbol{r}, \boldsymbol{p}, t) = \sum_{nm} f_{nm}(\boldsymbol{r}, \boldsymbol{p})\mathrm{e}^{-\mathrm{i}(\omega_n-\omega_m)(t-t_\circ)}\frac{1}{h^3}\iint f^*_{nm}(\boldsymbol{r}', \boldsymbol{p}') f_w(\boldsymbol{r}', \boldsymbol{p}', t_\circ)\mathrm{d}\boldsymbol{r}'\mathrm{d}\boldsymbol{p}'. \tag{17.19}$$

This equation gives the free coherent evolution of the WF, starting from its value at the initial time $t_\circ$, when the dynamics is described by the Hamiltonian (17.18).

Free-Electron Evolution

In the particular case of free electrons, ($V = 0$ in (17.18)), the eigenfunctions of the Hamiltonian are simply plane waves: $\phi_{\boldsymbol{k}}(\boldsymbol{r}) = (2\pi)^{-3/2}\exp(\mathrm{i}\boldsymbol{k}\boldsymbol{r})$, and the coefficients (17.5) become

$$f_{\boldsymbol{k}\boldsymbol{k}'}(\boldsymbol{r}, \boldsymbol{p}) = \frac{1}{(2\pi)^3}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar}\mathrm{e}^{\mathrm{i}\boldsymbol{k}(\boldsymbol{r}+\boldsymbol{s}/2)}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}'(\boldsymbol{r}-\boldsymbol{s}/2)}\,\mathrm{d}\boldsymbol{s},$$

or

$$f_{\boldsymbol{k}\boldsymbol{k}'}(\boldsymbol{r}, \boldsymbol{p}) = \mathrm{e}^{\mathrm{i}(\boldsymbol{k}-\boldsymbol{k}')\boldsymbol{r}}\delta(\boldsymbol{p}/\hbar - (\boldsymbol{k}+\boldsymbol{k}')/2).$$

If we apply these coefficients to the coherent evolution of the WF, as given by (17.19), a straightforward calculation yields

$$\boxed{f_w(\boldsymbol{r}, \boldsymbol{p}, t) = f_w(\boldsymbol{r} - (\boldsymbol{p}/m)(t-t_\circ)), \boldsymbol{p}, t_\circ)} \tag{17.20}$$

This result is of great physical relevance: *for free electrons, the WF evolves in time in exactly the same way as the distribution function of an ensemble of*

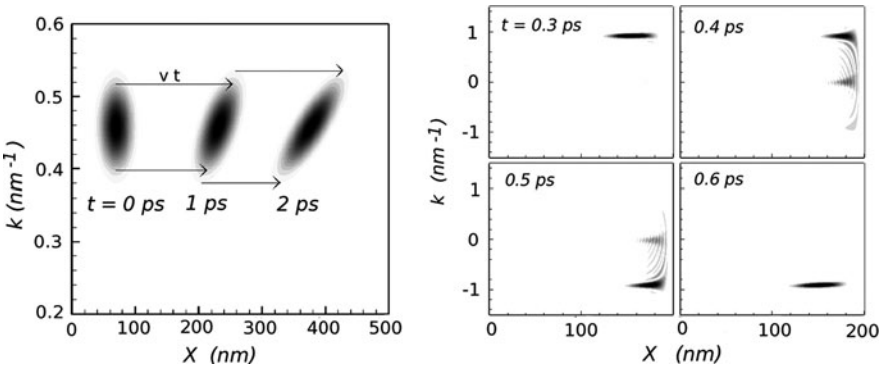


Fig. 17.1. *Left*: Free evolution in Wigner phase space of a WF for a Gaussian wave packet. Each point follows a classical trajectory. *Right*: Evolution of the WF of a classical wave packet hitting an infinite potential barrier. During the reflection, oscillatory components of the WF are present at intermediate momenta due to the correlation between incoming and outgoing components of the wavefunction [51]

classical particles. We shall see below that this result holds also in presence of linear and parabolic potentials.

Figure 17.1 shows the above evolution for the case of a wave packet [37]. Such an approach provides a very simple physical interpretation of the broadening of free wave packets: contributions of higher momenta move faster then contributions with lower momenta. It has been shown that also in the case of a particle hitting an infinite potential barrier, for time long enough after the scattering process, the evolution of the WF coincides with that of a classical distribution function [51], as shown in the right part of the same figure.

Scattering States

If a potential $V(\boldsymbol{r})$ is present in the Hamiltonian (17.18), the coherent evolution of the WF can be analyzed in exactly the same way as for free electrons, in the representation of the scattering states, defined in Appendix B.

17.4 Dynamical Equations of the Wigner Function

To derive a dynamical equation for the WF we start by differentiating its definition (17.10):

$$\frac{\partial}{\partial t} f_w(\boldsymbol{r}, \boldsymbol{p}, t) = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \frac{\partial}{\partial t} \left[\Psi(\boldsymbol{r} + \boldsymbol{s}/2, t)\Psi^*(\boldsymbol{r} - \boldsymbol{s}/2, t)\right] \mathrm{d}\boldsymbol{s}.$$

Then we apply the Schrödinger equation:

$$\frac{\partial}{\partial t} f_w = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \frac{1}{\mathrm{i}\hbar} \left[(\mathcal{H}\Psi(+))\Psi^*(-) - \Psi(+)(\mathcal{H}\Psi(-))^*\right] \mathrm{d}\boldsymbol{s}, \qquad (17.21)$$

where $\Psi(\pm) = \Psi(\boldsymbol{r} \pm \boldsymbol{s}/2, t)$. It is clear that the above expression is linear with respect to the Hamiltonian. Therefore, we may deal separately with different terms in $\mathcal{H}$.

Free Electrons

Let us consider first the kinetic term in (17.18):

$$\left.\frac{\partial}{\partial t} f_w\right|_{\circ} = \frac{\mathrm{i}\hbar}{2m} \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \left[(\nabla^2 \Psi(+)) \Psi^*(-) - \Psi_+ \nabla^2 \Psi_-^* \right] \mathrm{d}\boldsymbol{s}$$

$$= \frac{\mathrm{i}\hbar}{2m} \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} 4 \left[(\nabla_s^2 \Psi(+)) \Psi^*(-) - \Psi_+ \nabla_s^2 \Psi_-^* \right] \mathrm{d}\boldsymbol{s} \ . \quad (17.22)$$

Using the identity $\phi \nabla^2 \psi = \nabla \cdot (\phi \nabla \psi) - \nabla \phi \cdot \nabla \psi$ and integrating by parts, we obtain the dynamical equation of the WF for free electrons as

$$\frac{\partial}{\partial t} f_w + \frac{\boldsymbol{p}}{m} \nabla f_w = 0. \quad (17.23)$$

The above equation is identical to the Liouville (Boltzmann) equation for classical free particles. This result is consistent with the propagation of the WF as given in (17.20), which is solution of the above (17.23).

In the derivation of (17.23), it is assumed that the wavefunctions and their derivatives vanish at the integration limits. If, on the other hand, particles are confined by infinite potential barriers, such that the wavefunctions vanish but their derivatives do not, an extra term must be added in the equation [216].

Electrons Subject to a Potential $V(\boldsymbol{r})$

If a potential term $V(\boldsymbol{r})$ is present in the Hamiltonian, it can be elaborated in different ways in the analysis of the coherent dynamics of the WF. The corresponding term in (17.21) is

$$\left.\frac{\partial}{\partial t} f_w\right|_{V} = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \frac{1}{\mathrm{i}\hbar} \left[V(+) - V(-) \right] \Psi(+) \Psi^*(-) \, \mathrm{d}\boldsymbol{s}. \quad (17.24)$$

In the first type of elaboration, the Fourier integral of the potential is considered:

$$V(\boldsymbol{r}) = \int \tilde{V}(\boldsymbol{k}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}\boldsymbol{r}} \, \mathrm{d}\boldsymbol{k} = \int |\tilde{V}(\boldsymbol{k})| \mathrm{e}^{\mathrm{i}[\boldsymbol{k}\boldsymbol{r} + \phi(\boldsymbol{k})]} \mathrm{d}\boldsymbol{k},$$

where in the Fourier transform of $V(\boldsymbol{r})$ the phase has been explicitly indicated: $\tilde{V}(\boldsymbol{k}) = |\tilde{V}(\boldsymbol{k})| \mathrm{e}^{\mathrm{i}\phi(\boldsymbol{k})}$. If this Fourier integral is substituted into (17.24) and the

reality condition of $V(\boldsymbol{r})$ is taken into account, we obtain

$$\left.\frac{\partial}{\partial t} f_w(\boldsymbol{r},\boldsymbol{p},t)\right|_V = \frac{2}{\hbar}\int \mathrm{d}\boldsymbol{k}\int \mathrm{d}\boldsymbol{s}|\tilde{V}(\boldsymbol{k})|\sin[\boldsymbol{k}\boldsymbol{r}+\phi(\boldsymbol{k})] f_w(\boldsymbol{r},\boldsymbol{p}-\hbar\boldsymbol{k}/2,t).$$

This is the first expression we can obtain for the term of the dynamical equation for the WF due to a potential term in the Hamiltonian.

In the second, more common, type of elaboration, we insert a δ function in (17.24), and use its plane-wave representation:

$$\begin{aligned}\left.\frac{\partial}{\partial t} f_w\right|_V &= \int\int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar}\frac{1}{\mathrm{i}\hbar}\left[V(\boldsymbol{r}+\boldsymbol{s}/2)-V(\boldsymbol{r}-\boldsymbol{s}/2)\right]\\ &\quad\times\Psi(\boldsymbol{r}+\boldsymbol{s}'/2,t)\Psi^*(\boldsymbol{r}-\boldsymbol{s}'/2,t)\delta(\boldsymbol{s}-\boldsymbol{s}')\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}\boldsymbol{s}'\\ &= \int\int \frac{1}{h^3}\int \mathrm{e}^{\mathrm{i}(\boldsymbol{s}-\boldsymbol{s}')\boldsymbol{p}'/\hbar}\,\mathrm{d}\boldsymbol{p}'\mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar}\frac{1}{\mathrm{i}\hbar}\left[V(\boldsymbol{r}+\boldsymbol{s}/2)-V(\boldsymbol{r}-\boldsymbol{s}/2)\right]\\ &\quad\times\Psi(\boldsymbol{r}+\boldsymbol{s}'/2,t)\Psi^*(\boldsymbol{r}-\boldsymbol{s}'/2,t)\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}\boldsymbol{s}'\\ &= \frac{1}{h^3}\int\int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}'\boldsymbol{p}'/\hbar}\mathcal{V}_w(\boldsymbol{r},\boldsymbol{p}-\boldsymbol{p}')\Psi(\boldsymbol{r}+\boldsymbol{s}'/2,t)\Psi^*(\boldsymbol{r}-\boldsymbol{s}'/2,t)\,\mathrm{d}\boldsymbol{s}'\,\mathrm{d}\boldsymbol{p}',\end{aligned}$$

where we have put

$$\mathcal{V}_w(\boldsymbol{r},\Delta\boldsymbol{p}) = \frac{1}{\mathrm{i}\hbar}\int \mathrm{e}^{-\mathrm{i}\Delta\boldsymbol{p}\boldsymbol{s}/\hbar}\left[V(\boldsymbol{r}+\boldsymbol{s}/2)-V(\boldsymbol{r}-\boldsymbol{s}/2)\right]\mathrm{d}\boldsymbol{s}. \tag{17.25}$$

In conclusion:

$$\left.\frac{\partial}{\partial t} f_w(\boldsymbol{r},\boldsymbol{p},t)\right|_V = \frac{1}{h^3}\int \mathcal{V}_w(\boldsymbol{r},\boldsymbol{p}-\boldsymbol{p}')f_w(\boldsymbol{r},\boldsymbol{p}',t)\,\mathrm{d}\boldsymbol{p}'. \tag{17.26}$$

Note that this scattering term is not local in $V(\boldsymbol{r})$, since $\mathcal{V}_w(\boldsymbol{r},\boldsymbol{p}-\boldsymbol{p}')$ depends on the values of V in points different from $\boldsymbol{r}$. Thus, in absence of dephasing processes the effect of $V(\boldsymbol{r})$ extends to infinity. When instead phase-breaking collisions are present, the nonlocality of $\mathcal{V}_w$ is expected to be relevant only up to regions where the electron correlation is different from zero.

Collecting the above results (17.23) and (17.26), we obtain the coherent quantum transport equation for electrons subject to a potential $V(\boldsymbol{r})$:

$$\boxed{\frac{\partial}{\partial t} f_w + \frac{\boldsymbol{p}}{m}\nabla f_w = \frac{1}{h^3}\int \mathcal{V}_w(\boldsymbol{r},\boldsymbol{p}-\boldsymbol{p}')f_w(\boldsymbol{r},\boldsymbol{p}',t)\,\mathrm{d}\boldsymbol{p}'} \tag{17.27}$$

We observe here the similarity of this quantum transport equation with the classical Boltzmann transport equation. $\mathcal{V}_w$ is a real function; furthermore, it is antisymmetric with respect to $\Delta\boldsymbol{p}$ so that the integral term works at the same time as scattering "in" and scattering "out". In fact, the value of the WF in a point $(\boldsymbol{r},\boldsymbol{p})$, increases (decreases) the value of $f_w(\boldsymbol{r},\boldsymbol{p}+\Delta\boldsymbol{p})$ and decreases (increases) of the same amount the value of $f_w(\boldsymbol{r},\boldsymbol{r}-\Delta\boldsymbol{p})$.

17.4.1 Moyal Expansion

We have noted that (17.27) resembles the Boltzmann equation with a scattering integral. However, when the particles of the ensemble are subject to a potential energy field $V(\mathbf{r})$, its effect in classical transport is described by an acceleration term in the l.h.s. of the Boltzmann equation. The connection between (17.27) and the classical equivalent can be seen very explicitly by expanding in powers of $\boldsymbol{s}$ the potential $V(\pm)$ in (17.24) [316]:

$$\left.\frac{\partial}{\partial t} f_w\right|_V = \int \mathrm{e}^{-\mathrm{i}\boldsymbol{s}\boldsymbol{p}/\hbar} \times \\ \times \frac{1}{\mathrm{i}\hbar}\left[\sum_{\lambda=0}^{\infty} \frac{1}{\lambda!} \frac{\partial^\lambda V(\boldsymbol{r})}{\partial \boldsymbol{r}^\lambda} \left(\frac{\boldsymbol{s}}{2}\right)^\lambda - \sum_{\lambda=0}^{\infty} \frac{1}{\lambda!} \frac{\partial^\lambda V(\boldsymbol{r})}{\partial \boldsymbol{r}^\lambda} \left(-\frac{\boldsymbol{s}}{2}\right)^\lambda\right] \Psi(+)\Psi^*(-)\,\mathrm{d}\boldsymbol{s}.$$

For even λ, the two terms in the square brackets are opposite and cancel, while for odd λ they are equal. Furthermore, the factors $\boldsymbol{s}$ can be obtained by differentiating the WF with respect to $\boldsymbol{p}$. The result is

$$\left.\frac{\partial}{\partial t} f_w\right|_V = \frac{2}{\mathrm{i}\hbar} \sum_{\lambda=\mathrm{odd}}^{\infty} \frac{1}{\lambda!} \frac{\partial^\lambda V(\boldsymbol{r})}{\partial \boldsymbol{r}^\lambda} \left(\frac{\hbar}{-2\mathrm{i}}\right)^\lambda \frac{\partial^\lambda}{\partial \boldsymbol{p}^\lambda} f_w.$$

By inserting this expansion into (17.27) and taking into account that the first derivative of $V(\boldsymbol{r})$ yields the force $\boldsymbol{F} = -\nabla V$, which can be taken to the l.h.s., we obtain

$$\frac{\partial}{\partial t} f_w + \frac{\boldsymbol{p}}{m} \nabla f_w + \boldsymbol{F} \nabla_p f_w = \sum_{\lambda=3,5,\ldots} \frac{1}{\lambda!} \frac{\partial^\lambda V(\boldsymbol{r})}{\partial \boldsymbol{r}^\lambda} \left(\frac{\mathrm{i}\hbar}{2}\right)^{\lambda-1} \frac{\partial^\lambda}{\partial \boldsymbol{p}^\lambda} f_w. \quad (17.28)$$

It is very easy to recognize here the l.h.s. as the classical Liouvillian, i.e., the l.h.s. of the BE, while the r.h.s. provides the quantum corrections. This is expressed as a series of powers of $\hbar$ multiplied by the successive derivatives of the potential energy. The first correction, of the order of $\hbar^2$ involves the third derivative of V. Thus, as anticipated, in the cases of a constant force and of a harmonic oscillator the dynamical evolution of the WF is the same as that of a distribution function in classical statistical mechanics. The essential difference between classical and quantum theory is, here, that in the former case, for a single particle, its representative point describes one precise orbit in phase space, while in the quantum case, even for a single particle, the WF is described by a distribution of points following classical trajectories. Such a distribution is determined by the initial condition and is compatible with the uncertainty relation. For the harmonic oscillator this situation is illustrated in Fig. 17.2. Figure 17.3 shows the WF for the 8-th eigenstate of the harmonic oscillator.

It must be noted, at this point, that to use the expansion in (17.28), the form of $V(\boldsymbol{r})$ has to be analytical in the whole region occupied by the WF.

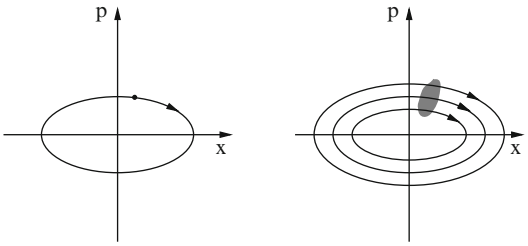


Fig. 17.2. Trajectory in phase space of a classical particle subject to a harmonic potential (*left*) compared to the trajectories of the points of the WF of a quantum particle subject to the same potential (*right*)

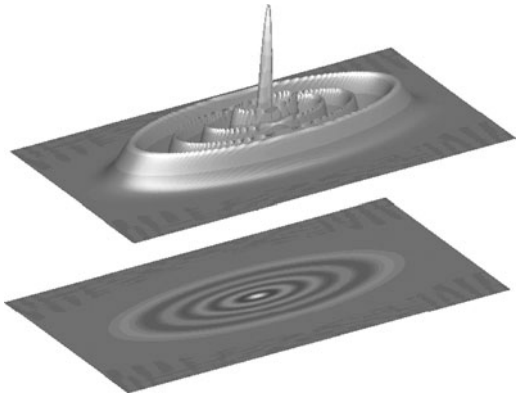


Fig. 17.3. Wigner function of the 8-th energy eigenstate of a harmonic oscillator. Courtesy of Paolo Bordone

Thus, potential discontinuities, such as those present when potential barriers are considered, cannot be treated with this equation. On the other hand, the same equation also shows that the classical approximation works better with smooth potential profiles, as expected.

17.5 Electron–Phonon Interaction

If our electron is interacting with the phonon gas, the state of the system is described by the electron state and the state of the crystal vibrations. A proper basis set in this case can be

$$|\boldsymbol{r}, n_{\boldsymbol{q}}\rangle,$$

where $\{n_{\boldsymbol{q}}\}$ is the set of occupation numbers of the phonon modes $\boldsymbol{q}$. Our state can then be described by the wavefunction

$$\Psi(\boldsymbol{r}, \{n_{\boldsymbol{q}}\}, t) = \langle \boldsymbol{r}, \{n_{\boldsymbol{q}}\} | \Psi(t) \rangle. \tag{17.29}$$

The matrix elements of the total density matrix are given by

$$\rho_{\rm tot}(\boldsymbol{r},\{n_{\boldsymbol{q}}\};\boldsymbol{r}'\{n'_{\boldsymbol{q}}\}) = \overline{\Psi(\boldsymbol{r},\{n_{\boldsymbol{q}}\})\Psi^*(\boldsymbol{r}',\{n'_{\boldsymbol{q}}\})}. \tag{17.30}$$

The electron WF will be generated by the reduced electron density matrix:

$$\rho(\boldsymbol{r},\boldsymbol{r}') = Tr_{ph}\left[\rho_{\rm tot}(\boldsymbol{r},\{n_{\boldsymbol{q}}\};\boldsymbol{r}'\{n'_{\boldsymbol{q}}\})\right] = \sum_{\{n_{\boldsymbol{q}}\}} \rho_{\rm tot}(\boldsymbol{r},\{n_{\boldsymbol{q}}\};\boldsymbol{r}'\{n_{\boldsymbol{q}}\}). \tag{17.31}$$

The Hamiltonian of the system is now

$$\mathcal{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r}) + \mathcal{H}_p + \mathcal{H}_{ep}, \tag{17.32}$$

where $\mathcal{H}_p$ and $\mathcal{H}_{ep}$ are the free-phonon and the e–ph interaction Hamiltonians, respectively:

$$\mathcal{H}_p = \sum_{\boldsymbol{q}} \left(\mathsf{a}^\dagger_{\boldsymbol{q}}\mathsf{a}_{\boldsymbol{q}} + \frac{1}{2}\right)\hbar\omega_q, \quad \mathcal{H}_{ep} = \sum_{\boldsymbol{q}} \mathrm{i}\hbar F(\boldsymbol{q})\left(\mathsf{a}_{\boldsymbol{q}}\mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} - \mathsf{a}^\dagger_{\boldsymbol{q}}\mathrm{e}^{-\mathrm{i}\boldsymbol{q}\boldsymbol{r}}\right). \tag{17.33}$$

Here, $\mathsf{a}_{\boldsymbol{q}}$ and $\mathsf{a}^\dagger_{\boldsymbol{q}}$ are the annihilation and creation operators for the phonon mode $\boldsymbol{q}$ with frequency ω_q, and $F(\boldsymbol{q})$ is a real function that depends on the e–ph interaction mechanism.

As well known, the trace operation in (17.31) does not commute with the Hamiltonian so that it is not possible to write a rigorous closed dynamical equation for the reduced density matrix. As a consequence, it is not possible to write a closed dynamical equation for the WF of electrons interacting with phonons. Here we shall use a different approach. We first define a generalized WF $f_w(\mathbf{r},\mathbf{p},\{n_{\boldsymbol{q}}\}\{n'_{\boldsymbol{q}}\})$ starting from the total density matrix in (17.30). Then, owing to the linearity of the WF with respect to the density matrix we recover the reduced electron WF as

$$f_w(\boldsymbol{r},\boldsymbol{p}) = Tr_{ph}\left[f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\}\{n'_{\boldsymbol{q}}\})\right] = \sum_{\{n_{\boldsymbol{q}}\}} f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\}\{n_{\boldsymbol{q}}\}). \tag{17.34}$$

The generalized $f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\}\{n'_{\boldsymbol{q}}\})$ contains a huge number of variables, but we shall see that, through sampling, an MC approach can handle this situation, and the trace over the phonon variables can be performed on the solution rather then on the equation.

To realize this project, let us first define the function

$$g(\boldsymbol{r},\{n_{\boldsymbol{q}}\},t) = \mathrm{e}^{\mathrm{i}\omega(\{n_{\boldsymbol{q}}\})(t-t_\circ)}\Psi(\boldsymbol{r},\{n_{\boldsymbol{q}}\},t), \tag{17.35}$$

where $t_\circ$ is the time of the initial condition, and

$$\hbar\omega(\{n_{\boldsymbol{q}}\}) = \sum_{\boldsymbol{q}} n_{\boldsymbol{q}}\hbar\omega_q$$

is the total energy of the free phonons in state $\{n_{\boldsymbol{q}}\}$. As we shall see, the exponential factor in (17.35) cancel the time dependence due to the free phonon dynamics. Its introduction in the definition of the generalized WF eliminates from the dynamical equation the free phonon term.

The generalized WF is now defined as

$$f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\},\{n'_{\boldsymbol{q}}\},t) \equiv \int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar} g\left(\boldsymbol{r}+\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\},t\right) g^*\left(\boldsymbol{r}-\boldsymbol{s}/2,\{n'_{\boldsymbol{q}}\},t\right) \mathrm{d}\boldsymbol{s}. \tag{17.36}$$

The dynamical equation is obtained in the usual way, first differentiating with respect to time and then applying the Schrödinger equation:

$$\begin{aligned}\frac{\partial}{\partial t} f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\},\{n'_{\boldsymbol{q}}\},t) &= i[\omega(\{n_{\boldsymbol{q}}\})-\omega(\{n'_{\boldsymbol{q}}\})] f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\},\{n'_{\boldsymbol{q}}\},t) \\ &+ \int \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar} \left[\left(\mathrm{e}^{\mathrm{i}\omega(\{n_{\boldsymbol{q}}\}(t-t_\circ)} \frac{\mathcal{H}}{\mathrm{i}\hbar} \Psi(+,\{n_{\boldsymbol{q}}\},t) \right) g^*\left(-,\{n'_{\boldsymbol{q}}\}\right) \right. \\ &\left. + g(+,\{n_{\boldsymbol{q}}\}) \left(\mathrm{e}^{-\mathrm{i}\omega(\{n'_{\boldsymbol{q}}\}(t-t_\circ)} \frac{1}{-\mathrm{i}\hbar} \left(\mathcal{H}\Psi(-,\{n'_{\boldsymbol{q}}\},t)\right)^* \right) \right] \mathrm{d}\boldsymbol{s}, \end{aligned} \tag{17.37}$$

where a short notation has been used, as in (17.21). The second term above is again linear in the Hamiltonian and each term in (17.32) can be treated separately.

The free-phonon Hamiltonian yields:

$$\begin{aligned}\left.\frac{\partial}{\partial t} f_w\right|_p &= \int \mathrm{d}\boldsymbol{s}\, \mathrm{e}^{-\mathrm{i}\mathbf{p}\mathbf{s}/\hbar} \\ &\times \left(\frac{\hbar\omega(\{n_{\boldsymbol{q}}\})}{\mathrm{i}\hbar} g(+,\{n_{\boldsymbol{q}}\}) + g(+,\{n_{\boldsymbol{q}}\}) \frac{\hbar\omega(\{n'_{\boldsymbol{q}}\})}{-\mathrm{i}\hbar} \right) g^*\left(-,\{n'_{\boldsymbol{q}}\}\right). \end{aligned} \tag{17.38}$$

This expression exactly cancels the first term in (17.37): as anticipated, the exponential factor in the definition (17.35) of g eliminates the free phonon dynamics.

As it regards the kinetic and potential terms of the dynamical equation, as expressed by (17.23) and (17.26), they are not changed by the new definition, since they do not involve phonon variables.

The term due to e–ph interaction is far more complicated than all the other ones and requires some attention. In (17.37), this term is

$$\begin{aligned}\left.\frac{\partial}{\partial t} f_w\right|_{ep} &= \frac{1}{\mathrm{i}\hbar} \int \mathrm{d}\boldsymbol{s}\, \mathrm{e}^{-\mathrm{i}\mathbf{p}\mathbf{s}/\hbar} \left[\left(\mathrm{e}^{\mathrm{i}\omega(\{n_{\boldsymbol{q}}\}(t-t_\circ)} \mathcal{H}_{ep} \Psi(+,\{n_{\boldsymbol{q}}\},t) \right) g^*\left(-,\{n'_{\boldsymbol{q}}\}\right) \right. \\ &\left. - g(+,\{n_{\boldsymbol{q}}\}) \mathrm{e}^{-\mathrm{i}\omega(\{n'_{\boldsymbol{q}}\}(t-t_\circ)} (\mathcal{H}_{ep}\Psi(-,\{n'_{\boldsymbol{q}}\},t))^* \right]. \end{aligned} \tag{17.39}$$

Now, remembering the expression of $\mathcal{H}_{ep}$ in (17.33), we have

$$\mathcal{H}_{ep}\Psi(\pm,\{n_{\boldsymbol{q}}\},t) = \sum_{\boldsymbol{q}'} \mathrm{i}\hbar F(\boldsymbol{q}') \left(\mathsf{a}_{\boldsymbol{q}'}\mathrm{e}^{\mathrm{i}\boldsymbol{q}'(\boldsymbol{r}\pm\boldsymbol{s}/2)} - \mathsf{a}^{\dagger}_{\boldsymbol{q}'}\mathrm{e}^{-\mathrm{i}\boldsymbol{q}'(\boldsymbol{r}\pm\boldsymbol{s}/2)}\right)\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\},t). \tag{17.40}$$

Here

$$\begin{aligned}\mathsf{a}_{\boldsymbol{q}'}\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\},t) &= \langle\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\}|\mathsf{a}_{\boldsymbol{q}'}|\Psi(t)\rangle \\ &= \langle\Psi(t)|\mathsf{a}^{\dagger}_{\boldsymbol{q}'}|\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\}\rangle^{*} = \langle\Psi(t)|\sqrt{n_{\boldsymbol{q}'}+1}|\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_1,\ldots n_{\boldsymbol{q}'}+1,\ldots\}\rangle^{*} \\ &= \sqrt{n_{\boldsymbol{q}'}+1}\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_1,\ldots n_{\boldsymbol{q}'}+1,\ldots\},t),\end{aligned}$$

and similarly

$$\mathsf{a}^{\dagger}_{\boldsymbol{q}'}\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\},t) = \sqrt{n_{\boldsymbol{q}'}}\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_1,\ldots n_{\boldsymbol{q}'}-1,\ldots\},t).$$

Substituting these expressions into (17.40), we obtain

$$\begin{aligned}\mathcal{H}_{ep}\Psi(\pm,\{n_{\boldsymbol{q}}\},t) &= \sum_{\boldsymbol{q}'} \mathrm{i}\hbar F(\boldsymbol{q}') \Big[\mathrm{e}^{\mathrm{i}\boldsymbol{q}'(\boldsymbol{r}\pm\boldsymbol{s}/2)}\sqrt{n_{\boldsymbol{q}'}+1}\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_1,\ldots n_{\boldsymbol{q}'}+1,\ldots\},t) \\ &\quad - \mathrm{e}^{-\mathrm{i}\boldsymbol{q}'(\boldsymbol{r}\pm\boldsymbol{s}/2)}\sqrt{n_{\boldsymbol{q}'}}\Psi(\boldsymbol{r}\pm\boldsymbol{s}/2,\{n_1,\ldots n_{\boldsymbol{q}'}-1,\ldots\},t)\Big].\end{aligned}$$

Thus, substituting into (17.39), we obtain, after straightforward calculations

$$\begin{aligned}\frac{\partial}{\partial t}f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\}\{n'_{\boldsymbol{q}}\},t)\bigg|_{ep} &= \sum_{\boldsymbol{q}'} F(\boldsymbol{q}') \\ &\times \Big\{\mathrm{e}^{\mathrm{i}(\boldsymbol{q}'\boldsymbol{r}-\omega_{\boldsymbol{q}'}(t-t_\circ))}\sqrt{n_{\boldsymbol{q}'}+1}f_w(\boldsymbol{r},\boldsymbol{p}-\hbar\boldsymbol{q}'/2,\{n_1,\ldots n_{\boldsymbol{q}'}+1,\ldots\},\{n'_{\boldsymbol{q}}\},t) \\ &- \mathrm{e}^{-\mathrm{i}(\boldsymbol{q}'\boldsymbol{r}-\omega_{\boldsymbol{q}'}(t-t_\circ))}\sqrt{n_{\boldsymbol{q}'}}f_w(\boldsymbol{r},\boldsymbol{p}+\hbar\boldsymbol{q}'/2,\{n'_1,\ldots n_{\boldsymbol{q}'}-1,\ldots\},\{n'_{\boldsymbol{q}}\},t) \\ &+ \mathrm{e}^{-\mathrm{i}(\boldsymbol{q}'\boldsymbol{r}-\omega_{\boldsymbol{q}'}(t-t_\circ))}\sqrt{n'_{\boldsymbol{q}'}+1}f_w(\boldsymbol{r},\boldsymbol{p}-\hbar\boldsymbol{q}'/2,\{n_{\boldsymbol{q}}\},\{n'_1,\ldots n'_{\boldsymbol{q}'}+1,\ldots\},t) \\ &- \mathrm{e}^{\mathrm{i}(\boldsymbol{q}'\boldsymbol{r}-\omega_{\boldsymbol{q}'}(t-t_\circ))}\sqrt{n'_{\boldsymbol{q}'}}f_w(\boldsymbol{r},\boldsymbol{p}+\hbar\boldsymbol{q}'/2,\{n_{\boldsymbol{q}}\},\{n'_1,\ldots n'_{\boldsymbol{q}'}+1,\ldots\},t)\Big\}.\end{aligned} \tag{17.41}$$

This is the term in the dynamical equation for the generalized WF due to the e–ph interaction. It mixes values corresponding to different phonon contents, preventing in this way the formulation of a closed equation for the reduced WF. Furthermore, the e–ph dynamics connects terms of the WF with only one of the phonon set of occupation numbers different by one unity. Thus, even if we are interested in the diagonal values of the WF, as indicated in (17.34), during e–ph interaction nondiagonal terms are involved. A complete, rigorous, microscopic (reversible) dynamical solution of the equation, which

involves a huge number of variables, is not possible in practice, and not even sought for. We shall see, however, that this form of the dynamical equation for the WF is suitable, at least in principle, for a numerical solution through an MC sampling of the possible phonon modes involved in the interaction with the electron.

The four terms appearing in the r.h.s. of (17.41) have a simple physical interpretation: e–ph interaction occurs as emission or absorption of a quantum of any mode $\boldsymbol{q}$ and this may appear in the state on the left or on the right of the bilinear expression that defines the density matrix and the WF in (17.10). Each elementary interaction or *vertex* changes only one of the two sets of variables of the WF; more precisely, one of the occupation numbers $n_{\boldsymbol{q}}$ is changed by one unity, as said above, and the electron-momentum variable of the WF is changed by half of the phonon momentum. For the completion of the electron transition a second vertex must occur where the momentum transfer is completed (real transitions) or the transferred half momentum is returned (virtual transitions). In the semiclassical limit, real and virtual transitions yield the scattering *in* and the scattering *out*, respectively.

The general equation for the WF of one electron interacting with phonons is then

$$\frac{\partial}{\partial t} f_w(\boldsymbol{r}, \boldsymbol{p}, \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t) + \frac{\boldsymbol{p}}{m} \nabla f_w + \boldsymbol{F} \nabla_{\boldsymbol{p}} f_w = \\ = \frac{1}{h^3} \int \mathrm{d}\boldsymbol{p}' \; \mathcal{V}_w(\boldsymbol{r}, \boldsymbol{p} - \boldsymbol{p}') f_w(\boldsymbol{r}, \boldsymbol{p}', \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t) + \frac{\partial}{\partial t} f_w \Big|_{ep} , \quad (17.42)$$

where $\boldsymbol{F}$ represents a "regular" constant or harmonic force; the potential term has been written in the form given by (17.26), and the last e–ph term is the one in (17.41).

In the following sections, we shall see how this equation can be solved, at least in principle, with a "particle simulation method". However, other methods have been used (see, for example [79, 155, 160, 221, 245, 282, 317, 321, 356, 412]) often ignoring phonon scattering or treating it in a semiclassical approximation. Scattering states have been used to treat exactly the potential term in mesoscopic systems with the inclusion of quantum phonon scattering to the lowest order [55].

17.6 Wigner Paths and MC Simulation

17.6.1 Integral Equation

To obtain an integral equation analogous to the classical Chambers equation, it is now convenient to make the transformation to the path variables introduced in Sect. 10.5.1. Defining

$$f^*(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*) = f(\boldsymbol{r}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*), \boldsymbol{p}(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*), t(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*)), \quad (17.43)$$

the Liouvillian on the l.h.s. of (17.42) becomes $\partial f_w^*/\partial t^*$, and the dynamical equation (17.42) becomes

$$\frac{\partial}{\partial t^*} f_w^*(\boldsymbol{r}^*, \boldsymbol{p}^*, \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t^*) = \frac{1}{h^3} \int \mathrm{d}\boldsymbol{p}' \; \mathcal{V}_w^*(\boldsymbol{r}^*, \boldsymbol{p}^* - \boldsymbol{p}', t^*) f_w(\boldsymbol{r}^*, \boldsymbol{p}', \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t^*) + \left. \frac{\partial f_w}{\partial t} \right|_{ep}^* .$$

This equation can be formally integrated, yielding

$$f_w^*(\boldsymbol{r}^*, \boldsymbol{p}^*, \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t^*) = f_w^*(\boldsymbol{r}^*, \boldsymbol{p}^*, \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t_\circ) + \int_{t_\circ}^{t^*} \mathrm{d}t' \frac{1}{h^3} \times \int \mathrm{d}\boldsymbol{p}' \; \mathcal{V}_w^*(\boldsymbol{r}^*, \boldsymbol{p}^* - \boldsymbol{p}', t') f_w(\boldsymbol{r}^*, \boldsymbol{p}', \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t') + \int_{t_\circ}^{t^*} \mathrm{d}t \left. \frac{\partial f_w}{\partial t} \right|_{ep}^* .$$

To return to the original variables, we observe that, given the three values $(\boldsymbol{r}^*, \boldsymbol{p}^*, t^*)$, the corresponding variables $(\boldsymbol{r}, \boldsymbol{p}, t)$ are the points that correspond to the orbit passing through $\boldsymbol{r}^*$ and $\boldsymbol{p}^*$ at time $t_\circ$, advanced, on the orbit, to the time $t^* = t$. At the starred values $(\boldsymbol{r}^*, \boldsymbol{p}^*, t_\circ)$, therefore, correspond the unstarred values $(\boldsymbol{r} = \boldsymbol{r}^*, \boldsymbol{p} = \boldsymbol{p}^*, t = t_\circ)$. It is then convenient to define the position of "running points" in phase space as

$$\overline{\boldsymbol{r}}(t') = \boldsymbol{r}(\boldsymbol{r}^*, \boldsymbol{p}^*, t') \; , \quad \overline{\boldsymbol{p}}(t') = \boldsymbol{p}(\boldsymbol{r}^*, \boldsymbol{p}^*, t'). \tag{17.44}$$

Then, in the original variables, the dynamical equation for the WF is the integral equation

$$\begin{aligned} f_w(\boldsymbol{r}, \boldsymbol{p}, \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t) &= f_w(\overline{\boldsymbol{r}}(t_\circ), \overline{\boldsymbol{p}}(t_\circ), \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t_\circ) \\ &+ \int_{t_\circ}^{t} \mathrm{d}t' \frac{1}{h^3} \int \mathrm{d}\boldsymbol{p}' \; \mathcal{V}_w(\overline{\boldsymbol{r}}(t'), \overline{\boldsymbol{p}}(t') - \boldsymbol{p}') f_w(\overline{\boldsymbol{r}}(t'), \boldsymbol{p}', \{n_{\boldsymbol{q}}\}, \{n'_{\boldsymbol{q}}\}, t') \\ &+ \int_{t_\circ}^{t} \mathrm{d}t' \left. \frac{\partial f_w(t')}{\partial t} \right|_{ep} . \end{aligned} \tag{17.45}$$

This equation has a physical interpretation identical to that of Chambers equation (10.38): the value of the WF at any point and time is given by the ballistic contribution coming from the initial time, plus all the contributions inserted by the interactions with the potential and the phonons into the right path at any time and then transferred to the point under consideration. With respect to Chambers equation (10.38), in the present equation (17.45) the exponential with the state lifetime is missing. It is implicit in the collision integral. However, it can be recovered, together with a formal "self-scattering" procedure, in the simulation of the WF [52, 379].

In the above equation, the initial time is taken to be $t_\circ$, independent of $\boldsymbol{r}$, $\boldsymbol{p}$ and t, assuming that the initial WF at $t_\circ$ is known everywhere. If, however,

the system is defined in a finite domain with known boundary conditions, and, for a given set of values $(\boldsymbol{r}, \boldsymbol{p}, t)$, $\boldsymbol{r}^*$ lies outside of this domain, the initial time $t_\circ$ must be substituted with the time t_b at which the "backward orbit" crosses the boundary.

17.6.2 Neumann Expansion and Wigner Paths

Equation (17.45) can be iteratively inserted into itself, thus yielding what is known as its Neumann expansion. The explicit expression of such a series would be very cumbersome, and it is useful to introduce a simpler, more concise formalism. Let us then rewrite (17.45) as

$$f = f_\circ + \mathbf{S}f, \tag{17.46}$$

where $f_\circ$ is the ballistic contribution, and $\mathbf{S}$ is the integral operator that contains all interactions. Substituting this equation into itself iteratively, we obtain

$$f = f_\circ + \mathbf{S}f_\circ + \mathbf{S}\mathbf{S}f_\circ + \ldots. \tag{17.47}$$

When this series converges (and we have good physical reasons to assume it does, see also [320]), it represents a formal solution to our transport problem.

Here, each operator $\mathbf{S}$ implies a double integration, one over time and one over momentum transfer. For example, if we limit ourselves to electron interaction with the potential $V(\boldsymbol{r})$ and omit the phonon variables, the first-order term is given by

$$\mathbf{S}f_\circ = \int_{t_i}^{t} \mathrm{d}t' \frac{1}{h^3} \int \mathrm{d}\boldsymbol{p}' \; \mathcal{V}_w(\overline{\boldsymbol{r}}(t'), \overline{\boldsymbol{p}}(t') - \boldsymbol{p}') f_w(\overline{\boldsymbol{r}}_1(t_i'), \overline{\boldsymbol{p}}_1(t_i'), t_i'), \tag{17.48}$$

where t_i is the greater between t_b and $t_\circ$, and $\overline{\boldsymbol{r}}_1(t_i)$ and $\overline{\boldsymbol{p}}_1(t_i)$ are the position and momentum at time $t = t_i$ of a classical particle (here a mathematical simulation particle) which at time t' is in $\overline{\boldsymbol{r}}(t')$ with momentum $\overline{\boldsymbol{p}}(t')$. Thus, the first-order integral above contains all the contributions to the WF $f_w(\boldsymbol{r}, \boldsymbol{p}, t)$ that can be considered coming from particles that start at $t = t_i$ with any position and momentum and at any time t' are scattered into the right trajectory such that the particle will be in $\boldsymbol{r}$ at t with momentum $\boldsymbol{p}$.

Higher order terms are similarly interpreted in perfect analogy to the classical case, as discussed in Sect. 14.3.3.

A graphic representation of the orbits that contribute to the first three terms of the Neumann expansion in (17.47) for the case of potential interaction is given in Fig. 17.4. We call these orbits Wigner paths. A path is split into several free-flight sections by the interaction points called vertices. In each free-flight section, the representative particle follows a classical trajectory; at each vertex its momentum changes in a discontinuous way.

When the term containing phonon interaction is considered, the procedure to obtain the Neumann expansion is in principle the same, but it becomes, in

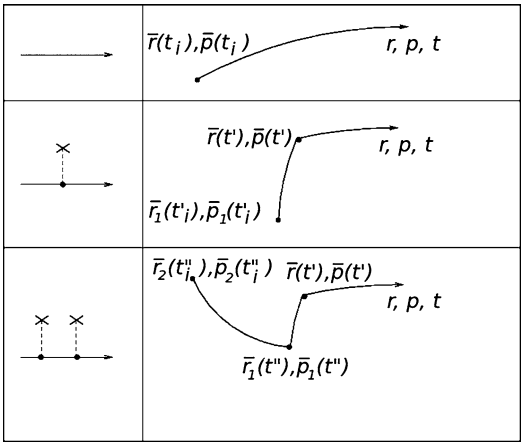


Fig. 17.4. Potential interactions and the corresponding Wigner paths. From *top*: ballistic term; one interaction, first order; two interactions, second order

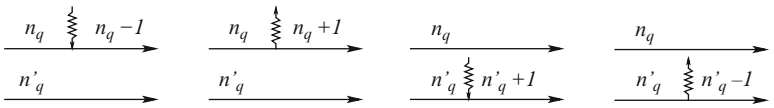


Fig. 17.5. Elementary vertices for e–ph interaction

practice, more complicated by the fact that each time four terms are inserted, according to the four terms in (17.41). Since the phonon occupation number can be changed in the first or second set of values in the arguments of the WF, the schematic representation of Fig. 17.4 must now distinguish the two cases, and this can be obtained by considering two lines, one for each set of arguments.[1] The four first-order terms of e–ph interaction can then be represented as in Fig. 17.5. Note that in the bottom line, which refers to the right phonon arguments in the WF, the operators act as hermitian conjugate of the ones acting on the upper line.

The second-order term of the Neumann expansion, considering only phonon interaction, would contain 16 terms. Let us remember, however, that each phonon vertex changes the occupation number of one mode on one set of arguments. Thus, if the initial value of the WF is supposed to be diagonal with respect to the phonon states (for example because we start with a non-interacting equilibrium state before switching on the e–ph interaction), and we are looking for a WF that is still diagonal (because we are interested in evaluating the trace over phonon variables), then each phonon mode must be involved twice. It can be emitted and reabsorbed in the first set of arguments (virtual emission), emitted in the first and in the second (real emission), and

[1] Note the equivalence with the two branches of the Keldish contour integral for the Green functions in Sect. 26.2.

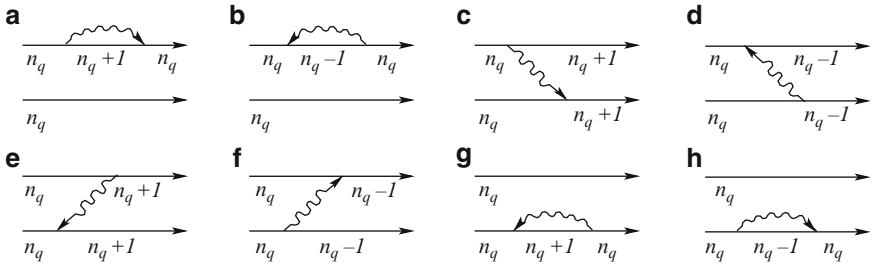


Fig. 17.6. 8 possible e–ph processes that leave the phonon bath in a diagonal state. (**a**) and (**g**) virtual emission, (**b**) and (**h**) virtual absorption, (**c**) and (**e**) real emission; (**d**) and (**f**) real absorption

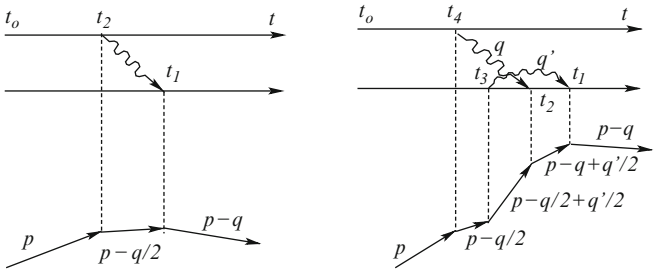


Fig. 17.7. *Left*: Example of real phonon emission with the corresponding Wigner path in real space. *Right*: Example of a multiple scattering formed by e real emission and a virtual absorption, and the corresponding Wigner path in real space [37]

so on. In this case, the number of terms reduces to 8, that in the Keldish symbolism, are given by the 8 terms in Fig. 17.6.

Examples of a second (real emission) and of a fourth (real emission plus virtual absorption) order term are given in Fig. 17.7 with the corresponding Wigner paths.

In general, the Neumann series in (17.47) will contain mixed terms with interactions with the potential $V(\boldsymbol{r})$ and with the phonon field.

17.6.3 Monte Carlo Simulation

We saw in Chap. 14 that MC algorithms for semiclassical electron transport can be understood on the basis of a direct simulation of particle dynamics or of a more formal MC solutions of integral equations.

In the present case of quantum transport, however, real trajectories do not exist; the particles do not posses well-defined positions nor momenta, and the more formal approach to MC algorithm is compulsory. The concept of Wigner paths developed above, however, makes the MC simulation for the Wigner function very similar to the one devised intuitively for the simulation of the classical distribution function.

According to the basic ideas introduced in Sect. 14.3, the MC algorithms can be used to solve the dynamical equation of the WF through a sampling of the terms of its Neumann expansion in (17.47). For the sake of clarity, we start again by considering only kinetic and potential terms in the Hamiltonian, neglecting, for the time being, phonon interaction.

Let $\boldsymbol{r}$ and $\boldsymbol{p}$ be the point in phase space where the WF is to be evaluated at time t. With arbitrary probability P_n $(n = 0, 1, \ldots)$ we first select the term to be sampled in the sum in (17.47) and set a weight W equal to the value $1/P_n$.

If $n = 0$ has been selected, the ballistic term must be considered, i.e., the first term in (17.45). In such a case, no further random selection is required; we simply evaluate $\overline{\boldsymbol{r}}(t_i)$ and $\overline{\boldsymbol{p}}(t_i)$, the initial or boundary values of the phase-space variables of the selected path and "read" the value of the known WF at this point. This value, multiplied by the weight W is the estimate of our WF in $\boldsymbol{r}$ and $\boldsymbol{p}$ at time t.

If a value of n greater than zero has been selected, n potential interaction vertices must be considered. The first interaction time t_1 must be selected with arbitrary probability P_{t_1} between time t and the initial or boundary time of the first backward trajectory. The weight W must correspondingly be divided by P_{t_1}. Then a momentum transfer $\Delta\boldsymbol{p}_1$ must be selected with arbitrary probability $P_{\Delta\boldsymbol{p}_1}$ and the weight divided by $P_{\Delta\boldsymbol{p}_1}$; the new value of the "virtual-particle" momentum just before the interaction vertex is then evaluated as $\boldsymbol{p}_1 = \overline{\boldsymbol{p}}(t_1) - \Delta\boldsymbol{p}_1$, and the new "free flight" is considered. The process is repeated for all n vertices, and the initial or boundary value of the WF is used, multiplied by the total weight W and by the values of the integrands $\mathcal{V}_w$ at the vertex coordinates, in order to estimate the WF at the selected point.

When phonon interaction is considered, it must be included in the random selection of the vertices. In this case, however, we must take into account that if values of the WF diagonal with respect to the phonon occupation numbers are to be evaluated, and at the initial or boundary condition the WF is already diagonal, each phonon mode must be involved in two vertices, as discussed above. It is of course possible that between two phonon vertices corresponding to the emission or absorption of a quantum of mode $\boldsymbol{q}$ an electric field may act, or a potential vertex may be present, etc.

As it regards phonon interaction, other considerations are needed that will be discussed in the next sections.

Thus, the MC sampling of our series of multiple integrals corresponds to a random selection of a number of Wigner paths among the infinite possible ones that contribute to the determination of the WF. This process corresponds to a sort of Feynman path integral, with all the known disadvantages related to the slow convergence of the estimator: a very large number of paths has to be considered to have a reasonable value of the estimator, and the variance is larger when higher-order terns are considered. For the ballistic zero-order term, no statistics is necessary: one single path yields the exact value. However,

with regard to paths with several interaction vertices, the size of the necessary sample grows exponentially with the number of vertices [400].

The above procedure, where we choose first the values of $\boldsymbol{r}$ and $\boldsymbol{p}$ where to calculate the WF, is equivalent to choose first, in the set of sums in (14.6), which sum S_k to evaluate, and then apply the procedure for a given sum. As we have seen, this implies a backward simulation of the Wigner paths, starting from the final point, and reaching at the end the initial/boundary value of the known WF. This, however, is not the only possibility: we can select terms of the different sums in arbitrary ways, which implies a great flexibility of the method that should be used to make the simulation more efficient. If we select first the initial/boundary value of the Wigner path and then the successive free flights and interaction vertices, we perform a forward propagation.

We have observed above that terms of different orders require quite different sizes of statistical samples. Thus, it is much more efficient to select first the order of the paths, and then sample all possible paths of that order with a suitable sample size.

Path Multiplicity

The result of the numerical procedure is obtained by simulating a very large number of paths, each weighted by a suitable factor. In the determination of this factor it must be taken into account that in the case of e–ph interaction several terms of the sum in (17.47) are complex conjugate of each other and correspond to the same Wigner path [216]. As an example, Fig. 17.8 shows the four graphs that yield the path at the bottom of the figure. An analysis of the complex-conjugate terms which contribute to the same trajectory shows that the exponential factors in (17.41) give rise to a factor equal to

$$2\cos[\boldsymbol{q}(\boldsymbol{r}_1 - \boldsymbol{r}_2) - \omega_q(t_1 - t_2)], \tag{17.49}$$

where $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, t_1 and t_2 are the positions and the times of the two vertices of the process, $\boldsymbol{q}$ and ω_q are the wavevector and the frequency of the emitted/absorbed phonon [216].

Phonon Average

The simulation procedure described above refers to values of the WF for all possible values of the phonon occupation numbers $\{n_{\boldsymbol{q}}\}$ and $\{n'_{\boldsymbol{q}}\}$. In principle, in a backward simulation, we should first determine the values of $\{n_{\boldsymbol{q}}\}_t$ and $\{n'_{\boldsymbol{q}}\}_t$ at which the WF has to be evaluated and then read the WF at the values of $\{n_{\boldsymbol{q}}\}_{t_\circ}$ and $\{n'_{\boldsymbol{q}}\}_{t_\circ}$ resulting from the simulation. Even if we start from diagonal terms ($\{n_{\boldsymbol{q}}\}_{t_\circ} = \{n'_{\boldsymbol{q}}\}_{t_\circ}$) and evaluate the WF at diagonal terms ($\{n_{\boldsymbol{q}}\}_t = \{n'_{\boldsymbol{q}}\}_t$) this would imply a huge number of values for the WF and of the sampling. This is not necessary, however, as long as we are not interested in hot phonon effects [288, 289] so that we can assume that

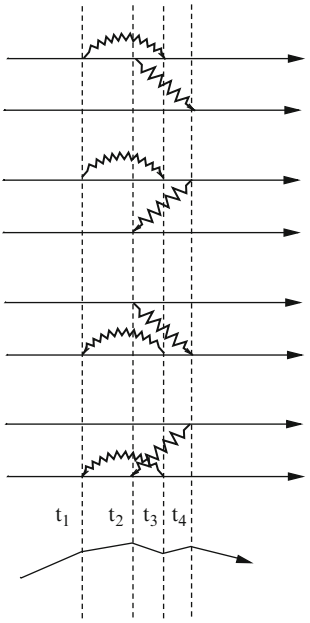


Fig. 17.8. Four diagrams corresponding to a multiple scattering with a virtual emission and a real emission contributing to the Wigner path at the bottom of the figure

the electron interacts with a phonon bath always in thermal equilibrium. In this case, in fact, we may "think" of sampling a given Wigner path with all possible values of the phonon occupation numbers, and the final process of average over the phonon distribution turns out to be equivalent to substitute each factor $n_{\boldsymbol{q}}$ (for absorption) and $(n_{\boldsymbol{q}}+1)$ for emission, with their average equilibrium number in each sampling path [73, 216].

17.7 Two-Time Wigner Function

The WF has been introduced in Sect. 17.2 as the Weyl–Wigner transformation of the density-matrix operator. The two functions g and g^* that appear in its generalized definition (17.36) are evaluated at the same time. If we extend even more the definition considering two different times for g and g^* and performing the Weyl–Wigner transformation also on the time variables, a new Wigner function is obtained which depends upon momentum and frequency independently, besides upon central position $\boldsymbol{r}$ and central time t [69]:

$$f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\},\{n'_{\boldsymbol{q}}\},t,\omega) = \int \mathrm{d}\boldsymbol{s}\, \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\boldsymbol{s}/\hbar} \times \int \mathrm{d}\tau\, \mathrm{e}^{-\mathrm{i}\omega\tau} g\left(\boldsymbol{r}+\boldsymbol{s}/2,\{n_{\boldsymbol{q}}\},t+\tau/2\right) g^*\left(\boldsymbol{r}-\boldsymbol{s}/2,\{n'_{\boldsymbol{q}}\},t-\tau/2\right), \quad (17.50)$$

and

$$f_w(\boldsymbol{r},\boldsymbol{p},t,\omega)=\sum_{\{n_{\boldsymbol{q}}\}} f_w(\boldsymbol{r},\boldsymbol{p},\{n_{\boldsymbol{q}}\},\{n_{\boldsymbol{q}}\},t,\omega).$$

It will be seen in the last part of the book that this extended definition of the WF can be introduced as the Fourier transform of the $G^<$ Green function [294].

The dynamical equation for the WF defined in (17.50) can be obtained following the same procedure already developed in Sects. 17.4 and 17.5. The Neumann expansion can again be performed; the concept of Wigner path is still applicable and allows us to apply the MC procedure based on the generation of Wigner paths to the calculation of $f_w(\boldsymbol{r},\boldsymbol{p},\{n_q\},\{n'_q\},t,\omega)$. The only requirement to be added in the sequence of random choices and operations which define a particular path, is that at each e–ph interaction vertex half of the phonon frequency (besides half of the phonon momentum) is either added or subtracted to the electron energy according to the selected process. Thus, when both vertices of an e–ph process are taken into account, the energy balance is completed and energy conservation is perfectly respected. However, the relation between momentum transfer and energy transfer is built up through the time integration (i.e., through sampling of different times, in a MC procedure) of the factor in (17.49). Since new vertices may occur before the time integration is completed, the classical relation between transferred momentum and transferred energy is not exactly recovered. In conclusion, *the relation between p and ω is not the classical one and not univocally determined: a collisional broadening appears because eigenstates of p are not eigenstates of the Hamiltonian; energy and momentum must be considered independent variables, but energy conservation is perfectly preserved at all times in each process.*

The formalism of the two-time WF, or of the WF depending on momentum and energy treated as independent variables, has been applied [207] to the quantum analysis of electrons interacting with polar optical phonons. Electron lifetimes, that is scattering rates, spectral functions, that is collisional broadening, and the dynamics of the polaron formation, have been analyzed with the MC Wigner–path technique [207].

17.8 Many-Particle Wigner Function

The original WF introduced by Wigner was already defined for a system of N particles. Here, we limit ourselves to the case of two particles, the generalization to any number of identical particles does not imply conceptual difficulties. Let us then define the WF for two particles, in analogy with (17.10), as

$$f_w(\boldsymbol{r}_1,\boldsymbol{r}_2,\boldsymbol{p}_1,\boldsymbol{p}_2)=\int \mathrm{d}\boldsymbol{s}_1 \int \mathrm{d}\boldsymbol{s}_2\ \mathrm{e}^{-\mathrm{i}\boldsymbol{s}_1\boldsymbol{p}_1/\hbar}\ \mathrm{e}^{-\mathrm{i}\boldsymbol{s}_2\boldsymbol{p}_2/\hbar}$$
$$\times \Psi(\boldsymbol{r}_1+\boldsymbol{s}_1/2,\boldsymbol{r}_2+\boldsymbol{s}_2/2)\Psi^*(\boldsymbol{r}_1-\boldsymbol{s}_1/2,\boldsymbol{r}_2-\boldsymbol{s}_2/2). \qquad (17.51)$$

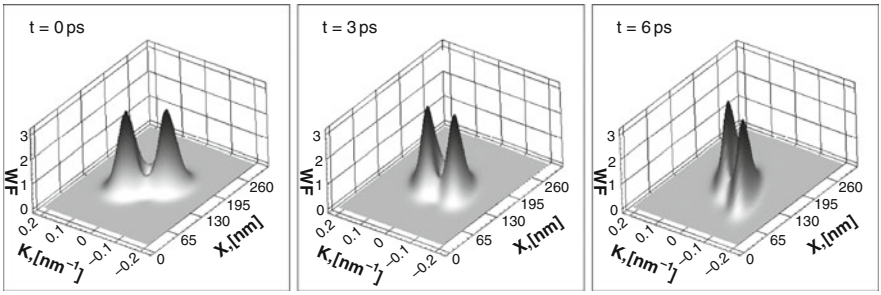


Fig. 17.9. WF of two identical fermions, integrated over position and momentum of one particle, at three different times [89, 90]

If we now assume that, as in our case, we are dealing with identical fermions, the antisymmetry of the wavefunction immediately implies that the WF is unchanged by exchanging both positions and momenta of the two particles:

$$f_w(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{p}_1, \boldsymbol{p}_2) = f_w(\boldsymbol{r}_2, \boldsymbol{r}_1, \boldsymbol{p}_2, \boldsymbol{p}_1). \tag{17.52}$$

If we integrate over $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, we immediately obtain

$$\int \mathrm{d}\boldsymbol{p}_1 \int \mathrm{d}\boldsymbol{p}_2 \; f_w(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{p}_1, \boldsymbol{p}_2) = |\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2)|^2, \tag{17.53}$$

as for the case of a single particle. An analogous result is obtained with the integration over momentum variables.

The dynamical theory of a many-particle WF can be developed in strict analogy with that of a single particle. Wigner paths of two simulative particles can be defined and used for the dynamical evolution. In this case we have to consider the possibility of a particle–particle interaction. In an interaction vertex the total momentum is conserved: $\Delta\boldsymbol{p}_1 = -\Delta\boldsymbol{p}_2$.

To visualize the consequence of the antisymmetry of fermion wavefunctions, the WF describing the free propagation of two electrons moving against each other is shown in Fig. 17.9. Electrons are described by a pair of one-dimensional antisymmetrized minimum-uncertainty wavepackets, not interacting apart from antisymmetry. The WF shown in the figure is integrated over position and momentum of one particle. The effect of the antisymmetry of the wavefunction can be seen in a cleft in the two Gaussian packets where they overlap [89, 90].

Part IV

Transport in Semiconductor Structures

18

Inhomogeneous and Open Systems: Electronic Devices

18.1 Inhomogeneous, Open Systems

The strong interest for the physics of semiconductors, and in particular for the charge transport properties of these materials, is mainly due to their use in electronic devices. The related technology has reached limits inconceivable when the adventure started, in 1948, with the invention of the transistor[1] [24, 416].

When we move from the study of homogeneous, theoretically infinite, bulk materials to the analysis of electronic devices, we must consider two essential points. On one side, the systems are not homogeneous, often made of parts realized with different materials. Furthermore, we are dealing with *open systems*, where electrons and/or holes can enter and leave the systems through *contacts* which connect them to external reservoirs, generally maintained at different potentials.

The general problem of electron transport in a device can be formulated as follows: find the distribution of charge density, current density, and electric potential, inside a system with given geometry of materials and doping concentrations, in the presence of a certain number of contacts, in general from two to four, kept at given potentials, through which electrons and/or holes can enter or leave the system.

The presence of regions with different concentrations of fixed (ionized dopants) and mobile charges implies that the transport equation must be solved self-consistently with the Poisson equation:

$$\boxed{\nabla \cdot (\varepsilon(\boldsymbol{r}) \nabla \phi(\boldsymbol{r})) = -\rho(\boldsymbol{r})} \tag{18.1}$$

[1] The name *transistor* was suggested by J.R. Pierce, as an adaptation of *transresistance* to the names of other devices, such as varistor and thermistor.

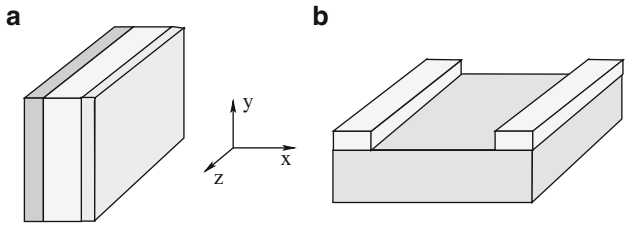


Fig. 18.1. The problem of modeling an electronic device is often simplified by reducing the number of essential dimensions on the basis of some symmetry of the device. In case (**a**), translational symmetry is assumed along the directions y and z so that the quantities of interest depend only upon x, and the problem becomes one-dimensional; in (**b**), the problem is two-dimensional, assuming symmetry along z

Here, $\phi(\boldsymbol{r})$ is the electric potential field, $\varepsilon(\boldsymbol{r})$ is the dielectric constant of the material, and $\rho(\boldsymbol{r})$ is the charge density given by

$$\rho(\boldsymbol{r}) = e(p - n + N_{\mathrm{D}} - N_{\mathrm{A}}), \tag{18.2}$$

where p, n, N_{D} ed N_{A} are the concentrations of holes, electrons, ionized donors, and ionized acceptors, respectively. The boundary conditions are given by the applied potentials on the electrode boundaries, and are supposed to be given by homogeneous Neumann conditions, i.e., zero normal derivative of the potential, $\nabla V|_n = 0$, at the insulating boundaries, on the basis of the continuity of the normal component of displacement field $\boldsymbol{D}$ and the higher dielectric constant of the device material with respect to the exterior.

In its general formulation, the problem is very difficult to solve and can be approached only with severe approximations and/or with numerical techniques. Often the problem is simplified by reducing the number of significant space dimensions, as illustrated in Fig. 18.1.

The problem becomes even more difficult when, with present-day technology, the linear dimensions of the device become comparable with the carrier coherence length (see below and Chap. 16), and a quantum treatment of the electron dynamics is necessary.

18.2 Self-Averaging Transport, Coherent Transport, and Intermediate Cases

When the physical system of interest, albeit inhomogeneous, may be decomposed in a number of relatively large homogeneous regions, within each of such regions charge carriers may encounter a large variety of different microscopic situations: they may be scattered by impurities with different impact parameters, or absorb and emit phonons of different branches with different momenta. In such a case, electrons behave in a manner similar to what they do in homogeneous materials. In each region, they sample, on average, the different possible occurrences. This situation is called *self-averaging regime*

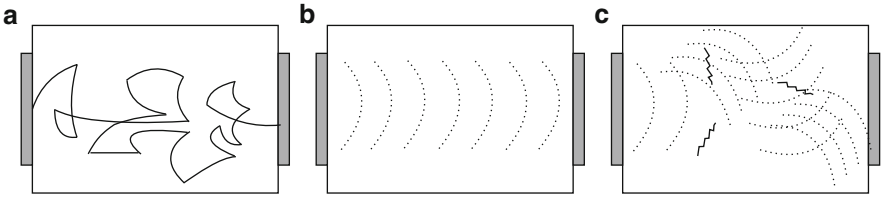


Fig. 18.2. Different transport regimes: (**a**) semiclassical, self-averaging; (**b**) coherent; (**c**) intermediate mesoscopic

(see part (a) of Fig. 18.2). In such a case, it is possible to define local transport properties, such as conductivity, mobility, diffusivity, etc., obtained by solving the Boltzmann transport equation. This is the transport regime always occurring during the first decades of semiconductor electronic devices.

In the opposite situation, when the dimensions of the system, or of a well-defined part of it, are so small to be comparable with the electron *coherence length*, i.e., the distance covered by the electrons during the coherence time defined in Sect. 16.1.3, the dynamics of the charge carriers is entirely described by the Schrödinger equation. This situation is called *coherent-transport regime* (part (b) of Fig. 18.2). It must be noted that for the realization of such a situation not only the systems dimensions must be very small, but also the temperature must be very low, in general, since phonon scattering interrupts the coherent dynamics of the electrons. As it regards impurities, on the contrary, we recall that a time-independent potential may be included in the Hamiltonian and does not destroy the carrier coherence. In fact, very interesting quantum effects can be observed in low-temperature coherent transport in presence of impurities, as discussed in Chap. 21. When also impurity scattering is absent, the resulting transport is said to be in the *ballistic regime*.

The situation that is probably most difficult to deal with is the intermediate case, when the system dimensions require a quantum description of the electron dynamics, but at the same time some scattering processes interrupt the coherent dynamics of the charge carriers inside the device (part (c) of Fig. 18.2). This situation, which may be called *mesoscopic regime*, i.e., intermediate between microscopic and macroscopic, requires not only a quantum treatment of dynamics, but also the inclusion of some phonon scattering events that interrupt the coherent dynamics of the electrons in the region of interest. Furthermore, in quantum terms the scattering processes cannot be seen as abrupt changes of the electron velocity along its path (the latter does not even exist). This regime must be treated with many-body techniques, typically using Green functions, introduced in the last part of this book, or with the Wigner function discussed in Chap. 17, which is, in fact, a particular Green function.

In the following pages, we shall discuss the problem of open systems formed by different parts with different physical properties. We shall assume, however,

to be always in the self-averaging regime with local dynamics described by the semiclassical Boltzmann transport equation (BTE).

18.3 pn Junctions

To begin with, let us consider the simple problem[2] of a *p–n junction*. It is a very important case, owing to its application to many electronic semiconductor devices. We shall assume a planar geometry so that, according to the considerations of the previous section, the problem is limited to one dimension. Furthermore, we shall make the *abrupt-junction approximation* which consists in assuming that a sharp surface separates the two parts of the junction.

The abrupt junction is a very simplified representation of the actual structure, and the descriptions that follow do not represent, of course, the processes realized for the fabrication of the corresponding devices. These technological aspects are outside of the scope of this book.

18.3.1 pn Junction at Equilibrium

Let us assume that we have two pieces of the same material, one n-doped and the other p-doped, not yet in contact. We also assume that the two dopant concentrations, and therefore also the mobile charge carriers at equilibrium, are uniform in the two single materials. Since we assume that the potential energy in the space around the system is constant and uniform, the valence and conduction band edges are the same in the two pieces, as shown in part (a) of Fig. 18.3.

In the p-type material, the one on the left in the figure, we have positive charges, i.e., holes, free to move, and negative fixed charges, the ionized acceptors. Here, the Fermi level[3] is close to the top of the valence band. In the n-type material, on the right, we have negative free electrons and fixed positive charges at the ionized donors, and the Fermi level is near the bottom of the conduction band.

As long as the two materials are kept at equilibrium and separate, the charges of opposite signs, mobile and fixed, neutralize each other, and the materials are neutral in each point. Poisson equation (18.1) is verified by a constant potential.

When the two pieces are put in contact (part (b) of Fig. 18.3), electrons free to move will diffuse from the n-type material into the adjacent material; here

[2] Not so simple, after all, if 120 pages of the classical book by Sze [431] and the entire volume II of the Addison–Wesley *Modular Series on Solid State Devices* [325] are devoted to this subject. Here, only the main ideas will be developed using the simplest possible model.

[3] For consistency with most of the device literature, in this chapter we call Fermi level what should be called electrochemical potential, according to the discussion in Sect. 8.6.

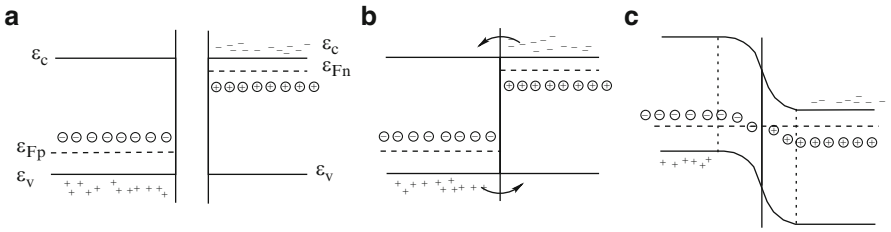


Fig. 18.3. Band diagram of a pn junction before (**a**), at (**b**), and after (**c**) the contact

they find available states in the valence band at energies below their original Fermi level, and occupy them, recombining with the holes. Thus, they leave unbalanced positive fixed charges on the right and negative fixed charges on the left. The holes will behave in the dual way, recombining with electrons on the right.

A dipole layer is thus formed that rises the electron potential energy of the p-type material and lowers that of the n-type material, until a new equilibrium situation is attained with a constant Fermi level throughout the device (part (c) of Fig. 18.3). In more elementary and qualitative terms, we may say that the transfer of charge from one piece to the other terminates when it is prevented by the electrostatic field. The fixed charges left unbalanced at the opposite sides of the interface have equal and opposite values owing to charge conservation and form a *space-charge region* also called *depletion region*, and constitute the dipole layer, which determines the potential difference between the two parts of the junction.

Far from the depletion region, the energy differences between the external surface of the materials, the Fermi levels, and the band edges are determined by the properties of the materials. Thus, the constant Fermi level implies that the potential difference between the two parts of the junction far from the interface is equal to the difference of the Fermi levels of the two separate materials. This difference is called *built-in potential* or *contact potential*:

$$V_{\mathrm{b}} = \epsilon_{Fn} - \epsilon_{Fp}.$$

The shape of the potential profile inside the space-charge region depends on the charge density according to the Poisson equation (18.1), which, in the simple one-dimensional (and fully depleted, see below) case,[4] yields two parabolic sections, as shown in part (c) of Fig. 18.3. The curvatures are determined by the charge densities of the dopants.

Let us make some quantitative considerations. With reference to part (a) of Fig. 18.4, take $x = 0$ at the interface of the two materials and call w_n and w_p

[4] Pay attention to the meaning of one- or two-dimensional Poisson equation: the charge density ρ must always be 3-D, otherwise the dimension of the potential (energy/charge) is lost. To be one-dimensional means that both V and ρ depends only upon one space variable.

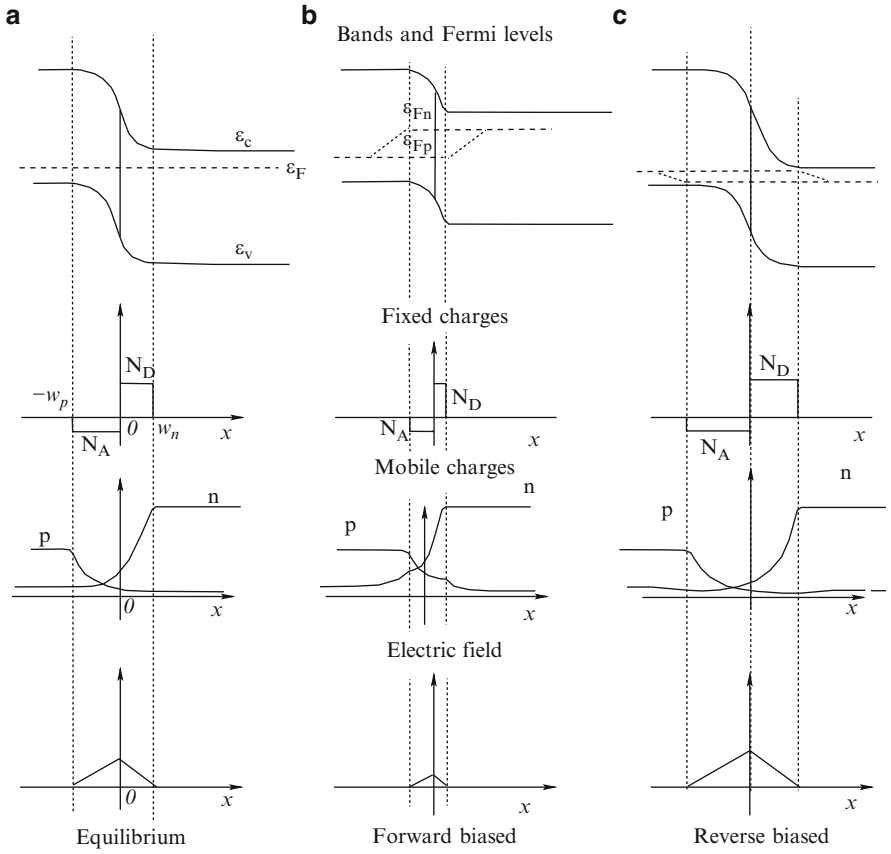


Fig. 18.4. Profile of bands and Fermi levels, fixed and mobile charges, and electric field in a pn junction at equilibrium (**a**), forward biased (**b**), and reverse biased (**c**). The diode is forward biased when the applied voltage $V > 0$ lowers the potential step and reduces the width of the depletion region. If $V < 0$, the potential step is higher, and the depletion regions becomes wider

the two lengths of the space-charge regions. Out of these regions, the electrical potential is constant from the left boundary to the position $x = -w_p$ and from $x = +w_n$ to the right boundary. The Poisson equation with constant ε

$$\frac{\mathrm{d}^2 V(x)}{\mathrm{d}x^2} = -\frac{1}{\varepsilon}\rho(x) \tag{18.3}$$

must then be solved in the space-charge region between $x = -w_p$ and $x = w_n$. In such a region the charge density is, according to the so-called *complete-depletion approximation*,

$$\rho(x) = \begin{cases} -eN_\mathrm{A} & -w_p < x < 0 \\ eN_\mathrm{D} & 0 < x < w_n \end{cases} . \tag{18.4}$$

The solution of the differential equation (18.3) is immediately obtained, separately in the two regions where ρ is supposed to be constant, with two successive integrations:

$$V(x) = \begin{cases} A_p + B_p\, x + \frac{eN_\mathrm{A}}{2\varepsilon}\, x^2 & -w_p < x < 0 \\ A_n + B_n\, x - \frac{eN_\mathrm{D}}{2\varepsilon}\, x^2 & 0 < x < w_n \end{cases}, \tag{18.5}$$

where A_i and B_i are constants to be determined by the boundary conditions in $x = -w_p$ and $x = w_n$ and by the continuity of the solution in $x = 0$. The boundary conditions can be taken as

$$V(-w_p) = 0\ , \quad V(w_n) = V_\mathrm{b}.$$

Poisson equation requires the continuity of $V(x)$ and of its derivative, since the second derivative if finite. Thus, at the left edge $x = -w_p$:

$$V(-w_p) = A_p + B_p(-w_p) + \frac{eN_\mathrm{A}}{2\varepsilon}(-w_p)^2 = 0,$$

$$V'(-w_p) = B_p + \frac{eN_\mathrm{A}}{\varepsilon}(-w_p) = 0.$$

From these two conditions, we obtain

$$B_p = \frac{A_p}{w_p} + \frac{eN_\mathrm{A} w_p}{2\varepsilon} \quad \text{and} \quad B_p = \frac{eN_\mathrm{A}}{\varepsilon} w_p.$$

From the comparison, we obtain the value of A_p:

$$A_p = \frac{eN_\mathrm{A} w_p^2}{2\varepsilon}.$$

If we replace these values of A_p e B_p in the expression for $V(x < 0)$, we obtain

$$V(x < 0) = \frac{eN_\mathrm{A} w_p^2}{2\varepsilon} + \frac{eN_\mathrm{A}}{\varepsilon} w_p x + \frac{eN_\mathrm{A}}{2\varepsilon} x^2 = \frac{eN_\mathrm{A}}{2\varepsilon}(x + w_p)^2.$$

If the same calculation is repeated at the boundary in $x = w_n$, the result is

$$V(x > 0) = V_\mathrm{b} - \frac{eN_\mathrm{D}}{2\varepsilon}(x - w_n)^2.$$

The continuity of the derivative of V at the interface yields

$$N_\mathrm{A} w_p = N_\mathrm{D} w_n, \tag{18.6}$$

which is the charge-neutrality condition. The continuity of V yields

$$V_\mathrm{b} = \frac{e}{2\varepsilon}\left(N_\mathrm{A} w_p^2 + N_\mathrm{D} w_n^2\right). \tag{18.7}$$

The two equations (18.6) and (18.7) are easily solved, and yield

$$w_n = \sqrt{\frac{2\varepsilon V_\mathrm{b}}{e}\frac{N_\mathrm{A}}{N_\mathrm{D}}\frac{1}{N_\mathrm{A}+N_\mathrm{D}}} \quad , \quad w_p = \sqrt{\frac{2\varepsilon V_\mathrm{b}}{e}\frac{N_\mathrm{D}}{N_\mathrm{A}}\frac{1}{N_\mathrm{A}+N_\mathrm{D}}}. \tag{18.8}$$

For the total length w of the space-charge region, we obtain

$$w = w_p + w_n = \left(1 + \frac{N_\mathrm{A}}{N_\mathrm{D}}\right) w_p = \sqrt{\frac{2\varepsilon V_\mathrm{b}}{e}\frac{N_\mathrm{A}+N_\mathrm{D}}{N_\mathrm{A}N_\mathrm{D}}}. \tag{18.9}$$

As a numerical example, if we take $\varepsilon = 11 \times 8.85 \times 10^{-12}$ F/m ($\mathrm{A^2 s^4/Kg\,m^3}$), $V_\mathrm{b} = 1\,\mathrm{V}$, $N_\mathrm{A} = 10^{22}\,\mathrm{m^{-3}}$, $N_\mathrm{D} = 10^{23}\,\mathrm{m^{-3}}$, and $e = 1.6 \times 10^{-19}$C, we obtain $w_n \approx 0.03\,\mu\mathrm{m}$, $w_p \approx 0.3\,\mu\mathrm{m}$.

18.3.2 pn Diode

If two metal contacts are attached at the edges of the pn system described above, on the basis of the same principles seen in the pn junction, potential differences are generated, at equilibrium, between the metals and the semiconductors, given by the differences of the Fermi levels of the isolated materials (see Sect. 18.5 below). If the two contacts are made of the same metal, at equilibrium the various contact potentials sum up to zero and the two external contacts are at the same potential. If a potential difference is then applied between the two contacts, it drops almost completely across the depletion region of the pn junction where the resistance is higher because of the absence of mobile charge carriers. In this section, we shall show that a pn junction in this situation behaves as a diode: it is a good conductor if the potential difference tends to reduce the voltage drop across the depletion region (*forward bias*), while it conducts very little in the opposite case (*inverse bias*), when the applied potential tends to increase the voltage drop across the depletion region.

If the potential difference is applied at the contacts, the device is no longer in equilibrium, and the Fermi level is no longer well defined. A *quasi Fermi level*, sometimes called *imref*, can be defined locally, based on the carrier concentration [417, 431], and in the pn junction it is defined separately for electrons and holes. The quasi Fermi level of the holes ϵ_{Fp} is almost constant in the p region and in the depletion region, and increases in the n region by an amount given by the applied potential, while the quasi Fermi level ϵ_{Fn} of the electrons is almost constant in the n region and in the depletion region, and decreases by the same amount in the p region, as shown in Fig. 18.4.

Let V be the applied voltage, and consider first the case of a forward bias, shown in part (b) of Fig. 18.4. The considerations made in the previous section on the application of the Poisson equation are still valid, but the potential drop across the space-charge region is no more V_b, but $V_\mathrm{b} - V$, less than the previous one. Thus, the space-charge region becomes thinner. In full-depletion approximation, (18.8) are still valid with $V_\mathrm{b} - V$ in place of V_b.

In equilibrium conditions, drift and diffusion currents cancel each other and no net current is present. If the diode is forward biased, the electrons in the n-type material continue to diffuse toward the other part of the junction, where they can find available states with lower energies. The diffusion current is no more equilibrated by the drift current. Electrons that reach the region of mobile charge in the p-type material soon recombine, and in stationary state an equal number of holes enter the diode from the left contact (i.e., electrons jump into the external circuit). Similarly, holes of the p region diffuse through the depletion region and recombine, so that new electrons enter from the negative contact.

To evaluate the current, first consider that the total current, constant along the device, is given by the sum of electron and hole currents, separately functions of position, because of recombinations and generations. In the space-charge region, however, owing to the low carrier densities, generations and recombinations are very small, given, in the full-depletion approximation, by their intrinsic thermal values. Hole and electron currents are therefore constant along the depletion region. We may thus calculate the electron current at the edge of the depletion region where they are minority carriers, and the hole current at the opposite edge. In these points, the electric field is very small and the drift current due to minority carriers is negligible. Thus, we may evaluate the currents from the diffusion equations neglecting the drift term. Let us then consider the diffusion equation (12.2) of Chap. 12 for holes in steady-state condition, with the addition of generation and recombination terms, g and r, respectively:

$$\frac{\partial p}{\partial t} = 0 = D_p \frac{\partial^2 p}{\partial x^2} + g - r. \tag{18.10}$$

The recombination must be proportional to the product of the concentrations with a proportionality constant that depends upon temperature:

$$r = A(T)np.$$

In thermal equilibrium, generation g_{th} and recombination r_{th} rates must equilibrate each other, with the concentrations given by $n_\circ$ e $p_\circ$ for that type of material at that particular temperature. In the n-type material

$$g_{\text{th}} = r_{\text{th}} = A(T)n_{n\circ}p_{n\circ},$$

where $n_{n\circ}$ and $p_{n\circ}$ are the equilibrium concentrations of electrons and holes, respectively, in the n region. However, the generation is not influenced by free carriers around, and in (18.10) we may consider it equal to its value at equilibrium. The difference at the r.h.s. of (18.10) is then given by

$$g - r = A(T)n_{n\circ}p_{n\circ} - A(T)np. \tag{18.11}$$

If we consider a *low-injection regime*, $\Delta n \ll n_\circ$, n can be approximated by $n_\circ$ in (18.11), which becomes

$$g - r = -\frac{p - p_{n\circ}}{\tau_p} \quad , \quad \tau_p = \frac{1}{A(T) n_{n\circ}},$$

where τ_p is the *minority-carrier lifetime*. Equation (18.10) becomes

$$D_p \frac{\partial^2 p}{\partial x^2} - \frac{p - p_{n\circ}}{\tau_p} = 0. \tag{18.12}$$

A similar relation holds for electrons in the p region. A particular solution of (18.12) is $p = p_{n\circ}$, and the general solution of (18.12) is therefore

$$p(x) = p_{n\circ} + A_- \mathrm{e}^{-x/L_p} + A_+ \mathrm{e}^{x/L_p}, \tag{18.13}$$

where

$$L_p = \sqrt{D_p \tau_p}$$

is called the *minority-carrier diffusion length*. According to (12.11) of Chap. 12, this is the average distance covered by minority holes before recombination.

At this point, to get the solution for our specific problem, we need two boundary conditions which will determine the two constants A_- and A_+. At large x the hole concentration is that of equilibrium in the n region, $p_{n\circ}$, since at a far distant position all excess holes have recombined. This means that the constant A_+ in (18.13) must be zero. For the other boundary condition, let us consider that the concentration of the holes in $x = w_n$ is given by those that have passed the potential step V_b, given by

$$p(w_n) = p_{p\circ} \mathrm{e}^{-e(V_\mathrm{b} - V)/KT} = p_{n\circ} \mathrm{e}^{eV/KT},$$

where $p_{p\circ}$ is the hole concentration in equilibrium in the p region, and $p_{n\circ} = p_{p\circ} \mathrm{e}^{-eV_\mathrm{b}/K_\mathrm{B}T}$ since, at equilibrium, the holes in the n region are those that surmount a potential step V_b. From (18.13), this condition requires

$$p(w_n) = p_{n\circ} \mathrm{e}^{eV/KT} = p_{n\circ} + A_- \mathrm{e}^{-w_n/L_p},$$

or

$$A = p_{n\circ} \left(\mathrm{e}^{eV/KT} - 1 \right) \mathrm{e}^{w_n/L_p}.$$

If replaced in the general solution, this yields

$$p(x) = p_{n\circ} + p_{n\circ} \left(\mathrm{e}^{eV/KT} - 1 \right) \mathrm{e}^{w_n/L_p} \mathrm{e}^{-x/L_p},$$

or, for x in the bulk n region,

$$\Delta p_n(x) = p(x) - p_{n\circ} = p_{n\circ} \left(\mathrm{e}^{eV/KT} - 1 \right) \mathrm{e}^{-(x - w_n)/L_p}.$$

From the concentration as a function of position, we may calculate the diffusion current:

$$j_p(x) = -eD_p \frac{\mathrm{d}p}{\mathrm{d}x} = eD_p p_{n\circ} \left(\mathrm{e}^{eV/KT} - 1\right) \mathrm{e}^{-(x-w_n)/L_p} \frac{1}{L_p}.$$

In $x = w_n$, this is

$$j_p(w_n) = eD_p \frac{p_{n\circ}}{L_p} \left(\mathrm{e}^{eV/KT} - 1\right). \tag{18.14}$$

In the same way, the electron current due to diffusion in $x = -w_p$ at the opposite edge of the depletion region is found to be

$$j_n(-w_p) = eD_n \frac{n_{p\circ}}{L_n} \left(\mathrm{e}^{eV/KT} - 1\right). \tag{18.15}$$

As indicated above, the sum of the contributions in (18.14) and (18.15) yields the total current density through the diode:

$$\boxed{j = j_s \left(\mathrm{e}^{eV/KT} - 1\right), \quad j_s = eD_p \tfrac{p_{n\circ}}{L_p} + eD_n \tfrac{n_{p\circ}}{L_n}} \tag{18.16}$$

shown in Fig. 18.5. j_s is called the *reverse saturation current*. The derivation of (18.16) has been performed having in mind a forward bias, but it holds also for a reverse bias with a negative sign of V. It is called the *ideal diode equation* or *ideal rectifier equation*, also referred to as *Shockley equation*. The two contributions, of electrons and holes, may be very different if one of the two sides of the junction has a doping concentration much larger than the other one. It may be instructive to note that the exponential dependence of the current upon the forward bias is due to the fact that the latter lowers the voltage drop across the depleted region linearly, and the distribution function depends exponentially upon energy.

The ideal diode equation is the result of a series of approximations, besides the use of semiclassical dynamics, that must be remembered: abrupt metallurgical junction, complete-depletion approximation, one-dimensional quantities

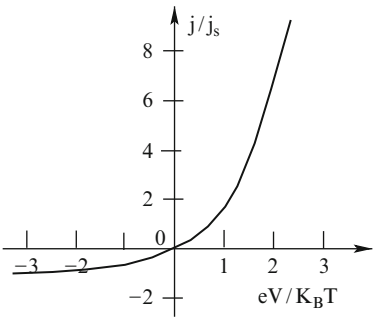


Fig. 18.5. Current-voltage characteristics of an ideal pn diode

(planar diode), negligible electric field outside the space-charge region, carrier densities at the boundaries in quasi equilibrium, low injection conditions, no generation-recombination current in the space-charge region.

In spite of such approximations, necessary for an analytical development of the theory, the experimental characteristics (I–V curves) of real pn diodes result to be in good agreement with the theoretical prediction of (18.16), as long as the geometric features of the devices justify the use of semiclassical dynamics.

A more complete analysis of the pn junction would require the study of many other aspects of the device, even in its ideal version, such as the transient behavior, switching times, capacitance, impedance to alternating signals. We are not going to discuss these points. Our purpose, here, was simply to show how the peculiar feature of semiconductors of having charge transport provided by carriers of opposite signs may lead to systems with very special electrical properties.

The elementary system of the pn junction is the fundamental block of a very large number of more complex electronic devices, from bipolar transistors, to light-emitting diodes, photovoltaic cells, etc. Excellent textbooks exist on this subject, such as [325, 431]. We shall limit ourselves, in what follows, to simplified qualitative descriptions of the main electronic devices.

18.3.3 Solar Cells

Solar cells, or *photovoltaic cells* were first developed in 1954 using Si pn junctions [101]. Subsequently, many other semiconductors have been employed both crystalline and amorphous, and even polymers, looking for the best trade-off between efficiency and cost. Given their enormous practical importance, a great amount of research is being devoted to their analysis, and entire scientific journals and, of course, many books are available in the literature (see, for example [324]).

When a pn junction is exposed to light, photons with energy $h\nu$ larger than the energy gap ϵ_g are absorbed, creating electron-hole pairs. The generated carriers diffuse, and electrons in the p region and holes in the n region that reach the space-charge region are accelerated by the electric field and contribute to the current, together with those generated directly in the space-charge region. The resulting current is a negative addition to the "dark" current of the diode, as shown in Fig. 18.6. The voltage at the contacts of the cell depends on the external load, and when the situation is that shown in the figure, with positive voltage and negative current, the absorbed power is negative, which means that we may obtain electrical power to charge a battery.

The side of the device exposed to the light must of course stop as little as possible of the impinging radiation. For this purpose, contacts made with very narrow stripes are processed as well as anti-reflection coating.

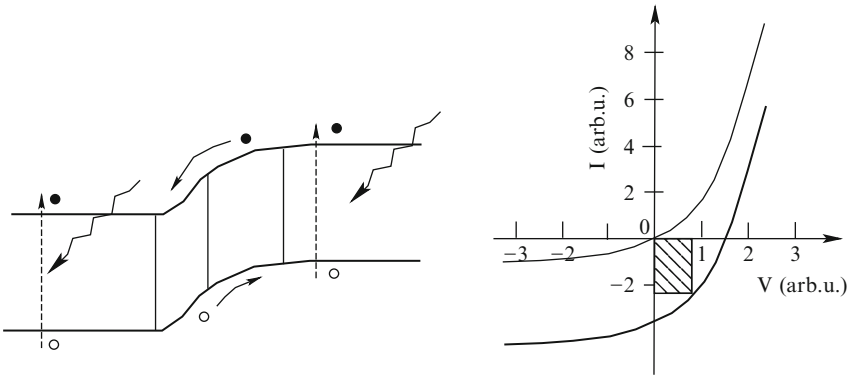


Fig. 18.6. A photovoltaic cell or solar cell is formed by a pn junction where a reverse current is produced by electron-hole pairs generated by photon absorption, separated by the built-in potential and collected at the contacts

The efficiency of solar cells, defined as the ratio between the electrical power obtained by the cell over the power absorbed by the cell, is in general rather low, around 20–25%. This means that solar cells are particularly useful where it would be difficult or too expensive to carry electric power obtained from other sources, as in space satellites and in isolated structures. Photovoltaic plants are becoming more and more popular, however, also for environmental reasons since they are particularly "clean".

In this respect, however, it is important to consider not only the energy that solar cells may generate without pollution, but also the energy that must be employed for their production, the so-called *energy pay-back*, and the impact on the environment due to their production, transportation, installation, and, finally, their disposal at the end of their lifecycles. It is today estimated [31] that the pay-back times, in terms of CO_2 equivalent emission and embodied energy, is about 3–4 years, compared to a life time around 15–30 years.

"Since all continents have sufficiently large areas covered by high average insolation, extensive worldwide utilization of solar energy can be expected in the future" [431]. Obviously, not only photovoltaic.

18.3.4 Light-Emitting Diodes

Solar cells convert light energy into electric energy. The opposite transformation is realized by the *light-emitting diodes* (LED) and semiconductor lasers. As for the latter, they will be treated briefly in Chap. 19, since they are realized mainly with quantum wells. LEDs, on the contrary, are special applications of the pn junctions treated in this section. More generally, electroluminescence is the generation of light as consequence of the application of an electric field to a material, excluding, of course, incandescence produced by Joule heating.

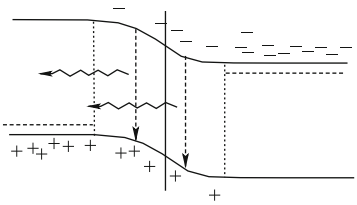


Fig. 18.7. A light-emitting diode, or LED, is formed by a forward biased pn junction where electron-hole radiative recombination emit photons with energy approximately equal to the semiconductor band gap

LEDs had been invented long ago (in the 1920s by Losev, in Russia), but were extensively developed only much later, in the 1960s. They are formed by forward-biased pn junctions designed in such a way as to maximize electron-hole radiative recombination processes, as shown in Fig. 18.7.

For reasons related to statistics and density of states, radiative recombinations occur mainly between electrons near the bottom of the conduction band and holes near the top of the valence band. The energy of the emitted photon is then approximately equal to the bandgap of the semiconductor in direct bandgap materials.

We saw in Chap. 7 that optical band-to-band transitions must be vertical, i.e., without change of electron momentum. In fact the total crystal momentum is conserved in the process; the momentum of the generated photon is (almost) zero, and the momenta of the electron and the hole which recombine are (almost) opposite. For such a reason, to have radiative direct transitions, it is necessary to have a direct-gap material, such as GaAs. In Si and Ge band-to-band transitions with an energy involved of the order of the gap must also involve a phonon that provides the necessary momentum transfer, and the process is much less efficient.

Radiative recombinations are competing in LEDs with nonradiative recombinations. In the latter, the energy released by the recombination is absorbed with some mechanism that does not involve photons, as in the Auger process (see, for example [372]), in which the extra energy is taken from another carrier. Auger recombination is in fact the reverse process of the impact ionization. Another nonradiative recombination is the Shockley–Read–Hall process (see, for example, [318]), in which the recombination occurs in impurity centers.

GaAs has a direct-gap with energy in the infrared range, while GaP has a gap well inside the visible range, but its gap is indirect. With alloys $GaAs_{1-x}P_x$ the gap may be increased above that of GaAs. For $x > 0.45$, however, the gap becomes indirect, since the bottom of X valleys decreases below that of the central valley. To increase the photon frequency without losing too much efficiency, it is possible to introduce radiative recombination centers, such as nitrogen in $GaAs_{1-x}P_x$ [431] which allow indirect-gap semiconductors to be good materials for LEDs.

GaAsP alloys use GaAs as substrate for direct band materials (red LEDs) and GaP substrate for indirect-gap materials (orange, yellow, and green LEDs). Many other materials are today used to produce LEDs with various colors, from infrared to ultraviolet.

White light can be obtained with the combination of LEDs emitting radiations of different frequencies (red, green, and blue). More common, today, is to obtain white light by using a bright blue or UV LED and convert its monochromatic light into a broad-spectrum light by means of a phosphor material.

After the photon emission by pair recombination, several phenomena may occur which degrade the efficiency of an LED, such absorption inside the device and partial or total reflections. Many technological expedients have been devised to overcome these problems.

Initially, LEDs were used mainly in digital displays. Since then, their use has steadily increased, and today they are replacing many of the more traditional light sources. It is conceivable that LEDs may soon replace light bulbs in domestic illumination.

18.4 The Bipolar Junction Transistor

The *bipolar junction transistor* (BJT) is an active device capable of current gain. It is formed by three parts made of semiconductors with alternately-different majority carriers, as shown in Fig. 18.8. There are, therefore, BJTs of pnp and npn types. Usually, the latter is preferred for technological reasons (higher gain and faster switching). A BJT consists of two successive pn junctions. All three regions have metallic contacts that keep them at controlled potential or current values. The three parts of a BJT are named *emitter*, *base* (the central part), and *collector*. Charge carriers can enter or exit from the various contacts, depending on the function performed by the transistor. The potentials and currents are named V_E, V_B, V_C, I_E, I_B, I_C, respectively, while the potential differences are named V_{EB}, V_{EC}, and V_{BC}.

Each of the two junctions can be forward or reverse biased. Thus, four possible operation regimes exist:

1. The first junction, E–B, is forward biased, and the second, B–C, is reverse biased. It is the *active region* of operation, used when the transistor is employed for amplification purposes.

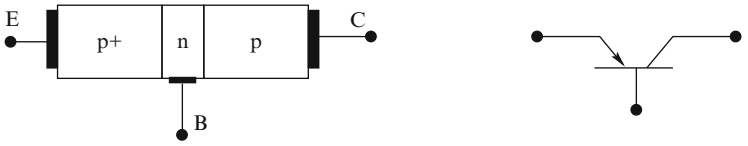


Fig. 18.8. Bipolar pnp junction transistor and its symbol in circuit design

2. Both junctions are forward biased. It is the *saturation region* of operation. It is used in logic circuits as "on", highly conductive, state since V_{CE} is small and I_C is large.
3. Both junctions are reverse biased. It is used as "off" state in logic circuits, with I_C very small even with relatively large V_{CE}.
4. The E–B junction is reverse biased, while the collector junction is forward biased. It is called the *inverted region* of operation.

There are also different modes of operation in the electronic circuits. Since two input and two output connections are necessary, while the BJT has only three terminals, one of these must be used for both input and output. In other words, one of the three terminals must be used as reference for the other two. There are, therefore, three different modes of operation used for different functions: common emitter, common base, and common collector. The output characteristics are different in the various configurations. Generally speaking, however, BJTs are now used only for particular applications as, for example, in ECL (Emitter-Coupled Logic).

The functioning of the BJT is based on the same principles seen in the description of the junction in the previous section. There are, however, important differences due to the different boundary conditions, the presence of several currents, and the interaction between the two junctions.

Let us briefly consider the case of a pnp BJT in the active mode of operation. The emitter region is doped with a carrier concentration much higher than that of the base, and the latter, in turn, is more doped than the collector. Furthermore, the base thickness must be much smaller than the minority-carrier diffusion length. Thus, almost all the holes that diffuse into the base through the first direct-biased junction, cross the base and reach the collector. This is the reason for the names "emitter" and "collector". A first important parameter of the BJT is the ratio between the hole current, which is emitted into the base and that reaching the collector, or

$$\alpha = I_C/I_E.$$

In good narrow-base devices, this ratio is close to unity.

When a potential difference is applied to the pnp transistor, between collector and emitter, which forward biases the E–B junction, holes injected from the emitter reach the base after surmounting a potential step which is reduced with respect to the equilibrium value. They then diffuse through the narrow base; most of them reach the B-C interface and are collected by the collector. They constitute the main part of both emitter and collector currents. The base current is in this case vary small.

If, in a common-emitter configuration, a base current is injected, it reduces the E–B potential step so that a large current from the emitter is generated. Owing to the narrow base, this results in a large collector current. The net result is a strong amplification from the base current to the collector current.

The second important parameter that characterizes the BJT is, in fact, the current gain, defined as

$$\beta = I_C/I_B.$$

In good BJTs, it reaches values of the order of 100.

18.5 Metal–Semiconductor Junctions, Schottky Barrier Diode

It was already mentioned in the previous sections that the electrical connections of semiconductor devices with the external circuits are realized with metallic contacts. Metal–semiconductor junctions are therefore of fundamental importance. They can be studied with the same physical principles used in the pn junctions, taking into account that the density of ions (corresponding to ionized donors in n-type semiconductors) is extremely high in metals. Several cases must be considered depending on the position of the Fermi level in the metal with respect to its position in the semiconductor. Figure 18.9 illustrates schematically the different possibilities of interfaces of a metal with an n-type semiconductor, before and after contact, for the different relative positions of the Fermi levels.

When the two materials are separate, the electrostatic energy of an external electron near their surfaces is the same in proximity of the two materials. This is the *vacuum energy level*. The difference between this energy and the Fermi level in the metal ϵ_{Fm} is the *work function* of the metal and is the energy necessary to extract an electron from the metal. Similarly, the work function of a semiconductor is the energy difference between the vacuum level and the Fermi level in the semiconductor. The difference between the vacuum level and the bottom of the conduction band of a semiconductor ϵ_c is the *electron affinity* of the semiconductor.

When the two materials are put in contact, because of the different Fermi levels a charge transfer occurs which generates an electrostatic counter-field. This, at equilibrium, prevents further charge transfer, and the Fermi level is constant throughout the entire structure, as shown in Fig. 18.9. Owing to the large charge density in the space-charge region in the metal, the curvature of the parabolic potential profile is much higher than in the semiconductor, and the built-in potential drops almost entirely in the semiconductor.

In part (a) of the figure, the Fermi level in the metal is higher than that of the separate semiconductor, at a level inside the conduction band. When the two materials are put in contact, on the metal side we have a very thin layer of positive ions left behind by the electrons that moved into the semiconductor. On the semiconductor side, we have an accumulation of electrons near the interface that form the *electron accumulation layer*. The double layer generates a potential difference equal to the difference between the Fermi levels of the isolated materials, as in the pn junction. In such a situation, electrons can

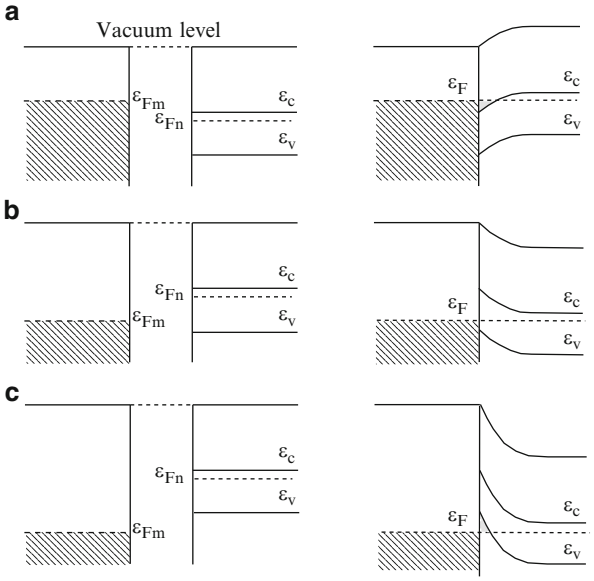


Fig. 18.9. Interface between a metal and an n-type semiconductors before (*left*) and after (*right*) contact, at equilibrium. When the materials are separate, the metal Fermi level may be at the level of the semiconductor conduction band (**a**), of the semiconductor gap (**b**), or of the semiconductor valence band (**c**). The triangular gray areas in the semiconductor correspond to an electron accumulation layer in (**a**) and to a hole inversion layer in (**c**)

easily flow from one part of the interface to the other and the interface is said to form an *ohmic contact*. In general, a contact may be defined "ohmic" when charge carriers can flow across it in both directions without having to pass a nonnegligible barrier. The current-voltage characteristic $I(V)$ in the origin is thus linear and symmetric.

In part (b) of Fig. 18.9, the Fermi level of the isolated metal is below that of the semiconductor, within the range of the energy gap. When the two materials are put in contact, electrons leave the semiconductor forming a *depletion region* in front of the interface. A potential barrier, called *Schottky barrier*, is formed. Electrons can pass the barrier by thermal excitation as in the thermionic effect, or by tunneling. At equilibrium, the current $J_\circ$ in one direction exactly cancels the current in the opposite direction.

If a potential difference is applied that rises the Fermi level of the semiconductor, the barrier which must be overcome by the electrons in the semiconductor to reach the metal decreases, while the barrier to be overcome in the opposite direction is not changed, as shown in Fig. 18.10b. The current in one direction increases exponentially while the current $J_\circ$ in the opposite direction is unaltered. If the voltage is applied in the opposite direction the electron flux from the semiconductor decreases exponentially and the small

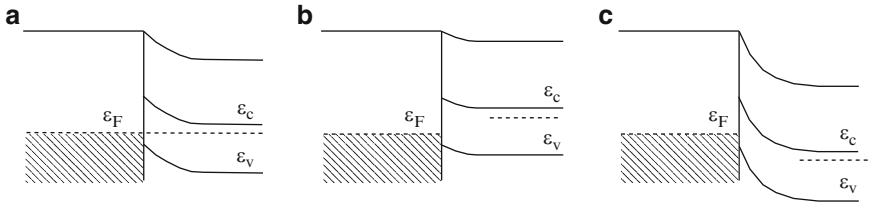


Fig. 18.10. Schottky interface in equilibrium (**a**), forward biased (**b**), and reverse biased (**c**)

current $J_\circ$ from the metal to the semiconductor remains the same. The system works as a rectifying diode, called *Schottky-barrier diode*. The net current as a function of the applied voltage is given by [325]

$$J = J_\circ \left(\mathrm{e}^{eV/K_\mathrm{B}T} - 1 \right).$$

In part (c) of Fig. 18.9, the Fermi level of the isolated metal is below the top of the valence band of the isolated n-type semiconductor. When the two materials are put in contact, the space-charge layer in the semiconductor is formed not only by the unbalanced ionized donors left behind by electrons that cross the interface to reach the metal, but also by some holes formed by electrons which leave the valence band. A *hole inversion layer* is formed at the interface.

18.6 Field-Effect Transistors

Field-effect transistors (FET) are based on the idea of controlling the conductance of a device between two contacts called *source* and *drain*, by means of an electric field obtained from the application of a potential difference at a third electrode, called *gate*. The idea is very old: FETs were proposed in patents by Lilienfeld in 1925 and by Heil in 1935. For technological reasons, however, they were not developed until after the BJT, and today they are by far the most used transistors in industrial microelectronics. There are three main families of FETs, called *junction field-effect transistor* (JFET), *metal-semiconductor-field-effect-transistor* (MESFET), and *metal-oxide-semiconductor-field-effect-transistor* (MOSFET). Each of these families contains several members with different names. In [431], all of them are classified and described with the necessary technological and functional details. Once again, we can give only a brief account of their structures and functioning principles.

JFET

The *junction field-effect transistor* (JFET) consists in a semiconductor slice, say of n type, sandwiched between two reverse-biased junctions formed by

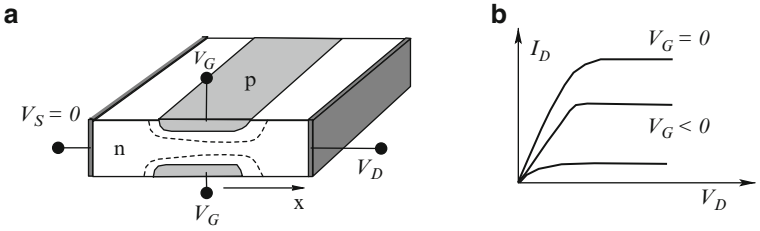


Fig. 18.11. JFET structure and characteristics

heavily-doped p-type material, as shown in Fig. 18.11. Here, V_S, V_D, and V_G indicate the *source voltage*, the *drain voltage*, and *gate voltage*, respectively.

Since the p-type semiconductor is much more doped than the n-type channel, in the expression (18.9) for the depletion width, N_D can be neglected with respect to N_A, yielding

$$w(x) = \sqrt{\frac{2\varepsilon V_b}{eN_D}(V_b - V_G + V(x))},$$

where the built-in potential has been replaced with the potential difference $V_b - V_G + V(x)$. The value of this potential corresponding to a depletion width equal to half of the channel width is called the *pinch-off* voltage, and is given by $V_P = eN_D a^2/2\varepsilon$, where $2a$ is the width of the device. The internal channel between the two depletion regions has the maximum width when the gate voltage V_G is zero (device normally on) and it is narrowed by negative values of V_G. Furthermore, for any V_G the channel is narrowed by V_D at increasing x, since $V(x)$ increases. Thus, at increasing drain voltages, the current through the device increases, but the channel between the two depletion regions becomes narrower, until pinch-off is reached at the drain side, and the current flowing through the depletion region saturates. In fact, a further increase in the drain voltage would shift the pinch-off position toward the source, thus increasing the length of the pinched-off region of the device. The current is essentially determined by the potential drop in the open part of the channel, which is not increased by an applied drain voltage in excess of the saturation voltage if the total channel is not too short. A detailed intuitive discussion of the saturation current in JFETs can be found in [343]. The $I(V)$ characteristics are qualitatively shown in part (b) of Fig. 18.11.

MESFET

The *Metal-Semiconductor-Field-Effect-Transistor* (MESFET), schematically shown in Fig. 18.12, is based on the same principle of the JFET. The depletion layer controlling the source-drain current is now generated at the interface between a metal contact and a semiconductor, as described in Sect. 18.5. According to the nature of the gate interface, the MESFET may be conductive

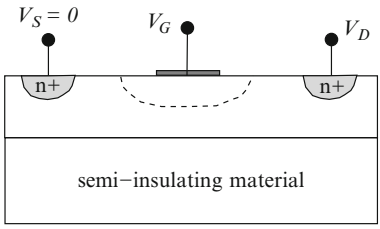


Fig. 18.12. MESFET structure

(*normally-on*) or not conductive (*normally-off*) at $V_G = 0$. In the normally-on MESFET, the depletion region leaves a conductive channel below the gate when $V_G = 0$, and a negative voltage must be applied to the gate to prevent the drain current. In the normally-off MESFET, the built-in potential of the gate junction is sufficient to totally deplete the channel region, and a positive V_G must be applied to obtain drain current.

MESFETs are mainly used for radiofrequency applications. Since electrons in GaAs have higher mobility than in Si, this material is preferred for very-high-speed applications. Furthermore, since it is not possible to grow an oxide layer on top of GaAs, the MESFET structure is preferred to the MOSFET one, described below.

MOSFET

A *Metal-Oxide-Semiconductor-Field-Effect-Transistor* (MOSFET) is schematically shown in Fig. 18.13. In its most common version, it is formed by a Si substrate of p type with a metallic gate (G) separated from the semiconductor by a thin layer of SiO_2 insulator. Source (S) and Drain (D) are formed with n+ Si diffusions inside the p-type material. With a positive voltage applied to the gate, a thin inversion layer of n type is formed at the interface between the p bulk material and the oxide, as shown in part (b) of the figure, in a complementary way with respect to the p channel seen in part (c) of Fig. 18.9 of Sect. 18.5. This inversion layer, called *channel*, is responsible for the conduction between source and drain.

The channel current is controlled by the gate voltage which may draw more or less charge into the channel. If the gate voltage V_G is below a threshold value V_T, the channel is not formed, and the device is in the off state. At increasing values of V_G above V_T, the conductance increases. The characteristics of the MOSFET are shown in part (c) of Fig. 18.13. At increasing drain voltage V_D, a first linear ohmic region is followed by a saturation current. The latter is due to the fact that at increasing V_D, the depletion region becomes wider near the drain end of the channel; when V_D is too large, the gate voltage is no more able to maintain a well-formed channel, and electrons tend to diffuse into the bulk before reaching the drain contact. This fact is equivalent to shortening the channel so that a current saturation occurs with a mechanism very similar to

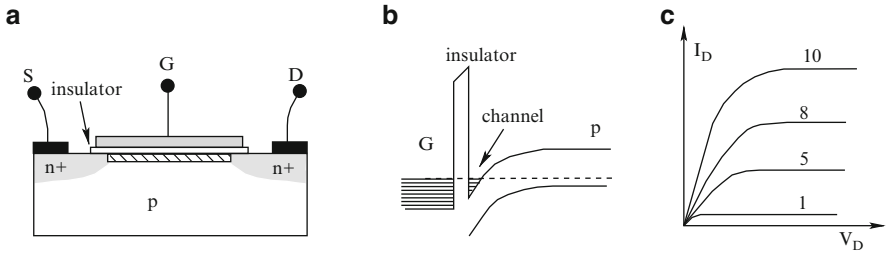


Fig. 18.13. Structure (**a**), energy diagram (**b**), and qualitative drain characteristics (**c**) of a MOSFET. Numbers in (**c**) indicate, in arbitrary units, the effective gate bias $V_G - V_T$

what happens in the MESFETs. Furthermore, in short channels of very small devices, the field reaches values where the electron drift velocity in silicon saturates, as seen in Sect. 15.1.

MOSFETs are used both for amplification and as digital switches. If a high enough V_D is applied, electrons in the channel may gain energy high enough to tunnel through the oxide. With this procedure, floating gates are used for nonvolatile memories whose bit value is defined by the charge captured in the floating gate.

The channel width may be very small, and level quantization in the triangular well shown in part (b) of Fig. 18.13 may occur. Thus, the channel of a MOSFET is a typical example of multi-sub-band two-dimensional electron gas, as discussed in next Chap. 19. The wavefunctions must almost vanish at the silicon-oxide interface. This means that the particle density is small near the interface, thus reducing the strength of surface-roughness scattering. In spite of this, such type of scattering is still the dominant mechanism in reducing the electron mobility in MOSFET channels [170, 479].

The level quantization and the shape of the wavefunctions normally to the interface has another important consequence. The depth of the triangular well, and therefore the energies of the eigenstates, depend upon the gate voltage. These levels are occupied by electrons up to the Fermi level at equilibrium. Out of equilibrium, the energy of the carriers is determined also by the transport phenomenon. However, the shape of the potential well is modified by the presence of the electron charge. Thus, the transport problem must be solved self-consistently with the shape of the well through Poisson and Schrödinger equations. The situation is complicated even more by the fact that the potential energy variation along the channel, as effect of the voltage applied between source and drain, modifies the shape of the triangular potential along the channel as well. As a consequence, the Schrödinger equation is not separable along the direction of the current and that orthogonal to the interface.

It may be concluded that the analysis of the MOSFET is extremely complex. It involves quantum physics, hot-electron effects, and complicated electrostatics geometries. Owing to its technological importance, a paramount research

effort has been devoted in the last decades to such a device, which may now be considered well understood in most of its features. These studies have also indicated the technological changes, which can be attempted to proceed in the performance improvement, miniaturization, reliability, and economy, which characterized the last decades.

Many variants of the traditional MOSFET described above have resulted from this research. Among the most important, let us mention:

1. *High-k devices.* To maintain a good capacitance (enough electrons in the channel) while the device is made smaller, the oxide should be made thinner, to the point that tunnel effect may produce a leakage current which degrades the performance of the MOSFET. Insulating materials with higher dielectric constants are employed to overcome this difficulty.
2. *Multigate and gate-all-around* transistors. To have a better control of the electrostatics of the channel, multiple-gate MOSFETs have been developed. In double-gate transistors, two gates are present on opposite sides of the device. In other designs (FinFETs), a gate embraces three sides of the device, or surrounds it completely.
3. SOI structures. In these devices, silicon is grown on top of an insulator (e.g., sapphire or, more often, quartz) with the purpose of reducing parasitic capacitances in deep submicron devices and thus improving their speed.

Other semiconductor devices will be described in Chap. 19, where low-dimensional structures are studied.

18.7 Device Simulation

In the previous sections, analytical solutions have been presented or simply outlined for the electrical behavior of some semiconductor structures. To obtain analytical solutions, many approximations had to be made, in particular on the geometry of the systems: planar geometries, abrupt surfaces, uniform dopings, and so on. For realistic device modeling, on the contrary, it is necessary to consider systems with much more complicated and detailed features. In such cases, it is possible to obtain accurate solutions only by means of numerical methods, which, thanks to modern computers, may be obtained with high accuracy.

Numerical solution are "exact" only in a limited sense. Often the starting equations are already results of approximate theories: semiclassical dynamics, drift-diffusion or hydrodynamic equations, Boltzmann equation with transition probabilities evaluated with first-order perturbation theory. Other sources of errors may still be present not due to physical approximations, but to numerics: discretization of the equations in finite-difference methods, finite convergence in self-consistent methods, etc.

In the decades of activity of the author of this book, the numerical simulation of electronic devices has gone a long way. In the 1960s, technologists

were smiling, and perhaps only for politeness did not laugh helplessly, when they heard that some theoreticians were trying to predict the behavior of a semiconductor device, even a simple one, based on theoretical equations. The distance between the actual device, obtained with secret recipes, and the idealized system of the theoreticians was too big for any useful comparison. Later it became clear that numerical simulations, although not yet able to give correct I–V curves of a device, were however useful to understand what was happening inside it, and therefore to imagine how to modify the design parameters to obtain the wanted results. Nowadays, no electronic industry would start the design of a new device without a previous, deep analysis of its foreseen functional behavior by means of numerical simulations. Several commercial simulation programs exist and are currently used, with different features more or less useful for different purposes. Furthermore, the simulation of a device cannot ignore its behavior within the entire circuit it is working in.

The reason of this dramatic change is due both to theoreticians, who have been able to insert in their models more details that make their predictions more realistic, and to experimentalists, who have been able to construct "cleaner" devices, geometrically and chemically perfect, much more similar to the idealized systems built by the "simulators" inside their computer programs. Still in a paper of 2003 [1] A. Abramo asks himself: "As far as modeling is concerned, [...] has this trend deeply taken advantage of the corresponding improvement of [...] TCAD tools, or rather, have been the skills of many technology actors, instead, the motor of the revolution, regardless of - or better - marginally dependently on the evolution of predictive simulation abilities?"

Several books have been published on the numerical simulation of electronic devices [187,209,227,405]. Here, we will limit ourselves to introduce the main ideas of the most common methods: drift diffusion, hydrodynamic, and Monte Carlo. These are all based on semiclassical dynamics. Quantum simulators have also been developed [243], based on the approach of non-equilibrium Green functions, presented in the last part of this book.

18.7.1 Drift-Diffusion Models

Any device simulation begins with the definition of the geometry of the system, the different regions of interest, the doping profile and the contact regions.

The drift-diffusion approach to the simulation of electronic devices is based on the numerical solution of the first two moment equations of the BTE, discussed in Sect. 10.4, coupled to Poisson equation.

In general terms, we need to know the following quantities in each point of the device:

$$n(\boldsymbol{r},t)\ ,\quad p(\boldsymbol{r},t)\ ,\quad V(\boldsymbol{r},t)\ ,\quad \boldsymbol{j}_n(\boldsymbol{r},t)\ ,\quad \boldsymbol{j}_p(\boldsymbol{r},t), \tag{18.17}$$

which indicate the electron and hole concentrations, the electric potential, and the electron and hole current densities, respectively. The time dependence

is relevant when transient and/or a.c. operation are of interest. Boundary conditions must be taken into account as indicated at the beginning of the chapter.

In (18.17), it will be sufficient to obtain the first three quantities since currents can then be obtained by the drift-diffusion equations, which give the name to the method:

$$\boldsymbol{j}_n = -n(-e)\mu_n \nabla V - (-e)D_n \nabla n \quad , \qquad \boldsymbol{j}_p = -pe\mu_p \nabla V - eD_p \nabla p. \tag{18.18}$$

Here, the current density is the electrical current, including the charge, $e > 0$, and μ has the sign of the carrier charge. The differential equations which govern the quantities above are the continuity equations for the two types of carriers, electrons and holes, to which we add the generation and recombination terms,

$$\frac{\partial n}{\partial t} = -\frac{1}{(-e)} \nabla \cdot \boldsymbol{j}_n + G_n - R_n \quad , \qquad \frac{\partial p}{\partial t} = -\frac{1}{e} \nabla \cdot \boldsymbol{j}_p + G_p - R_p, \tag{18.19}$$

and Poisson equation, given in (18.1) and (18.2). If (18.18) are inserted into (18.19), together with Poisson equation they become three differential equations for the three unknown functions n, p e V. In such equations products of unknown functions appear: in the expression for the currents the concentrations are multiplied by the gradient of the potential. This means that the equations are nonlinear, and a self-consistent solution must be found: the potential V depends upon the concentrations, which depend on the currents which, in turn, depend upon the potential. In practical terms, this self-consistency is sought, in general, with an iterative approach. A charge distribution is first guessed by means of some simple physical considerations; from this, with the solution of Poisson equation, the electric field is calculated as a function of the position in the device; with this tentative field, the currents are evaluated with the drift-diffusion equations (18.18); the obtained currents are inserted into the continuity equations (18.19) and the same equations are solved to obtain $n(\boldsymbol{r})$ and $p(\boldsymbol{r})$. The resulting n and p will, in general, be different from the ones guessed at the beginning, so that the whole procedure is repeated until the concentrations obtained at the end of the cycle are "equal" to those obtained with the previous iteration.

Care must be put in establishing when convergence is attained, since at times the convergence is slow and it is possible to confuse a small variation between two successive cycles due to slow convergence, with the small variation as result of the reached precision. More generally, the decision to stop the iteration procedure must be taken on the basis of a convenient convergence criterion or convergence check. See, e.g., [346], or good books of numerical analysis.

Nonlinear effects, with respect to the applied field, can be taken into account in an approximate way, by replacing the mobilities and diffusivities that appear in (18.18) with values obtained by nonlinear bulk simulations, as

described in Chap. 14. This approximation, however, may not be satisfactory because of nonlocality effects: when the space variations of the field are fast, the mean electron velocity and energy at one point may be due to values of the electric fields at different points, so that mobility and diffusivity may be nonlocal quantities.

Solution of the Differential Equations

In the above description, we have used the expression "the equations are solved..." as if it were an obvious task. In reality, the numerical solution of differential equations is a whole branch of the important discipline known as *numerical analysis*. To the knowledge of this discipline, the ability to write computer programs must be added. This, also, has become a sophisticated discipline, with parallel and distributed programming, shared memory, and so on. Today, computer programming is assisted by a large amount of very powerful commercial software.

The numerical solution provides the values of the unknown functions in a given set of points which form a discretization of the device space, called *grid*, which must be set at the beginning of the calculation. The points defining the grid are called *nodes*; the poligons whose vertices are the nodes are called *elements*. Several types of grids can be used [209, 397]. The simplest grid is obtained with rectangular spacing and constant steps, as shown in Fig. 18.14. This type of grid is, however, inefficient since it devotes the same amount of computation whether the functions have fast or slow variations. If the constant grid step is chosen taking into account the need of the regions where the functions have the fastest variations, a large number of points will be required also in the regions where they are constant or almost constant. A first improvement can be obtained considering a rectangular grid with variable spacing, as shown in the figure. Also triangular grids can be considered. A *multigrid algorithm* can also be implemented to refine the solution with an iterative method.

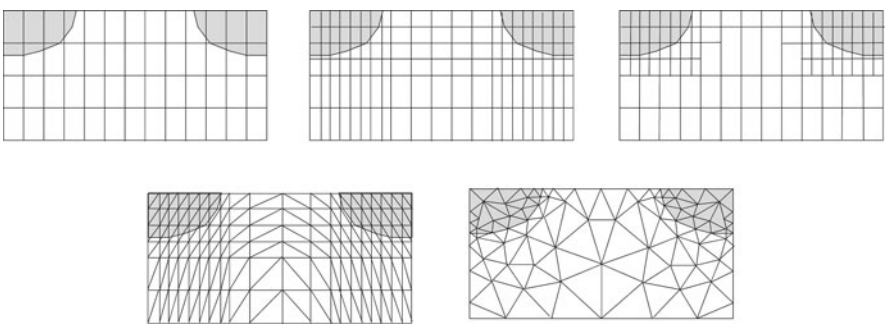


Fig. 18.14. Different types of grids for numerical device simulations

The derivatives of the functions that appear in the differential equations to be solved must be replaced by some discrete approximations. This can be obtained with the *finite-difference* method where the derivative is replaced by the difference quotient. The method is particularly simple when rectangular grids are used.

The finite-difference method, however, is not the only possible one. Often the *finite-element* method is used, where the unknown functions are given approximate analytical expressions, e.g., linear, over each element. The solution of the equations is then obtained through the determination of the parameters appearing in this analytical approximation.

18.7.2 Hydrodynamic Models

The drift-diffusion method described in the previous section is based on the equations obtained with the first two moments of the BTE, discussed in Sect. 10.4: the zero-order moment, which yields the conservation of the number of particles, and the first-order moment, which yields the current continuity and provide the drift-diffusion equation. In the *hydrodynamic model* more moments are used: the second-order moment, which describes the continuity of energy, and the third-order moment, which describes the continuity of energy flux.

There is no need to report here the explicit equations, which can be found in many papers and books such as, for example, in [153, 227, 291, 386].

Since the basic equations of the hydrodynamic approach have the same structure as the simpler drift-diffusion model, it is possible to incorporate the hydrodynamic equations into existing device-analysis codes, thus exploiting a number of robust solution schemes previously developed and tested.

The number of equations is larger with respect to the drift-diffusion scheme. Thus, the computational cost of the solution is higher. However, this higher cost is largely compensated by a better accuracy. In contrast with the drift-diffusion model, where the carrier temperature is kept at the same value as the lattice temperature, the hydrodynamic model provides the carrier temperature as a function of space and time, as part of the result of the analysis. In this way, a good part of the problems due to hot-carrier effects mentioned above are accounted for by hydrodynamic approaches. Also nonlocality of transport may be at least partially described by hydrodynamic models.

In any method based on moment equations of the BTE, a number of coefficients are needed, which must be obtained independently. In the drift-diffusion model, they are the mobilities and the diffusivities of the carriers. In the hydrodynamic models new parameters have to be added, such as the energy and energy-flux relaxation times. These parameters can be obtained either experimentally or by a full solution of the BTE, for example with bulk Monte Carlo simulations. It is also possible, by means of such parameters, to account for particular band structures, for example using tensor relaxation times.

18.7.3 Monte Carlo Simulations

The approaches to device simulation seen above make use of equations obtained from the BTE without an explicit evaluation of the carrier distribution function. They provide a good understanding of the device functioning but cannot describe with sufficient accuracy a number of phenomena that depend critically on such a distribution. Typically, effects occurring above a threshold energy, as impact ionization or oxide penetration, require a more detailed analysis of electron transport. This analysis is provided, still within the semiclassical approximation, by Monte Carlo (MC) simulations [187,209,227]. To account for such phenomena, full-band MC simulations (see Sect. 14.2.8) are, in general, necessary.

MC simulations require a high computational cost and are often performed for a deeper analysis of device designs already chosen on other considerations.[5]

In MC device simulations carrier trajectories are stochastically generated as described in Chap. 14. Since, however, the field profile is strongly dependent upon the position within the device, very short flight times must be generated, at the end of which it is stochastically decided whether a scattering event takes place. In the affirmative case, the scattering mechanism and the after-scattering state are determined as in the standard MC procedure.

Even though for steady-state analyses, it is possible to simulate single particles from their entrances into the device to their exits, ensemble MC simulations are more often used [272, 273], in which all considered carriers are simulated in parallel. In this way, it is possible to analyze transients and take into account, when necessary, particle–particle interactions, including the exclusion principle, as described in Chap. 14.

As it regards boundary conditions for the particle flow, particles can exit the simulation region at the drain or source contacts. To maintain neutrality conditions, at each time-step, the necessary number of electrons is replaced. The injection takes place in a uniform way along the cells adjacent to the contacts. When particles hit other surfaces of the simulated region, reflecting or periodic conditions may be applied, as described in [209]. To account for surface roughness, reflection at the boundaries of the device may be considered partially diffusive (see Sect. 19.2.2).

In ensemble MC simulations, when the number of particles in the device is too large, it is not possible to include all of them in the simulation. A smaller number of representative *superparticles* are then used. This, however, creates problems in the simulation of short-range electron–electron scattering [148].

[5] As already mentioned in Chap. 14, attempts have been made to introduce techniques that maintain the main positive features of MC simulation while keeping the amount of necessary computer time reasonable [226,250,395]. Another interesting possibility is to identify regions of the device where MC simulations are necessary, while in the rest of the device a faster simulation method is applied. The boundaries between the different regions must be treated with particular care.

The simulation of semiconductor devices with MC techniques requires, as in the other methods, the solution of the Poisson equation to obtain the spatial distributions of potential and electric field. For this purpose, a grid must again be properly defined. MC simulation, however, is "granular" in nature; on the contrary, the potential V and the electric field $\boldsymbol{E}$ are represented by values on the grid points. It is then necessary to map discrete values defined at the grid points onto the corpuscular distribution of carriers, and viceversa. This is done with the so-called "Particle-Mesh Method". The basic algorithm consists of the following steps: assign the charge to the mesh points; solve Poisson equation on the grid; compute the components of the electric field from the potential defined in the grid points; interpolate the field to the particle positions.

As it regards the Poisson equation, care must be put in the separation between short-range carrier–carrier interaction, to be treated as collisions, and long-range Coulomb interaction dealt with by the Poisson equation [2].

Finally, it is important to note that charge density fluctuations are damped by the self-consistent field and may give rise to plasma oscillations. It is thus necessary to solve Poisson equation, in the simulation, at very short time intervals, to avoid that a charge fluctuation acts for an unphysical time, producing unphysical instabilities. For concentrations of the order of $10^{20}\,\mathrm{cm}^{-3}$ in silicon, as can be found in the drain region of a MOSFET, a time-step for the application of the Poisson solver well below $10^{-15}\,\mathrm{s}$ must be used, since the plasma period is of this order of magnitude [148].

As a final important note, let us consider that in modern mesoscopic devices self-averaging, as defined at the beginning of this chapter, is not realized. For example, the number of impurities present in a device may be small, and their specific positions may influence the behavior of any real device in a particular way. This fact produces a large variance in the device characteristics from sample to sample, as discussed, for example, in [17, 327, 488].

19

Low-Dimensional Structures

In the years 1970s, the physics of semiconductors went through a profound revolution with the introduction of low-dimensional structures. The standard reference is that of Esaki and Tsu of 1970 [132], who proposed the fabrication of superlattices to realize negative-differential-conductivity devices and Bloch oscillations. Their suggestion of manufacturing new systems by means of epitaxial growth of heterostructures opened up a new field of research that produces new and exciting results still today. From the theoretical point of view, the backbone of the literature has been the review paper by Ando et al. [12].

In what follows the envelope-function theory in the effective-mass approximation will be used for the determination of the electronic states in semiconductor structures. For the analysis of electron-transport properties, this method is often considered acceptable. For other purposes, in particular related to optoelectronics, other continuous or atomistic methods may be more convenient (see, for example, [62, 158, 359, 422, 465]).

19.1 Epitaxial Heterostructures

The epitaxial growth of a material consists in the deposition of atoms which continue the crystal structure of a substrate. The substrate must of course have the appropriate crystal symmetry and a lattice constant very close to that of the growing layer (*epilayer*). The most common techniques of epitaxial growth are the *molecular beam epitaxy* (MBE), see for example [449], and the *metal-organic chemical vapor deposition* (MOCVD), see for example [428]. The epilayer and the substrate constitute, therefore, a unique crystal structure formed with two distinct substances, as shown in part (a) of Fig. 19.1.

In Fig. 19.2, several cubic compound semiconductor materials are indicated with their band gaps and lattice constants. It can be seen, as the most important example, that GaAs and AlAs have practically the same lattice constant (a variation of less than 0.15%) and different band gaps. If AlAs is grown epitaxially on top of a GaAs crystal, a single crystal structure is obtained with

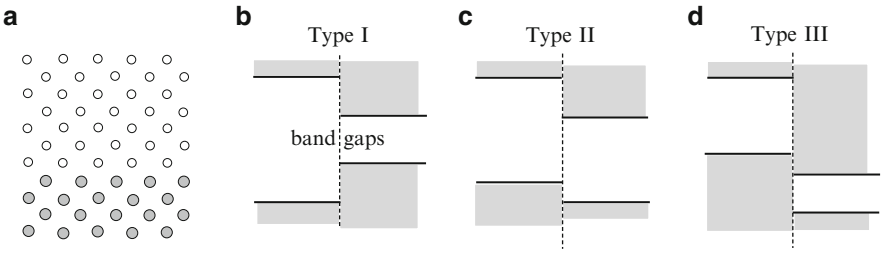


Fig. 19.1. (**a**) The epilayer and the substrate constitute a unique crystal structure formed with two distinct substances. In type-I heterostructures (**b**), the band gap of one material is contained inside the band gap of the other material; if the two band gaps overlap partially (**c**), the heterostructure is of type II; if the two gaps do not overlap (**d**), the heterostructure is of type III

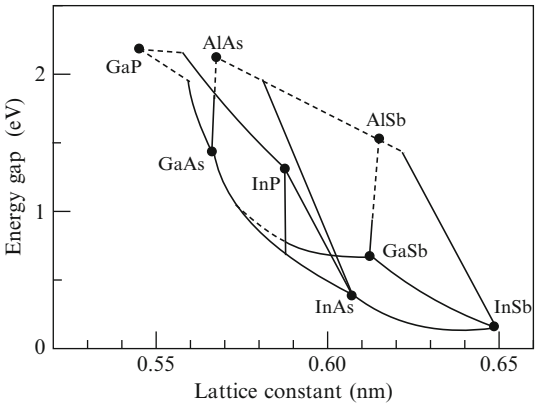


Fig. 19.2. Lattice constants and band gaps of some cubic semiconductors. *Full lines* indicate direct-gap materials; *dashed lines* indicate indirect-gap materials. (Adapted from [32])

the band gap that changes abruptly from the value of GaAs (1.42 eV) to that of AlAs (2.16 eV).

By growing epitaxially a semiconductor material on top of a crystal of a different semiconductor, we obtain a semiconductor *heterostructure*. The relative positions of the energy gaps of the two materials constitute the *band offset* of the heterojunction. This depends not only on the properties of the two bulk materials, but also on the microscopic properties of the interface [22]. Figure 19.1 shows the three possible types of heterostructures with respect to the band offset.

Besides the band offset, there will also be some band bending to line up, at equilibrium, the Fermi levels of the two substances, as shown in Fig. 19.3. As in the case of the junctions discussed in Chap. 18, free charges move from one side of the structure to the other, where they find available states of lower

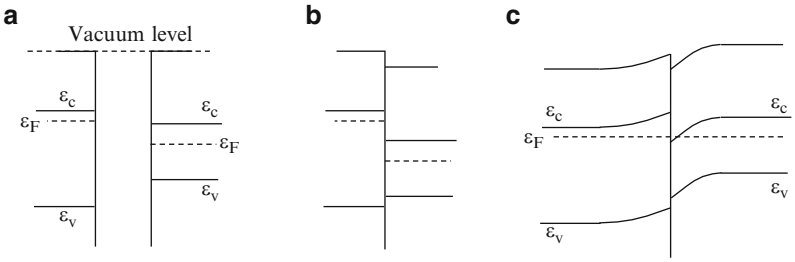


Fig. 19.3. Separate different semiconductor materials have the same vacuum energy level (**a**); if they are put in contact, or one is grown epitaxially on top of the other, the band offset is set up (**b**); free carriers move from one part of the heterostructure to the other until the Fermi level is constant throughout the system (**c**)

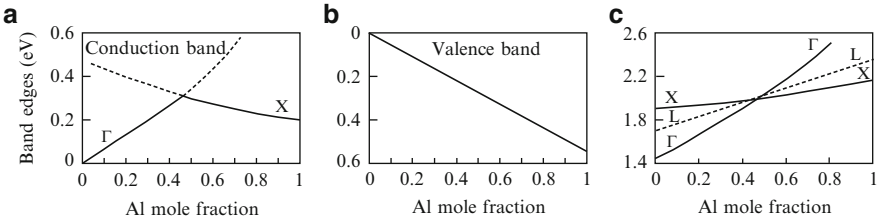


Fig. 19.4. (**a**) and (**b**) Energy alignment of the conduction and valence bands, respectively, of an $Al_xGa_{1-x}As$ alloy, in contact with GaAs. (**c**) Conduction-band edges in the alloy $Al_xGa_{1-x}As$ with respect to the top of the valence band [236]

energies, until the electrostatic field prevents further transfer, and the Fermi level is constant throughout the system. The curvatures of the band bending are again determined by the density of the fixed charges in the space-charge regions.

Alloys

We have just seen that GaAs and AlAs have the same lattice constant and different energy gaps. These two materials form a solid solution over the entire composition range. The band structure of the alloy is intermediate between the band structures of the two components, as shown in Fig. 19.4. The minimum at Γ of the conduction band of $Al_xGa_{1-x}As$ increases as the fraction x of Al increases from zero. On the contrary, the minimum at X of the conduction band decreases as x increases. Thus, the gap in the alloy changes from the direct gap of 1.42 eV of GaAs for $x = 0$ to the indirect gap of 2.16 eV of AlAs for $x = 1$.

From the foregoing arguments, it follows that it is possible to grow a GaAs/AlGaAs epitaxial heterostructure, varying the potential step at the interface by varying the aluminum content of the alloy.

The case of the $Al_xGa_{1-x}As$ alloy has been chosen here as the most important example, but many other alloys are used in heterostructures, also allowing some strain in the epilayer due to a variable content of the alloy components as the distance from the interface increases.

19.2 Quantum Wells

Let us consider a structure formed with two adjacent heterojunctions, such that a layer of a semiconductor material A is inserted between two layers of a second semiconductor B. Let us also assume that the band offset is such that the bottom of the valence band of A is below that of B. This system, shown in Fig. 19.5, is named a *quantum well* (QW). If the external layers are sufficiently thick or the potential step formed by the conduction bands is sufficiently high, such that tunneling out of the well is negligible, electrons are confined in one direction (z in the figure) and free to move along the other two orthogonal directions, i.e., in the x-y plane parallel to the interfaces. In such conditions, electrons form a *two-dimensional electron gas* (2DEG). A similar situation can be obtained with holes, if the internal material has the top of the valence band higher than that of the outer materials. In Fig. 19.5, the three types of electron QWs are shown, corresponding to the three types of band offsets.

As seen in Chap. 18, 2DEGs can be obtained also in the triangular potential wells formed at the interfaces of appropriate junctions, as in the MOSFETs (see Fig. 18.13).

19.2.1 Electron States

To obtain the electronic states in a QW, we assume that the planes are infinite and use the envelope function with the effective-mass approximation. In such a case, the Hamiltonian is separable, and the time-independent Schrödinger equation is

$$-\frac{\hbar^2}{2m}\nabla^2\psi(x,y,z)+V(z)\psi=\epsilon\psi, \tag{19.1}$$

where $V(z)$ is the potential that forms the QW, and m is the effective mass in the internal layer. (The fact that m changes from one material to the other

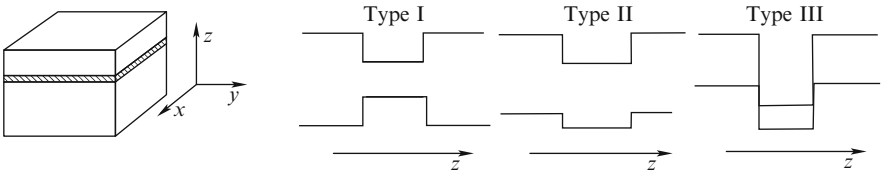


Fig. 19.5. A quantum well (QW) structure, with the three possible types of potential profiles

will be discussed below.) The electron dynamics along the x-y plane is that of free particles, and the eigenfunctions can then be written as

$$\psi(x, y, z) = \frac{1}{\sqrt{S}} \mathrm{e}^{\mathrm{i}k_x x} \mathrm{e}^{\mathrm{i}k_y y} \zeta(z), \tag{19.2}$$

where S is a normalization area of the QW. Substitution into (19.1) yields, after simplifications,

$$-\frac{\hbar^2}{2m} \frac{\mathrm{d}^2\zeta}{\mathrm{d}z^2} + V(z)\zeta(z) = \epsilon_n \zeta(z), \tag{19.3}$$

where ϵ_n is the electron energy associated with the dynamics orthogonal to the planes, and the total energy is given by

$$\boxed{\epsilon = \epsilon_n + \tfrac{\hbar^2}{2m}\left[k_x^2 + k_y^2\right]} \tag{19.4}$$

Before proceeding with the analysis of the transverse Schrödinger equation (19.3), let us make some important considerations that are independent of the particular shape of the well potential profile.

Subbands and Density of States

The energy given by (19.4) is shown in Fig. 19.6b as a function of k_x and n. It is given by several parabolas (paraboloids if also k_y is considered) called *subbands*, one for each value of the orthogonal energy ϵ_n. At low concentrations (low Fermi level) and low temperatures, electrons populate only the lowest subband; they do not participate in the dynamics orthogonal to the planes, and the dynamics is that of a real 2DEG. If, on the contrary, the temperature is not too low or the concentration is high enough, several subbands must be considered, and electron transport is influenced also by intersubband transitions. Note that the spacing between the subbands increases as the width of the well diminishes, so that, to eliminate completely the z degree of freedom, very narrow wells must be realized, kept at low temperatures.

The density of states in three-dimensional $\boldsymbol{k}$-space was derived in Chap. 8 (see (8.11)) and is given by

$$g_3(k) = 2\frac{V}{(2\pi)^3},$$

where V is the volume of the crystal. The resulting density of states in energy is (cf. (8.12))

$$g_3(\epsilon) = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \sqrt{\epsilon}.$$

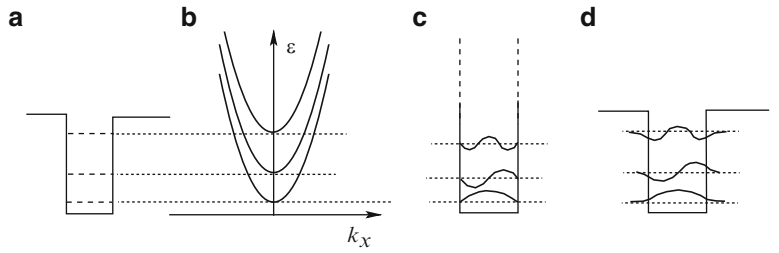


Fig. 19.6. (**a**) Potential well and energy levels; (**b**) Subbands in a QW; (**c**) Shapes of wavefunctions in the infinite square-well approximation; (**d**) Shapes of wavefunctions in a finite potential well

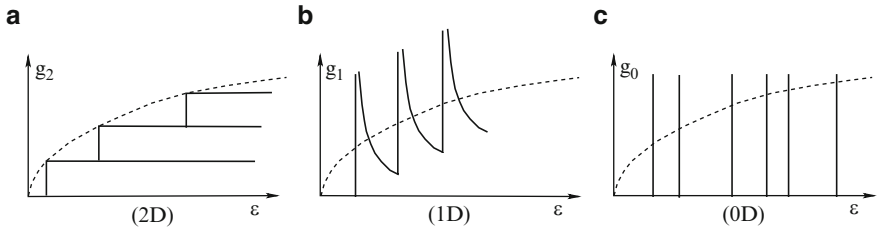


Fig. 19.7. Qualitative behavior of the density of states in low-dimensional systems, compared with the 3D case (*dashed line*). (**a**) Quantum well, (**b**) Quantum wire, (**c**) Quantum dot (QD)

In QWs, the two-dimensional density of states in $\boldsymbol{k}$ space is

$$g_2(k) = 2\frac{S}{(2\pi)^2} \ ,$$

where S is the area of the well. The surface in the k_x-k_y plane between the energies ϵ and $\epsilon + \mathrm{d}\epsilon$ is $\mathrm{d}S_k = 2\pi k\,\mathrm{d}k = 2\pi m/\hbar^2\,\mathrm{d}\epsilon$, and therefore the two-dimensional density of states in energy is

$$g_2(\epsilon) = 2\frac{S}{(2\pi)^2}\frac{2\pi m}{\hbar^2} = S\frac{m}{\pi\hbar^2}. \tag{19.5}$$

This value is independent of energy. However, at increasing energy, each time ϵ crosses an eigenvalue of the well the states of the new subband are added, and the density of states increases of a quantity given by (19.5), as shown in Fig. 19.7a.

Orthogonal States and Self Consistency

Let us now consider the dynamics of the electrons orthogonal to the planes, i.e., in the direction z, along which the motion of the electrons is confined.

Equation (19.3) is the Schrödinger equation for a particle in a one-dimensional confining potential. If the confining potential of the well can be approximated by an infinite potential well, the eigenvalues and eigenfunctions are those presented in Appendix B. The latter are sine and cosine functions, as shown also in Fig. 19.6c, while the eigenvalues are

$$\epsilon_n = \frac{\hbar^2}{2m}\frac{\pi^2}{d^2}n^2, \quad n = 1, 2, \ldots,$$

where d is the width of the well.

If we move from the simple infinite square well to a more general confining potential, two important points must be considered. The first problem is that the potential that confines the electrons inside the well is not infinite and the wavefunction penetrates somewhat inside the barriers, as shown in Fig. 19.6d. Let us remember that the wavefunction $\zeta(z)$ we are dealing with here, is the envelope wavefunction, developed in Chap. 7, and that we are assuming that this function envelops a well-defined Bloch state, as described by (7.25). This is not true in the present case, where the Bloch functions and the bands are different in the different materials. The theory of the envelope function has been revised to deal with this new situation, [80, 81]. The result is that with a position-dependent effective mass $m(z)$, the Schrödinger equation for the envelope function must be written as

$$-\frac{\hbar^2}{2}\frac{\mathrm{d}}{\mathrm{d}z}\frac{1}{m(z)}\frac{\mathrm{d}}{\mathrm{d}z}\zeta(z) + V(z)\zeta(z) = \epsilon_n\zeta(z). \tag{19.6}$$

Note that since the second term of the l.h.s. and the term at the r.h.s. are finite, if the effective mass has a discontinuity at the interface, the same must be true for the derivative of the envelope function. Equation (19.6) must then be solved with the boundary condition given by the continuity of the quantities

$$\zeta(z) \quad \text{and} \quad \frac{1}{m(z)}\frac{\mathrm{d}}{\mathrm{d}z}\zeta(z).$$

The criterion for the validity of the approximations that lead to (19.6) is that the envelope function must be slowly varying on the scale of a lattice constant.

Usually, the solution is obtained numerically, also because of second problem anticipated above: the selfconsistency of the potential with the presence of electrons in the well. In fact, the potential profile determines the transverse electron states and, therefore, the subbands. Statistics determines how many of such states and bands are occupied by electrons, and, in turn, the presence of electron charges modifies the potential profile. As we have already seen in Chap. 18, this self-consistency is controlled by Poisson equation. In the present case, the dielectric constant is function of position, since it is different in the different materials, as is the effective mass. Poisson equation (18.1) reads

$$\frac{\mathrm{d}}{\mathrm{d}z}\varepsilon(z)\frac{\mathrm{d}}{\mathrm{d}z}\phi(z) = -\left[\rho(z) + (-e)\sum_n N_n|\zeta_n(z)|^2\right], \tag{19.7}$$

where $\phi = V/(-e)$, ρ is the fixed charge density due to ionized impurities, and N_n is the effective number of electrons occupying the orthogonal state ζ_n. N_n is determined by the electron distribution in the n-th subband.

At equilibrium, the electron distribution in each subband is given by the Fermi–Dirac distribution with the two-dimensional density of states in (19.5):

$$N_n \propto \int_{\epsilon_n}^{\infty} g_2(\epsilon) \frac{1}{\mathrm{e}^{\frac{\epsilon-\mu}{K_B T}} + 1} \mathrm{d}\epsilon,$$

where μ is the electrochemical potential. g_2 is given by (19.5) and is energy independent. It was already noted, however, that in Poisson equation the charge density must always be three-dimensional. In evaluating the electron charge density in z, also the wavefunction along the plane xy is to be considered. Its squared amplitude is simply given by $1/S$, so that the S in front of g_2 cancels, as it should for physical consistency. The integral is easily evaluated, and the result is

$$N_n = \frac{m}{\pi\hbar^2} K_B T \ln\left(1 + \mathrm{e}^{\frac{\mu-\epsilon_n}{K_B T}}\right).$$

The simultaneous solution of (19.6) and (19.7) yields the self-consistent potential of the QW and its electron eigenfunctions.

Similar considerations hold also for other structures, such as quantum wires and dots.

When high fields are applied to the QW, the electron distribution gets out of equilibrium and Schrödinger, transport, and Poisson equations must be solved iteratively to reach self-consistency, as discussed in Chap. 18.

19.2.2 Transport

Electron transport in QWs [12, 33, 142, 143, 176, 310, 352, 355, 421] has much in common with bulk transport as well as many significant differences. The electron dynamics orthogonal to the interfaces, where confinement occurs, gives rise, as we have seen, to localized states. Transport occurs, therefore, along the plane of the 2DEG. Along such directions, electron transport is again described by Boltzmann equation with semiclassical dynamics, when appropriate, and Fermi golden rule for the scattering processes. There are, however, relevant differences due to the reduced dimensionality. Let us review the main of such differences. They will be dealt with in the following of the section:

1. Only two continuous coordinates define the electron position, and two momentum components define its velocity. A further quantum number indicates the orthogonal state, or the subband, of the electron.
2. The 2D density of states alters the effect of the scattering mechanisms with respect to the 3D case studied in Chap. 9. Within a given band, the density of states is reduced by the reduced dimensionality, but the third dimension is recovered by inter-subband transitions.

3. With the modulation-doping technique (see below) charge carriers can be kept far from the impurities which generated them. In this way, it is possible to reduce appreciably the ionized-impurity scattering and obtain extremely high electron mobilities at low temperatures.
4. Lattice vibrations contain confined modes, such as interface and slab modes, besides extended modes.
5. Surface-roughness may be relevant and must be added to the other scattering mechanisms.
6. In electron–phonon interaction, momentum is conserved only along the directions parallel to the interfaces. Along the normal direction, momentum is not a good quantum number since the walls of the well participate in momentum balance.

In spite of the above differences, the transport properties of a QW would be very similar to those of the bulk material, where it not for the effect of modulation doping.

Phonon Scattering

Lattice vibrations in low-dimensional structures and their interaction with charge carriers have been extensively studied (see, for example, [310,384,483]). They contain extended modes as well as confined modes, such as interface and slab modes. Often, for simplicity, it is assumed that the lattice vibrations are the same as in bulk materials. This can be a reasonable approximation when the materials forming the heterostructure are similar, owing to the epitaxial continuity of the crystal lattice. Accounting for interface and slab phonons did not generate particular differences in the results of electron transport in a GaAs–AlGaAs QW, once the scatterings with all modes have been summed up [54].

To evaluate the transition rate between two electron states in a QW, we start from the Fermi golden rule (9.2) as in Chap. 9. Now the electron state is defined by the wavevector $\boldsymbol{k}_\parallel$ along the plane and the index n of the subband. The phonon state is still labeled by the variable c. The transition rate from a state $|\boldsymbol{k}_\parallel, n, c\rangle$ to a state $|\boldsymbol{k}'_\parallel, n', c'\rangle$ induced by a perturbation Hamiltonian $\mathcal{H}'$ is given by

$$P(\boldsymbol{k}_\parallel, n, c; \boldsymbol{k}'_\parallel, n', c') = \frac{2\pi}{\hbar} |\langle \boldsymbol{k}'_\parallel, n', c' | \mathcal{H}' | \boldsymbol{k}_\parallel, n, c \rangle|^2 \delta(\epsilon(\boldsymbol{k}'_\parallel, n', c') - \epsilon(\boldsymbol{k}_\parallel, n, c)). \tag{19.8}$$

The interaction Hamiltonian is again expanded as Fourier series, as in (9.3), and the matrix element, given by (9.4) for the bulk case, takes now the form

$$\frac{1}{\sqrt{V}} \sum_{\boldsymbol{q}} \langle c' | \mathcal{A}(\boldsymbol{q}, \boldsymbol{y}) | c \rangle \int \psi^*_{\boldsymbol{k}'\parallel}(\boldsymbol{r}_\parallel) \mathrm{e}^{\mathrm{i}\boldsymbol{q}_\parallel \boldsymbol{r}_\parallel} \psi_{\boldsymbol{k}_\parallel}(\boldsymbol{r}_\parallel) \, \mathrm{d}\boldsymbol{r}_\parallel \int \zeta^*_{n'}(z) \mathrm{e}^{\mathrm{i}q_z z} \zeta_n(z) \, \mathrm{d}z, \tag{19.9}$$

where $\boldsymbol{q}$ is the phonon wavevector, $\boldsymbol{q}_{\parallel}$ and q_z are its components parallel and orthogonal to the QW plane, $\mathcal{A}(\boldsymbol{q}, \boldsymbol{y})$ is defined in (9.3) and accounts for absorption, $\psi_{\boldsymbol{k}_{\parallel}}(\boldsymbol{r}_{\parallel})$ is the plane wave along the plane of the QW, and $\zeta_n(z)$ is the n-th orthogonal wavefunction.

The first integral in (19.9), elaborated as in Sect. 9.2, yields the conservation of the momentum component in the QW plane:

$$\boldsymbol{k}'_{\parallel} = \boldsymbol{k}_{\parallel} \pm \boldsymbol{q}_{\parallel} + \boldsymbol{G}_{\parallel}.$$

Scattering processes can again be *umklapp* ($\boldsymbol{G}_{\parallel} \neq 0$) and *normal* or *non umklapp* ($\boldsymbol{G}_{\parallel} = 0$). If the periodic part enveloped by the envelope function is considered, the overlap integral discussed in Sect. 9.2 is also present. The last integral in (19.9) is often named *form factor*. It can be evaluated analytically [352] in the infinite potential square well. More generally, it must be evaluated numerically, in the self-consistent procedure indicated above.

The form factor substitutes the momentum conservation in the normal direction. The wavefunctions $\zeta_n(z)$ correspond to given momentum uncertainties, responsible for the momentum nonconservation shown in Fig. 19.8. If wider QWs are considered, the orthogonal wavefunctions become closer to plane waves and the momentum nonconservation diminishes, becoming a rigorous conservation for infinitely wide wells (bulk).

The quasielasticity of the acoustic–phonon scattering in bulk semiconductors, discussed in Sect. 9.3.1, was derived from energy and momentum conservations in the transition. Since the momentum conservation is relaxed in QWs, also the elastic approximation for acoustic scattering must be revised carefully [354].

As in the bulk case, the square matrix element of $\mathcal{A}(\boldsymbol{q}, \boldsymbol{y})$ in (19.9) is proportional to the phonon occupation number N_q for absorption, or $(N_q + 1)$ for emission. $\mathcal{A}(\boldsymbol{q}, \boldsymbol{y})$ contains also the coupling between phonons and electrons and is different for the different interaction mechanisms (deformation

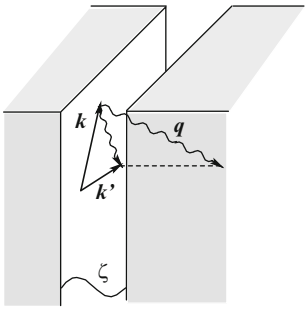


Fig. 19.8. Momentum non conservation in phonon scattering in QWs, where the walls of the well participate in momentum balance. The orthogonal states $\zeta(z)$ are not eigenstates of the momentum and contain a momentum uncertainty. Momentum nonconservation occurs within this uncertainty

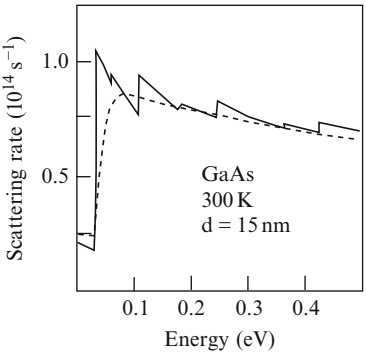


Fig. 19.9. Scattering rate of electrons in a GaAs quantum well interacting with polar optical phonons at 300 K (*continuous line*), compared with the bulk case (*dashed line*) [171]

potential, piezoelectric, polar). The same expressions given for bulk materials are used in QWs. Its integration over the phonon modes $\boldsymbol{q}$ yields the scattering rate, and the δ of energy conservation brings in the electron density of states. The latter is reflected in the transition rates, as shown in Fig. 19.9 for the case of polar optical modes in GaAs-AlGaAs QWs.

If several subbands are accessible to the electrons, intersubbands transitions must be considered among the other possible collision events, with the same procedures seen for nonequivalent intervalley scattering in bulk.

When the 2DEG is formed in a many-valley semiconductor, as in the silicon inversion layer of a MOSFET, the band structure of the bulk material must be projected on the plane of the interfaces. This involves specific band structures and effective masses for any particular interface orientation (see, for example [425]). Intervalley and intravalley transitions, as well as intersubband, have to be considered in electron transport.

In a QW formed by a silicon quantum layer between SiGe alloys, the Si layer is strained because of the different lattice constants of the two materials. If the content of Ge in the alloy is small, and the Si layer is thin enough, the silicon layer maintains its crystal integrity. As seen in Chap. 15, in strained Si the different band minima are differently shifted in energy, electrons occupy the lowest valleys, and with an appropriate choice of the crystal orientation a small effective mass may be realized along the transport direction. In this way, it is possible to obtain higher carrier mobilities with respect to relaxed bulk silicon [201].

Ionized-Impurity Scattering

Also ionized-impurity scattering in low-dimensional structures is treated in close analogy with the bulk theory with the Fermi golden rule. Impurities are again considered located in random positions to avoid correlations.

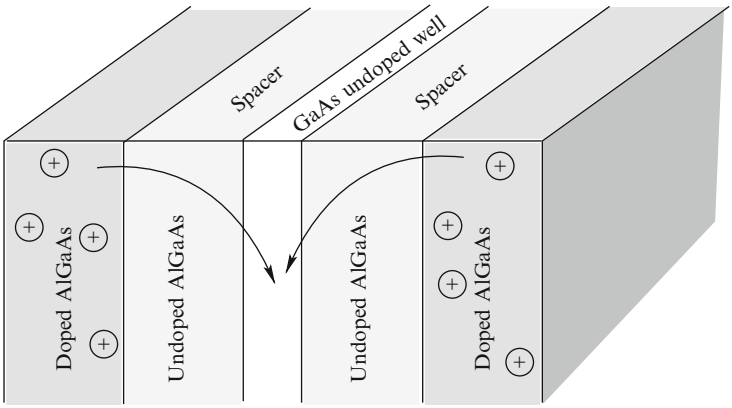


Fig. 19.10. With the modulation-doping technique charge carriers are provided by far-away dopants, thus reducing ionized-impurity scattering

As it regards the Coulomb potential, however, its screening must be evaluated accounting for the presence of electrons in several subbands [353]. Furthermore, if the materials forming the heterostructure have different permittivities, image charges must be used to account for the polarization charges at the interfaces. This is particularly true in the case of the 2DEG in the channel at the interface of the Si/SiO_2 of the MOSFET [12]. In a QW formed with materials with similar permittivities, such as GaAs/AlGaAs systems, the image charge is usually neglected.

Modulation Doping

Charge carriers can be present in the well as effect of appropriate doping. If the donors or acceptors are located in the central layer, i.e., in the well, the mobility of the carriers is reduced by ionized-impurity scattering with the ions that provide them, particularly at low temperature. If, however, only the external materials are doped, carriers will diffuse and "fall" inside the well, where they will remain and move more freely. This method to enhance electron mobility in 2DEGs is commonly used and is called *modulation doping* [119]. Ionized impurities still act as scatterers, but from a larger distance and, therefore, with minor effect. The efficiency of the modulation doping technique is amplified by the insertion of layers of undoped external material, called *spacers*, adjacent to the well, as shown in Fig. 19.10. In this way, the ionized impurities that provide the carriers are still more far away from the mobile charges in the well. With such a technique, very high mobilities have been obtained in GaAs at low temperatures, around $10^7\,\mathrm{cm}^2/\mathrm{Vs}$ for electrons [337] and $10^6\,\mathrm{cm}^2/\mathrm{Vs}$ for holes [300].

Surface-Roughness Scattering

If the interface of a heterostructure is not perfectly planar, its roughness is a source of possible electron transitions between eigenstates of the "perfect" Hamiltonian. This type of interaction becomes important in systems where the surface plays an important geometrical role, such as systems of reduced dimensionality.

On the contrary, it was just seen that in GaAs QWs extremely large mobilities can be obtained with the modulation doping technique, much higher than in bulk materials. This means that surface-roughness scattering is not very effective in such systems, as effect of both the extremely high quality reached with epitaxial MBE and the reduced particle density $|\psi|^2$ at the interface. The situation of the interface between the silicon channel and the oxide in MOSFETs is different. The interface, obtained by oxidation of the crystalline silicon, is never perfect, and roughness scattering is still one of the major causes of the electron mobility degradation.

In the simplest model of surface-roughness scattering, it is assumed, in a semiclassical approach, that when an electron hits the surface, it undergoes a diffuse reflection or a specular reflection with probabilities α and $1-\alpha$, respectively. The parameter α, determined empirically, is a measure of the interface roughness.

More accurate theory of this scattering mechanism can be found, for example, in [11, 12, 170, 391]. The scattering potential ΔV is assumed proportional to the fluctuation $\Delta(\boldsymbol{r})$ of the interface with respect to the ideal two-dimensional plane:

$$\Delta V = eF_s\Delta(\boldsymbol{r}),$$

where F_s is an average surface field. By assuming a Gaussian autocorrelation for $\Delta(\boldsymbol{r})$:

$$\langle\Delta(\boldsymbol{r})\Delta(\boldsymbol{r}-\boldsymbol{s})\rangle = \Delta^2 \mathrm{e}^{-s^2/L^2},$$

where Δ and L are the rms and the correlation length of $\Delta(\boldsymbol{r})$, respectively, a square matrix element between the states $\boldsymbol{k}$ and $\boldsymbol{k}' = \boldsymbol{k} \pm \boldsymbol{q}$ is found to be

$$|V_q|^2 = \pi \mathrm{e}^2 F_s^2 \Delta^2 L^2 \mathrm{e}^{-q^2L^2/4}.$$

This is then used in the Fermi golden rule to obtain the scattering rate due to surface roughness. Many improvements can of course be performed on the above theory to account for screening, image charge, different correlations, etc.

Hot-Electron Effects

As it regards hot electrons, in general, the new density of states is such that when the field is high enough to deviate from the linear-response regime, transport in a single band becomes unstable in 2D (as well as 1D) structures [374].

Steady state is attained only via intersubband (or intervalley) scattering. In contrast, threshold field for runaway, after the onset of intersubband scattering, is higher in GaAs QWs than in bulk materials, as a consequence of the sharp changes in the density of states [374].

A new hot-electron effect in QWs was proposed by Hess et al. [186]: when electrons in a high-mobility well are heated up by the field, they may reach enough energy to leave the well and populate the low-mobility lateral material. This phenomenon, called *real-space transfer* was later observed experimentally [235] and exploited to fabricate negative-differential-resistance devices with high peak-to-valley ratio [233, 292].

19.2.3 Multiple Quantum Wells

If a series of layers of different materials is fabricated, a *multiple quantum well* (MQW) may be obtained. By properly selecting materials, graded alloys, doping concentrations, and thicknesses, various potential profiles can be designed and realized. With this technology, called sometimes *band-gap engineering* [91], potential profiles can be realized that in the past where considered only textbook theoretical exercises.

Several applications of QWs and MQWs have been developed, (see, for example [92, 238]), some of which will be briefly described below.

19.3 Quantum Wires

Systems in which electrons are confined along two directions, say x and y, by a potential energy $V(x, y)$ and are free to move only along the third z direction are named *quantum wires* (QWR). There are several methods to fabricate QWRs, some of which are schematically shown in Fig. 19.11. In (a), a QWR is obtained by properly etching a QW. In (b), a *split gate* is deposited

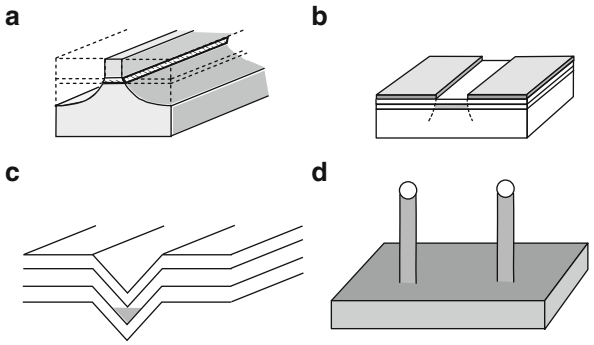


Fig. 19.11. Different techniques to fabricate quantum wires (**a**) Etching a QW; (**b**) Split gate; (**c**) V-grooved wells; (**d**) Catalyzed chemical vapor deposition

on top of a QW structure, and the application of a negative potential pushes away electrons from the 2DEG, except in a narrow region forming the QWR. In (c), QWRs are obtained with a localized two-dimensional electron state in the corner of a V-grooved well. A similar situation is obtained with a T-shaped well. Finally, in (d) nanoparticles, often gold, are located on a substrate, and then by chemical vapor deposition wires are grown below such particles which act as catalysts.

Silicon nanowires are of particular interest for their possible nanoelectronic applications. They have been obtained in single crystal form [311] with a technique similar to that shown in Fig. 19.11d. Their electronic properties have been studied, among others, in [485].

A very special type of quantum wire is that of carbon nanotubes. Owing to their particular structure and importance, they will be treated separately, in the next chapter.

Electron States

In QWRs, the electron dynamics along the wire z direction is that of free particles, and the eigenfunctions can then be written as

$$\psi(x,y,z) = \frac{1}{\sqrt{L}} \mathrm{e}^{\mathrm{i}k_z z} \zeta_n(x,y), \tag{19.10}$$

where L is a normalization length of the wire. The orthogonal states depend now upon two coordinates and are given by the eigenfunctions $\zeta_n(x,y)$ of the two-dimensional Schrödinger equation

$$\left\{ -\frac{\hbar^2}{2m} \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right] + V(x,y) \right\} \zeta_n(x,y) = \epsilon_n \zeta_n(x,y),$$

where ϵ_n is the n-th energy eigenvalue of the confining potential $V(x,y)$ orthogonal to the wire. The total energy is given by

$$\epsilon = \epsilon_n + \frac{\hbar^2 k_z^2}{2m}.$$

Subbands are still present and are shown in part (b) of Fig. 19.6, but this time the wavevector has only the component k_z. Electrons in QWRs form therefore a one-dimensional electron gas (1DEG).

The density of states in QWRs is obtained, for each subband, in the usual way: the one-dimensional density of states in k_z space is

$$g_1(k_z) = 2\frac{L}{2\pi}.$$

The length in the k axis between longitudinal energy ϵ and $\epsilon + \mathrm{d}\epsilon$ is $2\mathrm{d}k = (2/\hbar)\sqrt{m/2\epsilon}\,\mathrm{d}\epsilon$, where the extra factor 2 accounts for positive and

negative k_z. Thus, the one-dimensional density of states in energy, for each band is, including spin multiplicity,

$$g_1(\epsilon) = 2\frac{L}{2\pi}\frac{2}{\hbar}\sqrt{\frac{m}{2(\epsilon-\epsilon_n)}} = \frac{L}{\pi\hbar}\sqrt{2m}\frac{1}{\sqrt{\epsilon-\epsilon_n}}. \tag{19.11}$$

This density of states diverges at the bottom of each subband, i.e., at energies equal to the two-dimensional eigenvalues ϵ_n and then decreases, as shown in Fig. 19.7b. Collisional broadening (see Chaps. 9 and 16), however, smooths somewhat this very irregular density of states.

Transport

As it regards transport properties of QWRs, considerations similar to those made for QWs hold, in particular with respect to the possibility to use the standard Boltzmann equation for longitudinal transport and the need to solve Schrödinger, Boltzmann, and Poisson equations self-consistently. Once again, one must account for the peculiar density of states and for the presence of intersubband transitions.

As in the other semiconductor systems, electron transport in QWRs is dominated by ion and imperfection scatterings at lower temperatures and by phonons at higher temperatures.

The density of states in 1DEG could suggest an increased charge-carrier mobility at low temperatures [16, 150, 390]. In fact, if only the lowest subband is occupied by electrons, in one dimension scattering events can occur only as forward or backward collisions, without intermediate directions. At low temperatures, ionized impurities are in general the dominant scattering mechanism. Actually, if ions are present inside the wire, they may behave as traps or as insurmountable barriers according to the sign of their charge. With the modulation-doping technique, however, ionized impurities can be kept at a certain distance from the wire; a backward collision requires a large momentum transfer because of energy conservation and is not favored in Coulomb scattering. Thus, it was imagined that QWRs could have a very high mobility at low temperatures. Technological difficulties, however, make very hard to fabricate QWRs with ideal transport properties mainly because of surface-roughness scattering.

Impurity-limited mobility in QWRs has been studied, among others, in [275, 276], and boundary-roughness scattering, among others, in [314, 450]. Roughness scattering has often been studied in connection with the effect of a magnetic field because electron orbits are deflected by the magnetic field against the edges of the QWR [6, 7].

As it regards phonon scattering, confined phonon modes must be considered, of course, when free-standing QWRs are fabricated. Phonon scattering mechanisms and transport properties (linear and non linear) in wires has been studied in [50, 78, 239, 270, 277, 309, 319, 429, 430] and many others. The

considerations made above with respect to momentum nonconservation in phonon scattering in QWs are even more important in QWRs, where only one component of the crystal momentum is strictly conserved. This momentum nonconservation, as in QWs, is such that acoustic phonon scattering cannot be considered elastic, and it has been found in [309] that elastic or quasielastic approximations underestimate the electron mobility at low temperatures. The Boltzmann equation in QWRs with several types of scattering mechanisms has been studied, e.g., in [451].

Hot-phonon effects have been studied in [334] with a Monte Carlo technique.

Carrier–carrier scattering is of particular interest in QWRs. In an ideal one-dimensional system, electrons would not be able to pass each other because of Coulomb repulsion that diverges when the two carriers get close. QWRs, however, are, in reality, three-dimensional systems, where two degrees of freedom are quantized. Electron–electron scattering in QWRs has been considered, among others, in [136, 293].

In a perfect QWR, at so low temperatures that phonon scattering becomes negligible, electron transport is totally coherent, and the QWR becomes equivalent to a quantum point contact, studied in Chap. 21, devoted to the analysis of coherent transport in mesoscopic structures. Furthermore, in QWRs quantum interference effects, strong and/or weak localization discussed in Chap. 21 has been measured at low temperatures. See, for example [165, 446].

Laser emission has been observed in several types of QWRs. See, for example, [223, 232, 470].

Finally, let us mention that coupled QWRs have been proposed as elementary devices for quantum computation [27, 35, 36, 199, 357].

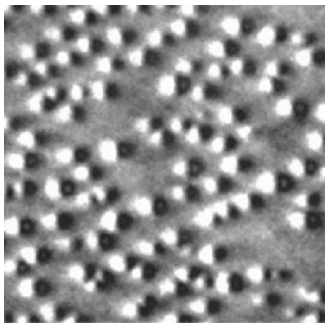
19.4 Quantum Dots

In *quantum dots* (QD), electrons are confined in all directions so that all space degrees of freedom are quantized. They are sometimes called artificial atoms, for the similarity of their electronic structure with that of atoms. See, for example, [377].

There are various methods to fabricate QDs. When epitaxial growth is performed with an appreciable lattice mismatch between the epilayer and the substrate, the resulting strain produces islands on top of the substrate. In this way, *self-assembled* QDs can be obtained [28, 125], as shown in Fig. 19.12. With clever technologies, it is also possible to induce a spatial order in self-organized QDs, both vertically and horizontally. See, for example, [466, 478]

Nanocrystals can be generated by colloidal synthesis of CdTe and CdSe (see Fig. 22.16), among other materials [481].

The fabrication of QDs is often based on a 2DEG in a QW, with appropriate gating design, as shown in Fig. 19.12. This technique allows the fabrication



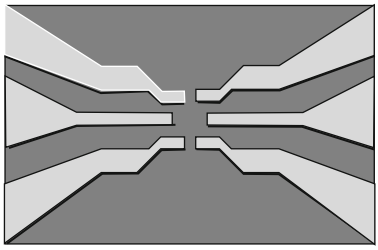


Fig. 19.12. *Left*: Self-assembled pyramid-shaped QDs (courtesy of National Institute of Standard and Technology); right: split-gate design on top of a 2DEG to obtain a QD

of complicated structures with several QDs and quantum wires electrically connected (see Fig. 22.15).

Since the wavefunction in a QD is confined in all directions, the energy spectrum is discrete. The density of states is given by a series of delta functions located at the energies of the eigenstates, as shown in Fig. 19.7. Eigenvalues and eigenfunctions can be obtained analytically only for special shapes of the dots, such as a 3D rectangular box with very high confining potential or a spherical dot [310]. In the latter case, the wavefunctions have an angular part given by the spherical harmonics, eigenfunctions of the angular momentum, as in atoms, and a radial part, obeying a one-dimensional Schrödinger equation. More often, numerical solutions are necessary.

Electrons in QDs are often accompanied by holes, and together they form excitons. Transitions occur between different levels of the exciton state including recombinations.

The frequency of the light emitted with photoexcitation depends upon the electronic energy levels and these depend on the size of the QDs. Thus, by controlling their size different colors are obtained. This effect is particularly evident in nanocrystals obtained with colloidal synthesis as shown in Fig. 22.16.

19.4.1 Transport: Coulomb Blockade

Transport through QDs is obtained with small contacts separated by energy barriers due to thin insulating layers through which tunneling may occur (see Fig. 19.13a).

When a charge Q is stored in a capacitor, the electrostatic energy of the system is given by

$$\epsilon(Q) = \frac{1}{2}\frac{Q^2}{C},$$

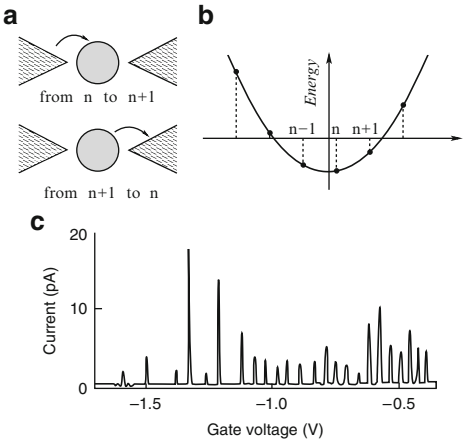


Fig. 19.13. Coulomb blockade. (**a**) Mechanism of conduction in a dot; (**b**) electrostatic energy as a function of the number of electrons in the dot; (**c**) Experimental results in [437] at 50 mK (although the electron temperature was estimated to be about 0.2 K)

where C is the capacitance of the system. For macroscopic capacitors, the change of such energy due to the addition of a single electron is negligible. In a QD, on the contrary, the capacitance is so small that this energy is comparable with the thermal energy, at least at very low temperatures. Let us then consider a QD separated by thin barriers from two contacts, as shown in Fig. 19.13a.

To have current through the QD, its charge must change of (at least) one unit, as shown in the figure, first from $n(-e)$ to $(n+1)(-e)$, and successively from $(n+1)(-e)$ back to $n(-e)$. If the QD is maintained at a potential V_d with a gate contact, the total electrostatic energy is

$$\epsilon(Q) = QV_d + \frac{1}{2}\frac{Q^2}{C}.$$

This energy is shown by the parabola in Fig. 19.13b. If the charge inside the QD were a continuous quantity, it could be possible to change it by infinitesimal amounts. Since, however, the charge can be only a multiple of the electron charge, only the values indicated by the dots in part (b) of the figure are realizable. For the current to flow, it is necessary that the system jumps back a forth from one of such points to the adjacent one. This is possible without energy exchange if two dots in the parabola are at the same height:

$$nqV_d + \frac{(nq)^2}{2C} = (n+1)qV_d + \frac{(n+1)^2q^2}{2C},$$

i.e., when the potential of the QD is kept at one of the values:

$$V_d = \frac{e}{2C}(2n+1). \tag{19.12}$$

It is therefore expected that the conductance, at very low temperature, is almost zero for a generic gate potential, and passes through a series of maxima when the dot voltage satisfies the condition in (19.12) for the various integers n. This phenomenon, called *Coulomb blockade*, is experimentally observed. Fig. 19.13c shows the result of Tarucha and coworkers [437].

The simplified above theory [456] considers the electrostatic energy of the dot, ignoring that inside the dot electrons have well-defined discrete available states. By studying the exact positions of the Coulomb blockade oscillations, it is possible to study the position of such energies and the shell filling of the QDs. The resulting *addition spectra* confirm the nature of "artificial atoms" of the QDs [437, 455].

When the condition (19.12) is satisfied, the current through the dot depends on the tunneling probability in and out the dot, and this depends upon the shape of the involved electron wavefunctions. This is the main cause of the different heights of the maxima in Fig. 19.13c.

Coulomb-blockade phenomena can be seen in many mesoscopic systems of single-electron electronics.

19.5 Superlattices

By alternating different material compositions, for example in an MBE apparatus, it is possible to fabricate a sample formed by a periodic arrangement of thin slabs of different materials, as shown in Fig. 19.14a, called a *superlattice*. The QW shown in Fig. 19.6 is repeated periodically, so that an artificial crystal is realized, whose lattice constant along the growth direction, called also "vertical" direction, is determined by the fabrication conditions and may vary from a single atomic layer to several tens of nanometers.

With superlattices we return to 3D physics, but the idea of fabricating a superlattice is what gave origin, in 1970, to the great adventure of low-dimensional structures.

19.5.1 Minibands

As it regards the electronic states in superlattices, the simplest electron potential model along the vertical direction is a sequence of square wells and

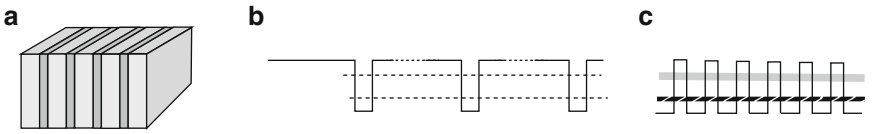


Fig. 19.14. (**a**) Superlattice. (**b**) Separate square-well potentials; the *dotted lines* indicate single-well energy levels. (**c**) Close square wells as in Kronig–Penney model; *shaded areas* indicate minibands

barriers, known as *Kronig–Penney model*, shown in Fig. 19.14. It was developed in 1930 [252] to find approximate energy eigenvalues and eigenfunctions of electrons in crystals. As a piecewise constant potential, in an effective-mass approximation, it yields wavefunctions which are combinations of sines and cosines in each region (see Appendix B). The continuity conditions of the wavefunctions and of their derivatives, combined with Bloch theorem, provide a solution for bands and gaps in quasi-closed form. See, for example, [342].

However, to get a more physical insight into this problem, we may make considerations very similar to those made in connection with the tight-binding approach to the band formation in crystals (see Sect. 6.4.1). If the wells were infinitely distant from each other, single electron states would be present in each well, degenerate with the equivalent states of the other wells, as shown in Fig. 19.14b. When the wells are brought close to each other, electron states of one well partially overlap with those of the adjacent wells and minibands are formed (Fig. 19.14c).

A linear combination of square well wavefunctions can be formed, obeying Bloch theorem, in a sort of LCAO or tight-binding theory. If coupling between only nearest-neighbor wells is considered, minibands are found whose amplitudes are given by [310]

$$\epsilon_i(k_z) = \epsilon_{oi} + s_i - 2|t_i| \cos(k_z d), \tag{19.13}$$

where ϵ_{oi} is the i-th energy level of the isolated well, s_i is the "shift integral", i.e., the mean value of the potential of a well in the i-th eigenstate of a neighboring well, t_i is the "transfer integral", given by the matrix element of the well potential between the two i-th eigenfunctions of the neighboring wells, and d is the superlattice constant. As it is evident from (19.13), the band width is given by $4|t_i|$. For GaAs/AlGaAs superlattices with well and barrier widths of the order of the nm, the bandwidths are of the order of tens of meV. This is the reason they are called "minibands". The band shape is shown in Fig. 19.15a.

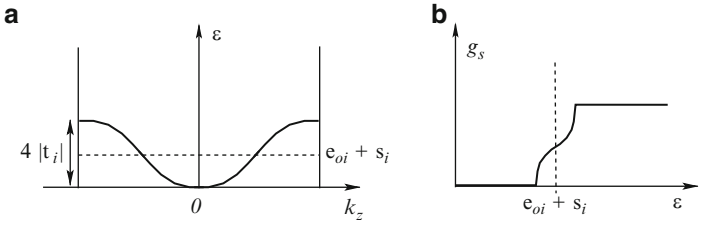


Fig. 19.15. (**a**) Shape of a miniband of a superlattice as given by (19.13) and (**b**) corresponding density of states [30]

Adding the energy along the planes, the total energy in the i-th miniband is given by

$$\epsilon_i = \epsilon_i(k_z) + \frac{\hbar^2}{2m}\left[k_x^2 + k_y^2\right].$$

The corresponding density of states is found in the usual way. Two successive plateaus of the two-dimensional density of states are joined by the density of states of the miniband, as shown in Fig. 19.15b [30].

19.5.2 Transport: Bloch Oscillations

Generally speaking, it is obvious that the conduction properties of superlattices is highly anisotropic, since along the so-called vertical direction, orthogonal to the interfaces, charge carriers must pass a series of potential barriers, while in the horizontal directions, parallel to the interfaces, they are free to move. In terms of band theory, we have seen that the minibands are very narrow; this implies a large effective mass along the vertical direction, and therefore a low mobility.

Thus, as it regards horizontal transport there is not much new in the semiclassical approximation of electron transport. Of course, the correct density of states and the intersubbands scattering must be considered.

In contrast, the vertical transport presents the interesting effect of *Bloch oscillations*, one of the reasons for which superlattices have been conceived. In the presence of an applied uniform and constant electric field $\boldsymbol{E}$, the electron wavevector changes according to

$$\hbar\dot{\boldsymbol{k}}(t) = (-e)\boldsymbol{E}, \tag{19.14}$$

as shown in Sect. 7.7. We have also seen in Sect. 8.1 that when the crystal momentum of an electron, under the action of an electric field, reaches the edge of the Brillouin zone (BZ), it is Bragg reflected and reenters the BZ from the opposite edge. Thus, in absence of scattering, an electron performs oscillations in both $\boldsymbol{k}$ space and real space, called Bloch oscillations (BO). In bulk materials, the free-flight times between collisions is never long enough to realize BOs.

In superlattices, however, the width of the BZ in the vertical direction is given by $2\pi/d$. Since the superlattice constant d may be much greater than the crystal lattice, the width of the superlattice BZ may be much smaller than that of the normal bulk. The period of the BO is easily found as the time necessary for an electron to cross the BZ:

$$T_B = (2\pi/d)/(eF/\hbar) = h/deF. \tag{19.15}$$

Thus in a superlattice, owing to the reduced BZ, Bloch oscillations can be realized with sufficiently perfect and pure samples and at sufficiently low temperatures. Their experimental observations require a very sophisticated technology, related to radiation emission in the THz frequency, and

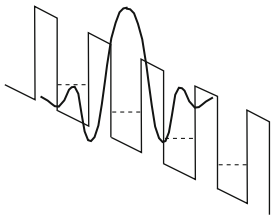


Fig. 19.16. Wannier–Stark ladder (*dashed levels*) and one Wannier–Stark function

could be obtained [279, 467] much after the first experimental realization of superlattices.

19.5.3 Wannier–Stark Ladder

When an electric field is applied along the vertical direction, the QWs forming a superlattice are tilted, each well having an energy profile shifted with respect to the neighboring ones, as shown in Fig. 19.16. If the field is weak enough, such that the energy difference eEd between two adjacent wells is much less than the miniband width, it may be treated in the semiclassical limit as described in the previous section.

If, on the contrary, the field is high, electron states start to become localized. Each of them has a maximum in a well. The reason for the localization induced by the field can be easily recognized as due to the fact that the single-well levels become misaligned, and the electron tunneling between adjacent wells is no more resonant (as is, instead, in absence of field).

These localized states are called *Wannier–Stark* (WS) states. The different WS states corresponding to the same well level, or miniband, have the same shape, with a maximum localized within a well as shown in Fig. 19.16. The energy difference between two adjacent WS levels is obviously eEd. These levels form the so-called *Wannier–Stark ladder*. The energy step of this ladder suggests the connection between the WS localization and the BO. In fact, if we consider the frequency of the latter and the associated energy, we find, from (19.15),

$$h\nu = \frac{h}{T_B} = deF,$$

the same value of the WS ladder step. We thus recognize the microwave radiation emitted by BOs as due to transitions between adjacent WS states and the WS localization as the quantization of the close orbits in space due to BOs.

If the electric field is too high, Zener transitions (see Sect. 7.3.3) may occur between adjacent WS states belonging to different minibands.

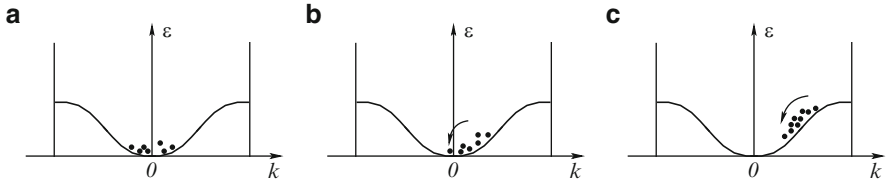


Fig. 19.17. Esaki–Tsu mechanism of negative differential conductivity (NDC) in superlattices. (**a**) At equilibrium, electrons occupy the bottom of the miniband. (**b**) With an applied electric field, electrons change their momentum according to semiclassical dynamics. Steady state is provided by the scattering mechanisms. (**c**) When the field is high, electrons occupy regions of the band where their velocities decreases at increasing energies

19.5.4 Negative Differential Conductivity

The original paper by Esaki and Tsu [132], which originated the physics of low-dimensional semiconductor structures, has the title "Superlattice and Negative Differential Conductivity (NDC) in Semiconductors".[1] NDC is of great interest to electronics since it provides a means to fabricate solid-state microwave generators through the formation of Gunn domains (see Sect.13.4). The mechanism envisioned by those authors is a very simple hot-electron effect. We know that when an electric field is applied to a semiconductor, electrons change their momentum according to the semiclassical dynamics (19.14). Scattering mechanisms tend to recover the equilibrium distribution so that, in steady-state condition the electron distribution function is shifted in the direction of the electric force, as shown in Fig. 19.17. The minibands of a superlattice are very narrow in energy, and the BZ, as seen above, is very narrow in momentum. Thus, it is easy for the electron distribution function to reach the region of the band with the inflection point, where the group velocity starts to decrease. As the electric field increases, more electrons reach that region of the band. Thus, the drift velocity decreases at increasing field strength, leading to a negative differential mobility and, as a consequence, to a NDC.

The effect has soon been observed experimentally [133, 420], and later applied in oscillators. See, for example, [175].

19.6 Applications

The possible applications of semiconductor heterostructures are countless in both electronics and optoelectronics. Here, only a few will be mentioned as examples.

[1] It seems appropriate to quote the prophetic last sentence of the abstract of that paper: *The study of superlattices and observations of quantum mechanical effects on a new physical scale may provide a valuable area of investigation in the field of semiconductors.*

High-Electron-Mobility Transistor

The *High-electron-mobility transistor* (HEMT) is a field-effect transistor with a structure similar to that of the MOSFET, described in Sect. 18.6, where the conducting channel is formed by a QW (typically an AlGaAs/GaAs/AlGaAs structure). Electrons are generated with the modulation-doping technique, so that their mobility is maintained very high.

Single-Electron Transistor

Single-electron transistors are based on the Coulomb-blockade effect in a QD, described above. With reference to Fig. 19.13, the two contacts function as source and drain. The gate is a contact which controls the voltage of the dot. Owing to the periodic variation of the conductance, Coulomb blockade makes such that the transistor can be switched on and off several times with a continuous monotonic variation of the gate voltage. The transistor turns on and off again every time an electron is added. Thus, it may be used to detect currents due to the transit of single electrons. For a review, see [234].

Quantum-Well Laser

Quantum-well lasers are the key components of CD players, laser printers, and optic communication systems. They are formed by a QW of type I (see Fig. 19.5) where one of the two lateral layers is n-doped and the other is p-doped. When current is forced across the well, electrons and holes meet and recombine in the well layer. The large concentration of carriers and the smaller number of states makes it easier to obtain the population inversion necessary for the lasing effect (Fig. 19.18).

The frequency of the emitted radiation is given by the energy difference between the localized states in the conduction and valence bands and therefore it may be changed by varying the well width.

Many technological variations have been developed starting from the simplified description above, often using multiple quantum wells and strained quantum wells (see, for example, [491]).

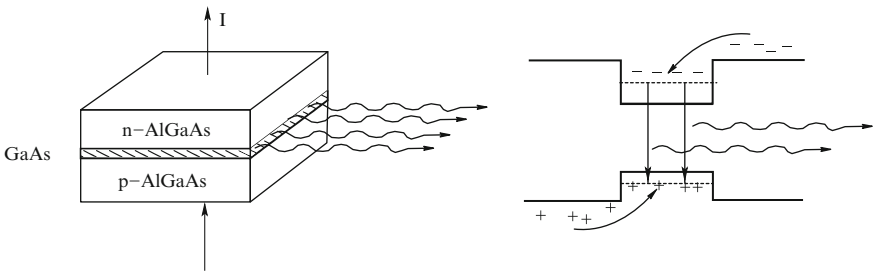


Fig. 19.18. Quantum-well laser

Quantum Cascade Laser

In general, semiconductor lasers are based on electron-hole recombination. On the contrary, in *quantum cascade lasers* only one type of carriers are involved [134]. This unipolar laser is formed by a series of multiple quantum wells. By properly designing the whole structure, and therefore the transition rates from one well to the next, a population inversion can be obtained which provides a lasing action.

20

Carbon Nanotubes

20.1 Introduction

Carbon nanotubes (CNTs) are the subject of the latest chapter of the long story of carbon fibers, initiated in the nineteenth century by the Edison's interest in filaments for light bulbs and continued during the twentieth century by space and aircraft industries, as carbon fibers form light-weight, very stiff materials. CNTs were discovered at NEC laboratories by Iijima in 1991 [197]. They present very peculiar physical properties and, in particular, a large variety of electrical behaviors, ranging from those of a semimetal with charge carriers mimicking massless Dirac particles [162], to semiconductors with band gaps varying from zero to about 1 eV, and even superconductors. For such reasons, they immediately generated a fascinating new field of solid-state physics, intensively studied for both basic research and possible industrial applications.

CNTs are produced with a variety of experimental techniques, such as arch discharge, chemical vapor deposition, laser vaporization. For reasons of space, we can only present here the basic ideas of the structure of CNTs and of their electronic properties, referring the interested reader to more specialized books, such as, for example, [224, 367, 389].

20.2 Structure

A single-walled CNT is a honeycomb lattice of carbon atoms rolled into a hollow cylinder with nanometric diameter and micrometric length.

To understand the structure and the physical properties of CNTs, it is convenient to start from the atomic structure of graphite, a stable solid phase of carbon.

Graphite

A carbon atom has two core electrons $1s^2$, and four valence electrons $2s^2 2p^2$. Since the energy difference between the $2s$ and $2p$ states is smaller than the

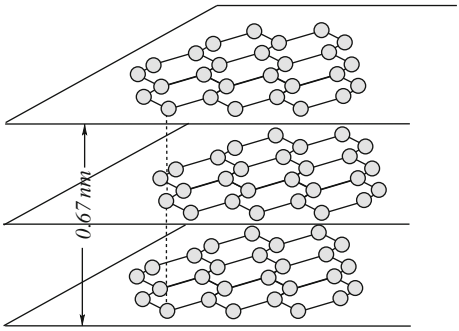


Fig. 20.1. Crystal structure of graphite

bonding energy of two neighboring C atoms, such orbitals form hybridized σ-bonds, σ^*-antibonds, π-bonds, and π^*-antibonds. Graphite is composed of layers of carbon atoms strongly bounded by σ bonds formed by $2s$, $2p_x$, and $2p_y$ orbitals in a honeycomb structure, as shown in Fig. 20.1. The different layers are weakly bounded by Van der Waals forces. The interlayer distance is $6.7/2\,\text{Å}$.

Graphene

An isolated single atomic layer of graphite is called *graphene*. Its hexagonal lattice structure and the unit cell, containing two atoms, are shown in Fig. 20.2, together with the reciprocal lattice and its Brillouin zone (BZ). The distance between two nearest carbon atoms (the side of the hexagon) is $a_{\text{cc}} = 1.42\,\text{Å}$. Then, the length of the unit vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, shown in Fig. 20.2, is

$$a = \sqrt{3}\, a_{\text{cc}} = 2.46\,\text{Å}.$$

The two unit vectors of the direct lattice are

$$a_{1x} = \frac{1}{2} a\sqrt{3}\,, \qquad a_{1y} = \frac{1}{2} a\,, \qquad a_{2x} = \frac{1}{2} a\sqrt{3}\,, \qquad a_{2y} = -\frac{1}{2} a.$$

To find the unit vectors of the reciprocal lattice of graphene, let us use the three-dimensional definition in Sect. 4.4, considering a third unit vector $\boldsymbol{a}_3$ of the direct lattice, orthogonal to the plane of graphene, of unit length, and directed into the plane of the figure. Then

$$\boldsymbol{b}_1 = 2\pi \frac{\boldsymbol{a}_2 \times \boldsymbol{a}_3}{\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3}, \qquad \boldsymbol{b}_2 = 2\pi \frac{\boldsymbol{a}_3 \times \boldsymbol{a}_1}{\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3}. \tag{20.1}$$

The unit vectors of direct and reciprocal lattices satisfy the condition

$$\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi \delta_{ij}, \quad i, j = 1, 2.$$

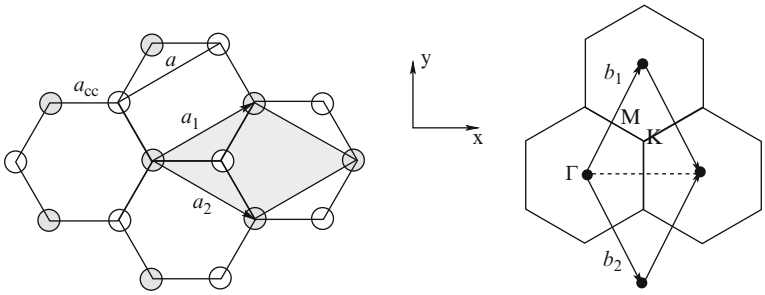


Fig. 20.2. Direct (*left*) and reciprocal (*right*) lattices of graphene. The shaded area shows the unit cell, with two atoms

The denominator in (20.1) becomes the area of the hexagon:

$$A = 6\frac{1}{2}a_{cc}\frac{1}{2}a_{cc}\sqrt{3} = \frac{3}{2}\sqrt{3}a_{cc}^2 = \frac{1}{2}\sqrt{3}a^2,$$

and the components of the unit vectors of the reciprocal lattice are

$$b_{1x} = \frac{2\pi}{a\sqrt{3}}, \qquad b_{1y} = \frac{2\pi}{a}, \qquad b_{2x} = \frac{2\pi}{a\sqrt{3}}, \qquad b_{2y} = -\frac{2\pi}{a},$$

shown in Fig. 20.2. The length of the basis vectors of the reciprocal lattice is

$$b = 2\pi\frac{a}{\frac{1}{2}\sqrt{3}a^2} = \frac{4\pi}{a\sqrt{3}}.$$

Graphene has been isolated in 2004 [330] and, owing to its very peculiar electrical and optical properties, is today the subject of very intense research (see, for example, [97, 162]).

Nanotubes

As said above, single-walled CNTs are single layers of graphene rolled up into a hollow cylinder (see Fig. 20.3). The diameters are of the order of nanometers.

The rolling direction of the graphene sheet determines the geometric properties, and, and we shall see, also the electrical properties of the CNT. This direction is defined by the pair of integer numbers (n, m) which indicate, in units of the direct lattice, the vector $\boldsymbol{C}_\mathrm{h}$ which joints two points of the graphene lattice that are brought to coincide after wrapping. $\boldsymbol{C}_\mathrm{h}$ is called *circumferential vector*, or *chiral vector*. As examples, the circumferential vectors (5,5), (5,0), and (2,7) are shown in Fig. 20.4. The axis of the CNT is orthogonal to the circumferential vector, and it can be easily seen that if the circumferential vectors are of types (n, n) and $(n, 0)$, the CNTs are not twisted, as shown in Fig. 20.5. They are *nonchiral tubes* and are named *armchair tubes*

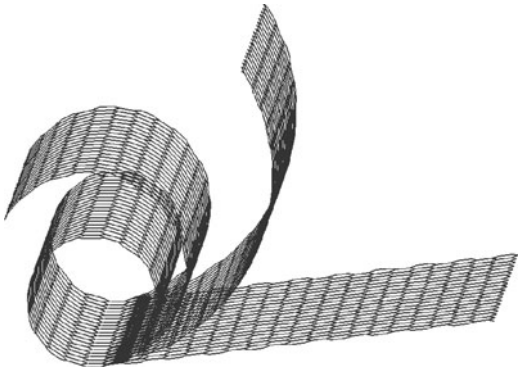


Fig. 20.3. Formation of a CNT from a graphene sheet. Courtesy of E. Piccinini

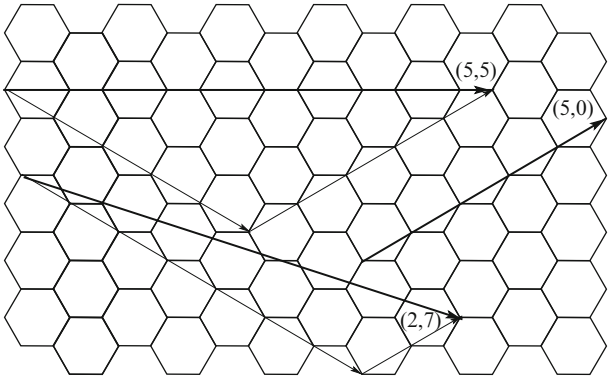


Fig. 20.4. Several circumferential vectors

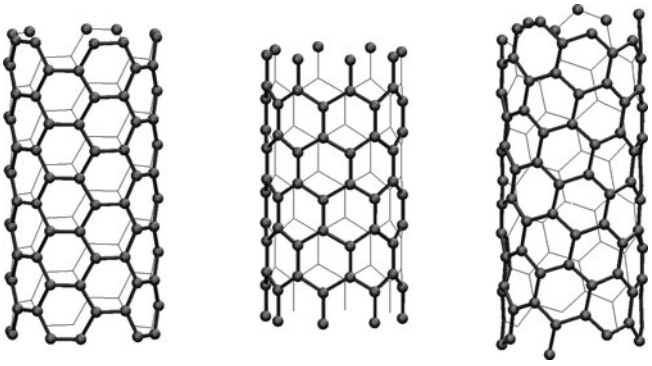


Fig. 20.5. Armchair (5,5), zigzag (8,0), and chiral (2,7) CNTs. Courtesy of E. Piccinini

and *zigzag tubes*, respectively, owing to the pattern formed by the atoms along the circumference. The other types of tubes are *chiral tubes*.

Multiwalled carbon nanotubes (MWCNTs) are formed by several coaxial tubes obtained from several rolled layers of graphene.

20.3 Electron States: Bands

To study the electron states in CNTs, it is convenient to start from graphene. The strong σ bonds hold the carbon atoms together in the graphene planes. Thus, they are responsible for most of the binding energy and mechanical properties of the graphene sheet. The orbital p_z, orthogonal to the planes, contribute to the week interaction between the graphene planes in graphite; they also connect the different walls in MWCNTs, and sometimes collect various tubes in bundles. Their major role for us, however, is the contribution to the formation of electron bands near the Fermi level and therefore available for electrical conduction. The electron bands of graphene are shown schematically in Fig. 20.6 [102]. The bands coming from the σ bonds are well below the Fermi level, and the bands coming from the σ^* antibonds are much higher. Thus, if only these bands were present, graphene would be an insulator. The conductivity of graphene is due to the bands coming from the π bonds and π^* antibonds, formed by the p_z orbitals, which touch each other at the Fermi level (see Fig. 20.6). Since the orbitals s, p_x and p_y, which form the σ bonds in the plane, do not couple with the orbitals p_z, it is possible to develop a simple tight-binding model [368] to obtain the bands π and π^* coming from these orbitals. An analytical result is obtained that depends only upon the parameter γ which describes the matrix element of the Hamiltonian between the states p_z of neighboring atoms:

$$\epsilon^{\pm}(k_x, k_y) = \pm\gamma\sqrt{1 + 4\cos\frac{\sqrt{3}k_x a}{2}\cos\frac{k_y a}{2} + 4\cos^2\frac{k_y a}{2}}. \tag{20.2}$$

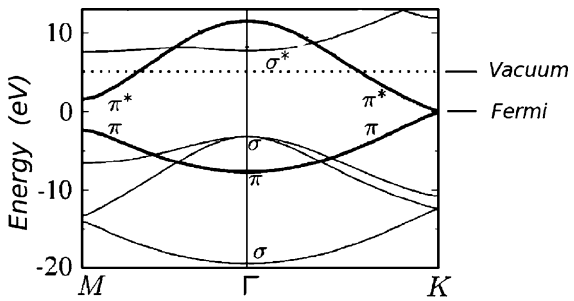


Fig. 20.6. Main bands of graphene [102]. The points Γ, M, and K are defined in Fig. 20.2

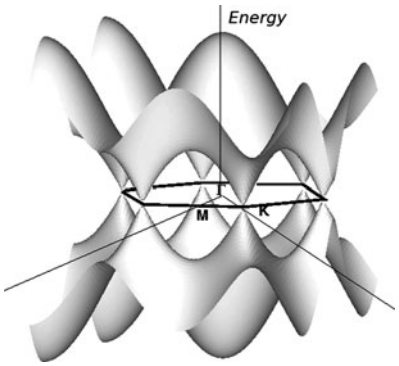


Fig. 20.7. π and π^* bands of graphene, as given in (20.2). The *dark hexagon* indicates the edge of the BZ

These bands are shown in Fig. 20.7. To have a better representation of their shapes, the two bands are shown in a region wider than the first BZ, whose limits are also shown together with the high-symmetry points.

Turning back to the CNTs, we may suppose, as first approximation, that the energies of the electron states are the same of the equivalent states in graphene, assuming that the surface bending does not change them in an essential way. The wavefunctions, however, must obey periodic boundary conditions along the circumferential direction, orthogonal to the tube axis. As a consequence, only some wavevectors $\boldsymbol{k}$ within the BZ are allowed, corresponding to wavelengths submultiples of the circumference. If we take as example the armchair (5,5) CNT, the length of the circumference is $C = 5a\sqrt{3}$, and therefore the possible transverse wavelengths and wavevectors are

$$\lambda_n = \frac{C}{n} = \frac{5a\sqrt{3}}{n} \ , \qquad k_n = \frac{2\pi}{\lambda_n} = \frac{2\pi n}{5a\sqrt{3}}.$$

Thus, the allowed wavevectors within the first BZ have a component orthogonal to the tube axis given by

$$0 \, , \ \pm\frac{2\pi}{5a\sqrt{3}} \, , \ \pm\frac{4\pi}{5a\sqrt{3}} \, , \ \pm\frac{6\pi}{5a\sqrt{3}} \, , \ \pm\frac{8\pi}{5a\sqrt{3}} \, , \ \frac{10\pi}{5a\sqrt{3}}.$$

The last value does not have the double sign since at the edge of the BZ the two corresponding wavevectors are equivalent. Accordingly, the wavevectors $\boldsymbol{k}$ allowed in a (5,5) CNT are those indicated by the vertical lines in Fig. 20.8. Similar considerations lead to the allowed $\boldsymbol{k}$ in a (8,0) CNT, also shown in Fig. 20.8.

From this figure, it is easy to understand why the electrical properties of CNTs depend upon the circumferential vector. In fact, the graphene bands touch each other in the point K at the Fermi level (see Fig. 20.6). If one of the lines representing the allowed $\boldsymbol{k}$s passes through the point K, i.e., if the

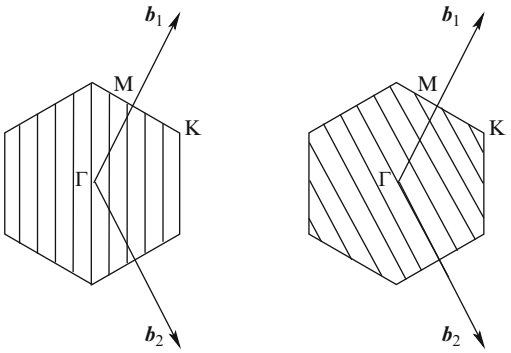


Fig. 20.8. Allowed wavevectors in a (5,5) CNT semimetal (*left*), and in a (8,0) CNT, semiconductor (*right*)

point K represents a possible wavevector, as in the case of (5,5) CNT, the bands of the CNT touch each other at the Fermi level, and the material is a semimetal. If, on the contrary, the lines do not pass through K, the bands have a direct gap and the material is a semiconductor, as in the case of the (8,0) CNT.

In general, (n, m) CNTs with $n = m$ or with $n - m$ multiple of 3 are conductors, the others are semiconductors [102].

It must be added, however, that the curvature of the graphene surface modifies the superposition integrals of the electronic orbitals and therefore the shape of the CNT bands, with respect to the simple theory described above, mainly in CNTs with small diameters [48].[1] On the opposite case, if the radius of the CNT is large, the lines of the allowed $\boldsymbol{k}$s are close to each other, and the possible gap becomes small. As the diameter increases, the situation becomes closer to that of graphene.

Considering now the CNT as a quantum wire, the component of $\boldsymbol{k}$ orthogonal to the axis of the tube describes the electron dynamics along the circumference and therefore the transverse state of the wire. We may thus consider the one-dimensional bands as a function of the component of $\boldsymbol{k}$ along the tube, and we shall have one such "subband" for each orthogonal state, according to the scheme developed in Chap. 19, with its density of state proportional to $1/\sqrt{\epsilon - \epsilon_n}$, as described in Sect. 19.3.

The translational symmetry along the axis of the CNT depends upon the circumferential vector; the unit cell of the tube contains more atoms than in graphene, and the BZ is reduced. Thus, even when the diameter of the CNT is sufficiently large ($>\sim$1 nm) to consider the energies of the electron states

[1] For example, the CNT (5,0) should be semiconductor according to the simple rule above, while it is a metal because of rehybridization of the π bonds produced by the small radius [280].

approximately equal to those of the equivalent states in graphene, the bands must be described in a *zone-folding scheme* [102, 389].

20.4 Electron Transport

Electron transport in CNTs presents aspects of great interest, owing to the large variety of its behavior. We have already seen that a CNT may have properties of a metal or of a semiconductor, according to its geometric properties. At low temperatures, it may even be a superconductor [246, 433, 435].

At room temperature, the resistivity is about $10^{-4} - 10^{-3}\,\Omega\,\mathrm{cm}$ in metallic nanotubes, and about $10^{1}\,\Omega\,\mathrm{cm}$ in semiconducting nanotubes. The semiconducting nanotubes exhibit a dependence of resistivity upon temperature consistent with energy gaps in the range of $0.1-0.3\,\mathrm{eV}$, coherent with the theoretical predictions of the previous section for the corresponding nanotube diameters.

CNTs exhibit high mobilities at low electric fields, exceeding the best semiconductors at room temperature.

Furthermore, they present the various transport regimes (coherent, ballistic, diffusive, and dissipative) discussed in Sect. 18.2. Thus, CNTs constitute ideal systems for both theoretical and experimental physicists to test the various models of transport in mesoscopic systems.

The linear dispersion of the graphene (and metallic CNTs) shown in Fig. 20.6 near the minimum of the conduction band is responsible for one of the most peculiar features of electron transport in such materials. A force applied to an electron modifies its crystal momentum according to the semiclassical equation (7.44). Its velocity, however, given by the slope of the band, does not change, mimicking in this way a massless relativistic Dirac particle with a constant velocity of the order of $10^{6}\,\mathrm{m/s}$.[2]

The great variety and complexity of electron-transport phenomena in CNTs imposes an oversimplification, for reasons of space and clarity, in the following pages.

Contacts

The first problem to solve in the study of electron transport in CNTs is that of contacts. It is not so trivial, after all, to make contacts on single conducting wires with diameters of the order of few nanometers [389]. This problem is, in general, solved by depositing CNTs on a substrate on which some metallic gates had been previously built or depositing first some CNTs on a substrate and then fabricating contacts on one CNT with scanning-microscopy

[2] Special consequences of this fact are, for example, an anomalous "integer" quantum Hall effect corresponding to a half-integer filling factor, predicted theoretically and experimentally verified [331, 486, 487], and an anomalous barrier penetration known as Klein paradox [424].

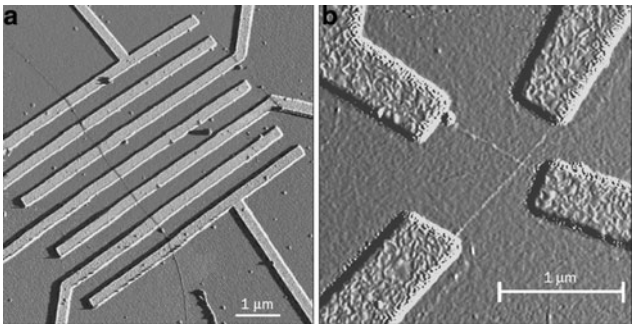


Fig. 20.9. AFM images of carbon nanotubes joining contacts. The devices are fabricated on a conducting substrate (gate) covered with an insulating oxide layer (from P.L. McEuen: Physics World, june 2000, p. 31)

techniques. In this way, CNTs are side contacted. Examples are shown in Fig. 20.9. The technology to wire CNTs, however, is in rapid evolution (see, for example, [224]). Note that, on this respect, it is necessary to know the intrinsic properties of the interface between metal contacts and CNT [8, 9].

Coherent Transport

The coherence length in CNTs is strongly energy dependent and, of course, quite sensitive to temperature, but even at room temperature it can be well over a micron. Thus, for samples sufficiently short and temperatures sufficiently low, electron transport is coherent and can be analyzed with Landauer theory presented in Chap. 21. Conductance quantization as well as weak localization and universal conductance fluctuations have been experimentally observed.

If the CNT is not very long, also the quantization of the longitudinal states may become appreciable, and the system shares some features with quantum dots, among which Coulomb blockade is of particular importance and has been observed experimentally [436].

Dissipative Transport

At higher temperatures, electron transport becomes dissipative, dominated by phonon scattering. This maintains equilibrium or at least, at high-field intensities, steady state, dissipating the power transferred from the field to the electrons.

To treat the phonon contribution to transport, it is of course necessary the knowledge of the dispersion of the various phonon branches. As it was done for the electron states, also for the lattice vibrations of the CNTs it is useful to start from the vibrations of graphene. Then the phonon modes $\boldsymbol{q}$ allowed by the periodicity along the circumference of the CNTs must be considered.

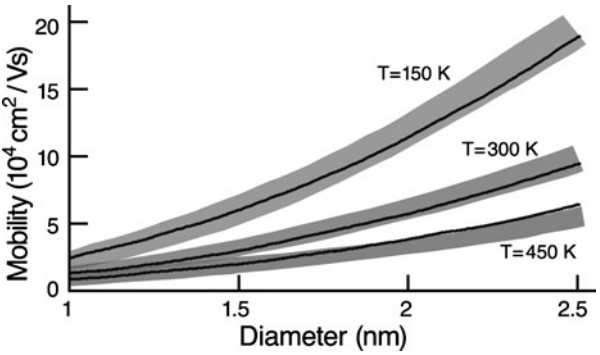


Fig. 20.10. Mobility of electrons in CNTs as a function of tube diameter at the indicated temperatures. The shaded areas are representative of many experimental data obtained in tubes of different chiralities; continuous lines are fitted to a single simple empirical expression [336]

Some of the modes, however, must be reinterpreted, in CNTs, as *breathing* modes or *twisting* modes. Since the unit cell in CNTs contains many atoms, the phonon dispersion curves become very complicated [123].

The electron interaction with phonons has the form of a deformation-potential type [297, 335].

Electron transitions induced by phonon scattering correspond to forward or backward scattering and involve phonon momenta that are close to $q = 0$ or to $\boldsymbol{q} = \boldsymbol{K}$ as effect of energy and momentum conservation.

As in bulk materials, low-energy acoustic phonons dominate electron transport at lower temperatures and fields. At high fields and/or temperatures optical phonons, phonons at zone boundaries, and phonons involved in interband transitions become more important [480].

The low-field mobility depends, as always, upon temperature, and increases at increasing nanotube diameter, as shown in Fig. 20.10. Also important hot-electron effects occur: the mobility is dramatically reduced by optical phonon emission, leading to saturation and negative differential mobility. In the latter case, the peak velocity has been found to be in the range between 2 and 5×10^5 m/s. It depends only slightly upon tube diameter, while the threshold field is more dependent on it, being lower in CNTs with larger diameters, as shown in Fig. 20.11.

Electron–Electron Interaction

Electron–electron interaction is particularly important in CNTs since their quasi-1D nature enhances the strength of Coulomb repulsion, especially in small-diameter CNTs. In metallic tubes, e–e interaction produces collective excitations, i.e., charge density fluctuations that compete, at low temperatures, with the superconducting state [102].

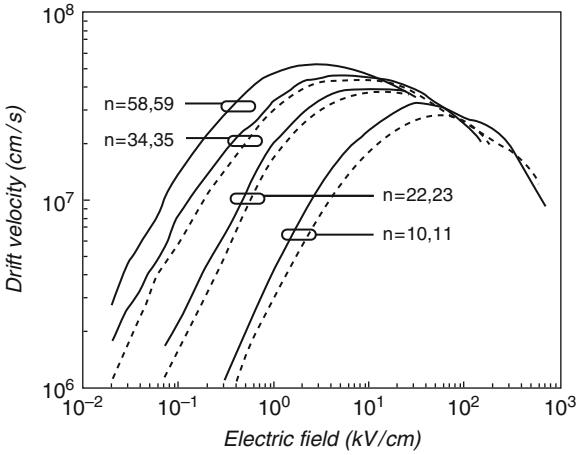


Fig. 20.11. Monte Carlo simulation of electron drift velocity vs electric field in a number of (n,n) zigzag CNTs [335]

Finally, in metallic QNTs electron correlations lead to a Luttinger-liquid behavior, characteristic of one-dimensional conductors [127].

Doping and Impurity Scattering

As in traditional bulk semiconductors, doping is important for the use of CNTs in electronic devices. Doping may be realized with substitutional nitrogen for *n*-type and boron for *p*-type. The carrier mean free path (of the order of 10^2 nm) decreases at increasing doping concentration because of impurity scattering, with a significant reduction in the electronic conduction. To avoid this effect, techniques have been developed in which doping is obtained by means of adsorption of various types of atoms or molecules at the surface of the CNT.

Magnetic Fields

As well known, the presence of a magnetic field changes significantly the electronic states in a material. If a magnetic field is applied parallel to the CNT axis, a number of phenomena emerge, such as band-degeneracy splitting [102] and Aharonov–Bohm-like quantum-phase interference, directly related to the magnetic flux inside the CNT.

Multiwalled Carbon Nanotubes

In MWCNTs, the first observed experimentally, several graphene sheets form coaxial tubes. The periodicities along the axis of the different walls can be

commensurate or incommensurate. Commensurate MWCNTs posses an overall periodicity, and it is therefore possible to describe the electron dynamics in terms of Bloch states, bands, and semiclassical dynamics. In *incommensurate* MWCNTs, electron dynamics is more complex with different time evolutions of the components of an electron wavepacket in the different walls. Furthermore, interwall transitions make transport a very complicated phenomenon.

Devices

Several electron devices of mesoscopic nature have been proposed, based on CNTs, such as low-resistance ballistic interconnects, probes for scanning spectroscopy, rectifying diodes, high-mobility field-effect transistors, light-emitters, single-electron detectors, and single-electron memories. Furthermore, intense infrared emission has been observed from electrically induced excitons in CNT field-effect transistors. Finally, by depositing on the surface or inserting inside the CNTs various types of ions or molecules, several chemical sensors have been obtained. CNTs have also been proposed as systems for hydrogen storage.

21

Coherent Transport in Mesoscopic Structures

Coherent transport has been defined in Sect. 18.2 as one described by Schrödinger equation, without phase-breaking collisions. It may be realized in both semiconducting and metallic systems. In such conditions, conductivity is not a well-defined quantity, since no local relation such as $j = \sigma E$ exists. In linear-response regime, it is possible to define a conductance between two contacts, as $G = I/V$, and a resistance $R = 1/G$, where I is the current and V the potential difference between the two contacts, but the resistance is not given by a local resistivity ρ, which is not defined.

21.1 Landauer–Büttiker Theory of Transport

Let us consider a system formed by a conductor S connected to a number of leads, as shown in Fig. 21.1. The leads are in contact with reservoirs maintained at electric potentials V_μ, where μ is the label, which identifies the lead. The contacts are supposed to be *reflectionless*, meaning that the electrons can enter the reservoir from the lead without appreciable reflection. This is justified by the fact that in the reservoirs the density of states is much higher than in the leads. The opposite is not true; from the reservoirs to the leads the reflection probability is very large, and this gives rise to *contact resistance*.

Our purpose is to find the conductance of S, i.e., the current going through each lead as a function of all the V_μ, under coherent-transport, steady-state, conditions.

The theory that will be presented in this section was formulated by Landauer in 1957 [262, 263] for a conductor with two terminals and extended by Büttiker [83] for the general case of n terminals.

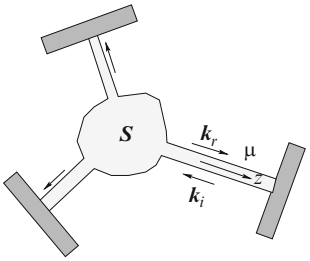


Fig. 21.1. A conductor with several leads described by Landauer–Büttiker theory. A scattering state is an eigenstate of the total Hamiltonian corresponding to an incident wave in one channel of one lead, and outgoing waves in all channels of all leads

Scattering Matrix and Transmission Coefficients

The basic idea is to consider incoming and outgoing wavefunctions in the quantum wires that form the leads, and treat the problem with the general scattering theory.

The set of energy eigenfunctions of the total system, conductor plus leads, contain states with incoming and outgoing terms in each lead, besides possible states localized in the conductor decaying exponentially (*evanescent states*) in the leads.

A transverse state in a quantum wire that forms a given lead is called a *channel*.

It is convenient, for our purpose, to consider the *scattering states*, eigenstates of the total Hamiltonian formed by waves coming toward the conductor in one channel of one lead, waves reflected in all channels of the same lead and waves transmitted in all channels of the other leads, as shown in Fig. 21.1. The shapes of such states inside the conductor are not relevant for the present general discussion. They will be relevant, of course, when specific solutions for a well-defined conductor are sought. This problem will be studied in the last Chap. 28.

Let us indicate the various leads with the Greek letters $\mu,\ \nu, \ldots$; the channels, i.e., the orthogonal states in each wire, corresponding to the subbands of the wires, are indicated with Latin letters $a,\ b, \ldots$; finally, the first Greek letters $\alpha,\ \beta, \ldots$ indicate the terns like (μ, a, k) which specify the longitudinal state k in channel a of lead μ.

The wavefunctions in the channel a, far from the conductor, are given by (cf. (19.10))

$$\psi_{ak}(x, y, z) = \frac{1}{\sqrt{L}} \mathrm{e}^{\mathrm{i}kz} \zeta_a(x, y), \tag{21.1}$$

where L is a normalization length of the wire, and a reference frame of the wire is used with z along the outgoing direction of the wire, as shown in Fig. 21.1; $\zeta_a(x, y)$ is the orthogonal eigenfunction defining channel a. In a

simple effective-mass approximation, the energy associated with the above wavefunction is

$$\epsilon_{ak} = \epsilon_a + \frac{\hbar^2 k^2}{2m},$$

where ϵ_a is the energy associated to the orthogonal state. The group velocity in a channel is given by $v_{\rm g} = \hbar k/m$, while the density of states, including spin multiplicity, is $g(\epsilon) = Lm/\pi\hbar^2 k$, half of the expression in (19.11) to account for only one direction of k. Thus, for each channel, the product of the group velocity by the density of states, per unit length of the wire, is a universal constant, independent of the material and the shape of the wire:

$$\boxed{\tfrac{1}{L} g(\epsilon) v_{\rm g}(\epsilon) = \tfrac{2}{h}} \tag{21.2}$$

The simplicity of this result is at the basis of most of the results of this chapter.

The total wavefunction in lead μ, far from the conductor, may be written as a linear combination of the wavefunctions of the different channels and k states in that lead:

$$\psi_\mu = \sum_{a\in\mu,k_i} A_{\alpha_i} \sqrt{\frac{L}{v_{\alpha_i}}} \psi_{a\,k_i} + \sum_{a\in\mu,k_r} B_{\alpha_r} \sqrt{\frac{L}{v_{\alpha_r}}} \psi_{a\,k_r}, \tag{21.3}$$

where α_i is the tern (μ, a, k_i) and v_{α_i} is the group velocity of state α_i. Negative and positive values of k have been separated and called k_i and k_r, respectively. The coefficients in the amplitudes have been chosen in such a way that the entering particle flux of the state α_i is given by

$$|A_{\alpha_i}|^2 \frac{L}{v_{\alpha_i}} \int |\psi_{\alpha_i}|^2 v_{\alpha_i}\, {\rm d}x\, {\rm d}y = |A_{\alpha_i}|^2,$$

where (21.1) has been used, and the orthogonal wavefunction is taken normalized to unity. This is similar for the outgoing states.

From the linearity of the Schrödinger equation, the coefficients B are proportional to the coefficients A:

$$B_\alpha = \sum_\beta \mathcal{S}_{\alpha\beta} A_\beta, \tag{21.4}$$

where $\mathcal{S}_{\alpha\beta}$ is called the *scattering matrix*. For the conservation of the number of particles, it is necessary that the sum of the incoming fluxes is equal to the sum of the outgoing fluxes, or

$$\sum_\alpha |A_\alpha|^2 = \sum_\beta |B_\beta|^2 = \sum_\beta \sum_{\gamma\eta} \mathcal{S}^*_{\beta\gamma} \mathcal{S}_{\beta\eta} A^*_\gamma A_\eta,$$

or

$$\sum_\alpha A^*_\alpha A_\alpha = \sum_{\gamma\eta} (\mathcal{S}^\dagger \mathcal{S})_{\gamma\eta} A^*_\gamma A_\eta,$$

where $\mathcal{S}^\dagger$ is the hermitian conjugate of $\mathcal{S}$. Since this relation must be true for arbitrary values of the coefficients A, it is necessary that $\mathcal{S}^\dagger\mathcal{S} = \mathcal{I}$, i.e., the scattering matrix is unitary:

$$\mathcal{S}^\dagger = \mathcal{S}^{-1}.$$

Furthermore, time-reversal symmetry requires that the state that proceeds backward in time (taking the final state as initial state with opposite k, the system follows the same evolution with opposite time direction) is described by the complex conjugate wavefunction. Thus it must be $A^*_\alpha = \sum_\beta S_{\alpha\beta} B^*_\beta$. Combining with (21.4), we obtain

$$A^*_\alpha = \sum_\beta S_{\alpha\beta} \sum_\gamma S^*_{\beta\gamma} A^*_\gamma,$$

which requires

$$\mathcal{S}^* = \mathcal{S}^{-1}.$$

If a magnetic field $\boldsymbol{B}$ is present, it changes sign for time reversal, so that the above becomes

$$\mathcal{S}(\boldsymbol{B})^* = \mathcal{S}(-\boldsymbol{B})^{-1}.$$

The probability that an electron entering the conductor from a state β exits from α is easily found by considering the scattering state with only A_β different from zero and equal to unity in (21.4) and taking the square modulus of the resulting B. The result is the *transmission coefficient* from β to α:

$$\boxed{T_{\alpha\beta} = |S_{\alpha\beta}|^2} \tag{21.5}$$

The same result would be obtained by considering fluxes instead of single electrons. The sum of the probabilities over all possible final outgoing states must be one,[1] so that

$$\sum_\alpha T_{\alpha\beta} = 1 = \sum_\beta T_{\alpha\beta}, \tag{21.6}$$

where the second equality results from the unitarity of the scattering matrix.

Since for a given energy the longitudinal wavevector in any given channel is well defined (also its direction is well defined in connection with the transmission coefficient), the tern $\alpha \equiv (\mu, a, k)$ may be substituted by (μ, a, ϵ) in the transmission coefficients; $T_{\alpha\beta}$ may be written as $T_{ab}(\epsilon)$, and (21.6) becomes

$$\boxed{\textstyle\sum_a T_{ab}(\epsilon) = \sum_b T_{ab}(\epsilon) = 1} \tag{21.7}$$

For a given energy, $T_{ab}(\epsilon)$ indicates the transmission from a state entering the conductor from channel b with energy ϵ into a state leaving the conductor in channel a with the same energy.

[1] In coherent motion, the particle cannot be captured by a state localized inside the conductor because of energy conservation.

Conductance Coefficients

The particle current in channel a carried by electrons with energy between ϵ and $\epsilon + \mathrm{d}\epsilon$ is given by the density of particles with that energy times their group velocities times the section s of the wire. Taking for reference a wire of length L, the total number of entering states is $g_a(\epsilon)\mathrm{d}\epsilon$, so that the density of such particles is $f_a(\epsilon)g_a(\epsilon)\mathrm{d}\epsilon/Ls$, where $f_a(\epsilon)$ is the fraction of occupied entering states in channel a. The electrical current in channel a is then

$$I_a = (-e)\int f_a(\epsilon)g_a(\epsilon)\frac{1}{L}v_{\mathrm{g}}(\epsilon)\,\mathrm{d}\epsilon - \sum_b (-e)\int f_b(\epsilon)g_b(\epsilon)\frac{1}{L}v_{\mathrm{g}}(\epsilon)T_{ab}(\epsilon)\,\mathrm{d}\epsilon.$$

The first term is the current entering the conductor from channel a; the second term is the current leaving the conductor from the same channel, coming from all channels and transmitted into channel a. This second term contains also, when $b = a$, the electrons coming from channel a and reflected back into the same channel. Using (21.2), the above becomes

$$I_a = (-e)\frac{2}{h}\sum_b \int f_b(\epsilon)\left[\delta_{ab} - T_{ab}(\epsilon)\right]\,\mathrm{d}\epsilon = \frac{1}{(-e)}\sum_b \int f_b(\epsilon)\Gamma_{ab}(\epsilon)\,\mathrm{d}\epsilon, \quad (21.8)$$

where

$$\Gamma_{ab}(\epsilon) = \frac{2e^2}{h}\left[\delta_{ab} - T_{ab}(\epsilon)\right].$$

Let us now assume that the different reservoirs are maintained at different electrochemical potentials ϵ_μ. It is important to realize that in such conditions the electron population in each lead μ is not in equilibrium, since electrons coming from the different reservoirs belong to distribution functions with different electrochemical levels. However, since the contacts are supposed to be reflectionless, entering electrons are distributed according to the equilibrium distributions of their reservoirs, so that the distribution functions appearing in (21.8) can be written as the Fermi function

$$f_a = \frac{1}{\mathrm{e}^{\beta(\epsilon-\epsilon_a)}+1},$$

where ϵ_a is the electrochemical potential of the reservoir injecting electrons in channel a. Let us also define a reference electrochemical potential $\epsilon_\circ$, so that the electrochemical potentials of the various reservoirs can be written as

$$\epsilon_a = \epsilon_\circ - eV_a.$$

If the potential differences are sufficiently small, as in linear-response regime, the Fermi distributions can be expanded to first order as

$$f_a = f_\circ(\epsilon) - f'_\circ(\epsilon)eV_a,$$

where $f'_\circ(\epsilon)$ is the derivative of f with respect to the electrochemical potential, evaluated at $\epsilon_\circ$. Since $f_\circ$ in all channels does not produce current, the current in (21.8) becomes

$$I_a = \sum_b \int f'_\circ(\epsilon) V_b \Gamma_{ab}(\epsilon)\, \mathrm{d}\epsilon,$$

or

$$\boxed{I_a = \textstyle\sum_b G_{ab} V_b, \quad G_{ab} = \int f'_\circ \Gamma_{ab}(\epsilon)\, \mathrm{d}\epsilon} \tag{21.9}$$

The coefficients G_{ab} are the wanted conductance coefficients.

In case of temperatures so low that the Fermi distribution can be approximated by the step function, we have (remember that the derivative is made with respect to the electrochemical potential)

$$f'_\circ \approx \delta(\epsilon - \epsilon_\circ),$$

and the second of the (21.9) becomes

$$\boxed{G_{ab} = \Gamma_{ab}(\epsilon_\circ) = \tfrac{2e^2}{h}\left[\delta_{ab} - T_{ab}(\epsilon_\circ)\right]} \tag{21.10}$$

In the simplest case of a conductor with two leads with only one channel each, with $V_a = 0$ e $V_b = V$, the above becomes

$$\boxed{I = GV \quad \text{with} \quad G = \tfrac{2e^2}{h} T(\epsilon_\circ)} \tag{21.11}$$

where T is the transmission coefficient between the two leads. Note that for the symmetry properties of the scattering matrix found above, in absence of magnetic fields, $T_{ab} = T_{ba}$.

Equation (21.11) is written in terms of the potential difference between the reservoirs. It is reported as *Landauer* equation, although the equation originally given by Landauer was different, written in terms of the potential difference at the edges of the conductor. The difference is somewhat subtle, since, as indicated above, the electrochemical potential is not perfectly defined in the leads, where electrons are not distributed according to a single equilibrium distribution. This point has been widely discussed in the literature. The interest reader may consult [116] for a thorough discussion. Here, a simplified version of the arguments given in [85] and [143] will be presented. The situation is schematically illustrated in Fig. 21.2. On the basis of the total electron densities, (fictitious) Fermi distributions with Fermi levels ϵ_L and ϵ_R are attributed to the left and right leads, respectively, such that the total carrier density in the left lead is given by

$$\int g_\mathrm{L}(\epsilon)\left[f_a(\epsilon)(1 + R(\epsilon)) + f_b T(\epsilon)\right] \mathrm{d}\epsilon = 2\int g_\mathrm{L}(\epsilon) f_\mathrm{L}(\epsilon)\, \mathrm{d}\epsilon,$$

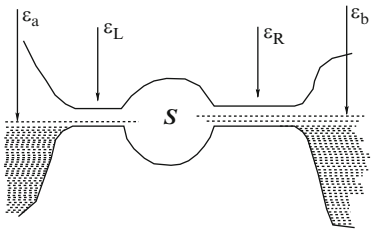


Fig. 21.2. A conductor with two terminals described by Landauer–Büttiker theory. During coherent conduction, electrochemical potentials are not well defined in the leads since incoming and outgoing electrons belong to different equilibrium distributions

where $R(\epsilon) = 1 - T(\epsilon)$ is the reflection coefficient, g_{L} is the density of states in the left lead, and the factor 2 in the r.h.s. recovers positive and negative wavevectors. Similarly, in the right lead

$$\int g_{\mathrm{R}}(\epsilon)\,[f_b(\epsilon)(1 + R(\epsilon)) + f_a T(\epsilon)]\,\mathrm{d}\epsilon = 2\int g_{\mathrm{R}}(\epsilon) f_{\mathrm{R}}(\epsilon)\,\mathrm{d}\epsilon.$$

Now let us assume the same density of states in the two leads and take the difference of the two above equations. Considering that the temperature is so low that the difference of the integrals is given by the integrand times the difference of the Fermi levels, the result is

$$(\epsilon_a - \epsilon_b)(1 + R - T) = 2(\epsilon_{\mathrm{L}} - \epsilon_{\mathrm{R}}),$$

or

$$(\epsilon_a - \epsilon_b) = (\epsilon_{\mathrm{L}} - \epsilon_{\mathrm{R}})/(1 - T).$$

If this result is inserted in place of V into (21.11), the result is the original Landauer equation:[2]

$$I = G(\epsilon_{\mathrm{L}} - \epsilon_{\mathrm{R}}) \quad \text{with} \quad G = \frac{2e^2}{h}\frac{T}{1-T}. \tag{21.12}$$

Note that if $T = 1$ the conductance in (21.12) diverges, the resistance vanishes, and the potential drop is entirely due to the contact resistance. The value of this contact resistance can be obtained from (21.11), which can be written as

$$G^{-1} = \frac{h}{2e^2} + \frac{h}{2e^2}\frac{1-T}{T}. \tag{21.13}$$

This may be interpreted as a series of two resistances: the first term, which remains when $T = 1$, is the contact resistance, and the second term is the resistance due to the finite transmission coefficient of the conductor.

[2] It is to remember, however, that in the leads electrochemical potentials are not well defined, and if we measure their electric potentials, the result depends on how the measurement is performed [116].

The energy furnished by the voltage difference is dissipated in the reservoirs when the electrons coming from the conductor thermalize.

If in the leads μ and ν several channels are active, they behave as conductors in parallel, and a transmission function may be defined as

$$\overline{T}_{\mu\nu}(\epsilon) = \sum_{a\in\mu}\sum_{b\in\nu} T_{ab}(\epsilon). \tag{21.14}$$

In the general case of multilead, multichannel leads at finite temperatures, the total current in lead μ is thus

$$\begin{aligned} I_\mu &= \frac{2(-e)}{h}\sum_{a\in\mu}\sum_{\nu}\sum_{b\in\nu}\int \left[T_{ba}(\epsilon)f_\mu(\epsilon) - T_{ab}(\epsilon)f_\nu(\epsilon)\right]\,\mathrm{d}\epsilon \\ &= \frac{2(-e)}{h}\sum\int \overline{T}_{\mu\nu}(\epsilon)\left[f_\mu(\epsilon) - f_\nu(\epsilon)\right]\,\mathrm{d}\epsilon, \end{aligned} \tag{21.15}$$

where it has been taken into account that since there is no current at equilibrium, the transmission functions must satisfy the sum rule:

$$\sum_{\nu}\overline{T}_{\mu\nu}(\epsilon) = \sum_{\nu}\overline{T}_{\nu\mu}(\epsilon).$$

21.2 Point Contacts

A *point contact* is a contact between two conductors with transverse dimension comparable with the wavelengths of the electrons which participate to charge transport. The electron wavefunctions transverse to the contact width are therefore quantized, as shown in Fig. 21.3.

If the point contact is seen as a conductor, Landauer equation (21.11) becomes particularly simple, since every electron that enters the contact exits on the other side, and therefore $T = 1$. Thus,

$$G = \frac{2e^2}{h}n, \tag{21.16}$$

where n is the number of active channels, i.e., of transverse states occupied by electrons. As the electron energy in the lower contact in Fig. 21.3 is increased,

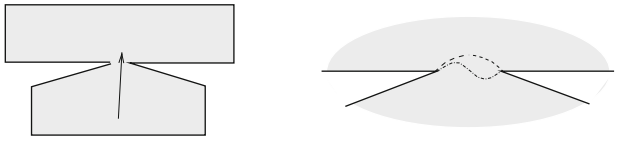


Fig. 21.3. Point contact – transverse wavefunctions of the first two channels

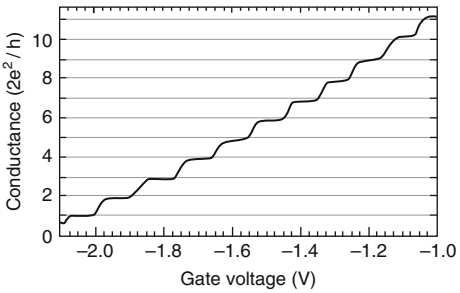


Fig. 21.4. Conductance in a point contact in units of $2e^2/h$, from [461]. The gate voltage is proportional to the difference of the Fermi levels of the two conductors connected by the point contact

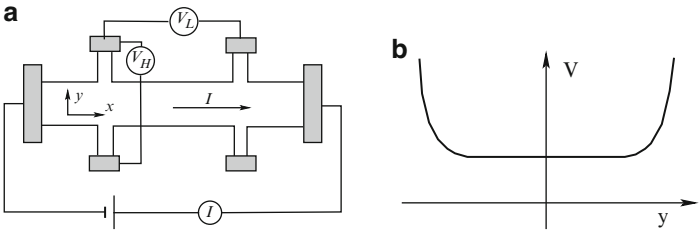


Fig. 21.5. Hall bar (**a**) and potential confining the electrons along the transverse direction (**b**). The Hall bar may have linear dimensions of the order of hundreds of microns

more channels become active and G increases of discrete quantities given by the *fundamental conductance unit* $2e^2/h$. In the right part of Fig. 21.3, the transverse wavefunctions of the first two channels are depicted.

Figure 21.4 shows experimental results in [461], where the quantization of the conductance of a point contact is clearly evident.

21.3 Quantum Hall Effect

In Sect. 11.3.4, we studied the classical Hall effect. With reference to Fig. 21.5a, we saw that if the current flows along the x direction with a magnetic field $\boldsymbol{B}$ along the direction z orthogonal to the Hall bar, a potential difference V_{H} is generated along the y direction, orthogonal to the direction of the current and to $\boldsymbol{B}$.

According to the simple theory developed in classical terms, all conductivity coefficients are proportional to the carrier concentration n. Thus, keeping constant the current, both Hall and longitudinal voltages should decrease, at increasing carrier concentration, as $1/n$. At high magnetic fields, however, quantum theory must consider the quantization of Landau levels, treated in Appendix D.

The Experimental Effect

In 1980, K.V. Klitzing and coworkers discovered a new effect [244], now known as *quantum Hall effect* (QHE), when the degenerate electron gas in the inversion layer of a MOSFET operated at low temperature in a strong magnetic field. Their results are shown in Fig. 21.6, where it can be seen that at increasing carrier concentration, the Hall voltage V_{H} presents a number of plateaux where the resistance assumes values given by

$$R_\nu = \frac{h}{e^2}\,\frac{1}{\nu}, \tag{21.17}$$

with ν integer. Corresponding to these plateaux, the potential drop V_{L} between two points along the Hall bar vanishes, showing a zero longitudinal resistance. The effect is present also if the applied magnetic field is varied, as shown in the right part of Fig. 21.6.

The effect was later found to be present in all kinds of materials containing a two-dimensional electron gas (2DEG) and also for fractional values of ν in (21.17) [452]. For this reason, two QHEs are now defined, and *integer quantum Hall effect* (IQHE) and a *fractional quantum Hall effect* (FQHE). Several books are available on the subject (see, for example, [154, 426, 482]).

It may immediately be noted that (21.17) corresponds to the conductance quanta in (21.16). As we shall shortly see, the factor 2 is missing because the magnetic field separates the channels with opposite spins so that the density of states does not contain the spin multiplicity.

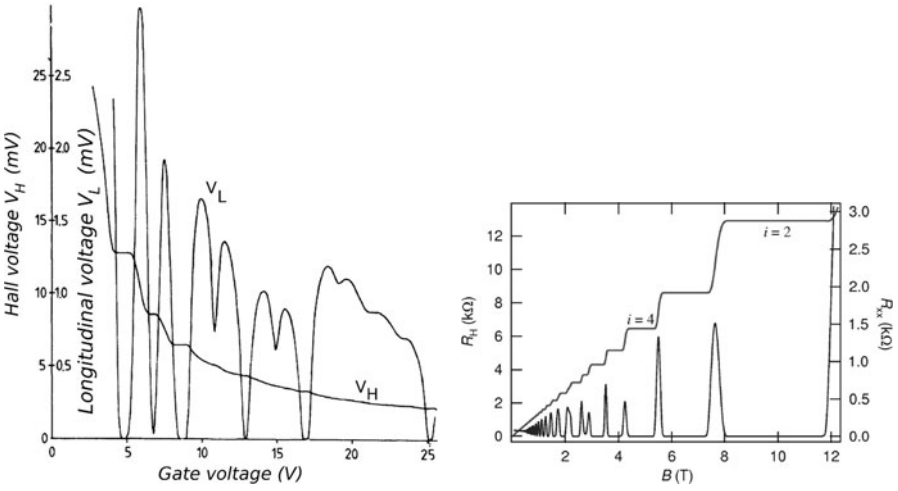


Fig. 21.6. Quantum Hall effect (QHE). *Left*: original results obtained in a MOSFET at 1.5 K [244]. The gate voltage controls the carrier concentration. *Right*: more recent results obtained in a GaAs/AlGaAs structure at 0.1 K varying the magnetic field [220]

The resistances of the plateaux are determined with such an experimental precision to become the definition of the resistance standard [220].

Edge States

It was already noted that at high magnetic fields electron transport must be treated in quantum terms. The presence of Planck constant h in the experimental results (21.17) confirms that the effect must be of a quantum nature.

Let us then consider the quantum theory of such phenomenon, assuming that the Hall bar shown in Fig. 21.5a is so thin that along the z direction, orthogonal to the Hall bar, the states are quantized, and only the lowest state is occupied at the low considered temperature, so that a perfect 2DEG is realized.

In Appendix D, it is shown that in the presence of a strong magnetic field, Landau levels are formed with energies $\epsilon_n = (n + 1/2)\hbar\omega_c$, with n integer, where ω_c is the cyclotron frequency eB/m. In an infinite sample, the number of degenerate Landau states per unit area is given by eB/h (see (D.4)). The unperturbed eigenstates are the Landau states (see (D.2)):

$$\psi_{n\alpha} = \mathrm{e}^{\mathrm{i}\alpha x}\psi_n(y - y_\circ),$$

where $y_\circ = -\hbar\alpha/qB$ is the y coordinate of the center of the Landau state.

Now we must consider that the sample has finite dimensions. Let $V(y)$ be the potential that confines the electrons along the y direction, transverse to the Hall bar, as in Fig. 21.5b. Far from the edges of the bar, Landau states can be described as in an infinite sample. However, when the centers of the clssical orbits get close to the rising part of the confining potential the situation becomes different. Figure 21.7 shows that this change is present also in classical terms: while close cyclotron orbits are present in the internal part of the bar, near the borders skipping orbits exist which determine an overall motion along the edges of the bar and of the leads.

To study this phenomenon in quantum terms, we consider the Schrödinger equation for the energy eigenstates of the 2DEG obtained from (D.1) in

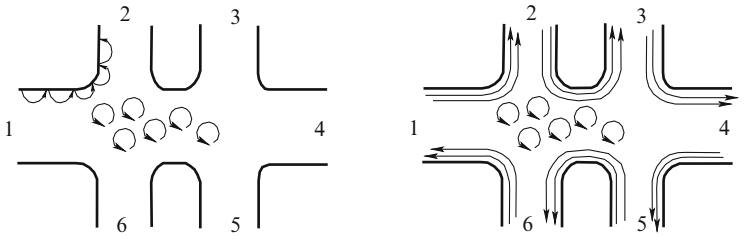


Fig. 21.7. Classical cyclotron and skipping orbits (*left*); Landau and edge states (*right*)

Appendix D, adding the confining potential $V(y)$:

$$\left[\frac{p_y^2}{2m} + \frac{1}{2}m\omega_c^2(y - y_\circ)^2 + V(y)\right]\phi(y) = \epsilon\phi(y). \tag{21.18}$$

Let us consider the first two terms in (21.18) as an unperturbed Hamiltonian and the last term, representing the confining potential, as a perturbation. According to first-order perturbation theory, when $y_\circ$ gets close to the edges, the energies of the eigenstates increase by an amount given by the average value of V evaluated in the unperturbed Landau states. Without detailed calculations, if the cyclotron radius is small enough that the confining potential can be assumed to be nearly constant on this scale, the energy correction is approximately the value of the confining potential in $y_\circ$. Thus, the spectrum of the eigenvalues, as a function of the wavevector α along the x direction is now given by

$$\epsilon_n(\alpha) \approx \left(n + \frac{1}{2}\right)\hbar\omega_c + V(y_\circ(\alpha)) \tag{21.19}$$

and takes the shape shown in Fig. 21.8.

The group velocity of a Landau state is given by

$$v_g(n,\alpha) = \frac{1}{\hbar}\frac{\partial}{\partial\alpha}\epsilon_n(\alpha) = \frac{1}{\hbar}\left(\frac{\partial}{\partial y_\circ}V(y_\circ)\right)\frac{\partial y_\circ}{\partial\alpha} = \frac{1}{eB}\frac{\partial V}{\partial y}\bigg|_{y_\circ}. \tag{21.20}$$

For Landau states well within the bar, where $y_\circ$ corresponds to the flat region of $\epsilon_n(\alpha)$ in Fig. 21.8, the drift velocity is zero, as for classical closed cyclotron orbits. When the state is near the border of the bar, the drift velocity in (21.20) is different from zero, as shown in Fig. 21.8. When $y_\circ$ approaches the top or the bottom edge of the bar the velocity is positive or negative, respectively. The resulting *edge states* [84], are shown in the right part of Fig. 21.7.

Fermi Energy and Landau-Level Filling

Let us assume that, for a given carrier concentration and a given magnetic field B, the n-th Landau level is partially occupied, while the lower levels

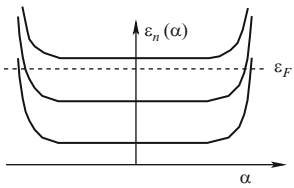


Fig. 21.8. Landau levels as a function of the wavevector α in presence of the confining potential. The Fermi energy indicated by the *dashed line* corresponds to a situation with the second Landau level totally filled, and the third one totally empty (filling factor = 2)

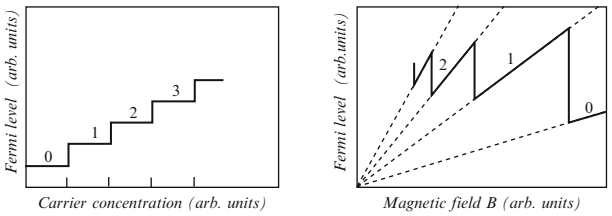


Fig. 21.9. Fermi level in a 2DEG bar as a function of carrier concentration (*left*) and of magnetic field (*right*). Numbers indicate the highest occupied Landau level, where the Fermi level is located

are totally occupied and the higher ones totally empty. In this situation, the Fermi level must coincide with the Landau level partially occupied.

If, keeping constant the magnetic field, the carrier concentration is increased, the n-th Landau level will eventually be totally occupied; electrons will begin to occupy the next Landau level, which will be the new position of the Fermi level. Thus, the Fermi level performs a series of jumps as shown in the left part of Fig. 21.9.

If, instead, starting from the initial situation indicated above, the magnetic field is increased keeping the carrier concentration constant, the n-th level partially occupied rises linearly with B according to (21.19) and so does the Fermi level. At the same time, however, each Landau level accepts more and more states according to (D.4), until eventually the lower levels empty the n-th one, and the Fermi level falls into the $(n-1)$-th as shown in the right part of Fig. 21.9. This oscillation of the Fermi energy as a function of the magnetic field is what generates also the Shubnikov–de Haas and de Haas–van Halphen oscillations mentioned in Sect. 11.4.

The expression *filling factor* is used to indicate the ratio between the number N_e of electrons in the 2DEG and the number N_ϕ of available states in each Landau level. For integer filling factors, the highest occupied Landau level is totally filled and the next one up is totally empty.

Integer QHE from Landauer–Büttiker Theory

On account of the foregoing, the integer QHE finds a simple and clear explanation in terms of edge states and Ladauer–Büttiker conduction theory. We have a conductor (the Hall bar) with six terminals, indicated in Fig. 21.7 with the numbers from 1 to 6. The active channels at the Fermi level are the filled edge states, and their number is the number of filled Landau levels, as shown in Fig. 21.8.

If the occupation of the Landau states is that shown in Fig. 21.8 where all the states of the highest occupied level well-inside the bar are filled, the electrons cannot be scattered away from their states since, at low temperature, they cannot get the necessary energy from the phonons to reach the

next Landau level. They cannot be scattered into edge states going along the different direction either, because such states are spatially separated. From Fig. 21.7, it is then clear that in absence of scattering the only transmission coefficients different from zero are, for all active channels,

$$T_{21} = T_{32} = T_{43} = T_{54} = T_{65} = T_{16} = 1.$$

Consider now the results in (21.10) applied to such a conductor, taking into account that at very low temperature and low applied voltages, all quantities are calculated at the Fermi energy. With the transmission coefficients above, the conductance coefficients, given by (21.10), are

$$G_{ii} = \frac{e^2}{h} n, \ i = 1, 2, \ldots 6,$$

where n is the number of active channels, and

$$G_{21} = G_{32} = G_{43} = G_{54} = G_{65} = G_{16} = -\frac{e^2}{h} n.$$

As mentioned before, in presence of a magnetic field, each channel has a density of state which is half of the one used in (21.2) because the two spin orientations correspond to different energies.[3]

In Hall experiments, the lateral contacts are used to measure the potential differences and no current is passing through them so that

$$I_2 = I_3 = I_5 = I_6 = 0.$$

In particular, considering the third lead, from (21.9),

$$0 = I_3 = \sum_\beta G_{3\beta} V_\beta = \frac{ne^2}{h} \sum_\beta (\delta_{3\beta} - T_{3\beta}) V_\beta = \frac{ne^2}{h} (V_3 - V_2),$$

which means $V_3 = V_2$, and similarly

$$V_1 = V_2 = V_3 = V; \qquad V_4 = V_5 = V_6 = V'.$$

[3] The energy separation between two successive Landau levels of equal spin orientation is given by $\Delta\epsilon = \hbar\omega_c$. The energy separation between two electrons in the same orbital state and opposite spins is given by $g\mu_B B$, where $g\mu_B$ is the magnetic moment associated with the electron spin, μ_B is the Bohr magneton $e\hbar/2m$, and g is a factor whose value is 2 for free electrons. Inside a crystal, however, owing to the spin-orbit coupling, the g-factor assumes different values for different materials [483]. For simplicity, here we shall consider Landau levels with single-spin density of states without considering their exact energy positions.

From the above, we obtain that no potential difference is measured between lateral leads on the same side of the bar, which means

$$\boxed{R_{xx} = 0} \tag{21.21}$$

If we now consider the current in leads 1 and 4, we obtain

$$I = I_4 = G_{44}V_4 + G_{45}V_5 = \frac{ne^2}{h}(V' - V).$$

The potential difference $V' - V$ is the measured Hall potential and corresponds to a resistance given by

$$\boxed{R_{xy} = \frac{h}{e^2 n}} \tag{21.22}$$

where n is the number of filled Landau levels. This is the result found experimentally, as indicated in (21.17).

Effect of Impurities in the Plateaux

In the previous discussion of the QHE, we have considered only the case of integer filling factors, i.e., situations where the highest occupied Landau level is totally filled, and the next one totally empty. When the electron concentration and the magnetic field are such that one Landau level is partially filled, imperfections should be able to scatter electrons from one cyclotron orbit to a different one, and from edge states to internal cyclotron states, producing the standard magneto-conductivity. In such a picture, plateaux should not be present where the filling factor moves from one integer value to the next one. The regions where the longitudinal resistance vanishes should reduce to points for integer values of the filling.

The presence of the plateaux can be understood, however, in terms of localized states induced by disorder [13]. Because of this disorder due, for example, to the presence of impurities, defects, and imperfections, not all Landau states have the same identical energy. The density of states, whuch would be a number of delta functions at the energies of the Landau levels in a perfect crystal, become a series of broadened peaks, as shown in Fig. 21.10. States at the borders of such peaks are localized states. They correspond to classical orbits near the bottom of the valleys or near the top of the hills of the energy profile in the Hall bar. At increasing filling factor, from an integer value to the next one, localized states are first occupied (see Fig. 21.10) which do not contribute to the conduction. Thus, the conduction does not increase, and a plateau is generated. Then extended states starts to be filled near the next Landau level, and the Hall conduction increases toward the next plateau and the longitudinal conduction is not zero. When the electrons start to occupy localized states of the higher energies of the new Landau level, the conduction stops increasing and a new plateau begins. From one plateau to the next also

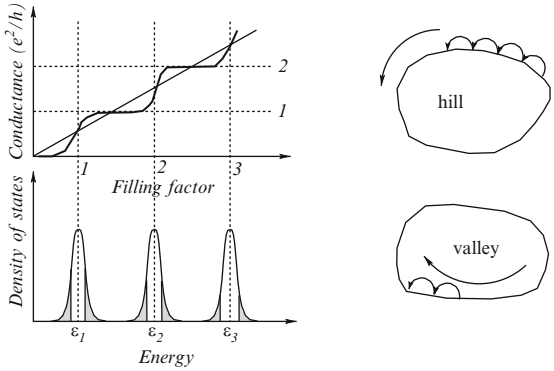


Fig. 21.10. Plateaux in the QHE are due to localized state around hills or valleys in the disordered 2D potential profile

edge states are less conductive (the transmission coefficients become less than one) because electrons can scatter from one side of the sample to the other side through the allowed energy states in the interior.

An Important Note

The striking precision reached by experiments in determining the value of the quantum of conduction is much greater than one would expect from the theoretical explanation given here, based on approximations not rigorously justified, such as single-particle Hamiltonian, effective mass approximation, etc. All this suggests that a deeper explanation of the QHE must exist, as is in fact the case. Explanations based on topological properties of the systems, gauge invariance and Berry phase can be found in the indicated general references. See also [19, 264, 265].

Fractional Quantum Hall Effect

The FQHE is more complicated to explain than the IQHE, and it requires to take into account electron–electron interaction with many-body techniques. In fact, it has been shown that at very low temperatures many-body eigenstates of the Hamiltonian with e–e interaction exist with lower energy with respect to uncorrelated products of single particle eigenstates. Electrons repel each other via Coulomb interaction and tend to form an ordered state, called Wigner crystal. These eigenstates correspond to fractional filling factors, so that the plateaux of the Hall conductance appear with fractional ν factors in (21.17). We shall not enter in more detail into this effect, clearly beyond the scope of this textbook.

21.4 Aharonov–Bohm Oscillations

In 1959, Y. Aharonov and D. Bohm published a paper titled *Significance of Electromagnetic Potentials in the Quantum Theory* [5], in which they showed theoretically that in particular systems the electromagnetic potentials influence electron transport, even though the electrons (more precisely, the squared modulus of the wavefunctions) never reach regions of space where the electromagnetic fields are present. The effect, named since *Aharonov–Bohm effect*, has been experimentally observed in gold rings [469] and in semiconductor parallel quantum wells [117], and is now very often present in magneto-transport in mesoscopic systems. It consists in current oscillations due to the interference of different possible electron paths in the presence of a magnetic field. These interference phenomena arose a great interest also because of other manifestations, such as the so-called *quantum corrections*, *weak localization*, and *universal conduction fluctuations* that we shall see below. However, the primary interest of the authors of the original paper was rather on the fact that such phenomena required a new view on the role of electromagnetic potentials in quantum mechanics. It may be interesting to reread a few lines from the conclusions of that paper:

In classical mechanics [...] *the potentials have been regarded as purely mathematical auxiliaries, while only the field quantities were thought to have a direct physical meaning.*

In quantum mechanics, the essential difference is that the equations of motion of a particle are replaced by the Schrödinger equation for a wave. This Schrödinger equation is obtained from a canonical formalism, which cannot be expressed in terms of the fields alone, but which also requires the potentials. [...] *The Lorentz force* $[e\boldsymbol{E} + (e/c)\boldsymbol{v} \times \boldsymbol{H}]$ *does not appear anywhere in the fundamental theory, but appears only as an approximation holding in the classical limit. It would therefore seem natural at this point to propose that, in quantum mechanics, the fundamental physical entities are the potentials, while the fields are derived from them by differentiations.* [...] *Of course, our discussion does not bring into question the gauge invariance of the theory.*

Turning now to the content of the Aharonov–Bohm effect, let us consider a mesoscopic system formed by a conducting ring connected to electron reservoirs by two leads, as illustrated in Fig. 21.11. The ring encloses a magnetic field $\boldsymbol{B}$, uniform in an area S that does not touch the ring. When electrons reach the split point E from the left lead their wavefunctions are split into two

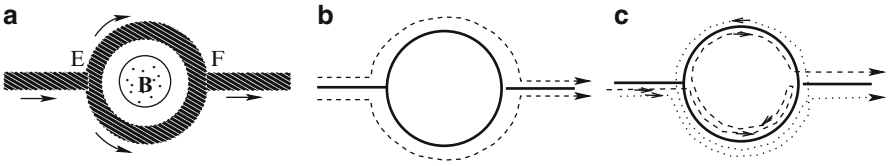


Fig. 21.11. Aharonov–Bohm ring

parts, assuming, for simplicity, that the reflection coefficient at the point E, named a *beam splitter*, is zero. One part of the wavefuction travels in the upper half ring, the other in the lower one. Let us assume that the wavefunction is split into two equal parts and that the propagation along the two paths is coherent, i.e., without energy-involving scattering events.

In Chap. 1, it was shown that the conjugate momentum of the position $\boldsymbol{r}$ is (cf. (1.28)) $\boldsymbol{p} = \boldsymbol{v} + q\boldsymbol{A}$. Thus, the phase change between two points along a trajectory $\boldsymbol{s}$ is the sum of a *dynamic phase* ξ, due to the particle velocity, and a *magnetic phase* ϕ, given by

$$\phi = \frac{q}{\hbar}\int_{\boldsymbol{s}} \boldsymbol{A}\cdot\,\mathrm{d}\boldsymbol{s}.$$

Note that this phase is not gauge invariant and therefore is not observable. When the two parts of the wavefunction reach the opposite point F of the ring, we have the superposition

$$\psi(F) = \psi_{\mathrm{u}}(F) + \psi_{\mathrm{l}}(F), \tag{21.23}$$

where $\psi_{\mathrm{u}}(F)$ and $\psi_{\mathrm{l}}(F)$ are the contributions of the parts of the wavefunction that traveled in the upper and lower parts of the ring, respectively. If we assume a conventional amplitude of the incoming wave equal to unity, these two contributions are given by

$$\psi_{\mathrm{u}}(F) = \frac{1}{\sqrt{2}}\mathrm{e}^{\mathrm{i}(\xi_{\mathrm{u}}+\phi_{\mathrm{u}})}, \qquad \psi_{\mathrm{l}}(F) = \frac{1}{\sqrt{2}}\mathrm{e}^{\mathrm{i}(\xi_{\mathrm{l}}+\phi_{\mathrm{l}})}.$$

The square modulus of (21.23)

$$|\psi(F)|^2 = \left|\frac{1}{\sqrt{2}}\mathrm{e}^{\mathrm{i}(\xi_{\mathrm{u}}+\phi_{\mathrm{u}})} + \frac{1}{\sqrt{2}}\mathrm{e}^{\mathrm{i}(\xi_{\mathrm{l}}+\phi_{\mathrm{l}})}\right|^2$$

contains an interference contribution proportional to

$$\cos(\xi_{\mathrm{u}} - \xi_{\mathrm{l}} + \phi_{\mathrm{u}} - \phi_{\mathrm{l}}). \tag{21.24}$$

Now

$$\phi_{\mathrm{u}} - \phi_{\mathrm{l}} = \frac{q}{\hbar}\int_{\mathrm{u}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{s} - \frac{q}{\hbar}\int_{\mathrm{l}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{s} = \frac{q}{\hbar}\oint_{\mathrm{ring}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{s} = \frac{q}{\hbar}\int_{\Sigma} \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{\sigma} = \frac{q}{\hbar}\Phi_B = \frac{q}{\hbar}BS,$$

where the curl theorem and the definition of the vector potential have been used. According to the above result, the current through the ring must show oscillations, by varying the flux of the magnetic field, with a period given by twice the quantum flux unit:

$$\Delta\Phi = \frac{h}{e} = 2\Phi_\circ, \qquad \Phi_\circ \equiv \frac{h}{2e}.$$

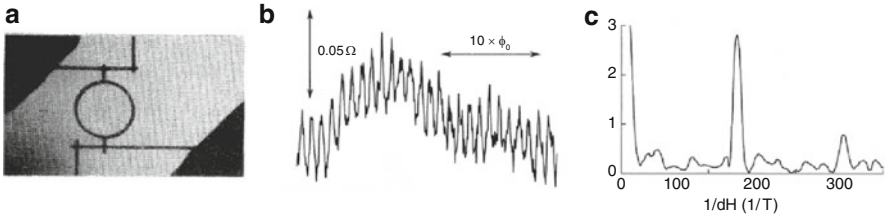


Fig. 21.12. (**a**): Aharonov–Bohm ring; (**b**): current oscillations; (**c**): Fourier transform of the current oscillations showing a major peak corresponding to the flux h/e and a second minor peak corresponding to the flux $h/2e$ [469]

Figure 21.12 shows experimental results [469], where the large peak in the Fourier transform of the current corresponds to the predicted periodicity. As noted by Aharonov and Bohm in the original paper, *the effect is evidently essentially quantum-mechanical in nature because it comes in the phenomenon of interference. We are therefore not surprised that it does not appear in classical mechanics.*

Several considerations must now be made.

1. The interference is produced by electrons which travel coherently along the two arms of the ring. Part of the electrons may suffer inelastic collisions along their motions. Thus, the oscillations will have an amplitude smaller than the total current.
2. Besides the magnetic phase difference, the two parts of the wavefunction traveling in the two arms of the ring have also the mechanical phase difference $(\xi_u - \xi_l)$ as shown in (21.24). Even if the two arms are macroscopically identical, such phase difference depends on the random positions of impurities which act as centers of elastic scattering. Such phase difference is the same for the different electrons crossing the ring so that it does not damage the oscillations seen above, but it is different in different experimental realizations of the ring. Thus, if an average is performed of the currents through many rings, even nominally identical, the oscillations disappear. The situation is different if we consider the interference between other possible electron paths. Let us assume that the coherence length is longer than the ring dimensions. In this case, we may consider the interference between the two electron paths shown in part (c) of Fig. 21.11. In this case, the difference between the mechanical phases cancels, while the difference between the magnetic phases doubles. In fact, the two half wavefunctions cover an identical arm of the ring plus a full ring in opposite directions. The mechanical phases are the same since when $\mathrm{d}\boldsymbol{s}$ changes sign in the integral also the velocity changes sign; the magnetic phases are opposite since $\mathrm{d}\boldsymbol{s}$ changes sign while $\boldsymbol{A}$ does not. The result is that the current oscillates with half of the previous period, i.e., with period $\Phi_\circ$. A small peak in the Fourier component in Fig. 21.12 can be seen at the corresponding frequency. Such

oscillations, independent of the impurity configurations, remain when an average is performed over many nominally identical systems.

21.5 Localization

In the previous section, we have seen that the Aharonov–Bohm effect is an interference phenomenon between different electron trajectories. Such interfering paths may be present also inside a conductor, due to phase-preserving scattering from impurities. At normal temperatures, these effects are not observable because of phase-breaking collisions. At very low temperature, on the contrary, interference paths produce a variety of interesting phenomena, such as *strong* and *weak localizations*, *quantum corrections* and *universal conduction fluctuations*, which will be presented in this and the following sections.

Following Datta [116], let us begin by considering the simplest case of a one-dimensional conductor with only one channel and two scatterers, as shown in Fig. 21.13.

Let T_1 and T_2 be the fractions of particles that are transmitted when hitting the two scatterers S_1 and S_2, respectively. Without considering interference phenomena, the total fraction of particles transmitted across the conductor is the fraction directly transmitted through the first and the second scatterer, given by the product T_1T_2, plus the fraction of particles that are transmitted through the first scatterer, reflected by the second, reflected again by the first, and finally transmitted across the second. This second contribution is $T_1R_1R_2T_2$, where $R_i = 1-T_i$ are the reflection coefficients. Then there is the fraction of particles which perform two round trips, before being finally transmitted across the two scatterers, three round trips, four, and so on. Thus, the total transmission coefficient is given by

$$T_{12} = T_1T_2 + T_1R_2R_1T_2 + T_1R_2R_1R_2R_1T_2 + \dots .$$

This is a geometric series whose sum is

$$T_{12} = \frac{T_1T_2}{1 - R_1R_2}. \tag{21.25}$$

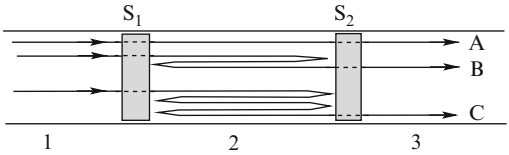


Fig. 21.13. Possible electron paths in a one-dimensional conductor with two scatterers, ignoring interference

From this expression, it is immediate to verify that

$$\frac{1-T_{12}}{T_{12}} = \frac{1-T_1}{T_1} + \frac{1-T_2}{T_2}. \tag{21.26}$$

This result is consistent with what was found in (21.13) as resistance due to the transmission coefficient across the conductor: the classical resistance due to the two scatterers is the sum of the series resistances. If we define ρ the resistance in unit of $h/2e^2$, we have, for any number of scatterers and neglecting interference,

$$\rho = \frac{1-T}{T} = \frac{1-T_1}{T_1} + \frac{1-T_2}{T_2} + \frac{1-T_3}{T_3} + \ldots = \rho_1 + \rho_2 + \rho_3 + \ldots.$$

As an interesting particular case, let us consider N identical, incoherent scatterers with transmission coefficients T':

$$\frac{1-T(N)}{T(N)} = N\frac{1-T'}{T'}.$$

From this,

$$T(N) = \frac{T'}{N(1-T')+T'},$$

or, if we call $D = L/N$ the mean distance between scatterers,

$$T(L) = \frac{L_\circ}{L+L_\circ}, \qquad L_\circ \equiv D\frac{T'}{1-T'},$$

or, finally,

$$\rho(L) = \frac{1-T(L)}{T(L)} = \frac{L}{L_\circ}. \tag{21.27}$$

This is the classical result of a resistance proportional to the length of the conductor.

Let us now consider the same one-dimensional conductor with two scatterers, assuming that the electron dynamics it completely coherent and, therefore, taking also into account the interference between the various paths. Figure 21.14 shows the complex amplitudes of wavefunctions in the three

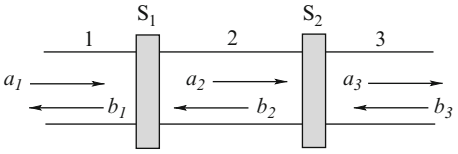


Fig. 21.14. Electron quantum states in a one-dimensional conductor with two scatterers

regions of the one-dimensional channel. To obtain the total transmission coefficient, we need a_3 as a function of a_1. Now, from the Schrödinger equation we may write

$$a_3 = t_2 a_2 + r_2 b_3, \quad a_2 = t_1 a_1 + r_1 b_2, \quad b_2 = r_2 a_2 + t_2 b_3. \tag{21.28}$$

For our purpose, we may take the solution corresponding to $b_3 = 0$, and from the above equations we obtain immediately

$$a_3 = t a_1, \quad t = \frac{t_1 t_2}{1 - r_1 r_2}.$$

The transmission and reflection coefficients are

$$R_i = |r_i|^2, \quad T_i = |t_i|^2, \quad r_i = \sqrt{R_i}\mathrm{e}^{\mathrm{i}\theta_i}.$$

The total transmission coefficient is therefore

$$T = \frac{T_1 T_2}{\left|1 - \sqrt{R_1 R_2}\mathrm{e}^{\mathrm{i}\theta}\right|^2} = \frac{T_1 T_2}{1 + R_1 R_2 - 2\sqrt{R_1 R_2}\cos\theta},$$

where $\theta = \theta_1 + \theta_2$ is the phase shift acquired in a round trip between the scatterers. The resistance of our conductor with two coherent scatterers is easily evaluated and results to be

$$\frac{1 - T}{T} = \frac{1 + R_1 R_2 - 2\sqrt{R_1 R_2}\cos\theta - T_1 T_2}{T_1 T_2}.$$

If we now average this result over the possible phases, we obtain

$$\left\langle \frac{1 - T}{T} \right\rangle = \frac{1}{2\pi} \int \frac{1 + R_1 R_2 - 2\sqrt{R_1 R_2}\cos\theta - T_1 T_2}{T_1 T_2}\, d\theta = \frac{1 + R_1 R_2 - T_1 T_2}{T_1 T_2}. \tag{21.29}$$

From this result, we may define an equivalent total quantum transmission coefficient T_q such that

$$\frac{1 - T_q}{T_q} = \frac{1 + R_1 R_2 - T_1 T_2}{T_1 T_2}.$$

Solving for T_q, we obtain

$$T_q = \frac{T_1 T_2}{1 + R_1 R_2}.$$

Comparison of this result with the incoherent equivalent (21.25) shows that the changed sign in the denominator makes the total transmission of coherent scatterers less than the incoherent transmission as effect of interference. Furthermore, from (21.29) we obtain, after some algebra,

$$\left\langle \frac{1 - T}{T} \right\rangle = \frac{1 - T_1}{T_1} + \frac{1 - T_2}{T_2} + 2\frac{1 - T_1}{T_1}\frac{1 - T_2}{T_2},$$

or

$$\rho_{12} = \rho_1 + \rho_2 + 2\rho_1\rho_2, \tag{21.30}$$

which confirms the larger resistance due to interference.

With the result in (21.30), it is not as simple as for the incoherent scatterers to extend the result to the case of N scatterers. We may, instead, add to a resistance ρ an infinitesimal contribution due to the addition of a length $\mathrm{d}L$ to the coherent conductor, to obtain a differential equation. We have again two series resistance to combine as in (21.30), one given by $\rho(L)$ and a second that, from (21.27) may be written as $\mathrm{d}L/L_\circ$. From (21.30), we obtain

$$\rho(L + \mathrm{d}L) = \rho(L) + \mathrm{d}L/L_\circ + 2\rho(L)\mathrm{d}L/L_\circ,$$

or

$$\frac{\rho(L + \mathrm{d}L) - \rho(L)}{\mathrm{d}L} = \frac{\mathrm{d}\rho}{\mathrm{d}L} = 1/L_\circ + 2\rho(L)/L_\circ,$$

with solution

$$\rho(L) = -\frac{1}{2} + Ce^{2L/L_\circ}.$$

For $L = 0$, we expect $\rho(0) = 0$ which yields the constant $C = 1/2$. Thus,

$$\rho(L) = -\frac{1}{2} + \frac{1}{2}e^{2L/L_\circ} = \frac{1}{2}\left[e^{2L/L_\circ} - 1\right],$$

or

$$\rho(L) = \frac{1}{2}\left[e^{2L/L_\mathrm{c}} - 1\right]. \tag{21.31}$$

The quantity $L_\mathrm{c} = L_\circ$ is the *localization length* for a single-channel conductor. If M channels are active in the one-dimensional conductor, the localization length becomes $L_\mathrm{c} = ML_\circ$. If the conductor becomes longer than or comparable with the localization length, yet remaining coherent, the resistance increases exponentially with the length of the conductor, not linearly as in the incoherent case. This phenomenon is called *strong localization* or *Anderson localization* [10]. It occurs when $\rho > 1$, i.e., when the resistance is bigger than the quantum resistance $h/2e^2 = 12.5\,\mathrm{k}\Omega$, while the electron dynamics remains coherent. In metals M is very large, and it is not possible to see the strong-localization effect. It is possible to observe it in one- and two-dimensional semiconductor structures. See, for example, [45, 165]. For a review, see [34].

21.6 Weak Localization – Quantum Corrections

In practice, it is difficult that the electron dynamics in a conductor remains coherent for distances long enough to realize the strong localization seen above.

If the length of a phase-coherent conductor is less than the localization length, coherent scattering between scatterers still produces some measurable

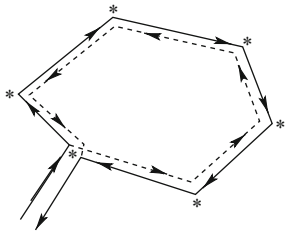


Fig. 21.15. Time reversal paths in weak localization. *Stars* indicate scattering impurities

effect, and the conductor is said to be in the *weak localization* regime. The effect on the conduction results in a *quantum correction*, which can be evaluated from the same (21.31) when $L \ll L_c$. In such a case (21.31) yields, to second order

$$\rho(L) \approx \frac{L}{L_c} + \left(\frac{L}{L_c}\right)^2 = \rho_{cl}\left(1 + \frac{L}{L_o}\right),$$

where the classical result (21.27) has been used. The above result shows that at low temperature, when the coherent length is greater than the mean distance between elastic collisions a quantum correction is present which enhances the conductor resistance.

In the foregoing this effect was studied in a one-dimensional conductor, where a scatterer produces always a backscattering. If we consider a 2- or 3-dimensional conductor, backscattering is associated with paths that starting from an electron position bring the electron into the same position. We expect that the phases of the various paths which bring an electron into its initial position are random so that interference effects cancel on the average. There are, however, special pairs of paths for which this is not true. They are paths obtained from one another by time-reversal symmetry, as the two ones shown in Fig. 21.15. Backscattering is therefore enhanced by interference phenomena. Thus, weak localization and quantum corrections are present also in two- and three-dimensional conductors, whenever the coherence length is longer that the mean free path of the electrons.

Note that a weak magnetic field breaks the constructive interference between the two paths that in absence of field enhances backscattering so that an increase in conductance is observed at very low temperatures when a magnetic field is applied.

21.7 Universal Conduction Fluctuations

The quantum interference between different electron paths at the origin of Aharonov–Bohm oscillations, localization, and quantum corrections, are also at the basis of the phenomenon known as *universal conductance fluctuations*,

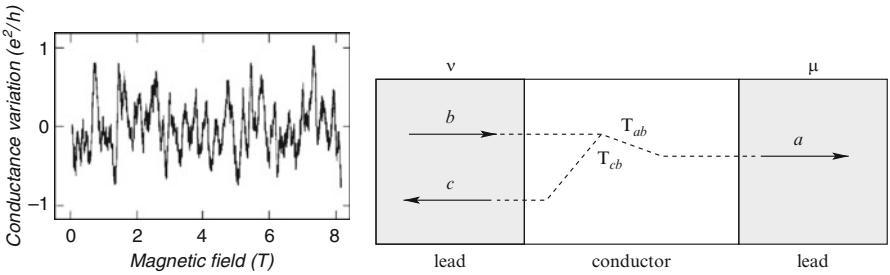


Fig. 21.16. *Left*: universal conductance fluctuations as a function of magnetic-field intensity [468]; *right*: transmission coefficients between different channels

observed in conductors at very low temperatures, when they are in weak localization regime. Conductance fluctuations are observed as a function of carrier concentration n or of the intensity B of an applied magnetic field. When n is changed, the Fermi energy is also changed, so that the wavelengths of conducting electrons and, therefore, the phases of interfering paths change; when B is changed, the magnetic flux inside close loops is also changed. Furthermore, from what was seen above also the Fermi level changes with B. Thus, again, the whole pattern of interfering paths change. Such fluctuations have always the same amplitude e^2/h, as shown in Fig. 21.16, and for this reason are called universal. Furthermore, they are reproducible for the same sample, while change from sample to sample, so that they are considered a sort of fingerprint of each sample.

From the foregoing, it is clear that to study the fluctuations in a sample obtained by varying the concentration or the magnetic field, we may study the fluctuations of the conductance in different, nominally equal, samples with different impurity configurations.

To evaluate the variance of the conductance, following [116,274], we apply the Landauer–Büttiker theory developed in Sect. 21.1. Let us label with ν and μ the input and output leads, respectively. According to (21.11), with T given by (21.14) for the case of many active channels at the Fermi level:

$$G = \frac{2e^2}{h}\overline{T}, \qquad \overline{T} = \sum_{a\in\mu}\sum_{b\in\nu} T_{ab}. \tag{21.32}$$

Let us define $\overline{R}$ the total reflection function, given by the sum of transmission coefficients from any channel of the input lead to any channel of the same lead:

$$\overline{R} = \sum_{c\in\nu}\sum_{b\in\nu} T_{cb}. \tag{21.33}$$

Since the sum of the transmission coefficients from a channel to any other channel is one,

$$\sum_{a\in\mu} T_{ab} + \sum_{c\in\nu} T_{cb} = 1,$$

(21.32) yields

$$G = \frac{2e^2}{h}\left(M - \overline{R}\right), \tag{21.34}$$

where M is the number of active channels in the input lead. Thus, the fluctuations of the conductance are given by the fluctuations of the reflection function $\overline{R}$.[4]

The reflection function $\overline{R}$ in (21.33) is the sum of the M^2 terms T_{cb}. If we assume that these transmission coefficients are uncorrelated, the variance of $\overline{R}$ is M^2 times the variance of any of such terms:

$$\left\langle \delta\overline{R}^2 \right\rangle = M^2 \left\langle \delta T_{cb}^2 \right\rangle = M^2 \left(\left\langle T_{cb}^2 \right\rangle - \left\langle T_{cb} \right\rangle^2 \right). \tag{21.35}$$

For a unit amplitude of the entering wave, the transmission coefficient is given by the square amplitude of the transmitted wave. In coherent dynamics, this amplitude is the sum of the amplitudes A_P corresponding to the different possible paths P leading from channel b to channel c (see right part of Fig. 21.16). Thus

$$\langle T_{cb} \rangle = \left\langle \left| \sum_P A_P \right|^2 \right\rangle = \left\langle \sum_P \sum_{P'} A_P A_{P'}^* \right\rangle = \left\langle \sum_P |A_P|^2 \right\rangle,$$

where it has been assumed that the different paths have random phases, and the average is taken over different samples or different Fermi levels and/or magnetic fields. Under the same hypotheses, the average of the squared transmission coefficient is given by,

$$\langle T_{cb}^2 \rangle = \left\langle \sum_P \sum_{P'} \sum_{P''} \sum_{P'''} A_P A_{P'}^* A_{P''} A_{P'''}^* \right\rangle.$$

With random phases these average products vanish, unless each amplitude is multiplied by its own complex conjugate, and this may happen in two possible ways:

$$\begin{aligned} \langle T_{cb}^2 \rangle &= \left\langle \sum_P \sum_{P'} \sum_{P''} \sum_{P'''} A_P A_{P'}^* A_{P''} A_{P'''}^* [\delta_{PP'}\delta_{P''P'''} + \delta_{PP'''}\delta_{PP'P''}] \right\rangle \\ &= 2 \sum_P \sum_{P'} \langle |A_P|^2 \rangle \langle |A_{P'}|^2 \rangle = 2 \langle T_{cb} \rangle^2. \end{aligned}$$

[4] It is interesting to note that from (21.32) the same result would be obtained by evaluating the variance of the transmission function. However, the variance of $\overline{R}$ is evaluated with the assumption that the reflection paths are uncorrelated, an hypothesis that cannot be applied to the much longer transmission paths [274].

Thus, from (21.34) and (21.35), the standard deviation of the conductance, i.e., the amplitude of its oscillations, is given by

$$\Delta G = \frac{2e^2}{h}\sqrt{\left\langle \delta\overline{R}^2 \right\rangle} = \frac{2e^2}{h} M \langle T_{cb} \rangle .$$

If the sample is much longer than the electron mean free paths, and the electrons hit several impurities before getting back to the input lead, then we may assume $\langle T_{cb} \rangle$ to be of the order of $1/M$, and therefore

$$\Delta G \approx \frac{2e^2}{h},$$

as found experimentally.

21.8 Resonant Tunneling Diode

In the conclusion of this chapter on coherent transport, let us consider the device known as *Resonant Tunneling Diode* (RTD). Since its introduction in 1974 [100], the RTD has become one of the emblems of nanoscopic quantum devices. It is also a sort of test case study for quantum transport theories and simulations of mesoscopic systems. We shall briefly review, here, the principles on which it is based.

An RTD is formed by a layered structure, as described in Chap. 19, where a layer of a high-mobility semiconductor material, typically undoped GaAs, is inserted between two layers of a high-mobility semiconductor material with higher potential energy (higher conduction-band minimum), typically undoped AlGaAs alloy (see Fig. 21.17a). Two contacts of low resistance material, typically doped GaAs, are situated at the two ends of the structure. The geometry of the structure is such that it can be modeled as a one-dimensional system, with the main axis orthogonal to the layers, while translational symmetry may be assumed along the two directions in the planes parallel to the layers. The potential profile along the direction of the main axis presents two barriers, a few nm wide and a few tenths of eV high, separated by a well of similar width, as shown in Fig. 21.17b. Scattering states can be considered, which are eigenstates of this potential profile formed by incident and reflected

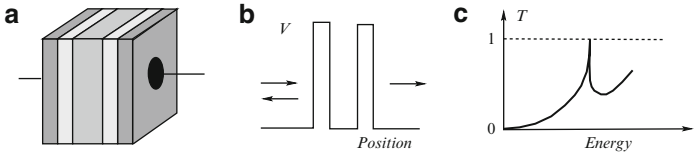


Fig. 21.17. Resonant tunneling diode (RTD): (**a**) structure; (**b**) potential profile and scattering state incoming from the left; (**c**) Transmission coefficient

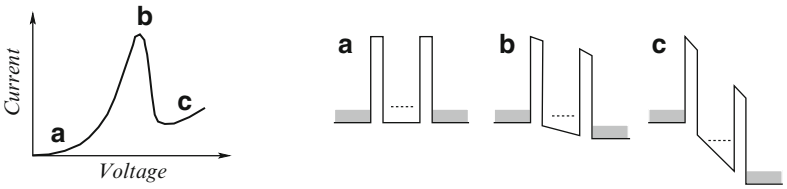


Fig. 21.18. Working principle of the RTD (see text). When electrons are injected at the resonant energy (*dotted line*), the current through the device is maximum

plane waves on one side of the structure (left in the figure) and only the transmitted wave on the other side. The transmission coefficient presents one or more resonances, with value one in the symmetric case, at given energies. In Fig. 21.17c, only one resonance is shown.

The mechanism of operation of the device is shown in Fig. 21.18. When a bias is applied to the device, as long as the carriers are injected into the system at energies below resonance, most of them are reflected by the potential barriers, owing to the small transmission coefficient, and the current is low. At increasing bias, the energies of the injected electrons become closer to the energy of the resonant state; the transmission coefficient and therefore the current increase, reaching a maximum and then decreasing again when the electrons are mainly injected above the resonant energy. The current vs voltage characteristics presents a negative differential conductivity with a high peak-to-valley ratio, a feature very valuable in electronic devices for microwave applications.

The coherent current through an RTD can be easily investigated by means of the Wigner function (WF). Let $|\phi_k\rangle$ be the scattering states described above. If each lead is in equilibrium at the given temperature with its own Fermi level, its density matrix is diagonal in this basis, and given by

$$\rho = \sum_k P_k |\phi_k\rangle\langle\phi_k|, \tag{21.36}$$

where P_k is the probability of finding a system of the statistical ensemble in the state $|\phi_k\rangle$, proportional to the Fermi (or Maxwell) distribution of the lead of the incident wave. The equilibrium WF in this case is then given by:

$$f_w(\boldsymbol{r},\boldsymbol{p}) = \sum_k P_k \int \mathrm{d}\boldsymbol{s}\ \mathrm{e}^{-\mathrm{i}\mathbf{p}\mathbf{s}/\hbar}\ \phi_k(\boldsymbol{r}+\boldsymbol{s}/2)\phi_k^*(\boldsymbol{r}-\boldsymbol{s}/2), \tag{21.37}$$

where $\phi_k(\mathbf{r})$ is the scattering-state wavefunction.

In Fig. 21.19, the WF of electrons in an RTD is shown, in absence of phonon interaction, at three values of the applied bias corresponding to the three operating conditions in Fig. 21.18. It is possible to see in the figure that some charge accumulates between the two barriers in resonant conditions.

Even if an RTD device is extremely small, the total length being of the order of tens of nm, phonon scattering is not completely negligible.

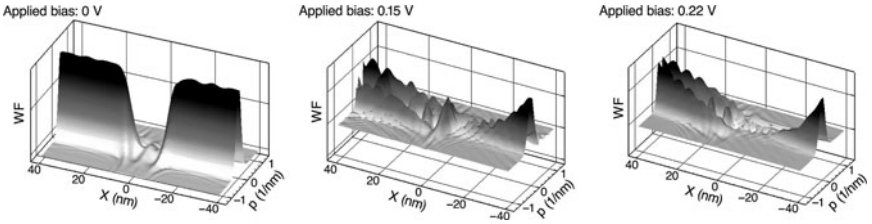


Fig. 21.19. WF in an RTD structure at three values of the applied bias corresponding to the three operating conditions indicated in Fig. 21.18

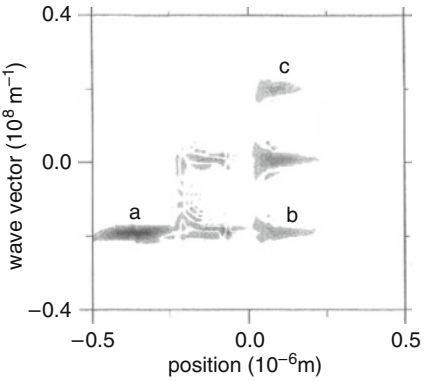


Fig. 21.20. Wigner function (WF) analysis of an e–ph scattering process while the electron is crossing a double barrier. See text [74]

Electron–phonon interaction in a WF formalism can be considered with different levels of rigor. In the simplest approach, a relaxation time toward a local equilibrium WF is used. A more microscopic approach consists of including in the dynamical equation for the WF a scattering integral as given in the Boltzmann transport theory. This means to consider the WF as completely equivalent to a semiclassical distribution function with respect to the scattering of electrons by phonons. In this approach, the completed collision limit is considered, and quantum effects, such as collision duration, intracollisional field effect, and collisional broadening are neglected. Owing to the very short transit time, it is hard to believe that these effects can be negligible. The size of the wavepacket of an electron is of the order of the device itself. Thus assuming pointlike transitions in space and time, as done in the semiclassical theory of scattering seems to be unreasonable.

Nevertheless, a rigorous quantum analysis of a single e–ph interaction has shown that the semiclassical approximation is not that unsatisfactory. In [74], an electron wave packet is considered approaching a double barrier with a average momentum corresponding to the resonant energy. When the packet starts to cross the barriers, a phonon emission begins that almost reverses the momentum of the electron. Figure 21.20 shows the WF corresponding to the

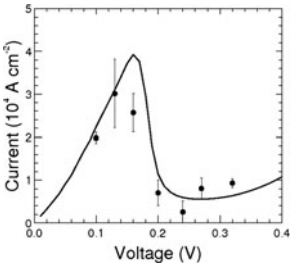


Fig. 21.21. Electron current as a function of the applied bias for the case of an RTD. Comparison between the results obtained using the ballistic WF (*solid line*) and the WF corrected by the affect of an e–ph scattering process (*dots*) [55]

electron after scattering of the transition (the state before scattering is not shown in the picture). Part (a) shows WF contributions coming from points of the WF scattered before they reached the barriers: they are at the left of the barrier and move with negative velocities. Part (b) represents points scattered after the original electron had crossed the barriers: they are at the right of the barriers and move with negative velocities toward the barriers. Part of them will cross the barriers and will enrich part (a) of the WF; part will be reflected by the barriers and form part (c) of the WF, which represents points that are scattered after the original electron has crossed the barriers and are then reflected by them. The other parts of the WF are correlations between those in a), b), and c). The picture is very similar to what would result if the WF represented not a single quantum particle, but an ensemble of classical particles (apart from the tunneling process).

Results for quantum e–ph interaction in an RTD are shown in Fig. 21.21. They have been obtained in [55] using electron scattering states and considering only one scattering process inside the device.

A bistability in the I(V) characteristic of an RTD was observed in [167] and interpreted as an intrinsic bistability due to charge accumulation in the quantum well. Theoretical calculations, including Poisson self-consistent field, have later confirmed the expectation of bistability at low temperature (see for example [39]), but phonon scattering tends to depress it until it should disappear completely at room temperature [144, 222].

Finally, the WF approach has recently been used to propose the existence of a special phenomenon of particular interest from the point of view of basic physics and in mesoscopic device applications. The phenomenon consists in a reduction of the scattering efficiency of the barriers due to the proximity of phase-breaking contacts [38, 140].

22

Semiconductor Photo Gallery

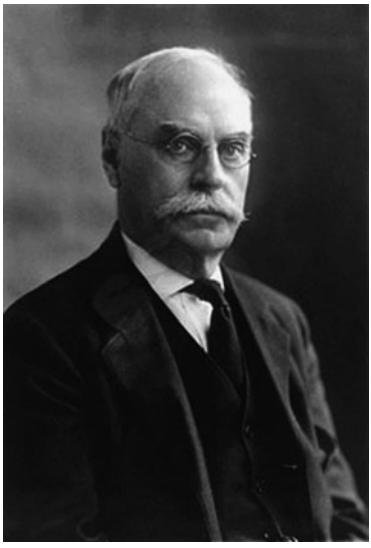


Fig. 22.1. *Left*: A. Volta, the inventor of the electric battery, was the first scientist to name materials that were neither good conductors nor good insulators as "materials with semiconductive nature". The modern use of the word "semicondutor" originated in the 1940s [404]. *Right*: E.H. Hall performed the measurements that brought to the concept of "hole" in semiconductors (Courtesy of D. Avnir)

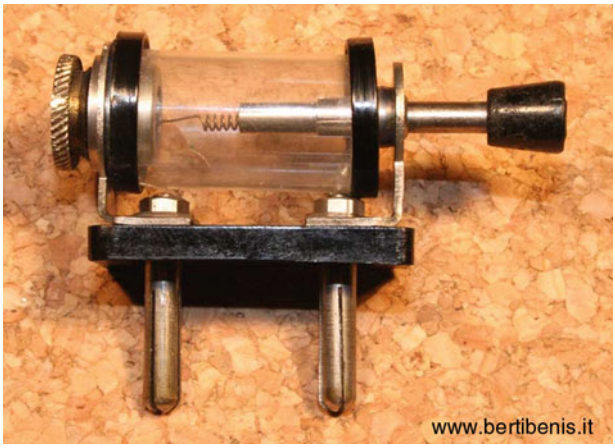


Fig. 22.2. After K.F. Braun discovered the point-contact rectifier effect, galena (PbS) was used in the first half of the twentieth century as a rectifying crystal (Reprinted from www.bertibenis.it)

Fig. 22.3. The first transistor from Bell Labs (Reprinted from www.porticus.org/bell/belllabs_transistor.html)

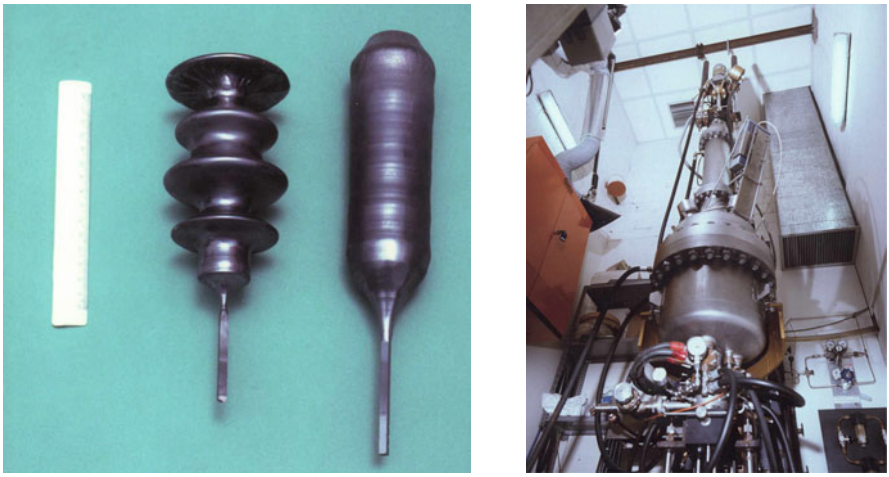


Fig. 22.4. *Left image*: GaAs single crystals grown at IMEM-CNR (Parma, Italy) by LEC (Liquid-Encapsulated Czochralski) techniques with (*on the right*) and without (*on the left*) automatic diameter control during growth; the cylindrical crystal on the right has a diameter of 5 cm. *Right image*: Czochralski apparatus for crystal growth (Courtesy of C. Paorici)

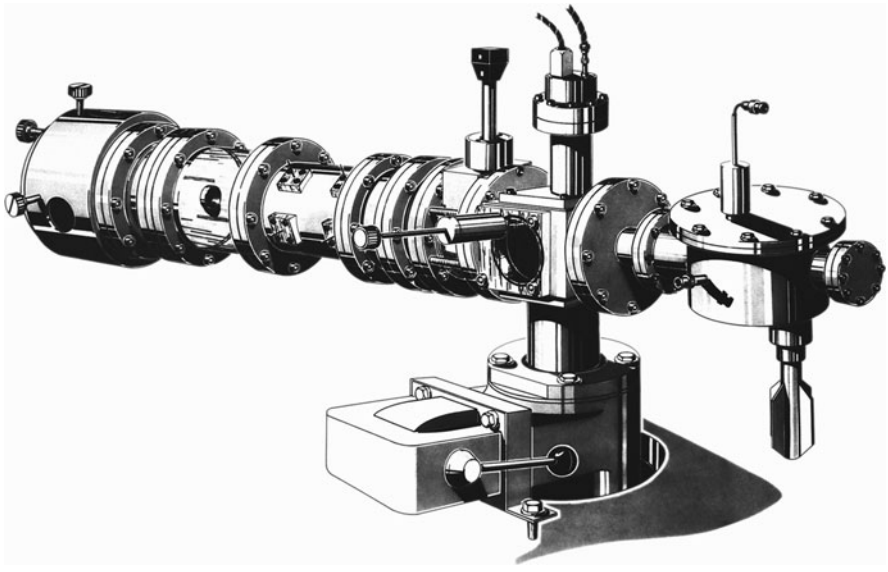


Fig. 22.5. Design of the electron gun used for measurements of mobility and diffusivity of charged carriers in semiconductors at the University of Modena, Italy

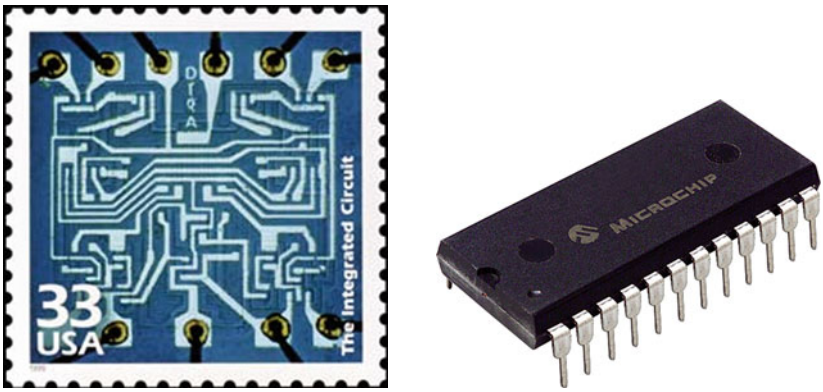


Fig. 22.6. Integrated circuits. The widespread diffusion of integrated circuits is so effective that they can be found in all of nowadays devices, even reproduced in a US stamp

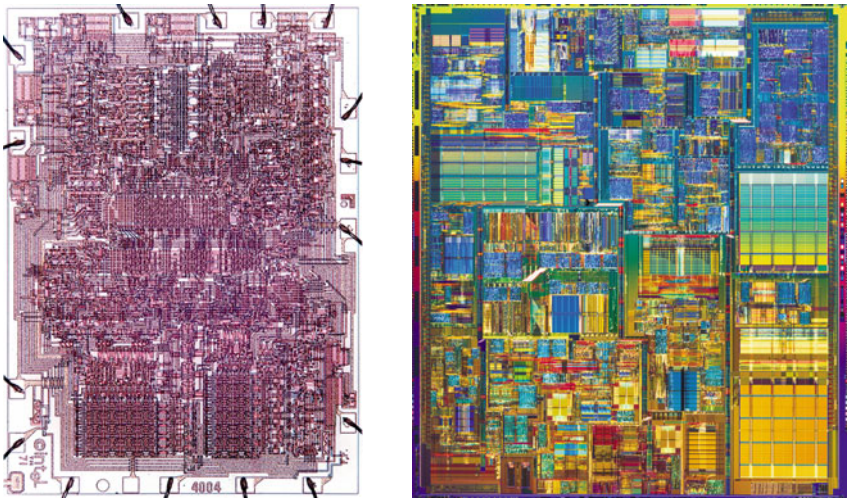
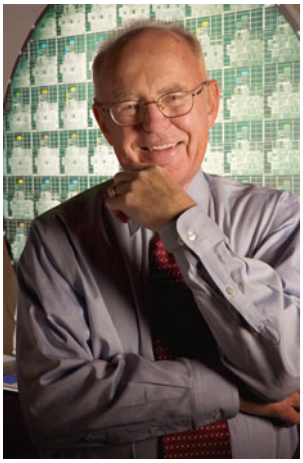


Fig. 22.7. *Left*: the Intel 4004 microprocessor was the first commercially available microprocessor released by Intel in 1971. It contained 2300 transistors on an area of 13.4 mm^2. *Right*: the first release of the Pentium 4 microprocessor, available in 2000. This microprocessor contained more than 41 million transistors on an area of 217 mm^2. Nowadays microprocessors have almost one billion transistors on a single chip (Images by courtesy of Intel Corporation)



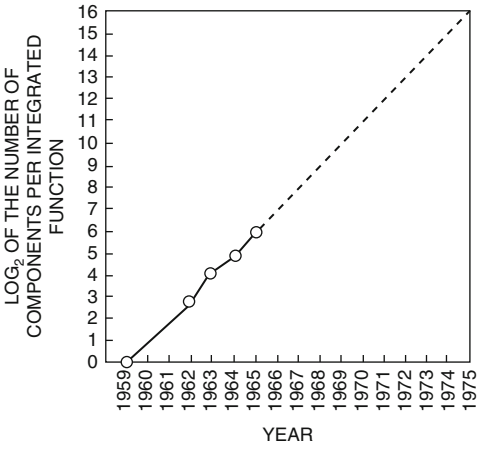


Fig. 2 Number of components per Integrated function for minimum cost per component extrapolated vs time.

Fig. 22.8. *Left*: Gordon E. Moore, Co-founder, Intel Corporation. *Right*: According to Moore's Law, the number of transistors in an integrated circuit doubles approximately every 2 years (Courtesy of Intel Corporation)

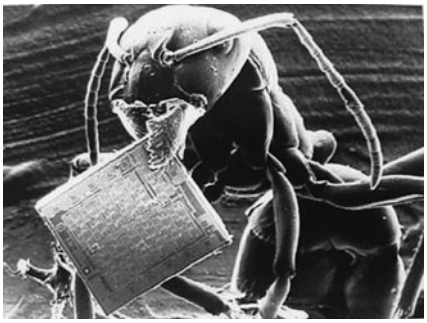


Fig. 22.9. SEM of an ant holding a microchip in its mandibles

Fig. 22.10. Solar cells are increasingly being used to generate "clean" energy, especially in places where it might be difficult to bring electrical power (*Right image* by courtesy of NASA)

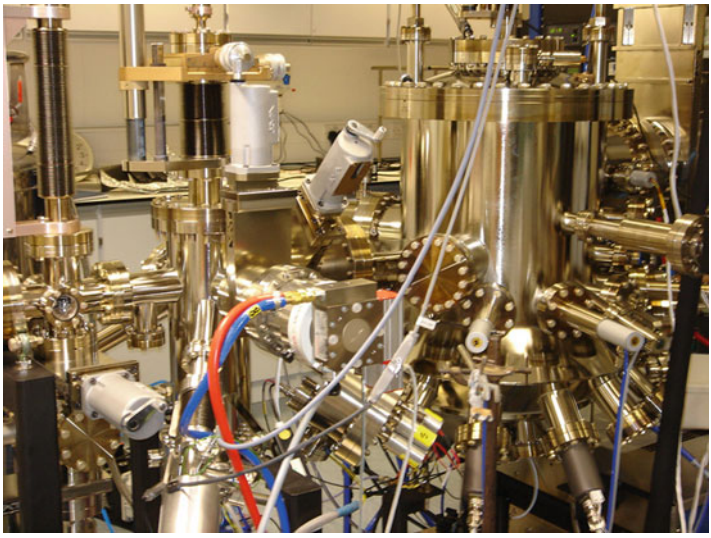


Fig. 22.11. MBE growth machine for single crystal thin-film epitaxy (Chemistry Research Laboratory, Oxford, UK)

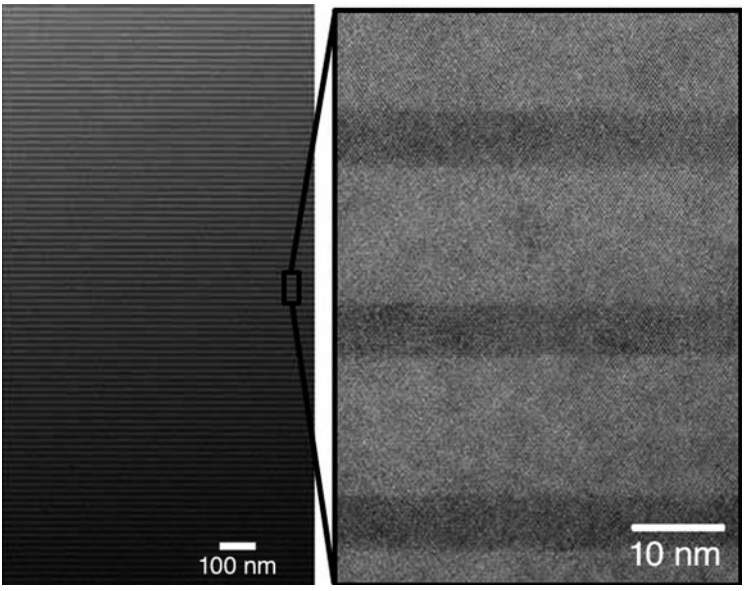


Fig. 22.12. Compound-semiconductor superlattice (Courtesy of M. Sugiyama, The University of Tokyo)

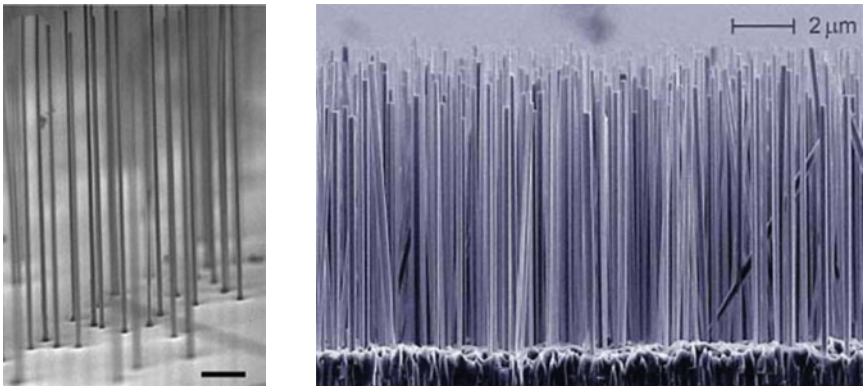


Fig. 22.13. *Left*: Scanning electron micrograph of vertically-grown silicon nanowires off a silicon substrate (courtesy of P. Yang, University of California, Berkeley); *Right*: An array of silicon nanowires (Courtesy of Prof. P. Yu, University of California, San Diego)

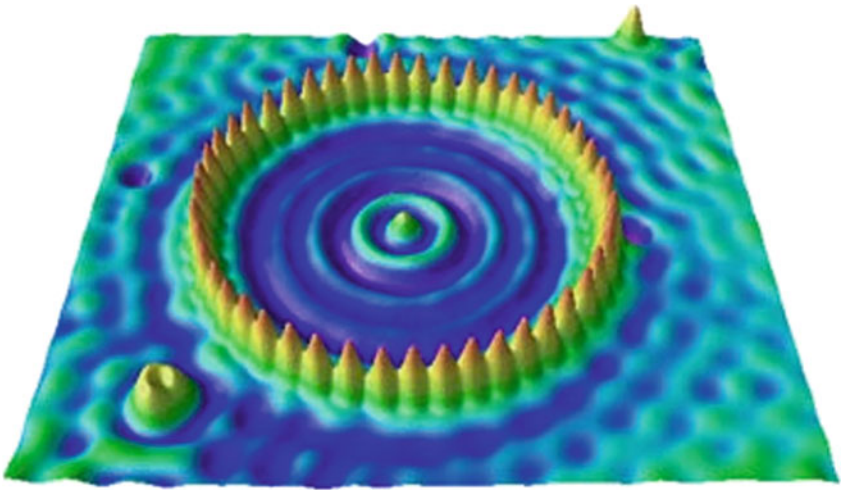


Fig. 22.14. Quantum corral: Fe atoms individually positioned on a Cu(111) surface using the tip of a scanning tunneling microscope. Inside the corral, the electron charge density of a surface electron state is detected (Image originally created by IBM Corporation)

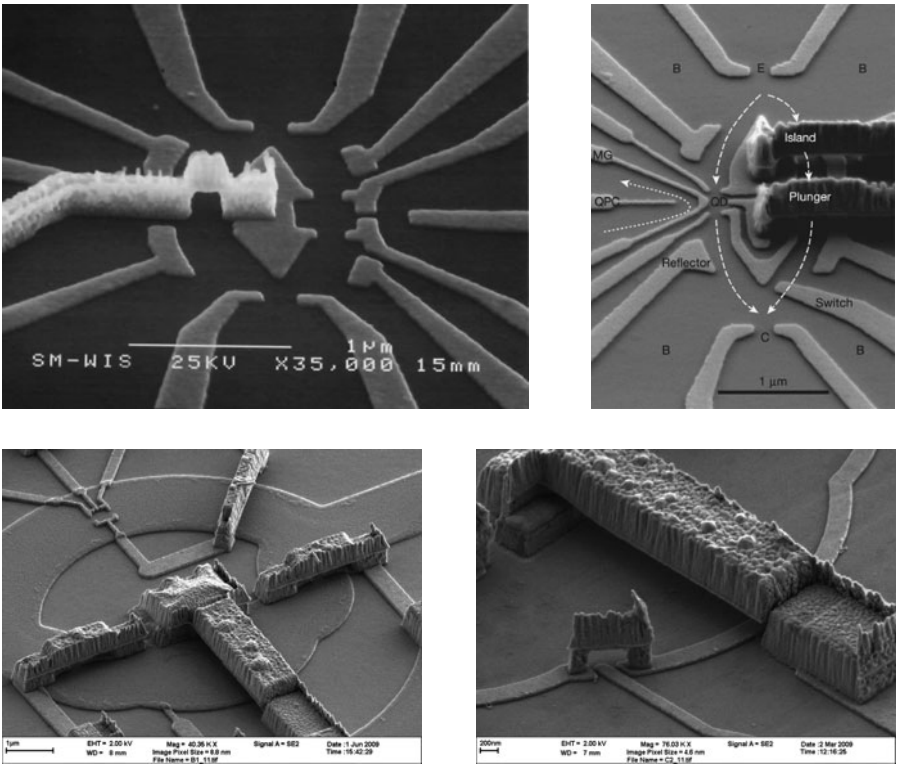


Fig. 22.15. With modern electron-beam lithography, it is possible to fabricate highly sophisticated nanoscopic systems with many quantum wires and quantum dots contacted in many different ways (SEM pictures reprinted by courtesy of M. Heiblum)

Fig. 22.16. CdSe quantum-dot particles in suspension. Dots of different sizes contain electron states with different energy levels and emit light of different colors (Printing by permission, copyright Prof. H. Weller, University of Hamburg, Germany)

Part V

Quantum Transport with Nonequilibrium Green Functions

23

Second-Quantization Formalism

The formalism of second quantization is essential for the description of systems where the number of particles may change as an effect of the dynamics. It is extremely useful, however, also when we deal with many-body systems where the number of particles is fixed.

In this chapter, we shall introduce the basic elements of the second-quantization formalism following the excellent text of Fetter and Walecka [145]. The interested reader may find there all the details of this theory.

23.1 Many-Particle Wavefunctions

Let us consider a system containing N identical particles. The identity of the particles requires a special symmetry of the wavefunction, as discussed in Sect. 2.6. Exchanging two particle variables, the wavefunction must remain unaltered for boson, or change sign for fermions:

$$\Psi(\ldots \boldsymbol{r}_i, \ldots, \boldsymbol{r}_j, \ldots) = \pm\Psi(\ldots \boldsymbol{r}_j, \ldots, \boldsymbol{r}_i, \ldots). \tag{23.1}$$

The upper and lower signs hold for bosons and fermions, respectively.

Let us now expand the wavefunction as combination of products of single-particle orthonormal functions $\psi_k(\boldsymbol{r})$:

$$\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) = \sum_{k_1, k_2, \ldots k_N} C(k_1, k_2, \ldots k_N)\psi_{k_1}(\boldsymbol{r}_1)\ldots\psi_{k_N}(\boldsymbol{r}_N)\ . \tag{23.2}$$

The coefficients C are obtained by projecting the wavefunction on the single-particle eigenfunctions:

$$C(k_1, \ldots, k_N) = \int \psi^*_{k_1}(\boldsymbol{r}_1)\ldots\psi^*_{k_N}(\boldsymbol{r}_N)\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)\mathrm{d}\boldsymbol{r}_1\ldots\mathrm{d}\boldsymbol{r}_N. \tag{23.3}$$

If the wavefunction is normalized to one, it can be immediately verified that

$$\sum_{k_1,k_2,\ldots k_N} |C(k_1,k_2,\ldots k_N)|^2 = 1. \tag{23.4}$$

With the explicit expression (23.3), we easily find that the coefficients C possess the same symmetry property (23.1) of the wavefunctions:

$$C(\ldots k_i \ldots k_j \ldots) = \pm C(\ldots k_j \ldots k_i \ldots). \tag{23.5}$$

Conversely, the above relation implies the symmetry property of the wavefunction in (23.1). Equation (23.5) also shows that in the case of fermions the coefficients C containing two equal k must be zero.

At this point, after giving the eigenfunctions (denumerable in Hilbert spaces) a definite, although arbitrary, order, we may use the symmetry properties of the coefficients C to reorder their labels in increasing order. For example, for bosons, put

$$C(312512\ldots) = C(11\ldots 22..3\ldots 5\ldots).$$

For fermions, no two equal indices may be present, and in reordering them we must change sign every time we invert two indices. For example,

$$C(31854\ldots) = (-1)^S C(13458\ldots),$$

where S is the order of the permutation, that is the number of exchanges necessary to bring the original order of the indices to coincide with the increasing order. We may then use a more concise and informative notation, using as arguments of the coefficients C the number of times that each k appears. For bosons:

$$C(312512\ldots) = C(11\ldots(n_1\text{times})\ 22..(n_2\text{times})\ldots) = \tilde{C}(n_1,n_2,\ldots).$$

For fermions only n_i equal to 0 or 1 are allowed:

$$C(31854\ldots) = (-1)^S \tilde{C}(10111001\ldots).$$

The number of different C that correspond to the same $\tilde{C}$ is given, for fermions, by the number of possible orders of N particles, i.e., $N!$ For bosons, the number of possible orders must be divided by the number of permutations of the identical particles in the same state. The number of different C corresponding to the same $\tilde{C}$ is therefore $N!/n_1!n_2!\ldots$, that includes also the case of fermions, in which case the denominator is always one. If we then define

$$f(n_1,n_2,\ldots) = \sqrt{\frac{N!}{n_1!n_2!\ldots}}\tilde{C}(n_1,n_2,\ldots), \tag{23.6}$$

the normalization condition (23.4) in terms of coefficients $\tilde{C}$ and f is

$$\sum_{n_1 n_2 \ldots} |\tilde{C}(n_1,n_2,\ldots)|^2 \frac{N!}{n_1!n_2!\ldots} = \sum_{n_1 n_2 \ldots} |f(n_1,n_2,\ldots)|^2 = 1.$$

23.1.1 Expansion in Symmetric Wavefunctions

In the expansion of a many-particle wavefunction in terms of products of single-particle basis functions, the coefficients f, defined in (23.6), multiply combinations of products which have already the correct symmetry property. Let us consider a simple example of a system with three bosons. If the wavefunction contains a contribution like $C(112)\psi_1(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_2(\boldsymbol{r}_3)$, it must contain also, with equal coefficients $C(112) = \tilde{C}(21)$, the products obtained by this one exchanging the particle variables, so that the wavefunction contains

$$\tilde{C}(21)\left[\psi_1(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_2(\boldsymbol{r}_3) + \psi_1(\boldsymbol{r}_1)\psi_2(\boldsymbol{r}_2)\psi_1(\boldsymbol{r}_3) + \psi_2(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_1(\boldsymbol{r}_3)\right].$$

Since by the definition in (23.6), $f(21) = \sqrt{3}\tilde{C}(21)$, the above can be written as

$$f(21)\Phi_{2,1}(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3),$$

where $\Phi_{2,1}(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$ is the symmetrized and normalized wavefunction,

$$= \frac{1}{\sqrt{3}}\left[\psi_1(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_2(\boldsymbol{r}_3) + \psi_1(\boldsymbol{r}_1)\psi_2(\boldsymbol{r}_2)\psi_1(\boldsymbol{r}_3) + \psi_2(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_1(\boldsymbol{r}_3)\right],$$

that corresponds to a situation where we can find two particles in state 1 and one in state 2.

If we consider an example with fermions, we can put the particles only in different states. Let us consider the example of a situation with three fermions distributed among four states: one in the first state, one in the second, and one in the fourth. The antisymmetrized, normalized, wavefunction is

$$\Phi_{1,1,0,1}(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3) =$$

$$= \frac{1}{\sqrt{6}}\left[\psi_1(\boldsymbol{r}_1)\psi_2(\boldsymbol{r}_2)\psi_4(\boldsymbol{r}_3) - \psi_1(\boldsymbol{r}_1)\psi_4(\boldsymbol{r}_2)\psi_2(\boldsymbol{r}_3) - \psi_2(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_4(\boldsymbol{r}_3)\right.$$
$$\left. + \psi_2(\boldsymbol{r}_1)\psi_4(\boldsymbol{r}_2)\psi_1(\boldsymbol{r}_3) + \psi_4(\boldsymbol{r}_1)\psi_1(\boldsymbol{r}_2)\psi_2(\boldsymbol{r}_3) - \psi_4(\boldsymbol{r}_1)\psi_2(\boldsymbol{r}_2)\psi_1(\boldsymbol{r}_3)\right].$$

This is defined as the *Slater determinant*:

$$\Phi_{1,1,0,1}(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3) = \frac{1}{\sqrt{6}}\begin{vmatrix} \psi_1(\boldsymbol{r}_1) & \psi_1(\boldsymbol{r}_2) & \psi_1(\boldsymbol{r}_3) \\ \psi_2(\boldsymbol{r}_1) & \psi_2(\boldsymbol{r}_2) & \psi_2(\boldsymbol{r}_3) \\ \psi_4(\boldsymbol{r}_1) & \psi_4(\boldsymbol{r}_2) & \psi_4(\boldsymbol{r}_3) \end{vmatrix}. \tag{23.7}$$

The general definition of properly symmetrized basic wavefunctions is, for bosons,

$$\Phi_{n_1 n_2 \ldots}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \sqrt{\frac{n_1! n_2! \ldots}{N!}} \sum_{k_1 k_2 \ldots}^{[n_1 n_2 \ldots]} \psi_{k_1}(\boldsymbol{r}_1) \ldots \psi_{k_N}(\boldsymbol{r}_N),$$

where the sum runs over the values of k_i such that k_1 appears n_1 times, k_2 n_2 times, and so on, and that $n_1 + n_2 + \ldots = N$. For fermions

$$\Phi_{n_1 n_2 \ldots}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \sqrt{\frac{1}{N!}} \sum_{k_1 k_2 \ldots}^{[n_1 n_2 \ldots]} (-1)^S \psi_{k_1}(\boldsymbol{r}_1) \ldots \psi_{k_N}(\boldsymbol{r}_N).$$

Here S represents the order of the permutation defined above. The most general wavefunction with appropriate symmetry property is therefore

$$\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \sum_{n_1 n_2 \ldots}^{[n_1 + n_2 + \ldots = N]} f(n_1, n_2, \ldots) \Phi_{n_1, n_2, \ldots}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N). \tag{23.8}$$

This last expression is valid for both bosons and fermions.

23.2 Vector Space of Many-Particle States

The correctly symmetrized wavefunctions Φ in (23.8) represent states with N identical particles distributed over given single-particle states. Let us indicate the corresponding state vectors as

$$|n_1, n_2, \ldots\rangle \ , \quad n_i = 0, 1, 2, \ldots \ , \quad n_1 + n_2 + \ldots = N, \tag{23.9}$$

and (23.8) in vector terms becomes

$$\boxed{|\Psi\rangle = \textstyle\sum_{n_1 n_2 \ldots} f(n_1, n_2, \ldots) |n_1, n_2, \ldots\rangle} \tag{23.10}$$

If, in the above sum, we limit ourselves to integers n_i with sum N, our formalism is totally equivalent to the standard quantum mechanics (called, in this context *first quantization*), though more compact and intuitive. We may use the expression (23.10), however, without the above restriction, letting n_i be any positive or zero integers. In this way, we obtain a significative extension of the theory, where, in a given system, the number of particles may change.

Let us then consider a vector space, called Fock space, defined by the linear combinations of the basis vectors in (23.9), where now the sum of the n_i is variable. The integers n_i are called *occupation numbers*. Orthonormalization and completeness in this new vector space are given by the relations

$$\langle n_1, n_2, \ldots | n_1', n_2', \ldots\rangle = \delta_{n_1 n_1'} \delta_{n_2 n_2'} \ldots \ , \quad \sum_{n_1, n_2, \ldots} |n_1, n_2, \ldots\rangle\langle n_1, n_2, \ldots| = 1,$$

respectively. State vectors are orthogonal if they contain a different number of particles or particles differently distributed among the single-particle states. In the completeness relation, the sum is extended to all possible occupation numbers, corresponding also to different sums N.

23.2.1 Creation and Annihilation Operators

Since we are now dealing with vectors representing states with different number of particles, we need operators that, when applied to a state vector, yield vectors representing states with a different number of particles. This is obtained with the *creation* and *annihilation* operators.

For boson, the annihilation operator a_k is defined by

$$\boxed{\mathsf{a}_k|n_1, n_2, \ldots, n_k, \ldots\rangle = \sqrt{n_k}\ |n_1, n_2, \ldots, n_k - 1, \ldots\rangle} \tag{23.11}$$

Its hermitian conjugate $\mathsf{a}_k^\dagger$ is the creation operator with the following effect

$$\boxed{\mathsf{a}_k^\dagger|n_1, n_2, \ldots, n_k, \ldots\rangle = \sqrt{n_k + 1}\ |n_1, n_2, \ldots, n_k + 1, \ldots\rangle} \tag{23.12}$$

as can be shown taking the scalar product of (23.11) with $\langle n_1, n_2, ..., n_k - 1, ...|$ and then the hermitian conjugate of the resulting expression. The product of the two above is the *number operator* n_k, that yields the occupation number:

$$\boxed{\mathsf{n}_k = \mathsf{a}_k^\dagger \mathsf{a}_k\ ,\quad \mathsf{n}_k|n_1, n_2, \ldots, n_k \ldots\rangle = n_k\ |n_1, n_2, \ldots, n_k, \ldots\rangle} \tag{23.13}$$

For fermions, the antisymmetry of the wavefunctions requires a special care. The single-particle states k_i must be put in a definite, though arbitrary, order; then, remembering that the occupation numbers can be only 0 or 1, the annihilation operator b_s is defined by

$$\boxed{\mathsf{b}_s|n_1, n_2, \ldots, n_s, \ldots\rangle = \begin{cases} (-1)^{S_s}|n_1, n_2, \ldots, n_s - 1, \ldots\rangle & if\ n_s = 1 \\ 0 & if\ n_s = 0 \end{cases}} \tag{23.14}$$

where

$$S_s = n_1 + n_2 + \ldots + n_{s-1}$$

is the number of occupied states before reaching the state s. The hermitian conjugate of b_s is the creation operator whose effect on the basis vectors is

$$\boxed{\mathsf{b}_s^\dagger|n_1, n_2, \ldots, n_s, \ldots\rangle = \begin{cases} (-1)^{S_s}|n_1, n_2, \ldots, n_s + 1, \ldots\rangle & if\ n_s = 0 \\ 0 & if\ n_s = 1 \end{cases}} \tag{23.15}$$

In the number operator $\mathsf{n}_s = \mathsf{b}_s^\dagger \mathsf{b}_s$, the phase factor cancels, yielding the same result as in (23.13).

From the above definitions, the commutation relations of the creation and annihilation operators are easily found. We have used the letter a for boson operators and b for fermion operators. We then use the letter c for the general

case, and the commutation relations are [1]

$$\boxed{\left[\mathsf{c}_k, \mathsf{c}_{k'}^\dagger\right]_\mp = \delta_{kk'} \qquad [\mathsf{c}_k, \mathsf{c}_{k'}]_\mp = \left[\mathsf{c}_k^\dagger, \mathsf{c}_{k'}^\dagger\right]_\mp = 0} \tag{23.16}$$

where, again, the upper sign holds for bosons and the lower one for fermions. Most of the properties of creation and annihilation operators can be derived from the above commutation relations, leading to their definitions in (23.11) to (23.15). They can be easily derived with the same arguments used for the equivalent operators of the quantum theory of the harmonic oscillation (see Appendix C). We simply state these properties, referring to [145] for the detailed derivations.

Bosonic Operators

1b. The eigenvalues of $\mathsf{n}_k = \mathsf{a}_k^\dagger \mathsf{a}_k$ are ≥ 0.
2b. If $|n\rangle$ is an eigenvector of n_k, $\mathsf{a}_k|n\rangle$ is either zero or is still an eigenvector of n_k belonging to the eigenvalue $(n-1)$.
3b. The eigenvalues of n_k are integers, positive or zero.
4b. $\mathsf{a}_k^\dagger|n\rangle$ is never zero and is eigenvector of n_k belonging to the eigenvalue $(n+1)$.

Fermionic Operators

1f. The square of both creation and annihilation fermion operators are zero: $\mathsf{b}_k^2 = {\mathsf{b}_k^\dagger}^2 = 0$.
2f. The fermionic number operator is idempotent: $\mathsf{n}_k^2 = \mathsf{n}_k$.
3f. The only eigenvalues of n_k are 0 and 1 (*exclusion principle*).
4f. If $|0\rangle$ is the eigenvector of n_k belonging to the eigenvalue 0, then $\mathsf{b}_k^\dagger|0\rangle$ is again eigenvector of n_k belonging to the eigenvalue 1: $\mathsf{b}_k^\dagger|0\rangle = |1\rangle$.
5f. $\mathsf{b}_k^\dagger|1\rangle = 0$.
6f. $\mathsf{b}_k|1\rangle = |0\rangle$.
7f. $\mathsf{b}_k|0\rangle = 0$.

23.2.2 Field Operators

Starting from the creation and annihilation operators of single-particle states described by the wavefunctions $\psi_k(x)$, let us define the *field operators*, for both bosons and fermions, as

$$\boxed{\Psi(\boldsymbol{r}) = \textstyle\sum_k \psi_k(\boldsymbol{r})\mathsf{c}_k \ , \qquad \Psi^\dagger(\boldsymbol{r}) = \sum_k \psi_k^*(\boldsymbol{r})\mathsf{c}_k^\dagger} \tag{23.17}$$

[1] $[\mathsf{a}, \mathsf{b}]_-$ is the usual commutator $[\mathsf{a}, \mathsf{b}] = \mathsf{ab} - \mathsf{ba}$, while $[\mathsf{a}, \mathsf{b}]_+$ is the anticommutator defined as $\mathsf{ab} + \mathsf{ba}$.

where the sum is over all single-particle states.[2]

It is immediate to obtain the commutation relations of the field operators, starting from the commutation relations of the creation and annihilation operators:

$$\boxed{[\Psi(\boldsymbol{r}),\Psi(\boldsymbol{r}')]_{\mp}=\left[\Psi^{\dagger}(\boldsymbol{r}),\Psi^{\dagger}(\boldsymbol{r}')\right]_{\mp}=0\ ,\quad \left[\Psi(\boldsymbol{r}),\Psi^{\dagger}(\boldsymbol{r}')\right]_{\mp}=\delta(\boldsymbol{r}-\boldsymbol{r}')} \tag{23.18}$$

Equations (23.17) can be easily inverted multiplying the first by $\psi^{*}_{k'}(\boldsymbol{r})$ and the second by $\psi_{k'}(\boldsymbol{r})$ and integrating. The result is

$$\mathsf{c}_k=\int\psi^{*}_k(\boldsymbol{r})\Psi(\boldsymbol{r})\mathrm{d}\boldsymbol{r}\ \ ,\quad \mathsf{c}^{\dagger}_k=\int\psi_k(\boldsymbol{r})\Psi^{\dagger}(\boldsymbol{r})\mathrm{d}\boldsymbol{r}. \tag{23.19}$$

Using these, we obtain the total number operator as

$$\begin{aligned}\mathcal{N}&=\sum_k\mathsf{n}_k=\sum_k\mathsf{c}^{\dagger}_k\mathsf{c}_k=\sum_k\int\psi_k(\boldsymbol{r})\Psi^{\dagger}(\boldsymbol{r})\mathrm{d}\boldsymbol{r}\int\psi^{*}_k(\boldsymbol{r}')\Psi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}'\\&=\int\int\delta(\boldsymbol{r}-\boldsymbol{r}')\Psi^{\dagger}(\boldsymbol{r})\Psi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}\ \mathrm{d}\boldsymbol{r}'=\int\Psi^{\dagger}(\boldsymbol{r})\Psi(\boldsymbol{r})\mathrm{d}\boldsymbol{r}=\int\mathsf{n}(\boldsymbol{r})\mathrm{d}\boldsymbol{r}.\end{aligned}$$

[2] As an exercise, we may show that a particle in $\boldsymbol{r}_1$, described by a single-particle wavefunction $\psi(\boldsymbol{r})=\delta(\boldsymbol{r}-\boldsymbol{r}_1)$ is obtained by the application of the creation operator $\Psi^{\dagger}(\boldsymbol{r}_1)$ to the vacuum. Following the path that led us to the occupation-number representation, let us expand the wavefunction as series of single-particle states:

$$\Psi(\boldsymbol{r})=\sum_k C(k)\psi_k(\boldsymbol{r}).$$

The coefficients are given by

$$C_k=\int\psi^{*}_k(\boldsymbol{r})\psi(\boldsymbol{r})\mathrm{d}\boldsymbol{r}=\int\psi^{*}_k(\boldsymbol{r})\delta(\boldsymbol{r}-\boldsymbol{r}_1)\mathrm{d}x=\psi^{*}_k(\boldsymbol{r}_1)=\tilde{C}(0,0,\ldots,1,0,\ldots),$$

with the 1 in the place of the state k. Equation (23.6) yields

$$f(0,0,\ldots,1_k,0,\ldots)=\tilde{C}(0,0,\ldots,1,0,\ldots)=\psi^{*}_k(\boldsymbol{r}_1),$$

and the symmetrized wavefunctions are the $\psi_k(x)$ themselves: $\Phi_{0,0,\ldots 1_k,0,\ldots}(x)=\psi_k(\boldsymbol{r})$. Finally, (23.10) yields

$$\begin{aligned}|\psi\rangle&=\sum_{n_1,n_2,\ldots}f(n_1,n_2,\ldots)|n_1,n_2,\ldots\rangle=\sum_k\tilde{C}_k(0,0,\ldots 1,\ldots)|0,0,\ldots 1,0,\ldots\rangle\\&=\sum_k\psi^{*}_k(\boldsymbol{r}_1)\mathsf{c}^{\dagger}_k|000\ldots\rangle=\Psi^{\dagger}(\boldsymbol{r}_1)|000\ldots\rangle,\end{aligned}$$

Q.E.D.

Here, $\mathsf{n}(\boldsymbol{r})$ is the *density operator*:

$$\boxed{\mathsf{n}(\boldsymbol{r}) = \Psi^\dagger(\boldsymbol{r})\Psi(\boldsymbol{r})} \tag{23.20}$$

analogous to number operator in (23.13).

23.3 From First to Second Quantization

An operator, say $\mathcal{A}$, that represents a physical quantity in ordinary quantum mechanics has a well-defined and intuitive form in second-quantization formalism. Starting from the application of $\mathcal{A}$ on the many-body wavefunction,

$$\mathcal{A}\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \Psi'(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N), \tag{23.21}$$

its effects can be traced on the coefficients C, $\tilde{C}$, f, and finally on the state vectors in occupation-number representation. The calculation is rather lengthy, and the reader is referred to [145] for details. The result, on the contrary, is very simple: if $\mathcal{A}$ is a single-particle operator,

$$\mathcal{A} = \sum_i \mathcal{A}(\boldsymbol{r}_i),$$

as, e.g., the kinetic or potential energy of the particles, its second quantization form, given a basis of single-particle states $|k\rangle$, is

$$\boxed{\mathcal{A} = \textstyle\sum_{kk'} \mathsf{c}_k^\dagger A_{kk'} \mathsf{c}_{k'}} \tag{23.22}$$

This form has an immediate physical interpretation: the effect of the operator $\mathcal{A}$ is to remove electrons from the various states and regenerate them into new states, i.e., to transfer electrons from a state to another with a "efficacy" given by the matrix element $A_{kk'} = \langle k|\mathcal{A}|k'\rangle$ of the operator between these two states.

In terms of field operators, the above expression is written as

$$\mathcal{A} = \int\int \Psi^\dagger(\boldsymbol{r}) A(\boldsymbol{r}, \boldsymbol{r}') \Psi(\boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}\, \mathrm{d}\boldsymbol{r}'. \tag{23.23}$$

Note, however, that when, in wave mechanics, we write an operator applied to a wavefunction as $\mathcal{A}(\boldsymbol{r})\phi(\boldsymbol{r})$, we actually mean

$$\mathcal{A}(\boldsymbol{r})\phi(\boldsymbol{r}) = \langle \boldsymbol{r}|\mathcal{A}|\phi\rangle = \int \langle \boldsymbol{r}|\mathcal{A}|\boldsymbol{r}'\rangle \mathrm{d}\boldsymbol{r}' \langle \boldsymbol{r}'|\phi\rangle = \int A(\boldsymbol{r}, \boldsymbol{r}')\phi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}'.$$

Thus (23.23) above can also be written as

$$\boxed{\mathcal{A} = \textstyle\int \Psi^\dagger(\boldsymbol{r})\mathcal{A}(\boldsymbol{r})\Psi(\boldsymbol{r})\ \mathrm{d}\boldsymbol{r}} \tag{23.24}$$

When interacting particles are considered, it is necessary to deal with operators that depend on the coordinates of two particles:

$$\mathcal{V} = \frac{1}{2} \sum_{i \neq j} V(\boldsymbol{r}_i, \boldsymbol{r}_j).$$

The same elaborations indicated above for single-particle operators may be repeated for two-particle operators, although in the latter case the development is much more involved. The analogous form for the two-particle operators in second quantization is obtained:

$$\boxed{\mathcal{V} = \tfrac{1}{2} \sum_{k,k',k'',k'''} \mathsf{c}_k^\dagger \mathsf{c}_{k'}^\dagger \langle k, k' | V | k'', k''' \rangle \mathsf{c}_{k'''} \mathsf{c}_{k''}} \tag{23.25}$$

with the same physical interpretation seen above for single-particle operators. Attention must be paid to the order of the operators.

In terms of field operators, the above expression is written as

$$\boxed{\mathcal{V} = \tfrac{1}{2} \int \int \Psi^\dagger(\boldsymbol{r}) \Psi^\dagger(\boldsymbol{r}') V(\boldsymbol{r}, \boldsymbol{r}') \Psi(\boldsymbol{r}') \Psi(\boldsymbol{r}) \, \mathrm{d}\boldsymbol{r} \, \mathrm{d}\boldsymbol{r}'} \tag{23.26}$$

23.4 Dynamics

Dynamics in second quantization is still described by the Schrödinger equation

$$\mathrm{i}\hbar \frac{\partial}{\partial t} |\Psi\rangle = \mathcal{H} |\Psi\rangle,$$

with second-quantization operators and vectors that may describe states with different numbers of particles. According to (23.22), if $\mathcal{H}_\circ$ is the Hamiltonian of noninteracting particles, and the states taken as basis are the eigenstates of the single-particle Hamiltonian with eigenvalues ϵ_k, then the total Hamiltonian becomes

$$\mathcal{H} = \sum_k \mathsf{c}_k^\dagger \mathsf{c}_k \, \epsilon_k \; + \; \mathcal{V}, \tag{23.27}$$

where the interaction term $\mathcal{V}$ is given by (23.25).

In Heisenberg or interaction pictures, the field operators $\Psi^\dagger$ and Ψ are time dependent. At equal times, the commutation relations in (23.18) are still valid, but at different times this is not true any more, since the field operators do not commute with the evolution operator.

23.5 Commutations at Different Times for Non-Interacting Particles

For non-interacting particles, the Hamiltonian in (23.27) reduces to

$$\mathcal{H}_\circ = \sum_\lambda \mathsf{c}_\lambda^\dagger \mathsf{c}_\lambda \, \hbar\omega_\lambda, \tag{23.28}$$

assumed to be time independent, where $\omega_\lambda = \epsilon_\lambda/\hbar$.

Let us begin with the evaluation of the commutator of the annihilation operator with the Hamiltonian, with the use of the basic commutators (23.16):

$$[\mathsf{c}_\lambda, \mathcal{H}_\circ] = \sum_\mu (\mathsf{c}_\lambda \mathsf{c}_\mu^\dagger \mathsf{c}_\mu - \mathsf{c}_\mu^\dagger \mathsf{c}_\mu \mathsf{c}_\lambda)\, \hbar\omega_\mu = \sum_\mu (\mathsf{c}_\lambda \mathsf{c}_\mu^\dagger \mathsf{c}_\mu \mp \mathsf{c}_\mu^\dagger \mathsf{c}_\lambda \mathsf{c}_\mu)\, \hbar\omega_\mu$$

$$= \sum_\mu [\mathsf{c}_\lambda, \mathsf{c}_\mu^\dagger]_\mp \mathsf{c}_\mu \, \hbar\omega_\mu = \sum_\mu \delta_{\lambda\mu} \mathsf{c}_\mu \, \hbar\omega_\mu, \tag{23.29}$$

or

$$[\mathsf{c}_\lambda, \mathcal{H}_\circ] = \mathsf{c}_\lambda \, \hbar\omega_\lambda \quad \text{and} \quad [\mathsf{c}_\lambda^\dagger, \mathcal{H}_\circ] = -\mathsf{c}_\lambda^\dagger \, \hbar\omega_\lambda. \tag{23.30}$$

Let us now move to the Heisenberg picture for noninteracting particles. This picture can be considered as the interaction picture, when the system is described by the total Hamiltonian in (23.27), if the unperturbed Hamiltonian is taken as the noninteracting Hamiltonian in (23.28), and the potential $\mathcal{V}$ is treated as perturbation. The operator will be recognized as given in this new picture since their time dependence is explicitly indicated. The commutation relations at equal times are unchanged so that the relations (23.30) are still valid in the new picture. The dynamical equation is

$$\mathrm{i}\hbar \dot{\mathsf{c}}_\lambda(t) = [\mathsf{c}_\lambda(t), \mathcal{H}_\circ].$$

The Hamiltonian is the same as in the Schrödinger picture since, being time independent, it commutes with the unperturbed evolution operator. We then have the equation of motion

$$\dot{\mathsf{c}}_\lambda(t) = -\mathrm{i}\omega_\lambda \mathsf{c}_\lambda(t),$$

with immediate solutions

$$\mathsf{c}_\lambda(t) = \mathsf{c}_\lambda \, \mathrm{e}^{-\mathrm{i}\omega_\lambda (t-t_\circ)} \, , \qquad \mathsf{c}_\lambda^\dagger(t) = \mathsf{c}_\lambda^\dagger \, \mathrm{e}^{\mathrm{i}\omega_\lambda (t-t_\circ)},$$

if $t_\circ$ is the initial time, when interaction and Schrödinger picture coincide.

From the equations of motion just found, we may obtain the commutation relations for the operators $\mathsf{c}_\lambda(t)$ e $\mathsf{c}_\lambda^\dagger(t)$ and the field operators at different times. Let us consider first the commutator

$$[\mathsf{c}_\lambda(t_1), \mathsf{c}_\mu(t_2)]_\mp = \mathrm{e}^{-\mathrm{i}\omega_\lambda (t_1 - t_\circ)} \mathrm{e}^{-\mathrm{i}\omega_\mu (t_2 - t_\circ)} [\mathsf{c}_\lambda, \mathsf{c}_\mu]_\mp = 0,$$

and similarly

$$[\mathsf{c}_\lambda^\dagger(t_1), \mathsf{c}_\mu^\dagger(t_2)]_\mp = 0.$$

Finally

$$[\mathsf{c}_\lambda(t_1), \mathsf{c}_\mu^\dagger(t_2)]_\mp = \mathrm{e}^{-\mathrm{i}\omega_\lambda(t_1-t_\circ)}\mathrm{e}^{\mathrm{i}\omega_\mu(t_2-t_\circ)}[\mathsf{c}_\lambda, \mathsf{c}_\mu^\dagger]_\mp = \delta_{\lambda\mu}\mathrm{e}^{-\mathrm{i}\omega_\lambda(t_1-t_2)}.$$

With the above results, we may obtain the analogous rules for the field operators defined in (23.17). In the Heisenberg picture of noninteracting particles, these definitions become

$$\Psi(\boldsymbol{r}, t) = \sum_\lambda \phi_\lambda(\boldsymbol{r})\, \mathsf{c}_\lambda(t) = \sum_\lambda \phi_\lambda(\boldsymbol{r})\, \mathsf{c}_\lambda \mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t_\circ)}, \tag{23.31}$$

$$\Psi^\dagger(\boldsymbol{r}, t) = \sum_\lambda \phi_\lambda^*(\boldsymbol{r})\, \mathsf{c}_\lambda^\dagger(t) = \sum_\lambda \phi_\lambda^*(\boldsymbol{r})\, \mathsf{c}_\lambda^\dagger \mathrm{e}^{\mathrm{i}\omega_\lambda(t-t_\circ)}. \tag{23.32}$$

The commutators are then

$$[\Psi(\boldsymbol{r}_1, t_1), \Psi(\boldsymbol{r}_2, t_2)]_\mp = \sum_{\lambda\mu} \phi_\lambda(\boldsymbol{r}_1)\phi_\mu(\boldsymbol{r}_2)\mathrm{e}^{-\mathrm{i}\omega_\lambda(t_1-t_\circ)}\mathrm{e}^{-\mathrm{i}\omega_\mu(t_2-t_\circ)}[a_\lambda, a_\mu]_\mp = 0. \tag{23.33}$$

Similarly,

$$[\Psi^\dagger(\boldsymbol{r}_1, t_1), \Psi^\dagger(\boldsymbol{r}_2, t_2)]_\mp = 0,$$

while

$$\begin{aligned}[\Psi(\boldsymbol{r}_1, t_1), \Psi^\dagger(\boldsymbol{r}_2, t_2)]_\mp &= \sum_{\lambda\mu} \phi_\lambda(\boldsymbol{r}_1)\phi_\mu^*(\boldsymbol{r}_2)\mathrm{e}^{-\mathrm{i}\omega_\lambda(t_1-t_\circ)}\mathrm{e}^{\mathrm{i}\omega_\mu(t_2-t_\circ)}[\mathsf{c}_\lambda, \mathsf{c}_\mu^\dagger]_\mp \\ &= \sum_{\lambda\mu} \phi_\lambda(\boldsymbol{r}_1)\phi_\mu^*(\boldsymbol{r}_2)\mathrm{e}^{-\mathrm{i}\omega_\lambda(t_1-t_\circ)}\mathrm{e}^{\mathrm{i}\omega_\mu(t_2-t_\circ)}\delta_{\lambda\mu},\end{aligned}$$

or

$$[\Psi(\boldsymbol{r}_1, t_1), \Psi^\dagger(\boldsymbol{r}_2, t_2)]_\mp = \sum_\lambda \phi_\lambda(\boldsymbol{r}_1)\phi_\lambda^*(\boldsymbol{r}_2)\mathrm{e}^{-\mathrm{i}\omega_\lambda(t_1-t_2)}\ . \tag{23.34}$$

Note that all the commutators evaluated in this section for noninteracting particles are c-numbers.

23.6 Field Operators in Momentum and Energy Space

In terms of field operators, the creation and annihilation operators of an electron in a single-particle state are given by (23.19). If we take plane waves as single-particle states, they become

$$\Psi(\boldsymbol{k}) = \frac{1}{(2\pi)^{3/2}} \int \Psi(\boldsymbol{r})\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}\boldsymbol{r} \quad , \quad \Psi^\dagger(\boldsymbol{k}) = \frac{1}{(2\pi)^{3/2}} \int \Psi^\dagger(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}\boldsymbol{r}. \tag{23.35}$$

From the above equations, we see that the field operator in $\boldsymbol{k}$ space is the Fourier transform of the field operator in $\boldsymbol{r}$ space.

Let us now consider the same operator in the Heisenberg picture of non-interacting particles and perform the Fourier transform also with respect to the time variable:[3]

$$\Psi(\boldsymbol{k},\omega)=\int\Psi(\boldsymbol{k},t)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t,\quad \Psi^{\dagger}(\boldsymbol{k},\omega)=\int\Psi^{\dagger}(\boldsymbol{k},t)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t. \tag{23.36}$$

The new operator $\Psi^{\dagger}(\boldsymbol{k},\omega)$ generates a particle with momentum $\hbar\boldsymbol{k}$ and energy $\hbar\omega$. To convince ourselves that it is actually so, let us apply our creation operator to a state vector $|\Phi_{\mathrm{H}}\rangle$ at $t+\mathrm{d}t$ and consider the resulting state vector in the Schrödinger picture:

$$\{\Psi^{\dagger}(\boldsymbol{k},t+\mathrm{d}t)|\Phi_{\mathrm{H}}\rangle\}_{\mathrm{S}}. \tag{23.37}$$

If we had applied the same operator to the same state at time t and let it evolve for the time interval $\mathrm{d}t$ we would have obtained the same effect except for the absence, in (23.37), of the evolution factor due to the presence of the extra particle between t and $t+\mathrm{d}t$. If the additional energy due to the extra particle is ϵ, its presence would correspond to a phase factor $\mathrm{e}^{-\mathrm{i}(\epsilon/\hbar)\mathrm{d}t}$ in the evolution. We would therefore have

$$\mathrm{e}^{-\mathrm{i}(\epsilon/\hbar)\mathrm{d}t}\left\{\Psi^{\dagger}(\boldsymbol{k},t+\mathrm{d}t)|\Phi_{\mathrm{H}}\rangle\right\}_{\mathrm{S}}=\left\{\Psi^{\dagger}(\boldsymbol{k},t)|\Phi_{\mathrm{H}}\rangle\right\}_{\mathrm{S}}.$$

The numerical factor commutes with the operator of the unitary transformation, and we have

$$\mathrm{e}^{-\mathrm{i}(\epsilon/\hbar)\mathrm{d}t}\Psi^{\dagger}(\boldsymbol{k},t+\mathrm{d}t)|\Phi_{\mathrm{H}}\rangle=\Psi^{\dagger}(\boldsymbol{k},t)|\Phi_{\mathrm{H}}\rangle,$$

or, since the state $|\Phi_{\mathrm{H}}\rangle$ is arbitrary,

$$\Psi^{\dagger}(\boldsymbol{k},t+\mathrm{d}t)=\mathrm{e}^{\mathrm{i}(\epsilon/\hbar)\mathrm{d}t}\Psi^{\dagger}(\boldsymbol{k},t).$$

This means that, if the extra particle adds an energy ϵ, then

$$\Psi^{\dagger}(\boldsymbol{k},t)\propto\mathrm{e}^{\mathrm{i}(\epsilon/\hbar)t}.$$

This shows that the Fourier component ω of the field operator corresponds to the creation of a particle with energy $\hbar\omega$.

[3] The different constants used in the Fourier transforms in (23.35) and in (23.36)are suggested by the normalized time-dependent plane-wave wavefunctions $[1/(2\pi)^{3/2}]\exp(\mathrm{i}\boldsymbol{k}\boldsymbol{r}-\mathrm{i}\omega t)$.

24

Introduction to Green Functions

The Green functions (GFs) are named after George Green, the British mathematician who first developed this concept in the 1830s. They provide a very powerful, formal framework for many physical problems.

As it regards semiconductor physics, until some time ago they were popular among theoretical researchers interested in general properties of rather idealized systems, but considered of little utility by physicists and engineers who wanted solutions to "real" problems and numbers to compare with experimental results. Recently, however, GFs entered convincingly the realm of practical research both because present experiments deal with particularly clean quantum systems and because new theoretical techniques have been developed that make GFs convenient also for practical calculations.

For reasons of space, we can give only a brief account on the GF technique, limiting ourselves to the basic ideas and to few important applications. The interested reader may find excellent books that deal at length with GF theory in general [145, 173, 198, 295, 313, 371], and with GFs for electron transport in semiconductors [116, 184, 228].

24.1 GFs from Differential Equations to Many-Body Theory

24.1.1 GF of Schrödinger Equation

The general idea of a GF is that of a quantity that indicates how a *cause*, or *source* in $\boldsymbol{r}'$ at time t' determines an *effect* in $\boldsymbol{r}$ at time t.

Let us start from the concept of GF as used in the theory of linear differential equations. Assume that the equation satisfied by the function $f(\boldsymbol{r}, t)$ is

$$\mathcal{A}\, f(\boldsymbol{r}, t) = s(\boldsymbol{r}, t), \tag{24.1}$$

where $\mathcal{A}$ is some linear differential operator, and $s(\boldsymbol{r}, t)$ represents the source term. The GF $g(\boldsymbol{r}, t, \boldsymbol{r}', t')$ is then defined as the function that transfers to

the point $(\boldsymbol{r}, t)$ the effect of the source in $(\boldsymbol{r}', t')$, so that the function f may be obtained as the integral

$$f(\boldsymbol{r}, t) = \int g(\boldsymbol{r}, t, \boldsymbol{r}', t') s(\boldsymbol{r}', t') \mathrm{d}\boldsymbol{r}' \mathrm{d}t'.$$

It is now possible to verify immediately that a function g that satisfies the following *Green equation* for the operator $\mathcal{A}$,

$$\mathcal{A}\, g(\boldsymbol{r}, t, \boldsymbol{r}', t') = \delta(\boldsymbol{r} - \boldsymbol{r}')\delta(t - t'), \tag{24.2}$$

is a good GF since, in this case,

$$\mathcal{A} \int g(\boldsymbol{r}, t, \boldsymbol{r}', t') s(\boldsymbol{r}', t')\, \mathrm{d}\boldsymbol{r}' \mathrm{d}t' = \int \delta(\boldsymbol{r} - \boldsymbol{r}')\delta(t - t')\, \mathrm{d}\boldsymbol{r}' \mathrm{d}t' s(\boldsymbol{r}', t') = s(\boldsymbol{r}, t), \tag{24.3}$$

and (24.1) is satisfied. If, however, we add to this function g, solution of (24.2), a solution of the associated homogeneous equation $\mathcal{A}\, g(\boldsymbol{r}, t, \boldsymbol{r}', t') = 0$, we obtain a new GF as good as the previous one. This freedom must be used to satisfy the boundary conditions, necessary to have a unique solution of (24.1). A general theory of GFs [313] considers also a new $g_b(\boldsymbol{r}, t, \boldsymbol{r}_b, t_i)$ that transfers to the point $(\boldsymbol{r}, t)$ the effect of the boundary-initial value of f in $(\boldsymbol{r}_b, t_i)$. In general, however, boundary-initial conditions can be treated as a special case of source, and we shall follow this line of reasoning directly in the case of the Schrödinger equation, of interest for us.

Before proceeding, note that the r.h.s. of (24.2) is the coordinate representation of the identity operator. This identifies the GF as the coordinate representation of the inverse $\mathcal{A}^{-1}$ or *resolvent* of the linear differential operator $\mathcal{A}$.

Let us now consider the time-dependent Schrödinger equation for a single particle:

$$\mathrm{i}\hbar \frac{\partial}{\partial t} \Psi(\boldsymbol{r}, t) - \mathcal{H}\Psi(\boldsymbol{r}, t) = 0. \tag{24.4}$$

Since this equation in homogeneous, the only source to be considered is that coming from the initial conditions. Let us then assume that we have a flash source limited to the initial time:

$$\mathrm{i}\hbar \frac{\partial}{\partial t} \Psi(\boldsymbol{r}, t) - \mathcal{H}\Psi(\boldsymbol{r}, t) = s_i(\boldsymbol{r})\delta(t - t_i). \tag{24.5}$$

If this equation is integrated over an infinitesimal time interval containing t_i, we obtain the discontinuity produced by the source

$$\mathrm{i}\hbar \left[\Psi(\boldsymbol{r}, t_{i+}) - \Psi(\boldsymbol{r}, t_{i-})\right] = s_i(\boldsymbol{r}). \tag{24.6}$$

In fact, searching for a limited solution, the singularity is attributed to the derivative, and the integral of the second term in (24.5) over an infinitesimal

interval is vanishingly small. This means that the flash source $s_i(\boldsymbol{r})\delta(t-t_i)$ generates a discontinuity as in (24.6). Thus, if we are interested in obtaining the wavefunction for $t > t_i$, we assume the wavefunction zero for times less than t_i and substitute the initial condition with a source that, from (24.5) and (24.6), is given by $\mathrm{i}\hbar\Psi(\boldsymbol{r}, t_i)\delta(t'-t_i)$. The solution for $t > t_i$ will then be given by

$$\Psi(\boldsymbol{r}, t > t_i) = \int g^{\mathrm{r}}(\boldsymbol{r}, t, \boldsymbol{r}', t')\mathrm{i}\hbar\Psi(\boldsymbol{r}', t_i)\delta(t'-t_i)\,\mathrm{d}\boldsymbol{r}'\mathrm{d}t', \tag{24.7}$$

where g^{r} is the solution of the Green equation of the Schrödinger operator

$$\boxed{\left(\mathrm{i}\hbar\frac{\partial}{\partial t} - \mathcal{H}\right) g(\boldsymbol{r}, t, \boldsymbol{r}', t') = \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')} \tag{24.8}$$

with the condition that it is zero for $t < t'$, and is called the *retarded Green function*.

Alternatively, if we are interested in considering t_i as "final condition", i.e., to obtain the wavefunction for $t < t'$, we assume that for later times the wavefunction is zero and, from (24.6), take the source as $-\mathrm{i}\hbar\Psi(\boldsymbol{r}, t)\delta(t-t_i)$ The corresponding GF, vanishing for $t > t'$, is the *advanced Green function* g^{a} that satisfies the same equation (24.8).

24.1.2 The Evolution Operator as Greenian

Equation (24.7), after the obvious integration over t', becomes

$$\Psi(\boldsymbol{r}, t > t_i) = \mathrm{i}\hbar \int g^{\mathrm{r}}(\boldsymbol{r}, t, \boldsymbol{r}', t_i)\Psi(\boldsymbol{r}', t_i)\,\mathrm{d}\boldsymbol{r}'. \tag{24.9}$$

If we analyze this expression, we easily realize that the GF for the Schrödinger equation has been already encountered, with a different name, in our elementary quantum theory. In fact, the evolution operator $\mathcal{U}(t, t_i)$, defined in Sect. 2.2, is such that

$$|\Psi(t)\rangle = \mathcal{U}(t, t_i)|\Psi(t_i)\rangle. \tag{24.10}$$

In coordinate representation, this becomes

$$\Psi(\boldsymbol{r}, t) = \int U(\boldsymbol{r}, t, \boldsymbol{r}', t_i)\Psi(\boldsymbol{r}', t_i)\mathrm{d}\boldsymbol{r}', \tag{24.11}$$

where $U(\boldsymbol{r}, t, \boldsymbol{r}', t')$ is the matrix element $(\boldsymbol{r}, \boldsymbol{r}')$ of the evolution operator $\mathcal{U}(t, t')$. By comparing (24.11) with (24.9), we realize that the evolution operator has exactly the meaning of the GF which carries the contribution of the source $\Psi(\boldsymbol{r}', t_i)$ to generate $\Psi(\boldsymbol{r}, t)$. More precisely,

$$g^{\mathrm{r}}(\boldsymbol{r}, t, \boldsymbol{r}', t') = \frac{1}{\mathrm{i}\hbar}U(\boldsymbol{r}, t, \boldsymbol{r}', t')\theta(t-t'), \tag{24.12}$$

and

$$g^{\mathrm{a}}(\boldsymbol{r},t,\boldsymbol{r}',t') = -\frac{1}{\mathrm{i}\hbar}U(\boldsymbol{r},t,\boldsymbol{r}',t')\theta(t'-t), \tag{24.13}$$

where the Heaviside functions θ (equal to zero for negative argument and to one for positive argument) have been inserted to fulfill the condition that the retarded GF vanishes for $t < t_i$ and the advanced GF vanishes for $t > t_i$. The reader is invited to verify, as exercise, that the above functions satisfy the Green equation (24.8).

24.1.3 Green Functions for a One-Particle System

In terms of second quantization, the single-particle state vector $|\boldsymbol{r}\rangle$ can be obtained with the application of the creation field operator $\Psi^\dagger(\boldsymbol{r})$ to the vacuum state, so that the evolution operator can be written as

$$U(\boldsymbol{r},t,\boldsymbol{r}',t') = \langle \boldsymbol{r}|\mathcal{U}(t,t')|\boldsymbol{r}'\rangle = \langle 0|\Psi(\boldsymbol{r})\mathcal{U}(t,t')\Psi^\dagger(\boldsymbol{r}')|0\rangle.$$

Let us write the field operators in the Heisenberg picture (again identified by the presence of the time argument), with $t_\circ$ as the reference time at which the Schrödinger and Heisenberg pictures coincide:

$$\langle 0|\mathcal{U}(t,t_\circ)\Psi(\boldsymbol{r},t)\mathcal{U}^\dagger(t,t_\circ)\mathcal{U}(t,t')\mathcal{U}(t',t_\circ)\Psi^\dagger(\boldsymbol{r}',t')\mathcal{U}^\dagger(t',t_\circ)|0\rangle.$$

The three internal evolution operators yield the identity. Furthermore, the evolved of the vacuum is still the vacuum,[1] and we could think of writing the retarded GF as:

$$\frac{1}{\mathrm{i}\hbar}\langle 0|\Psi(\boldsymbol{r},t)\Psi^\dagger(\boldsymbol{r}',t')|0\rangle\theta(t-t'). \tag{24.14}$$

The central part of this expression is the scalar product of the vectors

$$\Psi^\dagger(\boldsymbol{r}',t')|0\rangle \quad \text{and} \quad \Psi^\dagger(\boldsymbol{r},t)|0\rangle.$$

We can therefore think of giving to the expression above the following meaning: if a particle is created in $\boldsymbol{r}'$ at t' the expression in (24.14) gives the probability amplitude to find it in $\boldsymbol{r}$ at a later time t.

The above argument, however, although correct, is not satisfactory. The product of the field operators in (24.14) is averaged over the vacuum state, and it would not be possible to extend this expression for the GF to a many-body system. Furthermore, following the same lines of the reasoning, we could think of obtaining the propagator working on the state occupied by a particle with wavefunction Φ, by annihilating the particle at t' in $\boldsymbol{r}'$ and evaluating the probability amplitude that it will be missing in $\boldsymbol{r}$ at t. This would be given by

$$\langle \Phi_{\mathrm{H}}|\Psi^\dagger(\boldsymbol{r},t)\Psi(\boldsymbol{r}',t')|\Phi_{\mathrm{H}}\rangle.$$

[1] In fact, $\mathcal{U}_{\mathrm{H}}$ is given by the unity plus a series of powers of $\mathcal{H}$, all of which contain annihilation operators to the right and applied to the vacuum give zero.

To maintain the convention to use the creation operator with primed argument, let us take the complex conjugate

$$\langle \Phi_{\mathrm{H}} | \Psi^\dagger(\boldsymbol{r}', t') \Psi(\boldsymbol{r}, t) | \Phi_{\mathrm{H}} \rangle. \tag{24.15}$$

This should indicate the probability amplitude that a "hole", or the lack of particle, in $\boldsymbol{r}$ at t comes from the annihilation in $\boldsymbol{r}'$ at t'.

To evaluate the expression in (24.15) first, we observe that if $\Phi(\boldsymbol{r}, t)$ is the wavefuction of the system under examination,

$$|\Phi_{\mathrm{H}}\rangle = \mathcal{U}^{-1}(t, t_\circ) |\Phi_{\mathrm{S}}(t)\rangle = \mathcal{U}^{-1}(t, t_\circ) \int |\boldsymbol{r}\rangle \, \mathrm{d}\boldsymbol{r} \, \langle \boldsymbol{r} | \Phi_{\mathrm{S}}(t)\rangle,$$

where $|\Phi_{\mathrm{S}}(t)\rangle$ is the state of the system at time t in Schrödinger picture,

$$= \mathcal{U}^{-1}(t, t_\circ) \int \mathrm{d}\boldsymbol{r} \, \Phi(\boldsymbol{r}, t) \Psi^\dagger(\boldsymbol{r}) |0\rangle = \int \mathrm{d}\boldsymbol{r} \, \Phi(\boldsymbol{r}, t) \mathcal{U}^{-1}(t, t_\circ) \Psi^\dagger(\boldsymbol{r}) \mathcal{U}(t, t_\circ) |0\rangle,$$

since the evolved of the vacuum state is still the vacuum state. Then at any t

$$|\Phi_{\mathrm{H}}\rangle = \int \mathrm{d}\boldsymbol{r} \, \Phi(\boldsymbol{r}, t) \Psi^\dagger(\boldsymbol{r}, t) |0\rangle.$$

The expression in (24.15) may then be written as

$$\int \mathrm{d}\boldsymbol{\rho}' \Phi^*(\boldsymbol{\rho}', t') \int \mathrm{d}\boldsymbol{\rho} \, \Phi(\boldsymbol{\rho}, t) \langle 0 | \Psi(\boldsymbol{\rho}', t') \Psi^\dagger(\boldsymbol{r}', t') \Psi(\boldsymbol{r}, t) \Psi^\dagger(\boldsymbol{\rho}, t) | 0 \rangle. \tag{24.16}$$

The field operators evaluated now at the same time may be commutated for bosons, or anticommutated for electrons, applying (23.18). The result is a δ function plus the product of the same operators in reverse order, which, applied to the vacuum state yields zero. Thus, it remains

$$\int \mathrm{d}\boldsymbol{\rho}' \Phi^*(\boldsymbol{\rho}', t') \int \mathrm{d}\boldsymbol{\rho} \, \Phi(\boldsymbol{\rho}, t) \langle 0 | \delta(\boldsymbol{r}' - \boldsymbol{\rho}') \delta(\boldsymbol{r} - \boldsymbol{\rho}) | 0 \rangle,$$

or, remembering that we are evaluating (24.15),

$$\langle \Phi_{\mathrm{H}} | \Psi^\dagger(\boldsymbol{r}', t') \Psi(\boldsymbol{r}, t) | \Phi_{\mathrm{H}} \rangle = \Phi(\boldsymbol{r}, t) \Phi^*(\boldsymbol{r}', t'). \tag{24.17}$$

This result is not the propagator that we expected; it is, instead, the extension of the density matrix in coordinate representation for the pure state $|\Phi\rangle$, at different times: $\langle \boldsymbol{r} | \Phi(t) \rangle \langle \Phi(t') | \boldsymbol{r}' \rangle$.

This result should not surprise since, considering the case of $\boldsymbol{r} = \boldsymbol{r}'$ and $t = t'$, the product of field operators $\Psi^\dagger(\boldsymbol{r}) \Psi(\boldsymbol{r})$, i.e., the annihilation of a particle in $\boldsymbol{r}$ followed by its re-creation does not leave the state vector unaltered but multiplies it by its density operator $\mathsf{n}(\boldsymbol{r}) = |\Phi(\boldsymbol{r}, t)|^2$. Thus, for different arguments (24.17) yields the correlation of the wavefunction at the considered points and times, not the simple propagator. In other words, it

contains information on the state of the system, not simply on the dynamics. Similarly, the application of the reverse product $\Psi(\boldsymbol{r})\Psi^\dagger(\boldsymbol{r}')$, for $\boldsymbol{r}' \to \boldsymbol{r}$, does not leave the state unaltered but it multiplies it by $\delta(\boldsymbol{r}-\boldsymbol{r}') \pm \mathsf{n}(\boldsymbol{r})$ as it results from the commutation relations in (23.18). Again, by application of this product of field operators, we obtain information on the state of the system. However, from the above discussion we also have the indication of how to correct our theory: not one product or its reverse must be applied, but rather their difference or sum, i.e., their commutator or anticommutator, so that the information on the state of the system cancels and the propagator remains. This discussion suggests to write the GFs as

$$\begin{aligned} G^{\mathrm{r}}(\boldsymbol{r},t,\boldsymbol{r}',t') &= \tfrac{1}{\mathrm{i}\hbar}\langle \Phi_{\mathrm{H}}|[\Psi(\boldsymbol{r},t),\Psi^\dagger(\boldsymbol{r}',t')]_\mp|\Phi_{\mathrm{H}}\rangle\theta(t-t'), \\ G^{\mathrm{a}}(\boldsymbol{r},t,\boldsymbol{r}',t') &= -\tfrac{1}{\mathrm{i}\hbar}\langle \Phi_{\mathrm{H}}|[\Psi(\boldsymbol{r},t),\Psi^\dagger(\boldsymbol{r}',t')]_\mp|\Phi_{\mathrm{H}}\rangle\theta(t'-t). \end{aligned} \tag{24.18}$$

We have used the capital letter for the GFs, here, to mark the arrival point of the path which led us from the GFs defined in the theory of differential equations to the GFs of many-body theory.

According to the above discussion, for a single-particle system the commutators in (24.18) must yield the evolution operator in the coordinate representation. We shall now show that this is in fact the case. For a time-independent Hamiltonian, the evolution operator is given by (2.4). If we write it in the coordinate representation and use a set of orthonormal single-particles eigenstates $|\phi_\lambda\rangle$, we have

$$\begin{aligned} U(\boldsymbol{r},t,\boldsymbol{r}',t') &= \sum_{\lambda\mu}\langle\boldsymbol{r}|\phi_\lambda\rangle\langle\phi_\lambda|\mathrm{e}^{-\mathrm{i}\mathcal{H}(t-t')/\hbar}|\phi_\mu\rangle\langle\phi_\mu|\boldsymbol{r}'\rangle \\ &= \sum_{\lambda\mu}\phi_\lambda(\boldsymbol{r})\phi_\mu^*(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\mu(t-t')}\langle\phi_\lambda|\phi_\mu\rangle = \sum_{\lambda}\phi_\lambda(\boldsymbol{r})\phi_\lambda^*(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t')}. \end{aligned}$$

This is the expression (23.34) obtained in Sect. 23.5 for the commutator of field operators at different times in the case of independent particles, and therefore applicable to the present case of a single particle. This completes the identification of the above definition (24.18) of the single-particle GF with that (24.12) derived from the evolution operator.

24.1.4 Single-Particle Green Functions in Many-Particle Systems

The direct extension of the concepts discussed in the previous section to the case of a system with many interacting particles would lead to the definition of a GF as the matrix elements of an evolution operator of the whole system. This function, however, would be too complicated, too difficult to evaluate, and probably even useless. It is more important, for us, to know how a particle evolves interacting with all the other particles in the system. This concept must be understood, of course, in the light of the indistinguishable identical particles. The *single-particle Green function* is then defined as the probability

amplitude of finding a particle in $\boldsymbol{r}$ at time t if a particle was present in $\boldsymbol{r}'$ at t'. The above considerations made in the case of a single particle convince us that such functions must be defined as in (24.18), where now, however, many-particle states are considered:

$$\boxed{\begin{aligned} G^{\mathrm{r}}(\boldsymbol{r},t,\boldsymbol{r}',t') &= \tfrac{1}{\mathrm{i}\hbar}\langle[\Psi(\boldsymbol{r},t),\Psi^{\dagger}(\boldsymbol{r}',t')]_{\mp}\rangle\theta(t-t') \\ G^{\mathrm{a}}(\boldsymbol{r},t,\boldsymbol{r}',t') &= -\tfrac{1}{\mathrm{i}\hbar}\langle[\Psi(\boldsymbol{r},t),\Psi^{\dagger}(\boldsymbol{r}',t')]_{\mp}\rangle\theta(t'-t) \end{aligned}} \tag{24.19}$$

The mean values are intended over the state of interest or, when statistical problems are analyzed, over an ensemble. These new functions depend now on the presence of all particles in the system and there will not be simple equations of motion.[2]

It is immediate to verify that

$$G^{\mathrm{r}}(\boldsymbol{r},t,\boldsymbol{r}',t') = (G^{\mathrm{a}}(\boldsymbol{r}',t',\boldsymbol{r},t))^{*} \tag{24.20}$$

Besides the retarded and advanced GFs defined above, other GFs are used, which are, more properly, correlation functions with particular meanings.

The two terms of the (anti)commutators in the GFs above are called *G greater* and *G less*, defined by

$$\boxed{\begin{aligned} G^{>}(\boldsymbol{r},t,\boldsymbol{r}',t') &= \tfrac{1}{\mathrm{i}\hbar}\langle\Psi(\boldsymbol{r},t)\Psi^{\dagger}(\boldsymbol{r}',t')\rangle \\ G^{<}(\boldsymbol{r},t,\boldsymbol{r}',t') &= \pm\tfrac{1}{\mathrm{i}\hbar}\langle\Psi^{\dagger}(\boldsymbol{r}',t')\Psi(\boldsymbol{r},t)\rangle \end{aligned}} \tag{24.21}$$

From the discussion in the previous section, we may recognize the physical meaning of these two correlation functions: $G^{<}$ is the correlation of the amplitude of the particle presence in $(\boldsymbol{r},t)$ and in $(\boldsymbol{r}',t')$; $-G^{>}$ is the correlation of the available states (i.e., a holes) to generate particles in $(\boldsymbol{r},t)$ and in $(\boldsymbol{r}',t')$. Thus, the sum $G^{<}-G^{>}$ contains the correlation of both amplitudes and therefore the correlation of all states, occupied and empty, i.e., the dynamics, independent of the presence or absence of the particle.

The following simple relations are immediate:

$$G^{\mathrm{r}} = \theta(t-t')(G^{>}-G^{<}), \quad G^{\mathrm{a}} = -\theta(t'-t)(G^{>}-G^{<}). \tag{24.22}$$

[2] It is possible, in fact, to define two-particle GFs $G(\boldsymbol{r}_1,t_1,\boldsymbol{r}_2,t_2;\boldsymbol{r}'_1,t'_1,\boldsymbol{r}'_2,t'_2)$, three-particle GFs, and so on. A hierarchical set of equations can then be derived, where the first equation for the single-particle GF involves the two-particle GF; the second equation, for the two-particle GF involves the three-particle GF, and so on. This is known as BBGKY hierarchy (after Bogoliubov–Born–Green–Kirkwood–Yvon) already known in classical many-particle statistical physics. See, for example [369].

Subtracting these two equations, we obtain the relation

$$G^{>} - G^{<} = G^{\mathrm{r}} - G^{\mathrm{a}}. \tag{24.23}$$

If we require that the field operator on the right is applied at a minor time, we have the *time ordered* GF:[3]

$$G^{t}(\boldsymbol{r}, t, \boldsymbol{r}', t') = \frac{1}{\mathrm{i}\hbar}\langle \mathcal{T}\Psi(\boldsymbol{r}, t)\Psi^{\dagger}(\boldsymbol{r}', t')\rangle = \theta(t - t')G^{>} + \theta(t' - t)G^{<}, \tag{24.24}$$

where $\mathcal{T}$ is the *time-ordering operator*. It is define in such a way that, when applied to a product of time-dependent operators, it yields the product of the same operators, ordered in such a way that operators evaluated at minor times are on the right. Furthermore, each time two field operators must be exchanged to obtain the correct time order, a sign change must be inserted when dealing with fermions. Finally, the *anti-time ordered* GF is defined as

$$G^{\bar{t}}(\boldsymbol{r}, t, \boldsymbol{r}', t') = \theta(t' - t)G^{>} + \theta(t - t')G^{<}. \tag{24.25}$$

It is immediate to verify the following relations among the GFs defined above, to be added to the equations from (24.22) to (24.25):

$$G^{\mathrm{r}} = G^{t} - G^{<} = G^{>} - G^{\bar{t}}, \quad G^{\mathrm{a}} = G^{t} - G^{>} = G^{<} - G^{\bar{t}}. \tag{24.26}$$

24.2 Green Functions in Momentum and Energy Space

If the system is homogeneous,[4] the various GFs in real space depend upon $\boldsymbol{r}$ and $\boldsymbol{r}'$ only upon their difference, and we may define a $G(\boldsymbol{k}, t, t')$ as the following Fourier transform:

$$G(\boldsymbol{k}, t, t') = \frac{1}{(2\pi)^{3/2}} \int G(\boldsymbol{r}, t, \boldsymbol{r}', t')\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')}\mathrm{d}(\boldsymbol{r} - \boldsymbol{r}'), \tag{24.27}$$

where we have used the letter G without further specification to indicate any of the GFs described above.

If the system is also stationary, the GFs depend upon t e t' only via their difference, and we may define[5]

$$G(\boldsymbol{k}, \omega) = \int G(\boldsymbol{k}, t, t')\mathrm{e}^{\mathrm{i}\omega(t-t')}\mathrm{d}(t - t'). \tag{24.28}$$

[3] In (24.24), we recognize the reason for calling $G^{>}$ "greater" and $G^{<}$ "less": they coincide with G^{t} when t is greater or less than t', respectively.

[4] This means that we do not consider the presence of a periodic crystal potential. If the theory developed here is to be applied to semiconductors, a free-electron model must be applied.

[5] See note in Sect. 23.6.

It is also possible to define GFs or correlation functions by means of the field operators in momentum and energy space given in (23.35) and (23.36). The resulting GFs are of course strictly related to the above (24.27) and (24.28) when the system is homogeneous and/or stationary.

From the relation in (24.20) and the definition above, we obtain also

$$G^{\mathrm{r}}(\boldsymbol{k},\omega) = (G^{\mathrm{a}}(\boldsymbol{k},\omega))^{*}. \tag{24.29}$$

24.3 Equilibrium GFs for NonInteracting Particles

Noninteracting particles at equilibrium

If the system under examination is formed by noninteracting particles, the Hamiltonian is given by (23.28), where the single-particle states $|\phi_\lambda\rangle$ are the eigenstates of the single particle Hamiltonian. In a time-dependent picture, the field operators are given, as a function of time, by (23.31) and (23.32). We are therefore in the condition to write down all the various GFs for noninteracting particles, and we know that it will be enough to express $G^{>}$ and $G^{<}$ to obtain all the other ones from these. In what follows, the label "ni" is for "noninteracting". Let us start from $G^{>}$:

$$\begin{aligned} G^{>}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') &= \frac{1}{\mathrm{i}\hbar}\langle \Psi(\boldsymbol{r},t)\Psi^{\dagger}(\boldsymbol{r}',t')\rangle \\ &= \frac{1}{\mathrm{i}\hbar}\sum_{\lambda,\lambda'}\langle \mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}\rangle \phi_\lambda(\boldsymbol{r})\phi^{*}_{\lambda'}(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t_\circ)}\mathrm{e}^{\mathrm{i}\omega_{\lambda'}(t'-t_\circ)}. \end{aligned}$$

At equilibrium, the mean value of $\mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}$ is nonzero only for $\lambda = \lambda'$,[6] thus we have, using the standard commutation relations,

$$G^{>}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\sum_{\lambda}(1 \pm \langle n_\lambda\rangle)\phi_\lambda(\boldsymbol{r})\phi^{*}_{\lambda}(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t')}, \tag{24.30}$$

[6] In fact, taking as many-particle basis vectors the occupation eigenvectors $|\Phi_\mu\rangle$ in (23.9), which are eigenstates of the Hamiltonian for noninteracting particles, and taking into account that the equilibrium density matrix is diagonal in the energy eigenstates,

$$\langle \mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}\rangle = \mathrm{Tr}\{\rho \mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}\} = \sum_{\mu,\nu}\langle \Phi_\mu|\rho|\Phi_\nu\rangle\langle \Phi_\nu|\mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}|\Phi_\mu\rangle = \sum_{\mu}\rho_{\mu\mu}\langle \Phi_\mu|\mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}|\Phi_\mu\rangle.$$

If $\lambda \neq \lambda'$, the vector $\mathsf{c}_\lambda \mathsf{c}^{\dagger}_{\lambda'}|\Phi_\mu\rangle$ contains different occupation numbers than $|\Phi_\mu\rangle$, and the last matrix element is zero for any μ.

where $\langle n_\lambda \rangle = \langle \mathsf{c}_\lambda^\dagger \mathsf{c}_\lambda \rangle$. Similarly, we obtain

$$G^<_{\rm ni}(\boldsymbol{r},t,\boldsymbol{r}',t') = \pm\frac{1}{\mathrm{i}\hbar}\sum_\lambda \langle n_\lambda\rangle \phi_\lambda(\boldsymbol{r})\phi^*_\lambda(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t')}. \tag{24.31}$$

From the above, we obtain the other GFs using the relations found in the previous section 24.1.4:

$$G^{\rm r}_{\rm ni}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\theta(t-t')\sum_\lambda \phi_\lambda(\boldsymbol{r})\phi^*_\lambda(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t')}, \tag{24.32}$$

$$G^{\rm a}_{\rm ni}(\boldsymbol{r},t,\boldsymbol{r}',t') = -\frac{1}{\mathrm{i}\hbar}\theta(t'-t)\sum_\lambda \phi_\lambda(\boldsymbol{r})\phi^*_\lambda(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t'-t)}, \tag{24.33}$$

$$G^{t}_{\rm ni}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\sum_\lambda \{\theta(t-t') \pm \langle n_\lambda\rangle\}\phi_\lambda(\boldsymbol{r})\phi^*_\lambda(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t'-t)}, \tag{24.34}$$

$$G^{\bar{t}}_{\rm ni}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\sum_\lambda \{\theta(t'-t) \pm \langle n_\lambda\rangle\}\phi_\lambda(\boldsymbol{r})\phi^*_\lambda(\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\omega_\lambda(t'-t)}. \tag{24.35}$$

Note that, coherently with the general discussion of the previous chapter, $G^{\rm r}$ e $G^{\rm a}$ correspond to the propagation of noninteracting particles and therefore do not depend upon the state of the system, while the other GFs bring information on the correlations of the state of the system at different points and times. Furthermore, all these GFs depend upon t and t' only upon their difference, since we made the assumption of equilibrium.

Free particles at equilibrium

If we consider the case of free particles, we may choose plane waves as single-particle states; the product of wavefunctions in (24.30)–(24.35) become

$$\phi_k(\boldsymbol{r})\phi^*_k(\boldsymbol{r}') = \frac{1}{(2\pi)^3}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')},$$

and the various GF become (the label "pw" is for "plane waves")

$$G^>_{\rm pw}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar(2\pi)^3}\int (1 \pm \langle n_k\rangle)\mathrm{e}^{\mathrm{i}[\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')-\omega_k(t-t')]}\,\mathrm{d}\boldsymbol{k},$$

$$G^<_{\rm pw}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{\pm 1}{\mathrm{i}\hbar(2\pi)^3}\int \langle n_k\rangle\mathrm{e}^{\mathrm{i}[\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')-\omega_k(t-t')]}\,\mathrm{d}\boldsymbol{k},$$

and similarly for the other ones.

The Fourier transforms with respect to $(\boldsymbol{r}-\boldsymbol{r}')$, as indicated in (24.27), are now immediate and yield[7]

[7] The coefficients in front of these expressions depend on the normalizations of the wavefunctions and on the constant used in the definition of the Fourier transform. Different coefficients may be found in the literature.

$$G^{>}_{\rm pw}(\boldsymbol{k},t,t') = \frac{1}{{\rm i}\hbar(2\pi)^{3/2}}(1 \pm \langle n_k\rangle){\rm e}^{-{\rm i}\omega_k(t-t')}, \tag{24.36}$$

$$G^{<}_{\rm pw}(\boldsymbol{k},t,t') = \pm\frac{1}{{\rm i}\hbar(2\pi)^{3/2}}\langle n_k\rangle{\rm e}^{-{\rm i}\omega_k(t-t')}, \tag{24.37}$$

$$G^{\rm r}_{\rm pw}(\boldsymbol{k},t,t') = \frac{1}{{\rm i}\hbar(2\pi)^{3/2}}\theta(t-t'){\rm e}^{-{\rm i}\omega_k(t-t')}, \tag{24.38}$$

$$G^{\rm a}_{\rm pw}(\boldsymbol{k},t,t') = -\frac{1}{{\rm i}\hbar(2\pi)^{3/2}}\theta(t'-t){\rm e}^{-{\rm i}\omega_k(t-t')}, \tag{24.39}$$

$$G^{t}_{\rm pw}(\boldsymbol{k},t,t') = \frac{1}{{\rm i}\hbar(2\pi)^{3/2}}\{\theta(t-t') \pm \langle n_k\rangle\}{\rm e}^{-{\rm i}\omega_k(t-t')}, \tag{24.40}$$

$$G^{\bar{t}}_{\rm pw}(\boldsymbol{k},t,t') = \frac{1}{{\rm i}\hbar(2\pi)^{3/2}\{\theta(t'-t) \pm \langle n_k\rangle\}{\rm e}^{-{\rm i}\omega_k(t-t')}.} \tag{24.41}$$

If we now want to perform the Fourier transform also with respect to the time difference, we must pay attention to the step functions θ in connection with the convergence of the integrals:

$$\int_{-\infty}^{\infty}\theta(\tau){\rm e}^{{\rm i}(\omega-\omega_k)\tau}{\rm d}\tau = \int_{0}^{\infty}{\rm e}^{{\rm i}(\omega-\omega_k)\tau}{\rm d}\tau$$

does not converge for $\tau \to \infty$. As usual, let us introduce a convergence factor ${\rm e}^{-\gamma\tau}$, considering γ infinitesimal. The integral above becomes

$$\int_{0}^{\infty}{\rm e}^{{\rm i}(\omega-\omega_k+{\rm i}\gamma)\tau}{\rm d}\tau = \frac{-1}{{\rm i}(\omega-\omega_k+{\rm i}\gamma)}. \tag{24.42}$$

When $\theta(-\tau)$ is present, the opposite sign must be used in front of γ. With these precautions, the time Fourier transforms of (24.36)–(24.41) yield the following GFs as functions of $\boldsymbol{k}$ and ω for noninteracting free particles at equilibrium:

$$G^{>}_{\rm pw}(\boldsymbol{k},\omega) = \frac{1}{\sqrt{2\pi}}\frac{1}{{\rm i}\hbar}(1 \pm \langle n_k\rangle)\delta(\omega-\omega_k),\ G^{<}_{\rm pw}(\boldsymbol{k},\omega) = \pm\frac{1}{\sqrt{2\pi}}\frac{1}{{\rm i}\hbar}\langle n_k\rangle\delta(\omega-\omega_k), \tag{24.43}$$

$$G^{\rm r}_{\rm pw}(\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{3/2}}\frac{1}{\hbar}\frac{1}{\omega-\omega_k+{\rm i}\gamma},\ G^{\rm a}_{\rm pw}(\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{3/2}}\frac{1}{\hbar}\frac{1}{\omega-\omega_k-{\rm i}\gamma}, \tag{24.44}$$

and from (24.26) also $G^{t}_{\rm pw}(\boldsymbol{k},\omega)$ and $G^{\bar{t}}_{\rm pw}$ can be obtained. Again the retarded and advanced GFs do not depend upon the state of the system while the other ones do through the occupation numbers n_k. The equations of motion, that we will now find, will confirm this fact.

The equations of motion

From (24.30) to (24.35), it is easy to find the equations verified by the various GFs for noninteracting particles. In particular, from (24.30) we obtain

$$\mathrm{i}\hbar\frac{\partial}{\partial t}G^{>}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = \mathrm{i}\hbar\frac{1}{\mathrm{i}\hbar}\sum_{\lambda}(1\pm\langle n_\lambda\rangle)\phi_\lambda(\boldsymbol{r})\phi^*_\lambda(\boldsymbol{r}')(-\mathrm{i}\omega_\lambda)\mathrm{e}^{-\mathrm{i}\omega_\lambda(t-t')}.$$

This is the same expression we would obtain by application of the single-particle Hamiltonian $\mathcal{H}(\boldsymbol{r})$ to $G^{>}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t')$. Thus, the equation of motion for this GF is

$$\left\{\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}(\boldsymbol{r})\right\}G^{>}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = 0. \tag{24.45}$$

Similarly for $G^{<}$,

$$\left\{\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}(\boldsymbol{r})\right\}G^{<}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = 0. \tag{24.46}$$

When the same procedure is applied to the other GFs, the time derivative of the step function brings about an extra term with the $\delta(t-t')$. This delta eliminates the exponential factors in the sum over the single-particle states. The remaining sum yields a new $\delta(\boldsymbol{r}-\boldsymbol{r}')$, and the final result is

$$\left\{\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}(\boldsymbol{r})\right\}G^{(r,a)}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t'). \tag{24.47}$$

This result confirms the meaning of the $G^{(r,a)}_{\mathrm{ni}}$ as the GF of the time-dependent Schrödinger equation (cf. (24.8)).

Finally, from (24.26) it is immediate that the time ordered and anti-time ordered GFs verify the same equations as the retarded and advanced GFs, with a minus sign for the anti-time ordered.

Following the same procedure, we find the equations satisfied by the various GFs as functions of the primed variables:

$$\left\{\mathrm{i}\hbar\frac{\partial}{\partial t'}+\mathcal{H}(\boldsymbol{r}')\right\}G^{>}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = 0. \tag{24.48}$$

and

$$\left\{\mathrm{i}\hbar\frac{\partial}{\partial t'}+\mathcal{H}(\boldsymbol{r}')\right\}G^{(r)}_{\mathrm{ni}}(\boldsymbol{r},t,\boldsymbol{r}',t') = -\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t'). \tag{24.49}$$

The simple equations found above hold for noninteracting particles. Finding the analogous dynamical equations in presence of interactions is a major task of the application of GFs to electron transport. It will be the subject of the last chapters of this book.

24.4 Green Functions and Mean Quantities

It will be shown here how the GFs allow us to evaluate mean values of single-particle physical quantities in a system of many interacting particles. We saw in Sect. 23.3 that in second quantization formalism, a single-particle observable $\mathcal{A}$ is given by (23.23). Therefore, its mean value in the state $|\Phi(t)\rangle$ is, in Schrödinger picture,

$$\langle \mathcal{A}(t)\rangle = \int\int \langle \Phi(t)|\Psi^\dagger(\boldsymbol{r}')A(\boldsymbol{r}',\boldsymbol{r})\Psi(\boldsymbol{r})|\Phi(t)\rangle \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}'.$$

Moving to the Heisenberg picture, this becomes

$$\begin{aligned}\langle \mathcal{A}(t)\rangle &= \int\int A(\boldsymbol{r}',\boldsymbol{r})\langle \Phi|\Psi^\dagger(\boldsymbol{r}',t)\Psi(\boldsymbol{r},t)|\Phi\rangle \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}' \\ &= \pm \mathrm{i}\hbar \int\int A(\boldsymbol{r}',\boldsymbol{r})G^<(\boldsymbol{r},t,\boldsymbol{r}',t)\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}'. \end{aligned} \tag{24.50}$$

Now, we already noticed that in wave mechanics we write

$$\int A(\boldsymbol{r}',\boldsymbol{r})\psi(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \mathcal{A}(\boldsymbol{r}')\psi(\boldsymbol{r}').$$

However, if this rule is applied directly into (24.50), we lose the information that $\mathcal{A}$ must act on the first argument of $G^<$, so that we write

$$\langle \mathcal{A}(t)\rangle = \pm \mathrm{i}\hbar \int \lim_{r'\to r} \mathcal{A}(\boldsymbol{r})G^<(\boldsymbol{r},t,\boldsymbol{r}',t)\mathrm{d}\boldsymbol{r}. \tag{24.51}$$

An equivalent expression can be written with the time ordered GF. Since $G^<$ coincides with $G^{(t)}$ for $t < t'$, as can be seen in (24.24), the above can be written as

$$\langle \mathcal{A}(t)\rangle = \pm \mathrm{i}\hbar \int \lim_{r'\to r}\lim_{t'\to t_+} \mathcal{A}(\boldsymbol{r})G^t(\boldsymbol{r},t,\boldsymbol{r}',t')\mathrm{d}\boldsymbol{r}.$$

As an important example, let us consider the number N of particles corresponding to the first-quantization operator[8] $\mathcal{A} = 1$ with matrix elements $\delta(\boldsymbol{r}-\boldsymbol{r}')$. Using (24.51), we obtain

$$\langle N(t)\rangle = \pm \mathrm{i}\hbar \int G^<(\boldsymbol{r},t,\boldsymbol{r},t)\mathrm{d}\boldsymbol{r}.$$

This is consistent with a density given by

$$\langle n(\boldsymbol{r},t)\rangle = \pm \mathrm{i}\hbar G^<(\boldsymbol{r},t,\boldsymbol{r},t),$$

which is in fact the mean value of the density operator $\Psi^\dagger(\boldsymbol{r},t)\Psi(\boldsymbol{r},t)$.

[8] In fact in wave mechanics each wavefunction is an eigenfunction of the number of particles with eigenvalue 1.

24.5 Spectral Density

The quantity in (24.23) is defined, to within simple constants that depend on the coefficients used in the Fourier transforms, as the *spectral density*. For homogeneous and stationary systems, in $(\boldsymbol{k},\omega)$ representation, it is given by

$$A(\boldsymbol{k},\omega)=\sqrt{2\pi}\hbar\mathrm{i}\left[G^{>}(\boldsymbol{k},\omega)-G^{<}(\boldsymbol{k},\omega)\right]=\sqrt{2\pi}\hbar\mathrm{i}\left[G^{\mathrm{r}}(\boldsymbol{k},\omega)-G^{\mathrm{a}}(\boldsymbol{k},\omega)\right]. \tag{24.52}$$

To connect this definition with the original concept of spectral density, namely the probability that a particle in the state $\boldsymbol{k}$ can be found with an energy $\hbar\omega$ and vice versa, let us first observe that in the case of noninteracting free particles at equilibrium, (24.43) yields immediately

$$A(\boldsymbol{k},\omega)=\delta(\omega-\omega_k), \tag{24.53}$$

as expected.

More in general, let us express the GFs with their definitions in terms of field operators. Using (24.27) and (24.28) and the definitions in (24.21), we have

$$\begin{aligned}A(\boldsymbol{k},\omega)&=\frac{\mathrm{i}\hbar}{2\pi}\int\mathrm{d}(\boldsymbol{r}-\boldsymbol{r}')\int\mathrm{d}(t-t')\mathrm{e}^{-\mathrm{i}[\boldsymbol{k}(\boldsymbol{r}-\boldsymbol{r}')-\omega(t-t')]}\\&\quad\times\frac{1}{\mathrm{i}\hbar}\langle\Psi(\boldsymbol{r},t)\Psi^{\dagger}(\boldsymbol{r}',t')\mp\Psi^{\dagger}(\boldsymbol{r}',t')\Psi(\boldsymbol{r},t)\rangle\\&=\frac{1}{2\pi}\int\mathrm{d}(\boldsymbol{r}-\boldsymbol{r}')\int\mathrm{d}(t-t')\mathrm{e}^{-\mathrm{i}[\boldsymbol{k}(\boldsymbol{r}-\boldsymbol{r}')-\omega(t-t')]}\langle[\Psi(\boldsymbol{r},t),\Psi^{\dagger}(\boldsymbol{r}',t')]_{\mp}\rangle.\end{aligned}$$

If we perform the integration in $\mathrm{d}\omega$, we obtain a $\delta(t-t')$ that can be used for the time integration; the commutators at equal times yield the $\delta(\boldsymbol{r}-\boldsymbol{r}')$, and the final result is

$$\int A(\boldsymbol{k},\omega)\mathrm{d}\omega=\int\mathrm{d}(\boldsymbol{r}-\boldsymbol{r}')\mathrm{e}^{-\mathrm{i}\boldsymbol{k}(\boldsymbol{r}-\boldsymbol{r}')}\delta(\boldsymbol{r}-\boldsymbol{r}')=1.$$

The previous result in (24.53) is consistent with this general property.

We have previously seen that the difference $G^{>}-G^{<}$, in terms of $(\boldsymbol{r},t,\boldsymbol{r}',t')$ gives the dynamics of all states, occupied plus empty, from $\boldsymbol{r}',t'$ to $\boldsymbol{r},t$. Now we see that, coherently, when the system is homogeneous and stationary, the same difference, in terms of $\boldsymbol{k}$ and ω, yields the spectral density.

24.5.1 Relation Between $G^{<}$ and $G^{>}$ and the Spectral Density at Equilibrium

If we are dealing with an equilibrium situation, we may assume that the GFs depend only upon the time difference, and we may consider $G^{<}$ for $t'=0$

without loss of generality:

$$G^<(\boldsymbol{r},t,\boldsymbol{r}',0)=\pm\frac{1}{\mathrm{i}\hbar}\langle\Psi^\dagger(\boldsymbol{r}',0)\Psi(\boldsymbol{r},t)\rangle.$$

Move to the variable ω by Fourier transforming:

$$G^<(\boldsymbol{r},\boldsymbol{r}',\omega)=\pm\frac{1}{\mathrm{i}\hbar}\int\langle\Psi^\dagger(\boldsymbol{r}',0)\Psi(\boldsymbol{r},t)\rangle\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t.$$

Now we use the evolution operator for the operator Ψ:

$$G^<(\boldsymbol{r},\boldsymbol{r}',\omega)=\pm\frac{1}{\mathrm{i}\hbar}\int\langle\Psi^\dagger(\boldsymbol{r}',0)\mathrm{e}^{\frac{\mathrm{i}}{\hbar}\mathcal{H}t}\Psi(\boldsymbol{r},0)\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\mathcal{H}t}\rangle\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t.$$

To evaluate the ensemble average we use the grand-canonical ensemble with the equilibrium density matrix $\rho=\frac{1}{Z}\mathrm{e}^{-\beta(\mathcal{H}-\mu\mathcal{N})}$, where $\beta=1/KT$, μ is the chemical potential, $\mathcal{N}$ is the number operator, and Z is the partition function $Z=\mathrm{Tr}\left[\mathrm{e}^{-\beta(\mathcal{H}-\mu\mathcal{N})}\right]$. Thus our $G^<$ at equilibrium becomes

$$G^<_{\mathrm{eq}}(\boldsymbol{r},\boldsymbol{r}',\omega)=\frac{\pm1}{\mathrm{i}\hbar Z}\int\mathrm{Tr}\left[\mathrm{e}^{-\beta(\mathcal{H}-\mu\mathcal{N})}\Psi^\dagger(\boldsymbol{r}',0)\mathrm{e}^{\frac{\mathrm{i}}{\hbar}\mathcal{H}t}\Psi(\boldsymbol{r},0)\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\mathcal{H}t}\right]\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t.$$

Let us now insert a full basis $|n\rangle$ of the total Hamiltonian $\mathcal{H}$. Since we consider systems where $\mathcal{H}$ and $\mathcal{N}$ commute, we may insert a basis set of eigenstates $|n\rangle$ with energy ϵ_n and number of particles N_n:

$$\begin{aligned}
&G^<_{\mathrm{eq}}(\boldsymbol{r},\boldsymbol{r}',\omega)\\
&\quad=\frac{\pm1}{\mathrm{i}\hbar Z}\int\sum_{n,m}\langle n|\mathrm{e}^{-\beta(\mathcal{H}-\mu\mathcal{N})}\Psi^\dagger(\boldsymbol{r}',0)|m\rangle\langle m|\mathrm{e}^{\frac{\mathrm{i}}{\hbar}\mathcal{H}t}\Psi(\boldsymbol{r},0)\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\mathcal{H}t}|n\rangle\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t\\
&\quad=\frac{\pm1}{\mathrm{i}\hbar Z}\int\sum_{n,m}\mathrm{e}^{-\beta(\epsilon_n-\mu N_n)}\langle n|\Psi^\dagger(\boldsymbol{r}',0)|m\rangle\mathrm{e}^{\frac{\mathrm{i}}{\hbar}(\epsilon_m-\epsilon_n)t}\langle m|\Psi(\boldsymbol{r},0)|n\rangle\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t\\
&\quad=\frac{\pm2\pi}{\mathrm{i}\hbar Z}\sum_{n,m}\mathrm{e}^{-\beta(\epsilon_n-\mu N_n)}\hbar\delta(\epsilon_m-\epsilon_n+\hbar\omega)\langle n|\Psi^\dagger(\boldsymbol{r}',0)|m\rangle\langle m|\Psi(\boldsymbol{r},0)|n\rangle.
\end{aligned}\tag{24.54}$$

In the same way, for $G^>$ we have

$$\begin{aligned}
&G^>_{\mathrm{eq}}(\boldsymbol{r},\boldsymbol{r}',\omega)\\
&\quad=\frac{2\pi}{\mathrm{i}\hbar Z}\sum_{n,m}\mathrm{e}^{-\beta(\epsilon_n-\mu N_n)}\hbar\delta(\epsilon_n-\epsilon_m+\hbar\omega)\langle n|\Psi(\boldsymbol{r},0)|m\rangle\langle m|\Psi^\dagger(\boldsymbol{r}',0)|n\rangle.
\end{aligned}$$

In this last expression, first we exchange the dummy indices n and m; then we take into account that the matrix elements are different from zero only if $N_m=N_n-1$, so that we may substitute, in the exponent, one of these values

with the other one. Furthermore, we use the argument of the δ to substitute also ϵ_m with $\epsilon_n - \hbar\omega$. The result is

$$G^{>}_{\rm eq}(\boldsymbol{r},\boldsymbol{r}',\omega) = \frac{2\pi}{\mathrm{i}\hbar Z}\mathrm{e}^{-\beta(\mu-\hbar\omega)} \times \sum_{n,m} \mathrm{e}^{-\beta(\epsilon_n-\mu N_n)}\hbar\delta(\epsilon_m-\epsilon_n+\hbar\omega)$$
$$\times\langle m|\Psi(\boldsymbol{r},0)|n\rangle\langle n|\Psi^{\dagger}(\boldsymbol{r}',0)|m\rangle.$$

Comparing this with the similar expression (24.54) obtained for $G^{<}$, we have the relation

$$G^{>}_{\rm eq}(\boldsymbol{r},\boldsymbol{r}',\omega) = \pm\mathrm{e}^{\beta(\hbar\omega-\mu)}G^{<}_{\rm eq}(\boldsymbol{r},\boldsymbol{r}',\omega).$$

If we assume that the system is homogeneous and move to the k representation, we obtain

$$G^{>}_{\rm eq}(\boldsymbol{k},\omega) = \pm\mathrm{e}^{\beta(\hbar\omega-\mu)}G^{<}_{\rm eq}(\boldsymbol{k},\omega). \tag{24.55}$$

At this point, we use the definition of the spectral density in (24.52) to substitute $G^{>}$ and obtain

$$G^{<}_{\rm eq}(\boldsymbol{k},\omega) + \frac{1}{\sqrt{2\pi}\mathrm{i}\hbar}A(\boldsymbol{k},\omega) = \pm\mathrm{e}^{\beta(\hbar\omega-\mu)}G^{<}_{\rm eq}(\boldsymbol{k},\omega),$$

or

$$\boxed{G^{<}_{\rm eq}(\boldsymbol{k},\omega) = \pm\tfrac{1}{\sqrt{2\pi}\mathrm{i}\hbar}A(\boldsymbol{k},\omega)f_{\circ}(\omega)} \tag{24.56}$$

where

$$\boxed{f_{\circ}(\omega) = \tfrac{1}{\mathrm{e}^{\beta(\hbar\omega-\mu)}\mp 1}} \tag{24.57}$$

Furthermore, since

$$1 \pm f_{\circ} = \frac{\mathrm{e}^{\beta(\hbar\omega-\mu)}}{\mathrm{e}^{\beta(\hbar\omega-\mu)}\mp 1} = \mathrm{e}^{\beta(\hbar\omega-\mu)}f_{\circ},$$

equation (24.55) becomes[9]

$$G^{>}_{\rm eq} = \pm\mathrm{e}^{\beta(\hbar\omega-\mu)}G^{<}_{\rm eq} = \frac{1}{\sqrt{2\pi}\mathrm{i}\hbar}A(\boldsymbol{k},\omega)[1\pm f_{\circ}(\omega)]. \tag{24.58}$$

The above equations are consistent with the physical interpretation we have previously given of $G^{<}$ and $G^{>}$, and can be compared with (24.43) for noninteracting particles.

[9] Equation (24.58) appears often in the literature without the double sign for bosons or fermions. This is obtained, following [228], at the price of a different definitions of the Fourier transform for $G^{>}$ and $G^{<}$, as explicitly stated in p. 7 of [228].

25

Wick–Matsubara Theorems

In Chap. 23, it was shown that in second-quantization formalism observables are formulated in terms of field operators. In a perturbative expansion, a quantity is given as a series of successive powers of the perturbation Hamiltonian. This implies products of increasing numbers of field operators. Wick–Matsubara theorems deal exactly with such products. The Wick theorem is an operator identity and is useful in particular to find the electronic properties of a many-body system. The Matsubara version of the theorem deals with equilibrium averages and is particularly useful to study the dynamical evolution of electron GFs. Wick–Matsubara theorems are the basic tools for the definition and use of Feynman graphs.

25.1 Time-Ordered Products, Normal Products, and Contractions

Let $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ be field operators in interaction picture with the unperturbed Hamiltonian defined for noninteracting particles. The time-ordering operator $\mathcal{T}$, already seen in connection with the time-ordered GF, acts on a product of field operators, ordering them in such a way that their time arguments increase from right to left, i.e., in the order of their applications. More precisely, the *time-ordered product* of the operators $\mathcal{A}(t_a)$, $\mathcal{B}(t_b)$, $\mathcal{C}(t_c)$, ... is defined as

$$\mathcal{T}[\mathcal{A}(t_a)\mathcal{B}(t_b)\mathcal{C}(t_c)\ \ldots] = f(p)\mathcal{X}(t_x)\mathcal{Y}(t_y)\ \ldots,$$

where $\mathcal{X}(t_x)\mathcal{Y}(t_y)\ \ldots$ are the field operators $\mathcal{A}(t_a)\mathcal{B}(t_b)\ \ldots$ ordered as indicated above, and $f(p)$ is unity for bosons and $(-1)^p$ for fermions, where p is the number of permutations necessary to reach the final ordering starting from the original one.

The *normal product* of a number of field operators

$$\mathcal{N}[\mathcal{A}(t_a)\mathcal{B}(t_b)\mathcal{C}(t_c)\ \ldots] = f(p)\mathcal{X}(t_x)\mathcal{Y}(t_y)\ \ldots$$

is defined as the product of the operators ordered in such a way that all creation operators are at the left and all annihilation operators are on the right, kept in their original time order. $f(p)$ has the same meaning as above.

Furthermore, both $\mathcal{T}$ and $\mathcal{N}$ are defined as distributive, such that

$$\mathcal{T}[\mathcal{AB}\ldots + \mathcal{XY}\ldots] = \mathcal{T}[\mathcal{AB}\ldots] + \mathcal{T}[\mathcal{XY}\ldots],$$

and similarly for $\mathcal{N}$.

The *contraction* of two field operators $\mathcal{A}$ e $\mathcal{B}$ is defined as

$$\mathcal{A}^{\cdot}\mathcal{B}^{\cdot} = \mathcal{T}[\mathcal{AB}] - \mathcal{N}[\mathcal{AB}] \tag{25.1}$$

and therefore it is the correction that must be made if we substitute a time-ordered product of two field operators with their normal product. The contraction of two field operators of the same kind, both annihilation or both creation, vanishes. In fact, with obvious symbols,

$$\begin{aligned} &\text{if } t_1 > t_2 \qquad \Psi(\boldsymbol{r}_1,t_1)^{\cdot}\Psi(\boldsymbol{r}_2,t_2)^{\cdot} = \Psi(1)\Psi(2) - \Psi(1)\Psi(2) = 0,\\ &\text{if } t_1 < t_2 \qquad \Psi(\boldsymbol{r}_1,t_1)^{\cdot}\Psi(\boldsymbol{r}_2,t_2)^{\cdot} = \pm\Psi(2)\Psi(1) - \Psi(1)\Psi(2) = 0, \end{aligned}$$

because of (23.33). A similar result is obtained for two creation field operators:

$$\Psi^{\dagger}(\boldsymbol{r}_1,t_1)^{\cdot}\Psi^{\dagger}(\boldsymbol{r}_2,t_2)^{\cdot} = 0.$$

If, furthermore, the two field operators of a contraction are one creation and one annihilation, the contraction still vanishes if the annihilation operator is evaluated at a previous time:

$$\begin{aligned} &\text{if } t_1 > t_2 \qquad \Psi^{\dagger}(\boldsymbol{r}_1,t_1)^{\cdot}\Psi(\boldsymbol{r}_2,t_2)^{\cdot} = \Psi^{\dagger}(1)\Psi(2) - \Psi^{\dagger}(1)\Psi(2) = 0,\\ &\text{if } t_1 < t_2 \qquad \Psi(\boldsymbol{r}_1,t_1)^{\cdot}\Psi^{\dagger}(\boldsymbol{r}_2,t_2)^{\cdot} = \pm\Psi^{\dagger}(2)\Psi(1) - [\pm\Psi^{\dagger}(2)\Psi(1)] = 0. \end{aligned}$$

The only nonvanishing contractions are those of two operators with a creation evaluated at a previous time and an annihilation at a later time:
if $t_1 > t_2$,

$$\Psi(\boldsymbol{r}_1,t_1)^{\cdot}\Psi^{\dagger}(\boldsymbol{r}_2,t_2)^{\cdot} = \Psi(1)\Psi^{\dagger}(2) - [\pm\Psi^{\dagger}(2)\Psi(1)] = [\Psi(1),\Psi^{\dagger}(2)]_{\mp}\ ;$$

if $t_1 < t_2$,

$$\Psi^{\dagger}(\boldsymbol{r}_1,t_1)^{\cdot}\Psi(\boldsymbol{r}_2,t_2)^{\cdot} = \pm\Psi(2)\Psi^{\dagger}(1) - \Psi^{\dagger}(1)\Psi(2)] = -[\Psi^{\dagger}(1),\Psi(2)]_{\mp}.$$

Remember that at the end of Sect. 23.5 it was shown that for a noninteracting Hamiltonian, all the above commutators are c-numbers. Thus, we may conclude that *all contractions are c-numbers*.

Let us finally extend the definition of a contraction, when it appears within a normal product. The two contracted operators must first be brought to adjacent position within the product with the usual change of sign for fermions,

then the pair is brought outside of the product and contracted:

$$\mathcal{N}[\mathcal{A}\mathcal{B}^{\cdot}\mathcal{C}\ \ldots\ \mathcal{D}\mathcal{E}^{\cdot}\mathcal{F}\ \ldots\ ] = f(p)\mathcal{B}^{\cdot}\mathcal{E}^{\cdot}N[\mathcal{A}\mathcal{C}\ \ldots\ \mathcal{D}\mathcal{F}\ \ldots\ ]. \tag{25.2}$$

Similar definition holds for more than one contraction. The operators of the contractions are indicated with upper dots, or double and triple dots if more than one contraction are present.

Note that, according to the above definitions, exchanging two field operators inside a time-ordered or a normal product implies the standard commutation rule for bosons or fermions.

25.2 Wick Theorem

Following [145], for the demonstration of Wick theorem, we first consider the following lemma.

25.2.1 Lemma

Let $\mathcal{A}, \mathcal{B}, \ldots, \mathcal{C}, \mathcal{D}, \mathcal{X}$ be field operators in the interaction picture, with $\mathcal{X}$ evaluated at a time less or equal to the times of all the other ones. We then state that

$$\begin{aligned}\mathcal{N}[\mathcal{A}\mathcal{B}\ \ldots\ \mathcal{C}\ \mathcal{D}]\mathcal{X} = \mathcal{N}[\mathcal{A}\mathcal{B}\ \ldots\ \mathcal{C}\ \mathcal{D}\mathcal{X}] + \mathcal{N}[\mathcal{A}\mathcal{B}\ \ldots\ \mathcal{C}\ \mathcal{D}^{\cdot}\mathcal{X}^{\cdot}] \\ + \mathcal{N}[\mathcal{A}\mathcal{B}\ \ldots\ \mathcal{C}^{\cdot}\ \mathcal{D}\mathcal{X}^{\cdot}] + \ \ldots\ + \mathcal{N}[\mathcal{A}^{\cdot}\mathcal{B}\ \ldots\ \mathcal{C}\ \mathcal{D}\mathcal{X}^{\cdot}].\end{aligned} \tag{25.3}$$

The r.h.s. contains the normal product of all operators with $\mathcal{X}$ added at the end plus all normal products with the contractions of $\mathcal{X}$ with all the other ones.

For the demonstration of this lemma, let us distinguish several cases.

$\mathcal{X}$ is an Annihilation Operator

In this case, the demonstration is particularly simple since the first term in the r.h.s. is equal to the l.h.s., and all the other terms vanish since contain contractions of pairs where the annihilation $\mathcal{X}$ is present with the lower time.

$\mathcal{X}$ is a Creation and All the Others are Annihilations

Note that in this case the normal product on the l.h.s. is equal to the product itself, since it contains only annihilation operators. Let us then proceed by induction. When the normal product contains only one operator, the l.h.s. of (25.3) becomes

$$\mathcal{N}[\mathcal{A}]\mathcal{X} = \mathcal{A}\mathcal{X}.$$

As it regards the r.h.s., taking into account that the time of $\mathcal{X}$ is less or equal to that of $\mathcal{A}$ and applying the definition of contraction (25.2), it becomes

$$\mathcal{N}[\mathcal{A}\mathcal{X}] + \mathcal{N}[\mathcal{A}^{\cdot}\mathcal{X}^{\cdot}] = \mathcal{N}[\mathcal{A}\mathcal{X}] + \mathcal{A}^{\cdot}\mathcal{X}^{\cdot} = \mathcal{N}[\mathcal{A}\mathcal{X}] + \mathcal{T}[\mathcal{A}\mathcal{X}] - \mathcal{N}[\mathcal{A}\mathcal{X}] = \mathcal{A}\mathcal{X},$$

and the thesis is demonstrated. Next assume that the thesis is true when the normal product contains n field operators. Equation (25.3) is then true for n annihilation operators $\mathcal{A}\mathcal{B}\ldots\mathcal{C}\ \mathcal{D}$. Let us multiply the same equation on the left for an additional annihilation operator $\mathcal{E}$, evaluated at a time greater than that of $\mathcal{X}$:

$$\begin{aligned}\mathcal{E}\mathcal{N}[\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}]\mathcal{X} = \mathcal{E}\mathcal{N}[\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}] + \mathcal{E}\mathcal{N}[\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}^{\cdot}\mathcal{X}^{\cdot}] \\ + \mathcal{E}\mathcal{N}[\mathcal{A}\mathcal{B}\ldots\mathcal{C}^{\cdot}\mathcal{D}\mathcal{X}^{\cdot}] + \ldots + \mathcal{E}\mathcal{N}[\mathcal{A}^{\cdot}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}^{\cdot}]. \quad (25.4)\end{aligned}$$

At this point, it is convenient to consider the first term in the r.h.s. separately. In this term, $\mathcal{X}$ must be brought to the first place since it is a creation operator and all the other ones are annihilations:

$$\mathcal{I} \equiv \mathcal{E}\mathcal{N}[\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}] = \mathcal{E}f(p)\mathcal{X}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}.$$

Then, using the definition of contraction in (25.1) and taking into account that the time of $\mathcal{X}$ is less than that of $\mathcal{E}$, we may write

$$\begin{aligned}\mathcal{I} &= f(p)\mathcal{T}[\mathcal{E}\mathcal{X}]\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D} = f(p)\mathcal{E}^{\cdot}\mathcal{X}^{\cdot}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D} + f(p)\mathcal{N}[\mathcal{E}\mathcal{X}]\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D} \\ &= (f(p))^2\mathcal{N}[\mathcal{E}^{\cdot}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}^{\cdot}] + f(p+1)\mathcal{X}\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D} \\ &= \mathcal{N}[\mathcal{E}^{\cdot}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}^{\cdot}] + (f(p+1))^2\mathcal{N}[\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}] \\ &= \mathcal{N}[\mathcal{E}^{\cdot}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}^{\cdot}] + \mathcal{N}[\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}],\end{aligned}$$

where we used that $f(p) = \pm 1$.

In the other terms of the r.h.s. of (25.4) $\mathcal{X}$ is always contracted and therefore is not really internal to the normal product. Thus, $\mathcal{E}$ may be taken inside the normal product, being an annihilation operator as all the others, and left in the first place. The same thing can be done to the l.h.s. so that (25.4) becomes

$$\begin{aligned}\mathcal{N}[\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}]\mathcal{X} = \mathcal{N}[\mathcal{E}^{\cdot}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}^{\cdot}] + \mathcal{N}[\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}] \\ + \mathcal{N}[\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}^{\cdot}\mathcal{X}^{\cdot}] + \mathcal{N}[\mathcal{E}\mathcal{A}\mathcal{B}\ldots\mathcal{C}^{\cdot}\mathcal{D}\mathcal{X}^{\cdot}] \\ + \ldots + \mathcal{N}[\mathcal{E}\mathcal{A}^{\cdot}\mathcal{B}\ldots\mathcal{C}\mathcal{D}\mathcal{X}^{\cdot}].\end{aligned}$$

Q.E.D.

$\mathcal{X}$ is a Creation Operator and the Others are Creation and Annihilation

For this last case, we may assume that the operators $\mathcal{A}\mathcal{B}\ldots\mathcal{C}\mathcal{D}$ in the normal products of (25.3) are already normally ordered inside the square brackets.

If this is not the case, in fact, we may order them in both sides obtaining the same factor $f(p)$ in both of them. Once the operators are normally ordered, all creation operators on the left in the product inside the square brackets can be taken outside the brackets: the contractions that are lost contain two creation operators and are zero as shown in Sect. 25.1. Thus, the situation is reduced to the previous case, and the lemma is entirely demonstrated.

25.2.2 Wick Theorem

We are now in the condition to prove Wick theorem: *The time-ordered product of a number of field operators in the interaction picture is given by their normal product, plus the sum of the normal products of the same operators with all possible contractions*:

$$\mathcal{T}[\mathcal{ABC}\ldots\mathcal{XYZ}] = \mathcal{N}[\mathcal{ABC}\ldots\mathcal{XYZ}] + \mathcal{N}[\mathcal{A}\cdot\mathcal{B}\cdot\mathcal{C}\ldots\mathcal{XYZ}] + \mathcal{N}[\mathcal{A}\cdot\mathcal{BC}\cdot\ldots\mathcal{XYZ}] + \ldots + \mathcal{N}[\mathcal{A}\cdot\mathcal{B}\cdot\cdot\mathcal{C}\cdots\ldots\mathcal{X}\cdots\mathcal{Y}\cdot\cdot\mathcal{Z}\cdot]. \tag{25.5}$$

For the proof let us proceed again by induction. If the product has only two operators, the thesis coincides with the definition of contraction in (25.1). Consider then the time product of $n+1$ field operators:

$$\mathcal{T}[\mathcal{AB}\ldots\mathcal{XYZ}] = f(p)\mathcal{A}'\mathcal{B}'\ldots\mathcal{X}'\mathcal{Y}'\mathcal{Z}',$$

where $\mathcal{A}', \mathcal{B}', \ldots$ are the same operators $\mathcal{A}, \mathcal{B}, \ldots$ in the correct time order. We may now take the time-ordered product of the first n operators, since they are already ordered, and apply the theorem, assumed valid for a product of n operators:

$$= f(p)\mathcal{T}[\mathcal{A}'\mathcal{B}'\ldots\mathcal{X}'\mathcal{Y}']\mathcal{Z}' = f(p)\left\{\mathcal{N}[\mathcal{A}'\mathcal{B}'\ldots\mathcal{X}'\mathcal{Y}'] + \mathcal{N}[\mathcal{A}'\cdot\mathcal{B}'\cdot\ldots\mathcal{X}'\mathcal{Y}'] + \ldots\right\}\mathcal{Z}'.$$

At each term, we may now apply the lemma just proved, and reset the field operators in the original order, thus obtaining the equality required by the thesis.

Note that if we apply Wick theorem (25.5) to the mean value of a time-ordered product of fields operators to the vacuum state:

(a) If the product on the l.h.s. of (25.5) does not contain an equal number of creation and annihilation operators, its mean value on the vacuum state becomes the scalar product of states containing a different number of particles and vanishes. Only products with an equal number of creation and annihilation operators must be considered.
(b) The normal products contain annihilation operators on the right and applied to the vacuum yield zero. Thus only the totally contracted products remain.

(c) The contractions are c-numbers, different from zero only if formed by a creation operator and an annihilation operator evaluated at a later time, as shown in Sect. 25.1.

What is left is therefore the sum of the products of all possible contractions with the property indicated above:

$$\langle 0|\mathcal{T}[\mathcal{ABC}\ \ldots\ \mathcal{XYZ}]|0\rangle = (\mathcal{A}\cdot\mathcal{B}\cdot)(\mathcal{C}\cdot\mathcal{X}\cdot)\ \ldots\ +\ \ldots.$$

25.3 Wick–Matsubara Theorem

This theorem, due to Matsubara, is of a somewhat different nature with respect to that of the previous Wick theorem and concerns the mean value of a product of field operators in interaction picture on an equilibrium ensemble of states:

$$\langle \mathcal{A}(t_a)\mathcal{B}(t_b)\mathcal{C}(t_c)\ \ldots\rangle_{\mathrm{eq}} = \frac{1}{Z}\mathrm{Tr}\{\mathrm{e}^{-\beta\mathcal{H}_\circ}\mathcal{A}(t_a)\mathcal{B}(t_b)\mathcal{C}(t_c)\ \ldots\}, \tag{25.6}$$

where $\mathcal{H}_\circ$ is the unperturbed Hamiltonian of noninteracting particles, and Z is the partition function.

Note again that these mean values are different from zero only if the products contain an equal number of creation and annihilation operators. In fact, the states $|n_1, n_2, \ldots\rangle$ are eigenstates of $\mathcal{H}_\circ$, and using such basis we see that if the product in (25.6) does not contain an equal number of creation and annihilation operators, the mean value would contain only scalar products of states containing different numbers of particles and vanish.

Before giving the formulation of the theorem, we shall consider an example from which it will be easy to understand its general statement and validity.

Assume we want to evaluate the mean value at equilibrium of the product

$$\langle \Psi^\dagger(\boldsymbol{r}_1, t_1)\Psi(\boldsymbol{r}_2, t_2)\Psi^\dagger(\boldsymbol{r}_3, t_3)\Psi(\boldsymbol{r}_4, t_4)\rangle. \tag{25.7}$$

Using the time dependence of field operators given in (23.31) and (23.32) for noninteracting particles, this becomes

$$\sum_{\lambda\mu\nu\xi} \mathrm{e}^{\mathrm{i}\omega_\lambda(t_1-t_\circ)}\mathrm{e}^{-\mathrm{i}\omega_\mu(t_2-t_\circ)}\mathrm{e}^{\mathrm{i}\omega_\nu(t_3-t_\circ)}\mathrm{e}^{-\mathrm{i}\omega_\xi(t_4-t_\circ)} \times \phi^*_\lambda(\boldsymbol{r}_1)\phi_\mu(\boldsymbol{r}_2)\phi^*_\nu(\boldsymbol{r}_3)\phi_\xi(\boldsymbol{r}_4)\langle c^\dagger_\lambda c_\mu c^\dagger_\nu c_\xi\rangle. \tag{25.8}$$

First commute the first operator successively with those that follow, until it occupies the last place (in this derivation we keep terms which are obviously zero, to facilitate the extension to the general case):

$$\begin{aligned} c^\dagger_\lambda c_\mu c^\dagger_\nu c_\xi &= \{[c^\dagger_\lambda, c_\mu]_\mp \pm c_\mu c^\dagger_\lambda\}c^\dagger_\nu c_\xi = [c^\dagger_\lambda, c_\mu]_\mp c^\dagger_\nu c_\xi \pm c_\mu c^\dagger_\lambda c^\dagger_\nu c_\xi \\ &= [c^\dagger_\lambda, c_\mu]_\mp c^\dagger_\nu c_\xi \pm c_\mu\{[c^\dagger_\lambda, c^\dagger_\nu]_\mp \pm c^\dagger_\nu c^\dagger_\lambda\}c_\xi \\ &= [c^\dagger_\lambda, c_\mu]_\mp c^\dagger_\nu c_\xi \pm c_\mu[c^\dagger_\lambda, c^\dagger_\nu]_\mp c_\xi + c_\mu c^\dagger_\nu c^\dagger_\lambda c_\xi \\ &= [c^\dagger_\lambda, c_\mu]_\mp c^\dagger_\nu c_\xi \pm c_\mu[c^\dagger_\lambda, c^\dagger_\nu]_\mp c_\xi + c_\mu c^\dagger_\nu[c^\dagger_\lambda, c_\xi]_\mp \pm c_\mu c^\dagger_\nu c_\xi c^\dagger_\lambda. \end{aligned} \tag{25.9}$$

Then consider the mean value of the last term, using the cyclic property of the trace:

$$\langle c_\mu c_\nu^\dagger c_\xi c_\lambda^\dagger \rangle = \frac{1}{Z}\mathrm{Tr}\{\mathrm{e}^{-\beta H_\circ} c_\mu c_\nu^\dagger c_\xi c_\lambda^\dagger\} = \frac{1}{Z}\mathrm{Tr}\{c_\lambda^\dagger \mathrm{e}^{-\beta H_\circ} c_\mu c_\nu^\dagger c_\xi\}. \tag{25.10}$$

At this point, we must invert the first two operators. For this purpose, note that by using the second commutator in (23.30),

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\left\{\mathrm{e}^{-\beta\mathcal{H}_\circ} c_\lambda^\dagger \mathrm{e}^{\beta H_\circ}\right\} = \mathrm{e}^{-\beta H_\circ}[c_\lambda^\dagger, \mathcal{H}_\circ]\mathrm{e}^{\beta H_\circ} = -\left\{\mathrm{e}^{-\beta H_\circ} c_\lambda^\dagger \mathrm{e}^{\beta H_\circ}\right\}\hbar\omega_\lambda.$$

Solving the differential equation,

$$\left\{\mathrm{e}^{-\beta\mathcal{H}_\circ} c_\lambda^\dagger \mathrm{e}^{\beta H_\circ}\right\} = c_\lambda^\dagger \mathrm{e}^{-\beta\hbar\omega_\lambda}, \quad \text{or} \quad \mathrm{e}^{-\beta\mathcal{H}_\circ} c_\lambda^\dagger = c_\lambda^\dagger \mathrm{e}^{-\beta H_\circ}\mathrm{e}^{-\beta\hbar\omega_\lambda}.$$

Substitution of this result into (25.10) yields

$$\langle c_\mu c_\nu^\dagger c_\xi c_\lambda^\dagger \rangle = \frac{1}{Z}\mathrm{e}^{\beta\hbar\omega_\lambda}\mathrm{Tr}\{\mathrm{e}^{-\beta H_\circ} c_\lambda^\dagger c_\mu c_\nu^\dagger c_\xi\} = \mathrm{e}^{\beta\hbar\omega_\lambda}\langle c_\lambda^\dagger c_\mu c_\nu^\dagger c_\xi\rangle.$$

This result may now be inserted in the mean value of (25.9), and, remembering that all the present commutators are c-numbers, we obtain

$$\langle c_\lambda^\dagger c_\mu c_\nu^\dagger c_\xi\rangle = \frac{1}{(1 \mp \mathrm{e}^{\beta\hbar\omega_\lambda})}\left\{[c_\lambda^\dagger, c_\mu]_\mp\langle c_\nu^\dagger c_\xi\rangle \pm [c_\lambda^\dagger, c_\nu^\dagger]_\mp\langle c_\mu c_\xi\rangle + [c_\lambda^\dagger, c_\xi]_\mp\langle c_\mu c_\nu^\dagger\rangle\right\}. \tag{25.11}$$

The evaluation of the mean value of the product of four field operators has been reduced to the evaluations of three mean values of the products of two field operators. The same elaboration can now be repeated for these new products. For example, for the first product we obtain

$$\langle c_\nu^\dagger c_\xi\rangle = \frac{1}{(1 \mp \mathrm{e}^{\beta\hbar\omega_\nu})}[c_\nu^\dagger, c_\xi]_\mp. \tag{25.12}$$

In the following mean value, the first operator is an annihilation. In its elaboration, we must use first commutator in (23.30) instead of the second. The same result is obtained with an opposite sign in the exponent:

$$\langle c_\mu c_\xi\rangle = \frac{1}{(1 \mp \mathrm{e}^{-\beta\hbar\omega_\mu})}[c_\mu, c_\xi]_\mp, \tag{25.13}$$

and similarly for the last mean product. Equation (25.11) then becomes

$$\begin{aligned}\langle c_\lambda^\dagger c_\mu c_\nu^\dagger c_\xi\rangle = {} & \frac{[c_\lambda^\dagger, c_\mu]_\mp}{(1 \mp \mathrm{e}^{\beta\hbar\omega_\lambda})}\frac{[c_\nu^\dagger, c_\xi]_\mp}{(1 \mp \mathrm{e}^{\beta\hbar\omega_\nu})} \\ & \pm \frac{[c_\lambda^\dagger, c_\nu^\dagger]_\mp}{(1 \mp \mathrm{e}^{\beta\hbar\omega_\lambda})}\frac{[c_\mu, c_\xi]_\mp}{(1 \mp \mathrm{e}^{-\beta\hbar\omega_\mu})} + \frac{[c_\lambda^\dagger, c_\xi]_\mp}{(1 \mp \mathrm{e}^{\beta\hbar\omega_\lambda})}\frac{[c_\mu, c_\nu^\dagger]_\mp}{(1 \mp \mathrm{e}^{-\beta\hbar\omega_\mu})}.\end{aligned}$$

Using (25.12) and (25.13), the above becomes

$$\langle c_\lambda^\dagger c_\mu c_\nu^\dagger c_\xi\rangle = \langle c_\lambda^\dagger c_\mu\rangle\langle c_\nu^\dagger c_\xi\rangle \pm \langle c_\lambda^\dagger c_\nu^\dagger\rangle\langle c_\mu c_\xi\rangle + \langle c_\lambda^\dagger c_\xi\rangle\langle c_\mu c_\nu^\dagger\rangle.$$

Inserting back the numerical factor of (25.8), we immediately obtain

$$\begin{aligned}&\langle\Psi^\dagger(\boldsymbol{r}_1,t_1)\Psi(\boldsymbol{r}_2,t_2)\Psi^\dagger(\boldsymbol{r}_3,t_3)\Psi(\boldsymbol{r}_4,t_4)\rangle\\ &\quad= \langle\Psi^\dagger(1)\Psi(2)\rangle\langle\Psi^\dagger(3)\Psi(4)\rangle \pm \langle\Psi^\dagger(1)\Psi^\dagger(3)\rangle\langle\Psi(2)\Psi(4)\rangle\\ &\qquad+\langle\Psi^\dagger(1)\Psi(4)\rangle\langle\Psi(2)\Psi^\dagger(3)\rangle.\end{aligned}$$

Here vanishing terms are still present. The nonvanishing terms, those that contain pairs with a creation and an annihilation operator, are immediately related to noninteracting equilibrium GFs.

We can now state the general Wick–Matsubara theorem: *the equilibrium statistical average of a time-ordered product of field operators of non-interacting particles is given by the sum of equilibrium statistical average of all possible time-ordered pairs of the field operators. In case of fermions, a minus sign must be inserted in front of each product according to the number of interchanges needed to achieve the desired pairing.*

$$\langle\mathcal{T}[\mathcal{ABC}\ldots\mathcal{F}]\rangle_\circ = \pm\langle\mathcal{T}[\mathcal{AB}]\rangle_\circ\langle\mathcal{T}[\mathcal{CD}]\rangle_\circ\ldots \pm \langle\mathcal{T}[\mathcal{AC}]\rangle_\circ\langle\mathcal{T}[\mathcal{BD}]\rangle_\circ\ldots \pm \ \ldots. \tag{25.14}$$

Note that the mean value of the product of field operators is nonzero only if the product contains an equal number of creations and annihilations and therefore must contain an even number of operators, a necessary condition for (25.14) to be nonambiguous.

As it regards the proof of the theorem, if we change the order of the field operators in (25.14), the same change of sign occurs in both sides of the equation so that we may assume that the operators are already in the proper time order. At this point, the proof of the theorem follows exactly the same lines as in the example developed above, and it is not necessary to repeat it here. Only a number of dots (...) are often inserted in the proof to indicate that the number of operators may be any finite (even) number.

26

Perturbation Expansion of Green Functions: Feynman Diagrams and Dyson Equation

In this chapter, we shall see how it is possible to obtain a perturbation expansion of the Green functions that allows us to evaluate them, in principle, at times far from the noninteracting initial state. This perturbation expansion may be formulated in terms *Feynman diagrams*.

The resulting equation for the GFs, called *Dyson equation*, is in a sense the quantum equivalent of Boltzmann semiclassical equation, in its integral form. The Boltzmann equation contains, in the collision integral, the effect on the distribution function of the electron interactions with scattering agents, such as phonons and impurities. Similarly, the Dyson equation for the GFs contains the effect of the interactions in the *self-energy*. The Feynman diagrams provide a definition of such a function and a way to evaluate it, in principle, as an expansion in powers of the interaction Hamiltonian.

26.1 The Interaction Picture

Let us assume that at the initial time $t_\circ$ our system is in equilibrium with an unperturbed Hamiltonian $\mathcal{H}_\circ$, and that at $t_\circ$ the perturbation $\mathcal{H}'(t)$ is applied. We shall work in the interaction picture (see Sect. 2.2.4) and take $t_\circ$ as the reference time for the transformation between interaction and Schrödinger pictures.

We know that the transformation from the Schrödinger picture to the Heisenberg picture is obtained with the operator $\mathcal{U}^\dagger(t, t_\circ)$ that "evolves back" the vectors to the initial reference time $t_\circ$, where $\mathcal{U}(t, t')$ is the evolution operator due to the total Hamiltonian $\mathcal{H}$. The interaction picture is obtained from the Schrödinger picture by back evolving with the evolution operator $\mathcal{U}_\circ(t, t')$ due only to the unperturbed Hamiltonian $\mathcal{H}_\circ$:

$$|\Phi_{\mathrm{I}}(t)\rangle = \mathcal{U}_\circ^\dagger(t, t_\circ)|\Phi_{\mathrm{S}}(t)\rangle, \quad \Psi_{\mathrm{I}}(\boldsymbol{r}, t) = \mathcal{U}_\circ^\dagger(t, t_\circ)\Psi(\boldsymbol{r})\mathcal{U}_\circ(t, t_\circ). \tag{26.1}$$

Let us now find the evolution operator for the state vectors in the interaction picture:

$$|\Phi_{\mathrm{I}}(t)\rangle = \mathcal{U}_\circ^\dagger(t,t_\circ)|\Phi_{\mathrm{S}}(t)\rangle = \mathcal{U}_\circ^\dagger(t,t_\circ)\mathcal{U}(t,t')|\Phi_{\mathrm{S}}(t')\rangle$$
$$= \mathcal{U}_\circ^\dagger(t,t_\circ)\mathcal{U}(t,t')\mathcal{U}_\circ(t',t_\circ)|\Phi_{\mathrm{I}}(t')\rangle.$$

Thus, the evolution we looked for is

$$\mathcal{U}_{\mathrm{I}}(t,t') = \mathcal{U}_\circ^\dagger(t,t_\circ)\mathcal{U}(t,t')\mathcal{U}_\circ(t',t_\circ). \tag{26.2}$$

A differential equation for this operator can be easily found as follows:

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{U}_{\mathrm{I}}(t,t')$$
$$= \left(\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{U}_\circ^\dagger(t,t_\circ)\right)\mathcal{U}(t,t')\mathcal{U}_\circ(t',t_\circ) + \mathcal{U}_\circ^\dagger(t,t_\circ)\left(\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{U}(t,t')\right)\mathcal{U}_\circ(t',t_\circ)$$
$$= \mathcal{U}_\circ^\dagger(t,t_\circ)\left(-\mathcal{H}_\circ + \mathcal{H}\right)\mathcal{U}(t,t')\mathcal{U}_\circ(t',t_\circ) = \mathcal{U}_\circ^\dagger(t,t_\circ)\mathcal{H}'\mathcal{U}(t,t')\mathcal{U}_\circ(t',t_\circ).$$

Inserting $\mathcal{U}_\circ(t,t_\circ)$ times its inverse after $\mathcal{H}'$, we obtain

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{U}_{\mathrm{I}}(t,t') = \mathcal{H}'_{\mathrm{I}}\mathcal{U}_{\mathrm{I}}(t,t'), \tag{26.3}$$

where $\mathcal{H}'_{\mathrm{I}} = \mathcal{U}_\circ^\dagger(t,t_\circ)\mathcal{H}'\mathcal{U}_\circ(t,t_\circ)$ is the perturbation Hamiltonian in interaction picture. The solution of (26.3) can be written as[1]

[1] In fact, integration of (26.3) yields

$$\mathcal{U}_{\mathrm{I}}(t,t') = 1 + \frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathcal{H}'_{\mathrm{I}}(t_1)\mathcal{U}_{\mathrm{I}}(t_1,t')\,\mathrm{d}t_1,$$

since $\mathcal{U}_{\mathrm{I}}(t',t') = 1$. Substituting this expression into itself iteratively, we obtain

$$\mathcal{U}_{\mathrm{I}}(t,t') = 1 + \frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_1\mathcal{H}'_{\mathrm{I}}(t_1) + \frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_1\frac{1}{\mathrm{i}\hbar}\int_{t'}^{t_1}\mathrm{d}t_2\mathcal{H}'_{\mathrm{I}}(t_1)\mathcal{H}'_{\mathrm{I}}(t_2) + \ldots.$$

If the two operators $\mathcal{H}'_{\mathrm{I}}(t_1)$ and $\mathcal{H}'_{\mathrm{I}}(t_2)$ were commuting, the second integral could be extended to t with a factor $1/2$. Since, however, the perturbation Hamiltonians at different times do not commute, the integral can be extended with the prescription of keeping the correct time order. The same considerations hold for the successive terms of the above expansion, which, therefore, can symbolically be written as

$$\mathcal{U}_{\mathrm{I}}(t,t') = 1 + \frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_1\mathcal{H}'_{\mathrm{I}}(t_1) + \frac{1}{2}\frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_1\frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_2\mathcal{T}[\mathcal{H}'_{\mathrm{I}}(t_1)\mathcal{H}'_{\mathrm{I}}(t_2)]$$
$$+ \frac{1}{3!}\mathcal{T}\frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathcal{H}'_{\mathrm{I}}(t_1)\mathrm{d}t_1\frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_2\mathcal{H}'_{\mathrm{I}}(t_2)\frac{1}{\mathrm{i}\hbar}\int_{t'}^{t}\mathrm{d}t_3\mathcal{H}'_{\mathrm{I}}(t_3) + \ldots.$$

This expansion is equivalent to the exponential in (26.4).

$$\boxed{\mathcal{U}_{\mathrm{I}}(t,t') = \mathcal{T}\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_{t'}^{t}\mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau}} \tag{26.4}$$

where the time-ordering operator $\mathcal{T}$ indicates that in the series expansion that defines the exponential, the products of the Hamiltonians $\mathcal{H}'(\tau)$ at different times must be kept in the correct time order.

26.2 Contour Integration

Let us consider, as first example, the correlation function $G^{>}$. Its definition in (24.21) is given in Heisenberg picture. In Schrödinger picture, it becomes

$$G^{>}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\langle\Phi(t)|\Psi(\boldsymbol{r})\mathcal{U}(t,t')\Psi^{\dagger}(\boldsymbol{r}')|\Phi(t')\rangle,$$

with the following simple physical interpretation: given the many-body state, we generate an additional particle in $\boldsymbol{r}'$ at time t'; evolve the new state to time t, and evaluate the probability amplitude that an extra particle is found in $\boldsymbol{r}'$. We now transform the same quantity in the interaction picture, using (26.1) and (26.2), and obtain

$$G^{>}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\langle\Phi_{\mathrm{I}}(t)|\Psi_{\mathrm{I}}(\boldsymbol{r},t)\,\mathcal{U}_{\mathrm{I}}(t,t')\Psi_{\mathrm{I}}^{\dagger}(\boldsymbol{r}',t')|\Phi_{\mathrm{I}}(t')\rangle.$$

Now express the state vectors in terms of the evolutions of the corresponding initial quantities:

$$G^{>}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\langle\Phi_{\mathrm{I}}(t_{\circ})|\,\mathcal{U}_{\mathrm{I}}(t_{\circ},t)\Psi_{\mathrm{I}}(\boldsymbol{r},t)\,\mathcal{U}_{\mathrm{I}}(t,t')\Psi_{\mathrm{I}}^{\dagger}(\boldsymbol{r}',t')\,\mathcal{U}_{\mathrm{I}}(t',t_{\circ})|\Phi_{\mathrm{I}}(t_{\circ})\rangle. \tag{26.5}$$

According to the definition of the GFs, the mean value includes an ensemble average, and, consistently with our physical assumptions, the ensemble to be used in the average is at the unperturbed thermal equilibrium. With the evolution operator in the interaction picture given in (26.4), (26.5) yields

$$\begin{aligned} G^{>}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\langle&\mathcal{T}\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_{t}^{t_{\circ}}\mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau} \\ &\times\Psi_{\mathrm{I}}(\boldsymbol{r},t)\mathcal{T}\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_{t'}^{t}\mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau}\Psi_{\mathrm{I}}^{\dagger}(\boldsymbol{r}',t')\mathcal{T}\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_{t_{\circ}}^{t'}\mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau}\rangle. \end{aligned} \tag{26.6}$$

This expression can be seen as a single integral on the contour, introduced by Keldish and shown in Fig. 26.1, written as

$$G^{>}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\left\langle\mathcal{T}_{\mathrm{c}}\left\{\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_{\mathrm{c}}\mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau}\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi_{\mathrm{I}}^{\dagger}(\boldsymbol{r}',t')\right\}\right\rangle, \tag{26.7}$$

where now the *contour time-ordering* operator $\mathcal{T}_{\mathrm{c}}$ indicates that in the product of operators the time order of the contour must be kept.

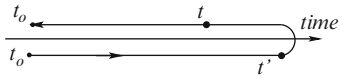


Fig. 26.1. Integration contour

t \ t'	forward	backward
forward	G^{t}	$G^{<}$
backward	$G^{>}$	$G^{\bar{t}}$

Fig. 26.2. Correspondence between the contour-ordered GF and the various GFs defined in Chap. 24

At this point, it is convenient to introduce a new GF, called *contour-ordered GF* and simply indicated as G, similar to the time ordered GF, where the time ordering is now that of the contour.

$$\boxed{G(\boldsymbol{r},t,\boldsymbol{r}',t') = \tfrac{1}{\mathrm{i}\hbar}\langle \mathcal{T}_{\mathrm{c}}[\Psi(\boldsymbol{r},t)\Psi^{\dagger}(\boldsymbol{r}',t')]\rangle} \tag{26.8}$$

Note that if both t and t' are in the forward part of the contour, the contour ordering is the same as the time ordering, and G coincides with $G^{(t)}$, while if both times are in the backward part of the contour, ordering is opposite to the chronological ordering, and G coincides con la anti-time-ordered GF. If t is on the forward path and t' on the backward path t comes before in the contour, the order of the field operators must be reversed in (26.8) and G coincides with $G^{<}$. Finally, if t is on the backward path and t' on the forward path no inversion is necessary, and G coincides with $G^{>}$. This last case is that represented in the Fig. 26.1. The correspondence between the new contour-ordered GF and the old ones is reported in Fig. 26.2.

Note that if $t > t'$ (or $t' > t$), t (or t') can be equally put in the forward or backward path of the contour. This fact does not create problems because, according to (24.24), G^{t} coincides with $G^{>}$ if $t > t'$ and with $G^{<}$ if $t < t'$. Similar considerations hold for $G^{\bar{t}}$.

The expression (26.7), without prescription on the location of the two times in the contour, becomes the general expression for G:

$$\boxed{G(\boldsymbol{r},t,\boldsymbol{r}',t') = \tfrac{1}{\mathrm{i}\hbar}\left\langle \mathcal{T}_{\mathrm{c}}\left\{\mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_c \mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau}\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')\right\}\right\rangle} \tag{26.9}$$

Fig. 26.3. Representation of the unperturbed (zero order) $G_\circ$

26.3 Perturbation Expansion and Feynman Diagrams, Potential Interaction

We are now ready to expand the contour evolution operator in (26.9) in powers of the perturbation Hamiltonian $\mathcal{H}'_{\mathrm{I}}$. To zero order, the evolution operator in the interaction picture is equal to one, and the GF reduces to

$$G^{(0)}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\langle \mathcal{T}_{\mathrm{c}}[\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')]\rangle = G_\circ(\boldsymbol{r},t,\boldsymbol{r}',t'). \tag{26.10}$$

This is simply the GF of the unperturbed system, called *free propagator*. It is represented graphically by the simplest Feynman diagram: a continuous line connecting the point $(\boldsymbol{r}',t')$ to $(\boldsymbol{r},t)$, as shown in Fig. 26.3.

The first-order term of (26.9) is written as

$$G^{(1)}(\boldsymbol{r},t,\boldsymbol{r}',t') = \frac{1}{\mathrm{i}\hbar}\left\langle \mathcal{T}_{\mathrm{c}}\left\{-\frac{\mathrm{i}}{\hbar}\int_c \mathcal{H}'_{\mathrm{I}}(\tau)\mathrm{d}\tau\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')\right\}\right\rangle. \tag{26.11}$$

To proceed, we must specify the type of interaction Hamiltonian we are dealing with. Let us begin with the simplest interaction, due to a time-independent potential field $V(\boldsymbol{r})$.

In second quantization formalism, the Hamiltonian due to the potential energy $V(\boldsymbol{r})$ in the interaction picture is written as (see 23.24)

$$\mathcal{H}'_{\mathrm{I}}(t) = \int \Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r},t)V(\boldsymbol{r})\Psi_{\mathrm{I}}(\boldsymbol{r},t)\,\mathrm{d}\boldsymbol{r}. \tag{26.12}$$

Equation (26.11) then becomes

$$\frac{1}{\mathrm{i}\hbar}\frac{1}{\mathrm{i}\hbar}\left\langle \mathcal{T}_{\mathrm{c}}\left\{\int_c \mathrm{d}t_1\int \mathrm{d}\boldsymbol{r}_1\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}_1,t_1)V(\boldsymbol{r}_1)\Psi_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')\right\}\right\rangle,$$

which implies the evaluation of products of field operators averaged over the initial equilibrium ensemble:

$$\left(\frac{1}{\mathrm{i}\hbar}\right)^2 V(\boldsymbol{r}_1)\langle \mathcal{T}_{\mathrm{c}}\{\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')\}\rangle,$$

where integration over $\boldsymbol{r}_1$ and t_1 is understood. We may now apply Wick–Matsubara theorem (25.14), which holds independently of the time ordering assumed as long as it is well defined, obtaining

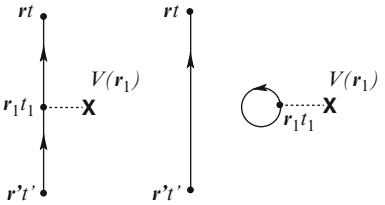


Fig. 26.4. First-order Feynman diagrams for potential interaction. *Continuous lines* indicate the unperturbed $G_\circ$; *dashed lines* with X indicate the potential interaction; the *circle* indicates the unperturbed $G_\circ$ with coincident initial and final points

$$\left(\frac{1}{\mathrm{i}\hbar}\right)^2 V(\boldsymbol{r}_1)\{\langle\mathcal{T}_\mathrm{c}[\Psi_\mathrm{I}^\dagger(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}(\boldsymbol{r}_1,t_1)]\rangle\langle\mathcal{T}_\mathrm{c}[\Psi_\mathrm{I}(\boldsymbol{r},t)\Psi_\mathrm{I}^\dagger(\boldsymbol{r}',t')]\rangle$$
$$\pm\langle\mathcal{T}_\mathrm{c}[\Psi_\mathrm{I}^\dagger(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}(\boldsymbol{r},t)]\rangle\langle\mathcal{T}_\mathrm{c}[\Psi_\mathrm{I}(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}^\dagger(\boldsymbol{r}',t')]\rangle$$
$$+\langle\mathcal{T}_\mathrm{c}[\Psi_\mathrm{I}^\dagger(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}^\dagger(\boldsymbol{r}',t')]\rangle\langle\mathcal{T}_\mathrm{c}[\Psi_\mathrm{I}(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}(\boldsymbol{r},t)]\rangle\}.$$

The last term is zero since it contains two creation, or two annihilation operators. The other terms contain unperturbed GFs:

$$V(\boldsymbol{r}_1)\{\pm G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r},t,\boldsymbol{r}',t')+G_\circ(\boldsymbol{r},t,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}',t'). \tag{26.13}$$

The Feynman diagrams corresponding to these two terms are shown, in opposite order, in Fig. 26.4. The point $(\boldsymbol{r}_1,t_1)$ where the interaction acts is called a *vertex*.

The second-order term of (26.9) is with the usual integration over repeated arguments $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, t_1 e t_2,

$$G^{(2)}=\frac{1}{\mathrm{i}\hbar}\left(\frac{1}{\mathrm{i}\hbar}\right)^2 V(\boldsymbol{r}_1)V(\boldsymbol{r}_2)\times\frac{1}{2}\langle\mathcal{T}_\mathrm{c}\{\Psi_\mathrm{I}^\dagger(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}(\boldsymbol{r}_1,t_1)\Psi_\mathrm{I}^\dagger(\boldsymbol{r}_2,t_2)$$
$$\times\Psi_\mathrm{I}(\boldsymbol{r}_2,t_2)\Psi_\mathrm{I}(\boldsymbol{r},t)\Psi_\mathrm{I}^\dagger(\boldsymbol{r}',t')\}\rangle. \tag{26.14}$$

With the same procedure above, this becomes

$$\begin{aligned}
&=\frac{1}{2}V(\boldsymbol{r}_1)V(\boldsymbol{r}_2)\times\\
&\times\{G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r}_2,t_2,\boldsymbol{r}_2,t_2)G_\circ(\boldsymbol{r},t,\boldsymbol{r}',t') && (a)\\
&\pm G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r},t,\boldsymbol{r}_2,t_2)G_\circ(\boldsymbol{r}_2,t_2,\boldsymbol{r}',t') && (b)\\
&\pm G_\circ(\boldsymbol{r}_2,t_2,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}_2,t_2)G_\circ(\boldsymbol{r},t,\boldsymbol{r}',t') && (c)\\
&+G_\circ(\boldsymbol{r}_2,t_2,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}',t')G_\circ(\boldsymbol{r},t,\boldsymbol{r}_2,t_2) && (d)\\
&+G_\circ(\boldsymbol{r},t,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}_2,t_2)G_\circ(\boldsymbol{r}_2,t_2,\boldsymbol{r}',t') && (e)\\
&\pm G_\circ(\boldsymbol{r},t,\boldsymbol{r}_1,t_1)G_\circ(\boldsymbol{r}_1,t_1,\boldsymbol{r}',t')G_\circ(\boldsymbol{r}_2,t_2,\boldsymbol{r}_2,t_2)\} && (f).
\end{aligned} \tag{26.15}$$

The corresponding Feynman diagrams are shown in Fig. 26.5.

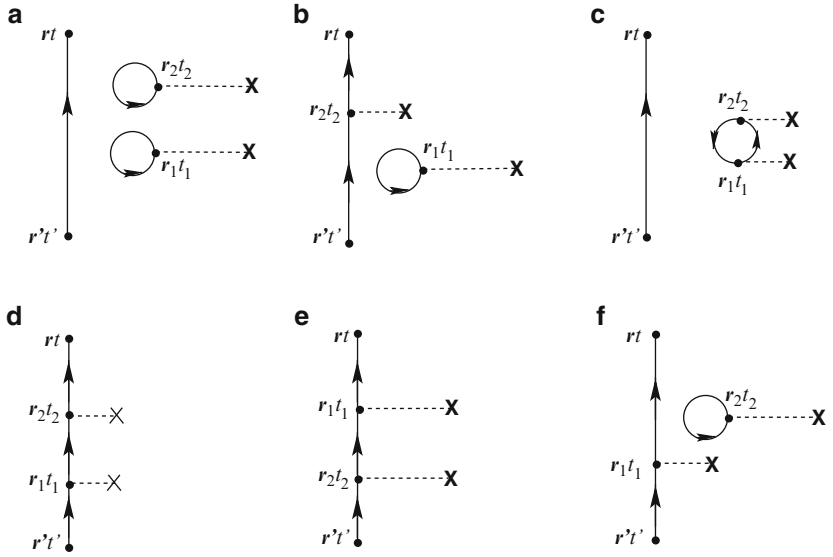


Fig. 26.5. Feynman diagrams for second-order potential interaction, as in (26.15)

26.3.1 Cancellation of Disconnected Diagrams

Note that some of the diagrams are entirely connected ((d) and (e) in Fig. 26.5), while others contain disconnected parts ((a), (b), (c), and (f) in the same figure)

If we consider all possible diagrams at all orders, for each connected diagram from $(\boldsymbol{r}', t')$ to $(\boldsymbol{r}, t)$, we obtain an infinite set of disconnected parts with all possible diagrams. Since each diagram represents a product, the common part connecting $(\boldsymbol{r}, t)$ to $(\boldsymbol{r}', t')$ can be collected as common factor, as indicated in Fig. 26.6. The separate part of the disconnected diagrams produces the GFs that do not connect $(\boldsymbol{r}, t)$ to $(\boldsymbol{r}', t')$ and therefore can be obtained as products of field operators of the type in (26.14), without $\Psi(\boldsymbol{r}, t)$ and $\Psi^\dagger(\boldsymbol{r}', t')$. This means that the series of diagrams inside the parentheses in Fig. 26.6 corresponds to the series expansion of the operator

$$\left\langle \mathcal{T}_c \left\{ \mathrm{e}^{-\frac{\mathrm{i}}{\hbar}\int_c H_I'(\tau)\mathrm{d}\tau} \right\} \right\rangle ,$$

as can be easily seen by eliminating the above-mentioned operators in (26.9). This, however, is the evolution operator along the contour that connects the initial condition with itself and therefore is equal to the identity.[2] All parentheses in Fig. 26.6 are unity and can be eliminated as factors. The final

[2] In some texts, the S matrix appears in the denominator of (26.9), since the contour connects times at $\mp\infty$. In such cases, the series of the disconnected parts of the diagrams yields the S matrix and cancels with the denominator.

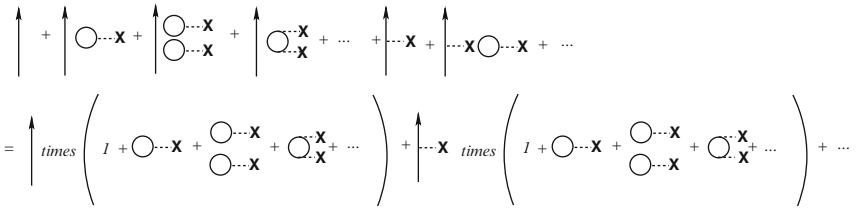


Fig. 26.6. Cancellation of the disconnected diagrams

result is that the only diagrams to be considered for the series expansion of the GF are those entirely connected.

26.3.2 Term Multiplicity

The connected diagrams of the second order (d) and (e) in Fig. 26.5 are equal, since they differ only for the interchange of dummy integration variables. Only one of them can therefore be retained multiplying its value by 2. This 2 cancels the 2 in the denominator in (26.15). Such a property is quite general: a connected diagram with m vertices appears $m!$ times since $m!$ is the number of combinations of the m dummy integration variables. This $m!$ cancels the same factor in the denominator of the terms of the series expansion of the evolution operator in (26.9).

The conclusion is that only the connected diagrams have to be considered and only once. This is the first of a number of algorithmic rules for the identification of all terms necessary for the evaluation of the GF by means of Feynman diagrams. The other rules refer to the signs and the numerical factors to be associated with each vertex of the diagram. For example, when a $G_\circ$ appears with equal positions and times, it is represented by a closed loop in the Feynman diagram, and a minus sign must be associated with each fermion closed loop. For the sake of brevity, the complete set of rules shall not be given here. By now, the basic idea should be clear to the reader. For practical purposes, we refer to more specialized books, such as, for example, [3, 145, 301].

Impurity Scattering

In the foregoing, we have considered the case of a potential $V(\boldsymbol{r})$ applied to the electrons in the crystal. When, however, this potential is due to a large number of impurities randomly distributed, but uniformly in a macroscopic scale, we are not interested, in general,[3] in a particular distribution of impurity

[3] This is certainly true in studying bulk properties of a material; it used to be true also in electronic devices until a few decades ago; it is no more true in present-day devices of nanometric dimensions, where the number of impurities may be very small and their specific positions may play a significant role. See, for example, [17, 327, 488].

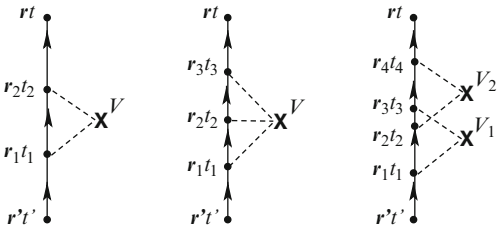


Fig. 26.7. Feynman diagrams for higher order impurity interactions

positions, but rather in the average effect over many possible configurations. In this average, the first-order term contains $\langle V(\boldsymbol{r})\rangle$ which would only lead to an irrelevant shift in energy and may be assumed to be zero. Higher-order terms yield nonzero contributions when the vertices refer to the same impurity. The diagrams to be considered must account for this fact and are represented as in Fig. 26.7. The interested reader may find a detailed discussion, for example, in [371].

26.4 Particle–Particle Interaction

If the perturbation Hamiltonian contains a two-particle term $V(\boldsymbol{r},\boldsymbol{r}')$, in place of (26.12), we now have

$$\mathcal{H}'_{\mathrm{I}}(t)=\frac{1}{2}\int\int \Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t)V(\boldsymbol{r},\boldsymbol{r}')\Psi_{\mathrm{I}}(\boldsymbol{r}',t)\Psi_{\mathrm{I}}(\boldsymbol{r},t)\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{r}'.$$

If we now expand the GF in (26.9) in series with this new Hamiltonian, the zero-order term is of course the same found for the previous case, shown in Fig. 26.3. The first-order terms are now given by

$$G^{(1)}(\boldsymbol{r},t,\boldsymbol{r}',t')=\frac{1}{\mathrm{i}\hbar}(1/2\mathrm{i}\hbar)\left\langle \mathcal{T}_{\mathrm{c}}\left\{\int_c \mathrm{d}t_1\int \mathrm{d}\boldsymbol{r}_1\int \mathrm{d}\boldsymbol{r}'_1\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}'_1,t_1)\right.\right.$$
$$\left.\left.\times V(\boldsymbol{r}_1,\boldsymbol{r}'_1)\Psi_{\mathrm{I}}(\boldsymbol{r}'_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')\right\}\right\rangle.$$

Then, we must evaluate the integrals of the products

$$\frac{1}{\mathrm{i}\hbar}\frac{1}{2\mathrm{i}\hbar}V(\boldsymbol{r}_1,\boldsymbol{r}'_1)\langle\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}'_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r}'_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r}_1,t_1)\Psi_{\mathrm{I}}(\boldsymbol{r},t)\Psi^{\dagger}_{\mathrm{I}}(\boldsymbol{r}',t')\rangle.$$

The Wick–Matsubara theorem transforms them into products, which are again recognized as free propagators joining interaction vertices,[4] represented as Feynman diagrams in Fig. 26.8.

[4] Each V in the expansion contains a factor $(1/\mathrm{i}\hbar)$. In the previous case of interaction with an external potential, these factors where exactly enough to identify

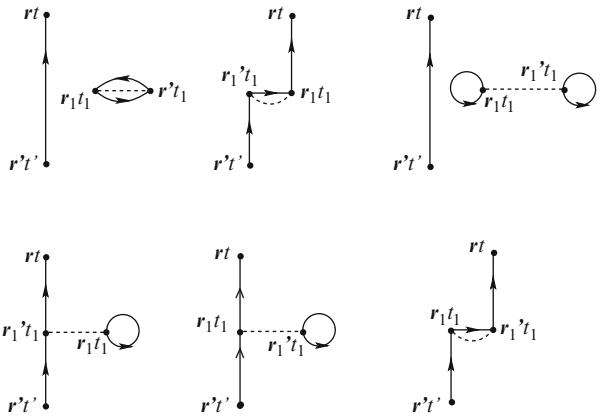


Fig. 26.8. First-order diagrams for particle–particle interaction. The *dashed lines* connecting two vertices represent the interaction potential

Once again the disconnected diagrams cancel, and the connected diagrams must be taken only once.

26.5 Electron–Phonon Interaction

We know that the interaction of electrons with phonons is of primary importance in the study of electron transport in semiconductors. In this section, we will briefly see how this interaction can be included in the electronic GF.

The lattice state can be described by the displacement of each atom with respect to its equilibrium position, given by (5.24). Here, it will be written as a displacement operator field, or *phonon field operator* as a function of the continuous variable $\boldsymbol{r}$:

$$\Phi(\boldsymbol{r}) = \sum_{\boldsymbol{q}} \mathbf{e}_{\boldsymbol{q}} \left(\frac{\hbar}{2MN\omega(\boldsymbol{q})} \right)^{\frac{1}{2}} \left\{ \mathsf{a}_{\boldsymbol{q}} \mathrm{e}^{\mathrm{i}\boldsymbol{q}\boldsymbol{r}} + \mathsf{a}_{\boldsymbol{q}}^{\dagger} \mathrm{e}^{-\mathrm{i}\boldsymbol{q}\boldsymbol{r}} \right\}, \tag{26.16}$$

where the polarization index has been included in the mode label $\boldsymbol{q}$. Note that according to this definition the phonon field is hermitian and contains both creation and annihilation operators. The free-phonon Hamiltonian and the commutation relations are (cf. (5.29) and (5.30))

$$\mathcal{H}_p = \sum_{\boldsymbol{q}} \hbar\omega(\boldsymbol{q}) \left(\mathsf{a}_{\boldsymbol{q}}^{\dagger} \mathsf{a}_{\boldsymbol{q}} + \frac{1}{2} \right), \quad \left[\mathsf{a}_{\boldsymbol{q}}, \mathsf{a}_{\boldsymbol{q}}^{\dagger} \right] = 1. \tag{26.17}$$

---

the products of two field operators with one free $G_\circ$. In this case, instead, each V, a two-point interaction, is accompanied by four field operators so that they are associated to two $G_\circ$. This means that at each diagram of order n a factor $(1/\mathrm{i}\hbar)^n$ must be associated with each nth-order term.

The total Hamiltonian is given by

$$\mathcal{H} = \mathcal{H}_{\mathrm{e}} + \mathcal{H}_{\mathrm{p}} + \mathcal{H}_{\mathrm{ep}}, \tag{26.18}$$

where $\mathcal{H}_{\mathrm{e}}$ is the free-electron Hamiltonian, $\mathcal{H}_{\mathrm{p}}$ the free-phonon Hamiltonian in (26.17), and $\mathcal{H}_{\mathrm{ep}}$ the electron–phonon Hamiltonian.

Following the usual procedure, the phonon field operator $\Phi(\boldsymbol{r})$ may now be considered in the interaction picture.[5] The development seen in the previous chapters for the electron fields can be repeated for the phonon fields. Since the unperturbed Hamiltonians of electrons and phonons act on different subspaces, no difficulty arises in the interaction picture. Phonon GFs $D(\boldsymbol{r}, t, \boldsymbol{r}', t')$ are defined in their various forms [295], up to the contour-order phonon GF, equivalent to (26.8).

In terms of field operators the interaction electron–phonon Hamiltonian is given by

$$\mathcal{H}_{\mathrm{ep}} = \sum_{\boldsymbol{q}} M(\boldsymbol{q}) \int \Psi^{\dagger}(\boldsymbol{r}) \Phi(\boldsymbol{r}) \Psi(\boldsymbol{r}) \mathrm{d}\boldsymbol{r}, \tag{26.19}$$

where the sum over the phonon modes $\boldsymbol{q}$ must be extended also over the phonon branches, and $M(\boldsymbol{q})$ is a numerical function of $\boldsymbol{q}$ that depends on the particular form of the considered interaction. Equation (26.19) is then inserted in the series expansion (26.11) and represented by Feynman diagrams. The first-order term contains single phonon operators, either creations or annihilations. Thus, its mean value over an equilibrium ensemble of noninteracting systems vanishes, and this is true for all odd-order terms.

If we consider the second-order term, to have nonvanishing mean values the product of two phonon operators, one creation and one destruction, must refer to the same mode, so that $\boldsymbol{q}$ and $\boldsymbol{q}'$ coming from the double application of the interaction Hamiltonian (26.19) must be equal and the double sum reduces to a single sum.

By application of the Wick–Matsubara theorem the phonon operators yield the phonon equilibrium GF $D_{\circ}$ in a series of terms analogous to those found for the second-order term when the electron–electron interaction is considered. If we represent such propagator with a wavy line, the resulting Feynman diagrams are similar to those in Fig. 26.8. Once again the disconnected diagrams cancel. The connected diagrams, to be counted only once, are shown in Fig. 26.9.

When several types of interaction are present, diagrams may combine them as indicated by the example in Fig. 26.10.

[5] In dealing with electron–phonon interaction, it may be more convenient to work in the $(\boldsymbol{k})$ representation. In this presentation, the $\boldsymbol{r}$ representation has been used for homogeneity with the previous cases of interactions.

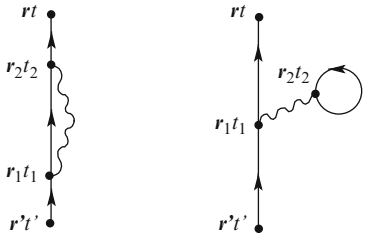


Fig. 26.9. Lowest-order electron–phonon diagrams. *Wavy lines* indicate phonon propagators

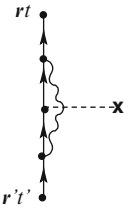


Fig. 26.10. Example of diagram with both potential and phonon interactions

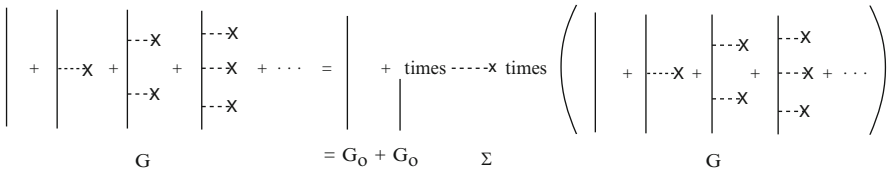


Fig. 26.11. Graphical representation of Dyson equation for the simple case of potential interaction

26.6 Self-Energy and Dyson Equation

Let us reconsider the Feynman diagrams for the interaction due, as the simplest example, to a potential scattering. Once eliminated the disconnected diagrams, they appear as in the left part of Fig. 26.11. The analytical expression can be obtained by collecting $G_\circ V$ on the left of the various contributions in (26.13), (26.15), and higher-order terms. With simplified notation, without the explicit indication of the variables, this is written as

$$G = G_\circ + G_\circ V G_\circ + G_\circ V G_\circ V G_\circ + \ldots = G_\circ + G_\circ V \left[G_\circ + G_\circ V G_\circ + \ldots \right].$$

It is now easy to realize that the content of the square brackets is the series expansion of the entire GF from $(\boldsymbol{r}_1, t_1)$ to $(\boldsymbol{r}', t')$, as it results also from Fig. 26.11. If, for this simple case, we define *self-energy* $\Sigma = V$, the above equation becomes the *Dyson equation*

$$G(\boldsymbol{r}, t, \boldsymbol{r}', t') = G_\circ(\boldsymbol{r}, t, \boldsymbol{r}', t') + G_\circ(\boldsymbol{r}, t, \boldsymbol{r}_1, t_1) \Sigma(\boldsymbol{r}_1) G(\boldsymbol{r}_1, t_1, \boldsymbol{r}', t'),$$

where it must be remembered that integration is understood over the repeated variables $\boldsymbol{r}_1$ e t_1.

In general, we call *self-energy* the sum of the factors represented in the Feynman diagrams by graphs which start and end on $G_\circ$ lines. More important is the concept of *irreducible self-energy*, often simply called self-energy and represented with $\Sigma(\boldsymbol{r}, t, \boldsymbol{r}', t')$, given by the sum of all possible factors represented by diagrams that connect two $G_\circ$ lines and cannot be separated in disconnected graphs by cutting a $G_\circ$ line inside it. If we represent the self-energy with a shadowed area, the complete GF can be represented as in Fig. 26.12.

We may now collect $G_\circ\Sigma$ on the left or $\Sigma G_\circ$ on the right, starting from the second term, and obtain the two equivalent forms of *Dyson equation*:

$$\boxed{G = G_\circ + G_\circ\Sigma G, \quad G = G_\circ + G\Sigma G_\circ} \tag{26.20}$$

These two equations are represented in Fig. 26.13.

Let us now multiply on the left (with the usual understood integrals) the first Dyson equation (26.20) by $(G_\circ^{-1} - \Sigma)$, obtaining

$$(G_\circ^{-1} - \Sigma)G = (G_\circ^{-1} - \Sigma)G_\circ + (G_\circ^{-1} - \Sigma)G_\circ\Sigma\, G = 1 - \Sigma(G_\circ - G + G_\circ\Sigma G) = 1.$$

This shows that the matrix $(G_\circ^{-1} - \Sigma)$ is the inverse of G, and we can write

$$G = \frac{1}{G_\circ^{-1} - \Sigma}\,, \tag{26.21}$$

showing that the self-energy is the correction to the inverse of G due to the interactions.

26.6.1 Matrix Formulation of G and of Dyson Equation

Following Craig [113], we shall develop now the connection between the contour ordered G and the various GFs defined previously. With reference to the correspondence shown in Fig. 26.2, let us define the matrix GF as

Fig. 26.12. Graphic representation of the full GF as sum of diagrams with free propagators and (irreducible) self-energies. The *double line* indicates G, simple line $G_\circ$, shaded area Σ

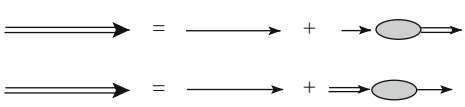


Fig. 26.13. Graphic representation of Dyson equation in the two forms in (26.20)

$$\tilde{G} = \begin{pmatrix} G^t & -G^< \\ G^> & -G^{\bar{t}} \end{pmatrix}. \tag{26.22}$$

The minus signs in the second column derive from the fact that these functions appear in time integrals, and in the backward part of the contour the integrations are performed in opposite directions.

The general definition of the self-energy Σ is based on the Feynman diagrams in the contour loop. We may then define several components of the self-energy depending on the positions of the connected times in the contour, according to correspondence indicated in Fig. 26.2, and then define [6]

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma^t & -\Sigma^< \\ \Sigma^> & -\Sigma^{\bar{t}} \end{pmatrix}. \tag{26.23}$$

As was done for G^t and $G^{\bar{t}}$, after Fig. 26.2, we note that if $t > t'$ (or $t' > t$), t (or t') can be equally put in the forward or backward path of the contour. This means that

$$\Sigma^t = \theta(t-t')\Sigma^> + \theta(t'-t)\Sigma^<, \quad \Sigma^{\bar{t}} = \theta(t'-t)\Sigma^> + \theta(t-t')\Sigma^<,$$

and therefore

$$\Sigma^t + \Sigma^{\bar{t}} = \Sigma^> + \Sigma^<. \tag{26.24}$$

Again, in analogy with the relations between the different GFs (24.26), let us define

$$\Sigma^{\mathrm{r}} = \Sigma^t - \Sigma^< = \Sigma^> - \Sigma^{\bar{t}}, \quad \Sigma^{\mathrm{a}} = \Sigma^t - \Sigma^> = \Sigma^< - \Sigma^{\bar{t}}, \tag{26.25}$$

where the second equal signs of the chains are obtained by application of (26.24).

In Fig. 26.14 we see that, according to the above definitions, we may write

$$G^t = G^t_\circ + G^t_\circ \Sigma^t G^t - G^t_\circ \Sigma^< G^> - G^<_\circ \Sigma^> G^t + G^<_\circ \Sigma^{\bar{t}} G^>. \tag{26.26}$$

Similarly we obtain, for the other components,

$$G^> = G^>_\circ + G^>_\circ \Sigma^t G^t - G^{\bar{t}}_\circ \Sigma^> G^t - G^>_\circ \Sigma^< G^> + G^{\bar{t}}_\circ \Sigma^{\bar{t}} G^>, \tag{26.27}$$

$$G^< = G^<_\circ + G^<_\circ \Sigma^{\bar{t}} G^{\bar{t}} - G^t_\circ \Sigma^< G^{\bar{t}} - G^<_\circ \Sigma^> G^< + G^t_\circ \Sigma^t G^<, \tag{26.28}$$

$$G^{\bar{t}} = G^{\bar{t}}_\circ + G^{\bar{t}}_\circ \Sigma^{\bar{t}} G^{\bar{t}} - G^>_\circ \Sigma^< G^{\bar{t}} + G^>_\circ \Sigma^t G^< - G^{\bar{t}}_\circ \Sigma^> G^<. \tag{26.29}$$

[6] Note that $\Sigma^<$ represents the contribution of the interactions corresponding, in the contour integral, to that of $G^<$ which is the correlation of the amplitude of particle presence in $(\boldsymbol{r}, t)$ and $(\boldsymbol{r}', t')$. Thus, $\Sigma^<$ corresponds to the scattering "in" in the classical limit. Similarly, $\Sigma^>$ contributes to the correlation of particle absence and corresponds to the scattering "out" in the classical limit.

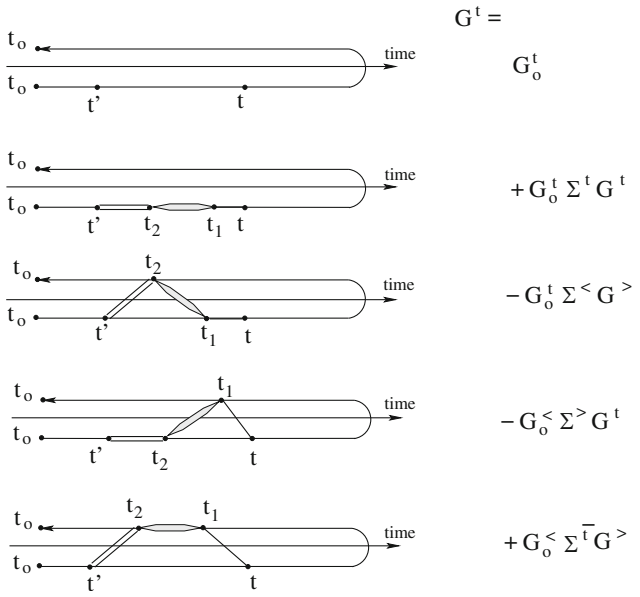


Fig. 26.14. Contributions to G^t in the contour. The *double line* indicates G, the simple line $G_\circ$, and the shaded area Σ. For G^t, t and t' must be in the forward branch of the contour; the other times, t_1 and t_2, are integrated over the entire contour

It is easy to verify that, with the definitions in (26.22) and in (26.23), the above relations can be written synthetically as

$$\boxed{\tilde{G} = \tilde{G}_\circ + \tilde{G}_\circ \tilde{\Sigma}\, \tilde{G}, \quad \tilde{G} = \tilde{G}_\circ + \tilde{G} \tilde{\Sigma}\, \tilde{G}_\circ} \tag{26.30}$$

The second equation is obtained from the first one, as usual, by inserting iteratively the equation into itself and collecting $\tilde{\Sigma}\tilde{G}_\circ$ on the right.

26.6.2 Dyson Equations for Separate GFs

From (26.26) to (26.29), it is now possible to obtain the Dyson equations for G^{r} and G^{a}. Start from the expression of G^{r} in terms of G^t and $G^<$ given in (24.26); then insert, for these quantities, the expressions (26.26) for G^t and (26.28) for $G^<$:

$$
\begin{aligned}
G^{\mathrm{r}} &= G^t - G^< = G^t_\circ - G^<_\circ + G^t_\circ \Sigma^t G^t - G^t_\circ \Sigma^< G^> - G^<_\circ \Sigma^> G^t + G^<_\circ \Sigma^{\bar{t}} G^> \\
&\quad - \left[G^<_\circ \Sigma^{\bar{t}} G^{\bar{t}} - G^t_\circ \Sigma^< G^{\bar{t}} - G^<_\circ \Sigma^> G^< + G^t_\circ \Sigma^t G^< \right] \\
&= G^{\mathrm{r}}_\circ + G^t_\circ \Sigma^t \left(G^t - G^< \right) - G^t_\circ \Sigma^< \left(G^> - G^{\bar{t}} \right) \\
&\quad - G^<_\circ \Sigma^> \left(G^t - G^< \right) + G^<_\circ \Sigma^{\bar{t}} \left(G^> - G^{\bar{t}} \right).
\end{aligned}
$$

Use again (24.26):

$$
\begin{aligned}
G^{\mathrm{r}} &= G^{\mathrm{r}}_\circ + G^t_\circ \Sigma^t G^{\mathrm{r}} - G^t_\circ \Sigma^< G^{\mathrm{r}} - G^<_\circ \Sigma^> G^{\mathrm{r}} + G^<_\circ \Sigma^{\bar{t}} G^{\mathrm{r}} \\
&= G^{\mathrm{r}}_\circ + \left[G^t_\circ \left(\Sigma^t - \Sigma^< \right) - G^<_\circ \left(\Sigma^> - \Sigma^{\bar{t}} \right) \right] G^{\mathrm{r}}.
\end{aligned}
$$

Now use (26.25) and obtain

$$
G^{\mathrm{r}} = G^{\mathrm{r}}_\circ + \left[G^t_\circ \Sigma^{\mathrm{r}} - G^<_\circ \Sigma^{\mathrm{r}} \right] G^{\mathrm{r}},
$$

or, finally, the Dyson equation for the retarded GF:

$$
\boxed{G^{\mathrm{r}} = G^{\mathrm{r}}_\circ + G^{\mathrm{r}}_\circ \Sigma^{\mathrm{r}} G^{\mathrm{r}} = G^{\mathrm{r}}_\circ + G^{\mathrm{r}} \Sigma^{\mathrm{r}} G^{\mathrm{r}}_\circ} \tag{26.31}
$$

The second form is again obtained by the first one inserting it into itself iteratively and then collecting $\Sigma^{\mathrm{r}} G^{\mathrm{r}}_\circ$ on the right.

The Dyson equation for G^{a} is obtained with the same procedure used for G^{r}, and the result is

$$
\boxed{G^{\mathrm{a}} = G^{\mathrm{a}}_\circ + G^{\mathrm{a}}_\circ \Sigma^{\mathrm{a}} G^{\mathrm{a}} = G^{\mathrm{a}}_\circ + G^{\mathrm{a}} \Sigma^{\mathrm{a}} G^{\mathrm{a}}_\circ} \tag{26.32}
$$

With the same procedure used to obtain (26.21) from (26.20) we obtain, from the above

$$
G^{\mathrm{r}} = \frac{1}{(G^{\mathrm{r}}_\circ)^{-1} - \Sigma^{\mathrm{r}}}, \quad G^{\mathrm{a}} = \frac{1}{(G^{\mathrm{a}}_\circ)^{-1} - \Sigma^{\mathrm{a}}}. \tag{26.33}
$$

Now, in the initial developments of the GFs it was seen (cf. (24.47)) that $G^{\mathrm{r,a}}_\circ$ are the inverse operator of $(\mathrm{i}\hbar\partial/\partial t - \mathcal{H})$. Thus we may write (26.33) as

$$
G^{\mathrm{r,a}} = \frac{1}{\mathrm{i}\hbar\partial/\partial t - \mathcal{H}_\circ - \Sigma^{\mathrm{r,a}}} \tag{26.34}
$$

and interpret the self-energy as a correction to the Hamiltonian due to the interactions. If this correction is inserted in the evolution operator, we understand that *the real part of Σ yields a shift of the energy eigenvalues and its imaginary part yields a lifetime of the energy eigenstates.*

Also $G^<$, $G^>$, G^t, and $G^{\bar{t}}$ obey Dyson-like equations. Since, however, they describe the evolution of the state of the system, their equations must involve

also G^{r} and G^{a} and are not as simple as the ones just found for G^{r} and G^{a}. Let us start with the search for the equation for $G^{<}$. The starting point will be the (26.28). According to what was just said, we must express all quantities that appear in (26.28) by means less, retarded, and advanced quantities, and this can be done by substituting $G^{t} = G^{\mathrm{r}} + G^{<}$, $\Sigma^{t} = \Sigma^{\mathrm{r}} + \Sigma^{<}$, $G^{\bar{t}} = G^{<} - G^{\mathrm{a}}$, $\Sigma^{\bar{t}} = \Sigma^{<} - \Sigma^{\mathrm{a}}$, and $\Sigma^{>} = \Sigma^{\mathrm{r}} + \Sigma^{\bar{t}} = \Sigma^{\mathrm{r}} + \Sigma^{<} - \Sigma^{\mathrm{a}}$, obtained by (24.26) and (26.25). After straightforward calculations, we obtain

$$(1 - G^{\mathrm{r}}_{\circ}\Sigma^{\mathrm{r}})G^{<} = G^{<}_{\circ}(1 + \Sigma^{\mathrm{a}}G^{\mathrm{a}}) + G^{\mathrm{r}}_{\circ}\Sigma^{<}G^{\mathrm{a}}.$$

Multiplying this by $(1 + G^{\mathrm{r}}\Sigma^{\mathrm{r}})$, after other simple algebra, we reach

$$(1+G^{\mathrm{r}}\Sigma^{\mathrm{r}})(1-G^{\mathrm{r}}_{\circ}\Sigma^{\mathrm{r}})G^{<} = (1+G^{\mathrm{r}}\Sigma^{\mathrm{r}})G^{<}_{\circ}(1+\Sigma^{\mathrm{a}}G^{\mathrm{a}}) + [G^{\mathrm{r}}_{\circ} + G^{\mathrm{r}}\Sigma^{\mathrm{r}}G^{\mathrm{r}}_{\circ}]\Sigma^{<}G^{\mathrm{a}}.$$

Finally, taking into account the second of (26.31), this becomes

$$G^{<} = (1 + G^{\mathrm{r}}\Sigma^{\mathrm{r}})G^{<}_{\circ}(1 + \Sigma^{\mathrm{a}}G^{\mathrm{a}}) + G^{\mathrm{r}}\Sigma^{<}G^{\mathrm{a}}. \tag{26.35}$$

Very similar calculations can be repeated for $G^{>}$, G^{t}, and $G^{\bar{t}}$. The resulting equations are

$$G^{>} = (1 + G^{\mathrm{r}}\Sigma^{\mathrm{r}})G^{>}_{\circ}(1 + \Sigma^{\mathrm{a}}G^{\mathrm{a}}) + G^{\mathrm{r}}\Sigma^{>}G^{\mathrm{a}}, \tag{26.36}$$

$$G^{t} = (1 + G^{\mathrm{r}}\Sigma^{\mathrm{r}})G^{t}_{\circ}(1 + \Sigma^{\mathrm{a}}G^{\mathrm{a}}) + G^{\mathrm{r}}\Sigma^{\bar{t}}G^{\mathrm{a}}, \tag{26.37}$$

$$G^{\bar{t}} = (1 + G^{\mathrm{r}}\Sigma^{\mathrm{r}})G^{\bar{t}}_{\circ}(1 + \Sigma^{\mathrm{a}}G^{\mathrm{a}}) + G^{\mathrm{r}}\Sigma^{t}G^{\mathrm{a}}. \tag{26.38}$$

26.6.3 $(\boldsymbol{k}, \omega)$ Representation

In case of homogeneous and stationary states, it is convenient to move to the $(\boldsymbol{k}, \omega)$ representation, since the convolution integrals become algebraic products,[7] and Dyson equations receive immediate solution. Equations (26.31)

[7] Let us consider, for simplicity, the simple case of two functions. The extension to more than two function is obvious. Assume that

$$f(\boldsymbol{r} - \boldsymbol{r}', t - t') = \int\int f_1(\boldsymbol{r} - \boldsymbol{r}'', t - t'')f_2(\boldsymbol{r}'' - \boldsymbol{r}', t'' - t')\mathrm{d}\boldsymbol{r}''\mathrm{d}t''$$

and write its Fourier transform,

$$\begin{aligned} F(\boldsymbol{k}, \omega) &= \frac{1}{(2\pi)^{3/2}} \int \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}(\boldsymbol{r}-\boldsymbol{r}'))}\mathrm{d}(\boldsymbol{r} - \boldsymbol{r}') \int \mathrm{e}^{\mathrm{i}\omega(t-t')}\mathrm{d}(t - t')f(\boldsymbol{r} - \boldsymbol{r}', t - t') \\ &= \frac{1}{(2\pi)^{3/2}} \int \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}(\boldsymbol{r}-\boldsymbol{r}''))}\mathrm{e}^{-\mathrm{i}(\boldsymbol{k}(\boldsymbol{r}''-\boldsymbol{r}'))}\mathrm{d}(\boldsymbol{r} - \boldsymbol{r}') \\ &\quad \int \mathrm{e}^{\mathrm{i}(\omega(t-t''))}\mathrm{e}^{\mathrm{i}\omega(t\text{“}-t')}\mathrm{d}(t - t') \\ &\quad \int\int f_1(\boldsymbol{r} - \boldsymbol{r}'', t - t'')f_2(\boldsymbol{r}'' - \boldsymbol{r}', t'' - t')\mathrm{d}\boldsymbol{r}''\mathrm{d}t''. \end{aligned}$$

and (26.32) in $(\boldsymbol{k},\omega)$ representation become

$$G^{\mathrm{r,a}}(\boldsymbol{k},\omega) = G^{\mathrm{r,a}}_{\circ}(\boldsymbol{k},\omega) + (2\pi)^3 G^{\mathrm{r,a}}_{\circ}(\boldsymbol{k},\omega)\Sigma^{\mathrm{r,a}}(\boldsymbol{k},\omega)G^{\mathrm{r,a}}(\boldsymbol{k},\omega),$$

immediately solved as

$$G^{\mathrm{r,a}}(\boldsymbol{k},\omega) = \frac{1}{\left(G^{\mathrm{r,a}}_{\circ}(\boldsymbol{k},\omega)\right)^{-1} - (2\pi)^3\,\Sigma^{\mathrm{r,a}}(\boldsymbol{k},\omega)}.$$

In place of the unperturbed GFs, we may substitute the expression (24.44) for noninteracting particles and obtain

$$G^{\mathrm{r,a}}(\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{3/2}\hbar} \quad \frac{1}{\omega - \omega_{\boldsymbol{k}} - (2\pi)^{3/2}\Sigma^{\mathrm{r,a}}(\boldsymbol{k},\omega)/\hbar}. \tag{26.39}$$

The converging factor γ is no more necessary since now the self-energy is present. The different numerical coefficients in the literature are related to the coefficients chosen in the Fourier transform.

Now, from the symmetry relation in (24.29) it is immediate to show that, for a stationary, homogeneous system, the self-energies verify the identical relation:

$$\Sigma^{\mathrm{a}}(\boldsymbol{k},\omega) = \left(\Sigma^{\mathrm{r}}(\boldsymbol{k},\omega)\right)^{*}.$$

If we define ω_s and Γ is such a way that

$$\Sigma^{\mathrm{r}}(\boldsymbol{k},\omega) = \frac{\hbar}{(2\pi)^{3/2}}(\omega_s - \mathrm{i}\Gamma), \quad \Sigma^{\mathrm{a}}(\boldsymbol{k},\omega) = \frac{\hbar}{(2\pi)^{3/2}}(\omega_s + \mathrm{i}\Gamma), \tag{26.40}$$

the GFs in (26.39) become

$$G^{\mathrm{r}}(\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{3/2}\hbar}\ \frac{1}{\sigma + \mathrm{i}\Gamma}, \quad G^{\mathrm{a}}(\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{3/2}\hbar}\ \frac{1}{\sigma - \mathrm{i}\Gamma}, \tag{26.41}$$

where

$$\sigma = \omega - \omega_{\boldsymbol{k}} - \omega_s. \tag{26.42}$$

From the comparison with the equivalent expressions (24.44) for free particles, we observe that the interaction shifts the real part of the poles of the GF of the quantity ω_s and introduces a spectral width Γ, i.e., a lifetime. From the analyicity considerations made to obtain (24.42), we expect Γ to be positive.

Note that from (26.25) and (26.40)

$$\Sigma^{\mathrm{r}} - \Sigma^{\mathrm{a}} = \Sigma^{>} - \Sigma^{<} = -2\mathrm{i}\frac{\hbar}{(2\pi)^{3/2}}\Gamma. \tag{26.43}$$

In place of t' and t'', let us use the new variables τ e τ', defined by $t'' = t - \tau$ and $t' = t'' - \tau'$, and similarly for the space variables, obtaining

$$F(\boldsymbol{k},\omega) = (2\pi)^{3/2} F_1(\boldsymbol{k},\omega)F_2(\boldsymbol{k},\omega).$$

From the definition in (24.52) and (26.41),

$$A = \sqrt{2\pi}\hbar \mathrm{i}(G^{\mathrm{r}} - G^{\mathrm{a}}) = \frac{1}{\pi} \quad \frac{\Gamma}{(\omega - \omega_{\boldsymbol{k}} - \omega_s)^2 + \Gamma^2}. \tag{26.44}$$

For vanishing ω_s and Γ the unperturbed spectral density in (24.53) is recovered.

From (26.41), it is immediate to find the important relations

$$A = 2h^2 G^{\mathrm{r}} \Gamma G^{\mathrm{a}} = 2h^2 G^{\mathrm{a}} \Gamma G^{\mathrm{r}}.$$

For later use, we need an expression for $\Sigma^<(\boldsymbol{k}, \omega)$ at equilibrium. From (26.35) we obtain, in $(\boldsymbol{k}, \omega)$ representation,

$$G^<(\boldsymbol{k}, \omega) = (1 + (2\pi)^3 G^{\mathrm{r}} \Sigma^{\mathrm{r}}) G^<_\circ (1 + (2\pi)^3 \Sigma^{\mathrm{a}} G^{\mathrm{a}}) + (2\pi)^3 G^{\mathrm{r}} \Sigma^< G^{\mathrm{a}}. \tag{26.45}$$

The first factor of the first term, using the results obtained above in (26.40), (26.41), and (26.42), becomes

$$1 + (2\pi)^3 G^{\mathrm{r}} \Sigma^{\mathrm{r}} = 1 + \frac{\omega_s - \mathrm{i}\Gamma}{\omega - \omega_{\boldsymbol{k}} - \omega_s + \mathrm{i}\Gamma} = \frac{\omega - \omega_{\boldsymbol{k}}}{\omega - \omega_{\boldsymbol{k}} - \omega_s + \mathrm{i}\Gamma}.$$

The third factor of the first term is the complex conjugate of the one just evaluated. Thus, the first product in (26.45) becomes, using (24.43) for $G^<_\circ$,

$$\pm \frac{1}{\sqrt{2\pi}} \frac{1}{\mathrm{i}\hbar} n_k \delta(\omega - \omega_k) \frac{(\omega - \omega_{\boldsymbol{k}})^2}{(\omega - \omega_{\boldsymbol{k}} - \omega_s)^2 + \Gamma^2}.$$

This expression is zero because of the delta and of the square in the numerator.[8] It remains, using (26.41),

$$G^<(\boldsymbol{k}, \omega) = (2\pi)^3 G^{\mathrm{r}} \Sigma^< G^{\mathrm{a}} = \frac{1}{\hbar^2} \frac{\Sigma^<}{(\omega - \omega_{\boldsymbol{k}} - \omega_s)^2 + \Gamma^2}.$$

Comparison with (24.56) yields, at equilibrium,

$$\Sigma^< = \pm \frac{\mathrm{i}\hbar}{\sqrt{2\pi}} [(\omega - \omega_{\boldsymbol{k}} - \omega_s)^2 + \Gamma^2] A(\boldsymbol{k}, \omega) n(\omega).$$

Finally, in place of A let us use its expression in (26.44), obtaining

$$\Sigma^<(\boldsymbol{k}, \omega) = \pm \frac{2\hbar}{(2\pi)^{3/2} \mathrm{i}} f_\circ(\omega) \Gamma. \tag{26.46}$$

Similarly, for $\Sigma^>$ using (26.36) and (24.43) for $G^>_\circ$,

$$\Sigma^>(\boldsymbol{k}, \omega) = \frac{2\hbar}{(2\pi)^{3/2} \mathrm{i}} [1 \pm f_\circ(\omega)] \Gamma. \tag{26.47}$$

As a consistency check, note that the difference between $\Sigma^>$ and $\Sigma^<$ just found reproduces (26.43).

[8] This vanishing of the first term of (26.45) is related to the physical meaning of the first terms of the (26.35)–(26.38): they represent the effect of the initial condition contained in the various $G_\circ$; in presence of interactions, this effect vanishes when steady state is reached.

26.7 Electron–Phonon Self-Energy

As it is clear from the previous sections, the self-energy is the quantity that, in the GF approach, carries the effect of the electronic interactions. As such it plays the role that in the Boltzmann transport equation is played by the scattering probabilities and it is therefore of great importance in the theory of electron transport. In particular, the self-energy due to electron–phonon interaction is essential for the application of nonequilibrium GFs to semiconductors. Self-energies, however, are hard to calculate in rigorous terms and often they are substituted by simplified models or first-order approximations.

The electron–phonon interaction Hamiltonian has been studied in Chap. 9 and reported in (26.19). To the lowest order, the diagrams that must be considered are those shown in Fig. 26.9. When we work in the momentum representation, however, the equivalent of the diagram on the right of Fig. 26.9 does not contribute because the initial and final electron momenta in each vertex are identical and only the matrix elements with $\boldsymbol{q} = 0$ would contribute. Since the interaction field of a phonon with the electrons is given by the space variation of the atom displacement, $\boldsymbol{q} = 0$ does not contribute [145].

As it regards the diagram on the left of Fig. 26.9, the corresponding self-energy contains the free-phonon GF, named D. Given the form of the interaction Hamiltonian, two terms contribute to the self-energy, one with the product $\langle \mathsf{a}_{\boldsymbol{q}} \mathsf{a}^{\dagger}_{\boldsymbol{q}} \rangle$ and one with the $\langle \mathsf{a}^{\dagger}_{\boldsymbol{q}} \mathsf{a}_{\boldsymbol{q}} \rangle$. Usually, they are considered as two terms of the phonon unperturbed GF, corresponding to the wavy line in Fig. 26.9. Two self-energies, namely $\Sigma^{<}$ and $\Sigma^{>}$, are in particular necessary in transport theory. To first order, and assuming the phonon population N_q at equilibrium, they are given by [296]

$$\Sigma^{<}(\boldsymbol{p}, \omega) = \sum_{q} M_q^2 [(N_q + 1) G^{<}(\boldsymbol{k} + \boldsymbol{q}, \omega + \omega_{\boldsymbol{q}}) + N_q G^{<}(\boldsymbol{k} - \boldsymbol{q}, \omega - \omega_{\boldsymbol{q}})]$$

and

$$\Sigma^{>}(\boldsymbol{k}, \omega) = \sum_{q} M_q^2 [N_q G^{>}(\boldsymbol{k} + \boldsymbol{q}, \omega + \omega_{\boldsymbol{q}}) + (N_q + 1) G^{>}(\boldsymbol{k} - \boldsymbol{q}, \omega - \omega_{\boldsymbol{q}})].$$

Since these self-energies are calculated by means of the GFs and the equations for the GFs contain the Σs, a self-consistent calculation is necessary.

Higher-order terms of the self-energy due to electron–phonon interaction are more difficult to evaluate, and most of the applications have used the lowest order, or Born, approximation.

27

Nonequilibrium Green Functions Applied to Transport: Quantum Boltzmann Equation

Nonequilibrium Green functions (NEGF) have been applied to electron transport in semiconductors along several lines, as described, for example, in [116, 184, 228, 295]. In this chapter and the following one, we shall treat two important applications. In this chapter, we shall derive the so-called *quantum Boltzmann equation* (QBE), derived by Kadanoff and Baym [228] and extended by Mahan and Hansch [179, 180, 298]. This equation is valid for weak applied potentials slowly varying in space and time and it is therefore especially useful for linear transport in homogeneous, stationary systems. In the following, final, chapter we shall report the method developed by Datta [116] for electron transport in open mesoscopic systems, especially useful for the analysis of modern quantum electronic devices.

We have seen in Chap. 17 that the Wigner function is particularly useful for the description of quantum transport. Since this function is a particular transform of $G^<$, we are interested in finding an equation that describes the dynamical behavior of this GF. However, from the development of Dyson equation seen in the previous chapter we know that the various GFs are strictly correlated to each other, so that it will not be possible to write a closed equation for just one of them. Two equations are necessary, since two of the GFs are defined independently, while the other ones can be expressed by means of these two. Furthermore, it must be remembered that the dynamics of the GFs is influenced by the interactions described by the self-energy, so that an independent derivation of Σ is in any case necessary.

Since the elaboration that follows is rather involved, it may be useful to give here, at the beginning of our trip, the path that will be followed in the rest of the chapter.

1. First, we shall derive equations of motion for $G^<$ and G^r in presence of a uniform electric field. The first one is what we are looking for, and the second must be coupled to the first, since G^r is the function that carries the information on the dynamics of the system, independently of its particular state.

2. We will then move to central and relative coordinates $(\boldsymbol{R}, T, \boldsymbol{s}, \tau)$, from the original variables $(\boldsymbol{r}, t, \boldsymbol{r}', t')$, and the equations found in (1) for $G^<$ and $G^{\rm r}$ will be transformed in terms of the new variables. This operation will allow us to distinguish between "macroscopic" dynamics and the dynamics that occurs inside the correlation length, or inside wavepackets. This change of variables will also be accompanied to the Fourier transformation with respect to $\boldsymbol{s}$ and τ leading to the function $G^<(\boldsymbol{R}, T, \boldsymbol{k}, \omega)$, i.e., to the "quantum distribution" Wigner function.
3. The next step is the introduction of the *gradient-expansion approximation* that simplifies the equations but greatly limits their validities. It assumes that the quantities of interest, namely the GFs and self-energies, do not change too rapidly as a function of R and T, so that they can be approximated by linear functions within the correlation length and time.
4. Finally, the approximated equations for $G^<$ and $G^{\rm r}$ obtained in (3) will be applied to the simplest case of steady-state conditions with a constant applied electric field.

In the following sections, we shall develop the project indicated above, step by step.

27.1 The Equations for $G^<(\boldsymbol{r}, t, \boldsymbol{r}', t')$ and $G^{\rm r}(\boldsymbol{r}, t, \boldsymbol{r}', t')$

Let us assume that an electric field $\boldsymbol{E}$ is present. For simplicity, we shall assume it constant in space and time. It will be included in the unperturbed single-particle Hamiltonian:

$$\mathcal{H}_{\circ \rm E} = \mathcal{H}_\circ + \mathcal{H}_{\rm E} \quad , \quad \mathcal{H}_{\rm E} = -q\boldsymbol{E} \cdot \boldsymbol{r},$$

where q is the particle charge, not to be confused with the phonon wavevector that in this chapter will not be considered.

Let us write the Dyson equation in matrix form (26.30) with explicit variables, where we recall that integration over repeated variables must be performed:

$$\tilde{G}(\boldsymbol{r}, t, \boldsymbol{r}', t') = \tilde{G}_\circ(\boldsymbol{r}, t, \boldsymbol{r}', t') + \tilde{G}_\circ(\boldsymbol{r}, t, \boldsymbol{r}_1, t_1)\tilde{\Sigma}(\boldsymbol{r}_1, t_1, \boldsymbol{r}_2, t_2)\tilde{G}(\boldsymbol{r}_2, t_2, \boldsymbol{r}', t'). \tag{27.1}$$

Then we apply the operator

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t} - \mathcal{H}_\circ(\boldsymbol{r}) - \mathcal{H}_{\rm E}(\boldsymbol{r})\right) \tag{27.2}$$

to the above equation. Since this operator acts on the variables $\boldsymbol{r}$ and t, it acts only on $\tilde{G}_\circ$, and, taking into account (26.22), (24.26), and (24.45)–(24.47), its effect is

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})\right)\tilde{G}_\circ(\boldsymbol{r},t,\boldsymbol{r}',t')=\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})\right)\begin{pmatrix}G^t_\circ & -G^<_\circ\\ G^>_\circ & -G^{\bar{t}}_\circ\end{pmatrix}$$
$$=\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_\circ(\boldsymbol{r})-\mathcal{H}_{\mathrm{E}}(\boldsymbol{r})\right)\begin{pmatrix}G^{(r)}_\circ+G^<_\circ & -G^<_\circ\\ G^>_\circ & -G^<_\circ+G^{(a)}_\circ\end{pmatrix}$$
$$=\begin{pmatrix}\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t') & 0\\ 0 & \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\end{pmatrix}=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\tilde{\mathcal{I}},$$

where $\tilde{\mathcal{I}}$ is the identity matrix. If we now apply the operator (27.2) to the Dyson equation (27.1) and use the above result, we obtain

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})\right)\tilde{G}(\boldsymbol{r},t,\boldsymbol{r}',t')$$
$$=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\tilde{\mathcal{I}}+\delta(\boldsymbol{r}-\boldsymbol{r}_1)\delta(t-t_1)\tilde{\mathcal{I}}\tilde{\Sigma}(\boldsymbol{r}_1,t_1,\boldsymbol{r}_2,t_2)\tilde{G}(\boldsymbol{r}_2,t_2,\boldsymbol{r}',t')$$
$$=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\tilde{\mathcal{I}}+\tilde{\Sigma}(\boldsymbol{r},t,\boldsymbol{r}_2,t_2)\tilde{G}(\boldsymbol{r}_2,t_2,\boldsymbol{r}',t'). \tag{27.3}$$

Let us write explicitly this equation for the element (1,2) of the matrix:

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})\right)G^<=\Sigma^tG^<-\Sigma^<G^{\bar{t}}. \tag{27.4}$$

According to our program, we need a second equation and we chose to look for the equation for $G^{\rm r}$. This function, however, does not appear as element of the matrix $\tilde{G}$, but it is simply related to $G^<$ and G^t by (24.26). Equation (27.3) yields, for the matrix element (1,1)

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})\right)G^t(\boldsymbol{r},t,\boldsymbol{r}',t')=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')+\Sigma^tG^t-\Sigma^<G^>.$$

Now we use (24.26) and (27.4), obtaining, after straightforward calculations,

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})\right)G^{\rm r}=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')+\Sigma^{\rm r}G^{\rm r}, \tag{27.5}$$

where we have also used $\Sigma^{\rm r}=\Sigma^t-\Sigma^<$.

Since our project envisages that we shall move to central and relative coordinates, we need the equations for $G^<$ and $G^{\rm r}$ also with respect to the primed variables $\boldsymbol{r}'$ and t'. For this purpose, we must repeat the derivation above, starting from the other form of the Dyson equation, i.e., $\tilde{G}=\tilde{G}_\circ+\tilde{G}\Sigma\tilde{G}_\circ$, and using (24.48) and (24.49) instead of (24.45) and (24.46). The resulting equations are

$$\left(-\mathrm{i}\hbar\frac{\partial}{\partial t'}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r}')\right)G^<(\boldsymbol{r},t,\boldsymbol{r}',t')=G^t\Sigma^<-G^<\Sigma^{\bar{t}}, \tag{27.6}$$

$$\left(-\mathrm{i}\hbar\frac{\partial}{\partial t'}-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r}')\right)G^{\rm r}(\boldsymbol{r},t,\boldsymbol{r}',t')=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')+G^{\rm r}\Sigma^{\rm r}. \tag{27.7}$$

27.2 The Equations for $G^<(\boldsymbol{R},\boldsymbol{T},\boldsymbol{k},\boldsymbol{\omega})$ and $G^{\rm r}(\boldsymbol{R},\boldsymbol{T},\boldsymbol{k},\boldsymbol{\omega})$

Summing and subtracting the two equations for $G^{\rm r}$, (27.5) e (27.7), we obtain

$$\left[\mathrm{i}\hbar\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial t'}\right)-(\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})+\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r}'))\right]G^{\rm r}(\boldsymbol{r},t,\boldsymbol{r}',t') = 2\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')+\Sigma^{(\rm r)}G^{\rm r}+G^{\rm r}\Sigma^{\rm r} \tag{27.8}$$

and

$$\left[\mathrm{i}\hbar\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial t'}\right)-(\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r}'))\right]G^{\rm r}(\boldsymbol{r},t,\boldsymbol{r}',t') = \Sigma^{\rm r}G^{\rm r}-G^{\rm r}\Sigma^{\rm r}. \tag{27.9}$$

At this point, we move to the central and relative coordinates used in the WF:

$$\begin{cases}\boldsymbol{R}=\frac{1}{2}(\boldsymbol{r}+\boldsymbol{r}')\ ,\ \boldsymbol{s}=\boldsymbol{r}-\boldsymbol{r}'\\ T=\frac{1}{2}(t+t')\ ,\quad \tau=t-t'\end{cases}\quad\text{or}\quad\begin{cases}\boldsymbol{r}=\boldsymbol{R}+\frac{1}{2}\boldsymbol{s}\ ,\ \boldsymbol{r}'=\boldsymbol{R}-\frac{1}{2}\boldsymbol{s}\\ t=T+\frac{1}{2}\tau\ ,\ \ t'=T-\frac{1}{2}\tau\end{cases}. \tag{27.10}$$

Note that the Jacobian is unity. For the above time derivatives, we have

$$\frac{\partial}{\partial t}+\frac{\partial}{\partial t'}=\frac{\partial}{\partial T}\ ,\qquad \frac{\partial}{\partial t}-\frac{\partial}{\partial t'}=2\frac{\partial}{\partial \tau}.$$

As it regards the terms with the Hamiltonian, it is useful to refer to an explicit expression for the Hamiltonian, and, for simplicity we assume the form

$$\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})=-\frac{\hbar^2}{2m}\nabla^2_{\boldsymbol{r}}-q\boldsymbol{E}\cdot\boldsymbol{r}.$$

With the above change of variables, we have

$$\nabla^2_{\boldsymbol{r}}=\frac{1}{4}\nabla^2_{\boldsymbol{R}}+\nabla^2_{\boldsymbol{s}}+\nabla_{\boldsymbol{R}}\cdot\nabla_{\boldsymbol{s}}\quad\text{and}\quad\nabla^2_{\boldsymbol{r}'}=\frac{1}{4}\nabla^2_{\boldsymbol{R}}+\nabla^2_{\boldsymbol{s}}-\nabla_{\boldsymbol{R}}\cdot\nabla_{\boldsymbol{s}}.$$

Thus,

$$\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})+\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r}')=-\frac{\hbar^2}{4m}\nabla^2_{\boldsymbol{R}}-\frac{\hbar^2}{m}\nabla^2_{\boldsymbol{s}}-2q\boldsymbol{E}\cdot\boldsymbol{R}$$

and

$$\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r})-\mathcal{H}_{\circ\mathrm{E}}(\boldsymbol{r}')=-\frac{\hbar^2}{m}\nabla_{\boldsymbol{R}}\cdot\nabla_{\boldsymbol{s}}-q\boldsymbol{E}\cdot\boldsymbol{s}.$$

Substituting the above results into (27.8) and (27.9), these become

$$\left[\mathrm{i}\hbar\frac{\partial}{\partial\tau}+\frac{\hbar^2}{8m}\nabla^2_{\boldsymbol{R}}+\frac{\hbar^2}{2m}\nabla^2_{\boldsymbol{s}}+q\boldsymbol{E}\cdot\boldsymbol{R}\right]G^{\rm r}(\boldsymbol{R},T,\boldsymbol{s},\tau) = \delta(\boldsymbol{s})\delta(\tau)+\frac{1}{2}\left[\Sigma^{\rm r}G^{\rm r}+G^{\rm r}\Sigma^{\rm r}\right] \tag{27.11}$$

and

$$\left[\mathrm{i}\hbar\frac{\partial}{\partial T}+\frac{\hbar^2}{m}\nabla_{\boldsymbol{R}}\cdot\nabla_{\boldsymbol{s}}+q\boldsymbol{E}\cdot\boldsymbol{s}\right]G^{\rm r}=\Sigma^{\rm r}G^{\rm r}-G^{\rm r}\Sigma^{\rm r}. \tag{27.12}$$

To perform the Fourier transform with respect to $\boldsymbol{s}$ and τ, multiply the two last equations by $1/(2\pi)^{3/2}\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\,\mathrm{e}^{\mathrm{i}\omega\tau}$ and integrate in $\mathrm{d}\boldsymbol{s}$ e $\mathrm{d}\tau$. Then, remembering that

$$\frac{1}{(2\pi)^{3/2}}\int G^{\rm r}(\boldsymbol{R},T,\boldsymbol{s},\tau)\,\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{s}}\mathrm{e}^{\mathrm{i}\omega\tau}\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau=G^{\rm r}(\boldsymbol{R},T,\boldsymbol{k},\omega),$$

and performing some integrations by parts, we transform the two equations above as

$$\begin{aligned}&\left[\hbar\omega+\frac{\hbar^2}{8m}\nabla^2_{\boldsymbol{R}}-\frac{\hbar^2k^2}{2m}+q\boldsymbol{E}\cdot\boldsymbol{R}\right]G^{\rm r}(\boldsymbol{R},T,\boldsymbol{k},\omega)\\&\quad=\frac{1}{(2\pi)^{3/2}}+\frac{1}{(2\pi)^{3/2}}\int\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\,\mathrm{e}^{\mathrm{i}\omega\tau}\frac{1}{2}\left[\Sigma^{\rm r}G^{\rm r}+G^{\rm r}\Sigma^{\rm r}\right]\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau\end{aligned}$$

and

$$\begin{aligned}&\left[\frac{\partial}{\partial T}+\frac{\hbar}{m}\boldsymbol{k}\cdot\nabla_{\boldsymbol{R}}+q\boldsymbol{E}\cdot\frac{1}{\hbar}\nabla_{\boldsymbol{k}}\right]G^{\rm r}(\boldsymbol{R},T,\boldsymbol{k},\omega)\\&\quad=\frac{1}{\mathrm{i}\hbar}\frac{1}{(2\pi)^{3/2}}\int\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\,\mathrm{e}^{\mathrm{i}\omega\tau}\left[\Sigma^{\rm r}G^{\rm r}-G^{\rm r}\Sigma^{\rm r}\right]\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau.\end{aligned}$$

Note that at the l.h.s. of the second equation we have the Liouvillian, as in Boltzmann equation. The r.h.s. with the self-energy is the contribution of the interactions (collisions). In this term, the Fourier transforms have not been substituted to the integrals for future application of the gradient expansion.

As it regards the first equation, which has clearly to do with the effect of the interactions on the energy, we see that the potential energy $-q\boldsymbol{E}\cdot\boldsymbol{R}$ of the particle in the position $\boldsymbol{R}$ is subtracted by the many-body energy $\hbar\omega$ of the system. Following Hänch e Mahan [180], we may simplify this expression by inserting the new frequency $\overline{\omega}$ by means of a change of variables. At present, the independent variables of $G^{\rm r}$ are the two "macroscopic" variables $\boldsymbol{R}$ e T and the two Fourier-conjugated variables $\boldsymbol{k}$ and ω, related to the momentum and energy of the particle. Let us then operate the transformation

$$\overline{\boldsymbol{R}}=\boldsymbol{R}\ \ ,\quad \overline{T}=T\ \ ,\quad \overline{\boldsymbol{k}}=\boldsymbol{k}\ \ ,\quad \overline{\omega}=\omega+\frac{1}{\hbar}q\boldsymbol{E}\cdot\boldsymbol{R}. \tag{27.13}$$

The only necessary derivatives are with respect to $\boldsymbol{k}$, unaltered, and with respect to $\boldsymbol{R}$, which becomes

$$\nabla_{\boldsymbol{R}}=\nabla_{\overline{\boldsymbol{R}}}+\frac{1}{\hbar}q\boldsymbol{E}\frac{\partial}{\partial\overline{\omega}}. \tag{27.14}$$

With this change of variables, the equations for G^{r} become

$$\left[\hbar\overline{\omega} + \frac{\hbar^2}{8m}\left(\nabla_{\boldsymbol{R}} + \frac{1}{\hbar}q\boldsymbol{E}\frac{\partial}{\partial\overline{\omega}}\right)^2 - \frac{\hbar^2k^2}{2m}\right] G^{\mathrm{r}}(\boldsymbol{R},T,\boldsymbol{k},\omega)$$
$$= \frac{1}{(2\pi)^{3/2}} + \frac{1}{(2\pi)^{3/2}}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\,\mathrm{e}^{\mathrm{i}\omega\tau}\frac{1}{2}\left[\Sigma^{\mathrm{r}}G^{\mathrm{r}} + G^{\mathrm{r}}\Sigma^{\mathrm{r}}\right]\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau \quad (27.15)$$

and

$$\left[\frac{\partial}{\partial T} + \frac{\hbar}{m}\boldsymbol{k}\cdot\nabla_{\boldsymbol{R}} + q\boldsymbol{E}\cdot\left(\frac{\boldsymbol{k}}{m}\frac{\partial}{\partial\overline{\omega}} + \frac{1}{\hbar}\nabla_{\boldsymbol{k}}\right)\right] G^{\mathrm{r}}(\boldsymbol{R},T,\boldsymbol{k},\omega)$$
$$= \frac{1}{\mathrm{i}\hbar}\frac{1}{(2\pi)^{3/2}}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\,\mathrm{e}^{\mathrm{i}\omega\tau}\left[\Sigma^{\mathrm{r}}G^{\mathrm{r}} - G^{\mathrm{r}}\Sigma^{\mathrm{r}}\right]\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau. \quad (27.16)$$

These equations for G^{r} are exact, but too complicated to be used in practice. It is necessary to introduce approximations. Shortly we shall discuss an approximation introduced by Kadanoff and Baym [228], known as *gradient expansion*. Before that, however, we must find the analogous equations for $G^{<}(\boldsymbol{R},T,\boldsymbol{k},\omega)$.

To find the equation for $G^{<}(\boldsymbol{R},T,\boldsymbol{k},\omega)$, we proceed in the same way as we did for G^{r}. First we sum and subtract the two equations for $G^{<}$, (27.4) and (27.6). Then we move to the central and difference variables, Fourier transform, and finally perform the transformation (27.13) to move from the total energy $\hbar\omega$ to $\hbar\overline{\omega}$, the energy diminished by the local potential energy due to the applied field. The resulting equations are

$$\left[\hbar\overline{\omega} + \frac{\hbar^2}{8m}\left(\nabla_{\boldsymbol{R}} + \frac{1}{\hbar}q\boldsymbol{E}\frac{\partial}{\partial\overline{\omega}}\right)^2 - \epsilon_{\boldsymbol{k}}\right] G^{<}(\boldsymbol{R},T,\boldsymbol{k},\omega)$$
$$= \frac{1}{(2\pi)^{3/2}}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\mathrm{e}^{\mathrm{i}\omega\tau}\frac{1}{2}\left[\Sigma^{t}G^{<} - \Sigma^{<}G^{(\bar{t})} + G^{t}\Sigma^{<} - G^{<}\Sigma^{(\bar{t})}\right]\mathrm{d}\boldsymbol{s}\mathrm{d}\tau \quad (27.17)$$

and

$$\left[\frac{\partial}{\partial T} + \frac{\hbar\boldsymbol{k}}{m}\cdot\nabla_{\boldsymbol{R}} + q\boldsymbol{E}\cdot\left(\frac{\boldsymbol{k}}{m}\frac{\partial}{\partial\overline{\omega}} + \frac{1}{\hbar}\nabla_{\boldsymbol{k}}\right)\right] G^{<}$$
$$= \frac{1}{\mathrm{i}\hbar}\frac{1}{(2\pi)^{3/2}}\int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\mathrm{e}^{\mathrm{i}\omega\tau}\left[\Sigma^{t}G^{<} - \Sigma^{<}G^{(\bar{t})} - G^{t}\Sigma^{<} + G^{<}\Sigma^{(\bar{t})}\right]\mathrm{d}\boldsymbol{s}\mathrm{d}\tau. \quad (27.18)$$

27.3 Gradient-Expansion Approximation

The gradient-expansion approximation consists in a first-order expansion of the GFs and self-energies, as functions of the "macroscopic" quantities $\boldsymbol{R}$ and T over distances $\boldsymbol{s}$ e τ. Thus, it may be applied if the quantities of interest are

slowly varying in space $\boldsymbol{R}$ and time T. Generally speaking, this approximation was good *when macroscopic systems were macroscopic*. Today, devices are built so small that this approximation is no more acceptable for them. For this reason, we devote next chapter to the Datta development of GFs in mesoscopic systems. The gradient-expansion approximation is, however, particularly useful to study electron transport in homogeneous and stationary systems.

Let us then consider one of the integrals in (27.15). A product that appears in the integrands is

$$\Sigma^{\mathrm{r}}(\boldsymbol{R}+\boldsymbol{s}/2, T+\tau/2, \boldsymbol{r}_1, t_1)\, G^{\mathrm{r}}(\boldsymbol{r}_1, t_1, \boldsymbol{R}-\boldsymbol{s}/2, T-\tau/2).$$

It is necessary to write the variables in this way because $\boldsymbol{R}$ e T are "external" variables, and $\boldsymbol{s}$, τ, $\boldsymbol{r}_1$, and t_1 are the integration variables.

To apply the gradient expansion, it is useful to perform another change of variables, which considers the time variable as a fourth component:

$$\boldsymbol{y} = (\boldsymbol{r}_1, t_1) - (\boldsymbol{R}, T) + (\boldsymbol{s}/2, \tau/2)\ , \qquad \boldsymbol{x} = (\boldsymbol{R}, T) + (\boldsymbol{s}/2, \tau/2) - (\boldsymbol{r}_1, t_1),$$

in place of $(\boldsymbol{s}, \tau)$ and $(\boldsymbol{r}_1, t_1)$. They are inverted as

$$(\boldsymbol{s}, \tau) = \boldsymbol{x} + \boldsymbol{y}\ , \qquad (\boldsymbol{r}_1, t_1) = (\boldsymbol{R}, T) + (\boldsymbol{y} - \boldsymbol{x})/2.$$

The product of the two functions above becomes

$$\Sigma^{\mathrm{r}}(\boldsymbol{R}+\boldsymbol{y}/2+\boldsymbol{x}/2, \boldsymbol{R}+\boldsymbol{y}/2-\boldsymbol{x}/2)\, G^{\mathrm{r}}(\boldsymbol{R}-\boldsymbol{x}/2+\boldsymbol{y}/2, \boldsymbol{R}-\boldsymbol{x}/2-\boldsymbol{y}/2).$$

The exponential in (27.15) is written as ($-\omega$ is considered as fourth component of $\boldsymbol{k}$)

$$\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\mathrm{e}^{\mathrm{i}\omega\tau} = \mathrm{e}^{-\mathrm{i}\boldsymbol{k}(\boldsymbol{x}+\boldsymbol{y})}.$$

Let us define G_{c} e Σ_{c} the same functions G e Σ in terms of the central and difference variables, i.e.,

$$G^{\mathrm{r}}(\boldsymbol{R}-\boldsymbol{x}/2+\boldsymbol{y}/2, \boldsymbol{R}-\boldsymbol{x}/2-\boldsymbol{y}/2) = G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}-\boldsymbol{x}/2, \boldsymbol{y})$$

and

$$\Sigma^{\mathrm{r}}(\boldsymbol{R}+\boldsymbol{y}/2+\boldsymbol{x}/2, \boldsymbol{R}+\boldsymbol{y}/2-\boldsymbol{x}/2) = \Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}+\boldsymbol{y}/2, \boldsymbol{x}).$$

The gradient expansion consists then in writing

$$G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}-\boldsymbol{x}/2, \boldsymbol{y}) \approx G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}, \boldsymbol{y}) - \frac{1}{2}\nabla_{\boldsymbol{R}} G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}, \boldsymbol{y}) \cdot \boldsymbol{x},$$

$$\Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}+\boldsymbol{y}/2, \boldsymbol{x}) \approx \Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}, \boldsymbol{x}) + \frac{1}{2}\nabla_{\boldsymbol{R}} \Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R}, \boldsymbol{x}) \cdot \boldsymbol{y}.$$

Now we insert these expressions in the first term of the integral in (27.15), symbolically named $I(\Sigma^{\mathrm{r}} G^{\mathrm{r}})$ and obtain, to first order,

$$
\begin{aligned}
I(\Sigma^{\mathrm{r}}G^{\mathrm{r}}) &\equiv \int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\,\mathrm{e}^{\mathrm{i}\omega\tau}\,\Sigma^{\mathrm{r}}G^{\mathrm{r}}\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau \\
&\approx \int \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{x}}\,\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{y}}\left\{\Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R},\boldsymbol{x})G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R},\boldsymbol{y}) - \frac{1}{2}\Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R},\boldsymbol{x})\nabla_{\boldsymbol{R}}G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R},\boldsymbol{y})\cdot\boldsymbol{x}\right. \\
&\quad \left. + \frac{1}{2}\nabla_{\boldsymbol{R}}\Sigma^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R},\boldsymbol{x})\cdot\boldsymbol{y}\,G^{\mathrm{r}}_{\mathrm{c}}(\boldsymbol{R},\boldsymbol{y})\right\}\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\ .
\end{aligned}
$$

The first term is already the product of the transforms which give G and Σ in terms of the central variables and of $\boldsymbol{k}$ and ω. The other terms can be obtained by differentiation. The entire integral becomes

$$
I(\Sigma^{\mathrm{r}}G^{\mathrm{r}}) = (2\pi)^3\left[\Sigma^{\mathrm{r}}(\boldsymbol{R},\boldsymbol{k})\,G^{\mathrm{r}}(\boldsymbol{R},\boldsymbol{k}) - \frac{\mathrm{i}}{2}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{R}}G^{\mathrm{r}} + \frac{\mathrm{i}}{2}\nabla_{\boldsymbol{R}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{k}}G^{\mathrm{r}}\right],
$$

where the arguments of the GF are always $(\boldsymbol{R},\boldsymbol{k})$, i.e., $(\boldsymbol{R},T,\boldsymbol{k},\omega)$. If we now move to the variable $\overline{\omega}$ we have, using (27.13) and (27.14)

$$
\begin{aligned}
I(\Sigma^{\mathrm{r}}G^{\mathrm{r}})/(2\pi)^3 &= \Sigma^{\mathrm{r}}(\boldsymbol{R},\boldsymbol{k})\,G^{\mathrm{r}}(\boldsymbol{R},\boldsymbol{k}) - \frac{\mathrm{i}}{2}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{R}}G^{\mathrm{r}} + \frac{\mathrm{i}}{2}\nabla_{\boldsymbol{R}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{k}}G^{\mathrm{r}} \\
&\quad + \frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\cdot\left[-\frac{1}{2}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\frac{\partial}{\partial\overline{\omega}}G^{\mathrm{r}} + \frac{1}{2}\frac{\partial}{\partial\overline{\omega}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{k}}G^{\mathrm{r}}\right]. \qquad (27.19)
\end{aligned}
$$

The integral just evaluated is necessary for the terms in (27.15) and (27.16). with the product $\Sigma^{\mathrm{r}}G^{\mathrm{r}}$. If we now consider the product in inverse order, we may repeat the same identical steps with the two functions exchanged. Thus, if we call $I(G^{\mathrm{r}}\Sigma^{\mathrm{r}})$ the new integral, we obtain, after some simplifications,

$$
\frac{1}{(2\pi)^3}I(\Sigma^{\mathrm{r}}G^{\mathrm{r}} + G^{\mathrm{r}}\Sigma^{\mathrm{r}}) = 2\Sigma^{\mathrm{r}}(\boldsymbol{R},\boldsymbol{k})\,G^{\mathrm{r}}(\boldsymbol{R},\boldsymbol{k})
$$

and

$$
\begin{aligned}
\frac{1}{(2\pi)^3}I(\Sigma^{\mathrm{r}}G^{\mathrm{r}} - G^{\mathrm{r}}\Sigma^{\mathrm{r}}) &= \{-\mathrm{i}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{R}}G^{\mathrm{r}} + \mathrm{i}\nabla_{\boldsymbol{R}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{k}}G^{\mathrm{r}} \\
&\quad + \frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\cdot\left[-\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\frac{\partial}{\partial\overline{\omega}}G^{\mathrm{r}} + \frac{\partial}{\partial\overline{\omega}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{k}}G^{\mathrm{r}}\right]\right\}.
\end{aligned}
$$

If we now substitute these results in (27.15) and (27.16), we obtain

$$
\begin{aligned}
&\left[\hbar\overline{\omega} - \epsilon_k + \frac{\hbar^2}{8m}\left(\nabla_{\boldsymbol{R}} + \frac{1}{\hbar}q\boldsymbol{E}\frac{\partial}{\partial\overline{\omega}}\right)^2\right. \\
&\qquad \left. - (2\pi)^{3/2}\Sigma^{\mathrm{r}}(\boldsymbol{R},T,\boldsymbol{k},\omega)\right]G^{\mathrm{r}}(\boldsymbol{R},T,\boldsymbol{k},\omega) = \frac{1}{(2\pi)^{3/2}} \qquad (27.20)
\end{aligned}
$$

and

$$\left\{\frac{\partial}{\partial T}+\boldsymbol{v_k}\cdot\nabla_{\boldsymbol{R}}+\frac{q}{\hbar}\boldsymbol{E}\cdot\left[\left(\boldsymbol{v_k}+\frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\right)\frac{\partial}{\partial\overline{\omega}}+\left(1-\frac{(2\pi)^{3/2}}{\hbar}\frac{\partial}{\partial\overline{\omega}}\Sigma^{\mathrm{r}}\right)\nabla_{\boldsymbol{k}}\right]\right\}G^{\mathrm{r}}$$
$$=-\frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\cdot\nabla_{\boldsymbol{R}}G^{\mathrm{r}}+\nabla_{\boldsymbol{R}}\Sigma^{\mathrm{r}}\cdot\frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}G^{\mathrm{r}},\qquad(27.21)$$

where $\epsilon_k=(\hbar^2k^2/2m)$ and $\boldsymbol{v_k}=\hbar\boldsymbol{k}/m$.

These are equations that, once Σ^{r} is known, describe the evolution of G^{r}, in gradient expansion approximation. As before, two equations are necessary since both the dynamics and the spectral density have to be determined.

To evaluate the probability distribution of physical quantities, it is however necessary to find the dynamical evolution of $G^<$, which carries the information on the state of the system. To apply the gradient expansion to the equations (27.17) and (27.18) for $G^<$, we follow the same procedure applied for G^{r}. This time we have to take care of four terms, which result to be given by

$$I(\Sigma^tG^<)\equiv\int\int\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\boldsymbol{s}}\mathrm{e}^{\mathrm{i}\omega\tau}\Sigma^tG^<\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}\tau$$
$$=(2\pi)^3\left\{\Sigma^t(\boldsymbol{R},\boldsymbol{k})\,G^<(\boldsymbol{R},\boldsymbol{k})-\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{k}}\Sigma^t\cdot\nabla_{\boldsymbol{R}}G^<+\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{R}}\Sigma^t\cdot\nabla_{\boldsymbol{k}}G^<\right.$$
$$\left.+\frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\cdot\left[-\frac{1}{2}\nabla_{\boldsymbol{k}}\Sigma^t\frac{\partial}{\partial\overline{\omega}}G^<+\frac{1}{2}\frac{\partial}{\partial\overline{\omega}}\Sigma^t\nabla_{\boldsymbol{k}}G^<\right]\right\};$$
$$I(\Sigma^<G^{\bar{t}})=(2\pi)^3\left\{\Sigma^<(\boldsymbol{R},\boldsymbol{k})\,G^{\bar{t}}(\boldsymbol{R},\boldsymbol{k})-\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{k}}\Sigma^<\nabla_{\boldsymbol{R}}G^{\bar{t}}+\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{R}}\Sigma^<\nabla_{\boldsymbol{k}}G^{\bar{t}}\right.$$
$$\left.+\frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\cdot\left[-\frac{1}{2}\nabla_{\boldsymbol{k}}\Sigma^<\frac{\partial}{\partial\overline{\omega}}G^{\bar{t}}+\frac{1}{2}\frac{\partial}{\partial\overline{\omega}}\Sigma^<\nabla_{\boldsymbol{k}}G^{\bar{t}}\right]\right\};$$
$$I(G^t\Sigma^<)=(2\pi)^3\left\{G^t(\boldsymbol{R},\boldsymbol{k})\,\Sigma^<(\boldsymbol{R},\boldsymbol{k})-\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{k}}G^t\nabla_{\boldsymbol{R}}\Sigma^<+\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{R}}G^t\nabla_{\boldsymbol{k}}\Sigma^<\right.$$
$$\left.+\frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\cdot\left[-\frac{1}{2}\nabla_{\boldsymbol{k}}G^t\frac{\partial}{\partial\overline{\omega}}\Sigma^<+\frac{1}{2}\frac{\partial}{\partial\overline{\omega}}G^t\nabla_{\boldsymbol{k}}\Sigma^<\right]\right\};$$
$$I(G^<\Sigma^{\bar{t}})=(2\pi)^{3/2}\left\{G^<(\boldsymbol{R},\boldsymbol{k})\,\Sigma^{\bar{t}}(\boldsymbol{R},\boldsymbol{k})-\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{k}}G^<\nabla_{\boldsymbol{R}}\Sigma^{\bar{t}}+\frac{\mathrm{i}}{2}\nabla_{\boldsymbol{R}}G^<\nabla_{\boldsymbol{k}}\Sigma^{\bar{t}}\right.$$
$$\left.+\frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\cdot\left[-\frac{1}{2}\nabla_{\boldsymbol{k}}G^<\frac{\partial}{\partial\overline{\omega}}\Sigma^{\bar{t}}+\frac{1}{2}\frac{\partial}{\partial\overline{\omega}}G^<\nabla_{\boldsymbol{k}}\Sigma^{\bar{t}}\right]\right\}.$$

Summing up and reordering the various terms, the r.h.s. of (27.17) becomes (the arguments are all $(\boldsymbol{R},\boldsymbol{k})$, i.e., $(\boldsymbol{R},T,\boldsymbol{k},\omega)$):

$$\frac{1}{2}(2\pi)^{3/2}\Big\{G^<\left(\Sigma^t-\Sigma^{\bar t}\right)+\Sigma^<\left(G^t-G^{\bar t}\right)$$
$$-\frac{\mathrm{i}}{2}\nabla_{\boldsymbol R}G^<\cdot\nabla_{\boldsymbol k}\left(\Sigma^t+\Sigma^{\bar t}\right)+\frac{\mathrm{i}}{2}\nabla_{\boldsymbol k}G^<\cdot\nabla_{\boldsymbol R}\left(\Sigma^t+\Sigma^{\bar t}\right)$$
$$+\frac{\mathrm{i}}{2}\nabla_{\boldsymbol k}\Sigma^<\cdot\nabla_{\boldsymbol R}\left(G^t+G^{\bar t}\right)-\frac{\mathrm{i}}{2}\nabla_{\boldsymbol R}\Sigma^<\cdot\nabla_{\boldsymbol k}\left(G^t+G^{\bar t}\right)$$
$$+\frac{\mathrm{i}}{\hbar}q\boldsymbol E\cdot\left[-\frac{1}{2}\nabla_{\boldsymbol k}\left(\Sigma^t+\Sigma^{\bar t}\right)\frac{\partial}{\partial\overline{\omega}}G^<+\frac{1}{2}\frac{\partial}{\partial\overline{\omega}}\left(\Sigma^t+\Sigma^{\bar t}\right)\nabla_{\boldsymbol k}G^<\right.$$
$$\left.+\frac{1}{2}\nabla_{\boldsymbol k}\Sigma^<\frac{\partial}{\partial\overline{\omega}}\left(G^t+G^{\bar t}\right)-\frac{1}{2}\frac{\partial}{\partial\overline{\omega}}\Sigma^<\nabla_{\boldsymbol k}\left(G^t+G^{\bar t}\right)\right]\Big\}. \tag{27.22}$$

To proceed we need the relations that express time e anti-time ordered GF as functions of the others. Using (24.26), we obtain

$$G^t-G^{\bar t}=G^{\mathrm r}+G^{\mathrm a}=2\mathcal R(G^{\mathrm r})\quad\text{and}\quad\Sigma^t-\Sigma^{\bar t}=\Sigma^{\mathrm r}+\Sigma^{\mathrm a}=2\mathcal R(\Sigma^{\mathrm r}). \tag{27.23}$$

where we have taken into account that $G^{\mathrm a}(\boldsymbol R,T,\boldsymbol k,\omega)$ is the complex conjugate of $G^{\mathrm r}(\boldsymbol R,T,\boldsymbol k,\omega)$, and similarly for the self-energies. As it regards the sums,

$$G^t+G^{\bar t}=G^<+G^>\quad,\quad\Sigma^t+\Sigma^{\bar t}=\Sigma^<+\Sigma^>.$$

Substituting the above results into (27.22), it becomes

$$(2\pi)^{3/2}\left\{G^<\mathcal R(\Sigma^{\mathrm r})+\Sigma^<\mathcal R(G^{\mathrm r})\right.$$
$$+\frac{\mathrm{i}}{4}\left[-\nabla_{\boldsymbol R}G^<\cdot\nabla_{\boldsymbol k}\Sigma^>+\nabla_{\boldsymbol k}G^<\cdot\nabla_{\boldsymbol R}\Sigma^>+\nabla_{\boldsymbol k}\Sigma^<\cdot\nabla_{\boldsymbol R}G^>-\nabla_{\boldsymbol R}\Sigma^<\cdot\nabla_{\boldsymbol k}G^>\right]$$
$$\left.+\frac{\mathrm{i}}{4\hbar}q\boldsymbol E\cdot\left[-\nabla_{\boldsymbol k}\Sigma^>\frac{\partial}{\partial\overline{\omega}}G^<+\frac{\partial}{\partial\overline{\omega}}\Sigma^>\nabla_{\boldsymbol k}G^<+\nabla_{\boldsymbol k}\Sigma^<\frac{\partial}{\partial\overline{\omega}}G^>-\frac{\partial}{\partial\overline{\omega}}\Sigma^<\nabla_{\boldsymbol k}G^>\right]\right\}.$$

The expression just found is the elaboration, in gradient-expansion approximation, of the r.h.s. of (27.17). The same calculations give to the r.h.s. of (27.18) the following form:

$$\frac{1}{\mathrm i\hbar}\frac{1}{(2\pi)^{3/2}}\int\mathrm e^{-\mathrm i\boldsymbol{ks}}\mathrm e^{\mathrm i\omega\tau}\left[\Sigma^tG^<-\Sigma^<G^{\bar t}-G^t\Sigma^<+G^<\Sigma^{\bar t}\right]\mathrm d\boldsymbol s\mathrm d\tau$$
$$=\frac{1}{\mathrm i\hbar}(2\pi)^{3/2}\left\{G^<\left(\Sigma^<+\Sigma^>\right)-\Sigma^<\left(G^<+G^>\right)-\mathrm i\nabla_{\boldsymbol R}G^<\cdot\nabla_{\boldsymbol k}\mathcal R(\Sigma^{\mathrm r})\right.$$
$$+\mathrm i\nabla_{\boldsymbol k}G^<\cdot\nabla_{\boldsymbol R}\mathcal R(\Sigma^{\mathrm r})-\mathrm i\nabla_{\boldsymbol k}\Sigma^<\cdot\nabla_{\boldsymbol R}\mathcal R(G^{\mathrm r})+\mathrm i\nabla_{\boldsymbol R}\Sigma^<\cdot\nabla_{\boldsymbol k}\mathcal R(G^{\mathrm r})$$
$$+\frac{\mathrm i}{\hbar}q\boldsymbol E\cdot\left[-\nabla_{\boldsymbol k}\mathcal R(\Sigma^{\mathrm r})\frac{\partial}{\partial\overline{\omega}}G^<+\frac{\partial}{\partial\overline{\omega}}\mathcal R(\Sigma^{\mathrm r})\nabla_{\boldsymbol k}G^<\right.$$
$$\left.\left.-\nabla_{\boldsymbol k}\Sigma^<\frac{\partial}{\partial\overline{\omega}}\mathcal R(G^{\mathrm r})+\frac{\partial}{\partial\overline{\omega}}\Sigma^<\nabla_{\boldsymbol k}\mathcal R(G^{\mathrm r})\right]\right\}.$$

The expressions just found are the elaborations, in gradient-expansion approximation, of the r.h.s. of (27.17) and (27.18). Substituting them into their equations, these become, after simple steps,

$$\left[\hbar\overline{\omega}+\frac{\hbar^2}{8m}\left(\nabla_{\boldsymbol{R}}+\frac{1}{\hbar}q\boldsymbol{E}\frac{\partial}{\partial\overline{\omega}}\right)^2-\epsilon_{\boldsymbol{k}}\right]G^<(\boldsymbol{R},T,\boldsymbol{k},\omega)$$
$$=(2\pi)^{3/2}\left\{G^<\mathcal{R}(\Sigma^{\mathrm{r}})+\Sigma^<\mathcal{R}(G^{\mathrm{r}})+\frac{\mathrm{i}}{4}\left[\Sigma^>,G^<\right]_P-\frac{\mathrm{i}}{4}\left[\Sigma^<,G^>\right]_P\right.$$
$$\left.+\frac{\mathrm{i}}{4\hbar}q\boldsymbol{E}\left[-\nabla_{\boldsymbol{k}}\Sigma^>\frac{\partial}{\partial\overline{\omega}}G^<+\frac{\partial}{\partial\overline{\omega}}\Sigma^>\nabla_{\boldsymbol{k}}G^<+\nabla_{\boldsymbol{k}}\Sigma^<\frac{\partial}{\partial\overline{\omega}}G^>-\frac{\partial}{\partial\overline{\omega}}\Sigma^<\nabla_{\boldsymbol{k}}G^>\right]\right\} \tag{27.24}$$

and

$$\left\{\frac{\partial}{\partial T}+\boldsymbol{v}_{\boldsymbol{k}}\nabla_{\boldsymbol{R}}+\frac{q\boldsymbol{E}}{\hbar}\left[\left(1-\frac{(2\pi)^{3/2}}{\hbar}\frac{\partial\Sigma_r^{\mathrm{r}}}{\partial\overline{\omega}}\right)\nabla_{\boldsymbol{k}}+\left(\boldsymbol{v}_{\boldsymbol{k}}+\frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}\Sigma_r^{\mathrm{r}}\right)\frac{\partial}{\partial\overline{\omega}}\right]\right\}G^<$$
$$=\frac{(2\pi)^{3/2}}{\mathrm{i}\hbar}\left\{G^<\Sigma^>-\Sigma^<G^>+\mathrm{i}\left[\mathcal{R}(\Sigma^{\mathrm{r}}),G^<\right]_P+\mathrm{i}\left[\Sigma^<,\mathcal{R}(G^{\mathrm{r}})\right]_P\right.$$
$$\left.+\frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\left[-\nabla_{\boldsymbol{k}}\Sigma^<\frac{\partial}{\partial\overline{\omega}}\mathcal{R}(G^{\mathrm{r}})+\frac{\partial}{\partial\overline{\omega}}\Sigma^<\nabla_{\boldsymbol{k}}\mathcal{R}(G^{\mathrm{r}})\right]\right\}. \tag{27.25}$$

where the Poisson brackets of two quantities A and B have been introduced, defined as

$$[A,B]_P=\nabla_{\boldsymbol{R}}A\nabla_{\boldsymbol{k}}B-\nabla_{\boldsymbol{k}}A\nabla_{\boldsymbol{R}}B.$$

Equations (27.24) and (27.25) are the *QBEs*, as given by Mahan e Hänsch. [179, 180, 294]. The unknown function to be obtained with their solutions is $G^<(\boldsymbol{R},T,\boldsymbol{k},\omega)$, i.e., the Wigner function, which can be used to evaluate physical quantities of interest to compare with experimental results. The equations, however, contain also the functions $G^>$ and G^{r}, and we know that only two of such functions are independent. Thus, the two equations (27.24) and (27.25) must be solved together with (27.20) and (27.21) with the help of the relations which connect the various GFs, indicated in Chap. 24. It is useful to repeat that such functions depend on $\boldsymbol{k}$ and ω separately (besides on $\boldsymbol{R}$ and T), and not only on $\boldsymbol{k}$, because the relation between $\boldsymbol{k}$ and ω is not known a priori as in the classical case, but also depends on the interaction among the particles.

As it regards the various self-energies Σ which appear in the equations, they must be evaluated separately since they contain information on the interaction mechanisms. But they depend, in turn, upon $G^<$ e $G^>$ and therefore must be evaluated self-consistently. This is not different, in principle, from what happens in the BE, since the scattering probabilities in the collision integrals may depend on the distribution function through, for example, screening effects. This fact makes transport equations, both semiclassical and quantum, intrinsically nonlinear. In metals, this effect is crucial; in semiconductors, this nonlinearity is often neglected, at low electron concentrations, assuming scattering probabilities or self-energies independent of the state of the system.

In any case, the situation described above for the determination of the necessary GFs and the complexity of the equations to be solved, make the

problem of quantum transport extremely difficult to be solved, also with the simplifying gradient-expansion approximation.

27.4 Equations for Linear Response in Homogeneous Systems in Steady State

The complicated situation described above is greatly simplified when we are looking for the linear response to an applied electric field $\boldsymbol{E}$ in a homogeneous system in steady-state condition. Equations (27.20) and (27.21) for G^{r} and (27.24) and (27.25) for $G^{<}$ take simpler forms, in fact, if the derivatives with respect to $\boldsymbol{R}$ and T vanish, and terms higher than linear in $\boldsymbol{E}$ are neglected. Equations (27.20) and (27.21) become

$$\left[\hbar\overline{\omega} - \epsilon_k - (2\pi)^{3/2}\Sigma^{\mathrm{r}}\right] G^{\mathrm{r}} = \frac{1}{(2\pi)^{3/2}}$$

and

$$\frac{q}{\hbar}\boldsymbol{E}\cdot\left[\left(\boldsymbol{v}_{\boldsymbol{k}} + (2\pi)^{3/2}\nabla_{\boldsymbol{k}}\Sigma^{\mathrm{r}}\right)\frac{\partial}{\hbar\partial\overline{\omega}} + \left(1 - (2\pi)^{3/2}\frac{\partial}{\hbar\partial\overline{\omega}}\Sigma^{\mathrm{r}}\right)\nabla_{\boldsymbol{k}}\right] G^{\mathrm{r}} = 0. \tag{27.26}$$

The first equation has immediate solution:

$$G^{\mathrm{r}}(\boldsymbol{k},\overline{\omega}) = \frac{1}{(2\pi)^{3/2}}\frac{1}{\hbar\overline{\omega} - \epsilon_k - (2\pi)^{3/2}\Sigma^{\mathrm{r}}}, \tag{27.27}$$

with the usual interpretation of the shift and broadening of the energy levels. We also remember that the independent variable $\overline{\omega}$ is now the total energy diminished by the local potential energy. The reader is invited to verify, as a simple exercise, that (27.27) satisfies also (27.26).

The two equations for $G^{<}$, (27.24) and (27.25), in the homogeneous, stationary state become, again to first order in $\boldsymbol{E}$

$$\begin{aligned}&\left[\hbar\overline{\omega} - \epsilon_{\boldsymbol{k}}\right] G^{<}(\boldsymbol{R},T,\boldsymbol{k},\omega) = (2\pi)^{3/2}G^{<}\mathcal{R}(\Sigma^{\mathrm{r}}) + (2\pi)^{3/2}\Sigma^{<}\mathcal{R}(G^{\mathrm{r}})\\ &+\frac{(2\pi)^{3/2}}{4\hbar}q\boldsymbol{E}\left[-\nabla_{\boldsymbol{k}}\Sigma^{>}\frac{\partial}{\partial\omega}G^{<} + \frac{\partial}{\partial\omega}\Sigma^{>}\nabla_{\boldsymbol{k}}G^{<} + \nabla_{\boldsymbol{k}}\Sigma^{<}\frac{\partial}{\partial\omega}G^{>} - \frac{\partial}{\partial\omega}\Sigma^{<}\nabla_{\boldsymbol{k}}G^{>}\right]\end{aligned}$$

and

$$\begin{aligned}&\frac{q\boldsymbol{E}}{\hbar}\left[\left(1 - \frac{(2\pi)^{3/2}}{\hbar}\frac{\partial}{\partial\omega}\mathcal{R}(\Sigma^{\mathrm{r}})\right)\nabla_{\boldsymbol{k}} + \left(\boldsymbol{v}_{\boldsymbol{k}} + \frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}\mathcal{R}(\Sigma^{\mathrm{r}})\right)\frac{\partial}{\partial\omega}\right] G^{<}\\ &= \frac{(2\pi)^{3/2}}{\mathrm{i}\hbar}\left\{G^{<}\Sigma^{>} - \Sigma^{<}G^{>} + \frac{\mathrm{i}}{\hbar}q\boldsymbol{E}\left[-\nabla_{\boldsymbol{k}}\Sigma^{<}\frac{\partial}{\partial\omega}\mathcal{R}(G^{\mathrm{r}}) + \frac{\partial}{\partial\omega}\Sigma^{<}\nabla_{\boldsymbol{k}}\mathcal{R}(G^{\mathrm{r}})\right]\right\}.\end{aligned} \tag{27.28}$$

Let us concentrate on the second equation. The terms containing the field $\boldsymbol{E}$ may be simplified taking into account that, looking for the linear response, what is multiplied by the electric field may be substituted by its equilibrium expression, given in the previous chapters. Real and imaginary parts of Σ^{r} are defined in (26.40), from which

$$1 - \frac{(2\pi)^{3/2}}{\hbar}\frac{\partial}{\partial\omega}\mathcal{R}(\Sigma^{\mathrm{r}}) = 1 - \frac{\partial\omega_s}{\partial\omega} = \frac{\partial\sigma}{\partial\omega}, \tag{27.29}$$

where also the definition of σ in (26.42) has been used. $G^<$ is related to the spectral density, at equilibrium, by (24.56). Taking this into account (with the sign for fermions), the first term in (27.28) can be written as

$$I \equiv \left(1 - \frac{(2\pi)^{3/2}}{\hbar}\frac{\partial}{\partial\omega}\mathcal{R}(\Sigma^{\mathrm{r}})\right)\nabla_{\boldsymbol{k}}G^< = \frac{\mathrm{i}f_\circ}{\sqrt{2\pi}\hbar}\nabla_{\boldsymbol{k}}A\frac{\partial\sigma}{\partial\omega}.$$

With the expression (26.44) for A, we obtain, after simple calculations

$$\nabla_{\boldsymbol{k}}A = \frac{1}{\pi}\left\{\frac{\sigma^2 - \Gamma^2}{(\sigma^2+\Gamma^2)^2}\nabla_{\boldsymbol{k}}\Gamma - \frac{2\sigma\Gamma}{(\sigma^2+\Gamma^2)^2}\nabla_{\boldsymbol{k}}\sigma\right\},$$

and we have

$$I = \frac{if_\circ}{\sqrt{2\pi}\hbar}\frac{\partial\sigma}{\partial\omega}\frac{1}{\pi}\left\{\frac{\sigma^2 - \Gamma^2}{(\sigma^2+\Gamma^2)^2}\nabla_{\boldsymbol{k}}\Gamma - \frac{2\sigma\Gamma}{(\sigma^2+\Gamma^2)^2}\nabla_{\boldsymbol{k}}\sigma\right\}.$$

For the second term in (27.28) we need

$$\frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}\mathcal{R}(\Sigma^{\mathrm{r}}) = \nabla_{\boldsymbol{k}}\omega_s \quad \text{and} \quad \frac{\partial G^<}{\partial\omega} = \frac{\mathrm{i}}{\sqrt{2\pi}\hbar}\frac{\partial}{\partial\omega}(Af_\circ),$$

obtained from (26.40), and (24.56), respectively. With the expression (26.44) for A, we obtain, after simple calculations,

$$\frac{\partial A}{\partial\omega} = \frac{1}{\pi}\frac{\sigma^2 - \Gamma^2}{(\sigma^2+\Gamma^2)^2}\frac{\partial\Gamma}{\partial\omega} - \frac{1}{\pi}\frac{2\sigma\Gamma}{(\sigma^2+\Gamma^2)^2}\frac{\partial\sigma}{\partial\omega}. \tag{27.30}$$

With the above results, for the second term in (27.28) we obtain

$$\begin{aligned} II &\equiv \left(\boldsymbol{v}_{\boldsymbol{k}} + \frac{(2\pi)^{3/2}}{\hbar}\nabla_{\boldsymbol{k}}\mathcal{R}(\Sigma^{\mathrm{r}})\right)\frac{\partial}{\partial\omega}G^< \\ &= \frac{\mathrm{i}}{\sqrt{2\pi}\hbar}\left(\boldsymbol{v}_{\boldsymbol{k}} + \nabla_{\boldsymbol{k}}\omega_s\right)\left(\frac{f_\circ}{\pi}\frac{\sigma^2 - \Gamma^2}{(\sigma^2+\Gamma^2)^2}\frac{\partial\Gamma}{\partial\omega} - \frac{f_\circ}{\pi}\frac{2\sigma\Gamma}{(\sigma^2+\Gamma^2)^2}\frac{\partial\sigma}{\partial\omega} + A\frac{\partial f_\circ}{\partial\omega}\right). \end{aligned}$$

In a similar way, for the third term multiplying the electric field in (27.28) we use (26.41) and (26.46) for fermions, obtaining

$$III \equiv \nabla_{\boldsymbol{k}} \Sigma^{<} \frac{\partial}{\partial \omega} \mathcal{R}(G^{\mathrm{r}})$$
$$= \frac{2\mathrm{i}\hbar}{(2\pi)^{3/2}} f_{\circ} \nabla_{\boldsymbol{k}} \Gamma \frac{1}{(2\pi)^{3/2}\hbar} \left\{ \frac{\Gamma^2 - \sigma^2}{(\sigma^2+\Gamma^2)^2} \frac{\partial \sigma}{\partial \omega} - \frac{2\Gamma\sigma}{(\sigma^2+\Gamma^2)^2} \frac{\partial \Gamma}{\partial \omega} \right\}.$$

Remember that $\boldsymbol{k}$ and ω are independent variables, so that $f_{\circ}$ is constant with respect to $\boldsymbol{k}$. Finally, for the fourth term we use again (26.46) and (26.41) and obtain

$$IV \equiv \frac{\partial}{\partial \omega} \Sigma^{<} \nabla_{\boldsymbol{k}} \mathcal{R}(G^{\mathrm{r}})$$
$$= \frac{2\mathrm{i}\hbar}{(2\pi)^{3/2}} \left(f_{\circ} \frac{\partial \Gamma}{\partial \omega} + \frac{\partial f_{\circ}}{\partial \omega} \Gamma \right) \frac{1}{\hbar (2\pi)^{3/2}} \left\{ \frac{\Gamma^2 - \sigma^2}{(\sigma^2+\Gamma^2)^2} \nabla_{\boldsymbol{k}} \sigma - \frac{2\Gamma\sigma}{(\sigma^2+\Gamma^2)^2} \nabla_{\boldsymbol{k}} \Gamma \right\}.$$

At this point, we must insert the results obtained above into (27.28). In doing so, we separate different kinds of terms. First we consider the terms proportional to the derivative $\partial\sigma/\partial\omega$. They yield $q\boldsymbol{E}/\hbar$ multiplied by the sum

$$\frac{\mathrm{i} f_{\circ}}{\sqrt{2\pi}} \frac{1}{\pi\hbar} \left\{ \frac{\sigma^2 - \Gamma^2}{(\sigma^2+\Gamma^2)^2} \nabla_{\boldsymbol{k}} \Gamma - \frac{2\sigma\Gamma}{(\sigma^2+\Gamma^2)^2} \nabla_{\boldsymbol{k}} \sigma \right\} + \frac{\mathrm{i}}{\sqrt{2\pi}\hbar} \nabla_{\boldsymbol{k}} \sigma \frac{f_{\circ}}{\pi} \frac{2\sigma\Gamma}{(\sigma^2+\Gamma^2)^2}$$
$$+ \frac{(2\pi)^{3/2}}{\hbar} \frac{2\mathrm{i}\hbar}{(2\pi)^{3/2}} f_{\circ} \nabla_{\boldsymbol{k}} \Gamma \frac{1}{(2\pi)^{3/2}\hbar} \frac{\Gamma^2 - \sigma^2}{(\sigma^2+\Gamma^2)^2} = 0,$$

where we have also taken into account that, from (26.42), $\boldsymbol{v}_{\boldsymbol{k}} + \nabla_{\boldsymbol{k}} \omega_s = -\nabla_{\boldsymbol{k}} \sigma$.

Proceeding in the same way, we find that also the coefficients of the terms which multiply the derivative $\partial\Gamma/\partial\omega$ sum up to zero.

Finally, the terms proportional to the derivative $\partial f_{\circ}/\partial\omega$ yield the contribution

$$\mathrm{i} \frac{q\boldsymbol{E}}{\hbar^2} \cdot \sqrt{2\pi} A^2 \left\{ -\Gamma \nabla_{\boldsymbol{k}} \sigma + \sigma \nabla_{\boldsymbol{k}} \Gamma \right\} \frac{\partial f_{\circ}}{\partial \omega}.$$

The full QBE (27.28) then becomes [180]:

$$\frac{q\boldsymbol{E}}{\hbar} \cdot A^2 \left\{ \sigma \nabla_{\boldsymbol{k}} \Gamma - \Gamma \nabla_{\boldsymbol{k}} \sigma \right\} \frac{\partial f_{\circ}}{\partial \omega} = 2\pi \left\{ \Sigma^{<} G^{>} - G^{<} \Sigma^{>} \right\}. \tag{27.31}$$

Note that the r.h.s. is formed by the "correlated amplitudes of available states" $G^{>}$ multiplied by the effect of scattering $\Sigma^{<}$. This product takes the role that in the semiclassical BE is played by the "scattering in" integral. This is diminished by the product of the "correlated amplitudes of present electrons" $G^{<}$ multiplied by the effect of the scattering $\Sigma^{>}$, playing the role of the "scattering out". Thus, the above equation may be directly compared with its corresponding classical equation (11.2).

The r.h.s of (27.31) may be given an alternative form using (24.52) and (26.43):

$$\frac{q\boldsymbol{E}}{\hbar}\cdot A^2\left\{\sigma\nabla_{\boldsymbol{k}}\Gamma-\Gamma\nabla_{\boldsymbol{k}}\sigma\right\}\frac{\partial f_\circ}{\partial\omega}=\mathrm{i}\sqrt{2\pi}\left[\frac{\hbar}{\pi}G^<\Gamma-\frac{1}{\hbar}\Sigma^<A\right]. \tag{27.32}$$

In this chapter, we have seen how the GF approach deals with electron transport in a system where fields vary slowly in time and space and, in particular, for the linear response in homogeneous, steady-state situations. The technique has been applied also for nonlinear transport (see, e.g., [25, 184, 217, 218, 237, 284]). For reasons of space, this problem will not be treated here.

28

Nonequilibrium Green Functions Applied to Transport: Mesoscopic Systems

In this last chapter, the GF technique will be applied to the study of the electrical conduction of mesoscopic systems. Such a problem was already treated in Chap. 21, where Landauer theory established a relation between the conductance and the transmission coefficients of electron wavefunctions at the different leads connected to the system. Here, it will be shown how the GFs may be evaluated and used for the determination of the transmission coefficients. The method was introduced in the 1970s by several authors [94,95,305] and more recently extended by Datta in a number of papers and presented in an excellent book [116] that quickly became a standard reference for this subject. This chapter is totally based on that book.

The approach is very general and can be applied to different types of systems, including devices made of single molecules. Furthermore, the effect of scattering agents, such as phonons and impurities, can be included.

When we analyze the conductance of a system of mesoscopic dimensions, the structure of its electron states becomes dominant. We cannot consider any more the continuum of states of a homogeneous "infinite" system. We must take into account explicitly the distribution of the energy levels of the conductor. In particular, to have conduction, it is essential that available states exist at energies around the Fermi level of the two contacts, as indicated in Fig. 28.1.

In fact, assume that we keep the two metal contacts at different electrochemical potentials: ϵ_{FL} in the left contact and ϵ_{FR} in the right one. If energy levels exist in the mesoscopic conductor $\boldsymbol{S}$ between these two values, electrons from the left contact tends to occupy them, while the right contact tends to keep them empty. As net effect, electrons are transferred from the left contact to the conductor $\boldsymbol{S}$ and from this to the right contact with a consequent current flow. In the figure the leads are also shown, which are supposed to be quantum wires. In the metal contacts, the states are extremely dense, owing to the their dimensions. This is true also for the leads because, while the orthogonal states are quantized, as discussed in Sect. 19.3, the energies of the longitudinal states vary practically as a continuum.

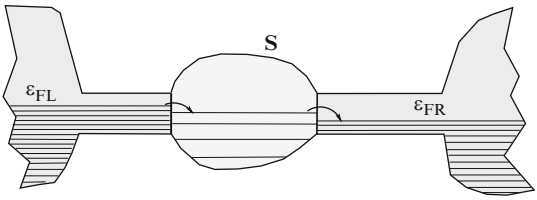


Fig. 28.1. Conduction in a mesoscopic system: electrons are injected into the conductor **S** from the source contact through a lead and reach the drain contact through a second lead. For this process, it is necessary that at least one electron state in the conductor is available between the electrochemical potentials of source and drain

To study the conductance, we must consider the Hamiltonian of the conductor $\boldsymbol{S}$, taking into account possible internal perturbations, such as phonons and impurities, and the presence of the contacts.

Following Datta [116], first a connection will be established between the GFs of the system and the scattering matrix S between the contacts. Then a method for the evaluation of the GFs will be presented, so that the S matrix may be obtained. From S, the transmission coefficients through the different leads, necessary for the determination of the conductance, will be obtained.

28.1 GFs for the Time-Independent Schrödinger Equation

Let us start from the equation of the retarded Green function $G^{\mathrm{r}}_{\epsilon}$ [1] for the time-independent Schrödinger equation:

$$[\epsilon - \mathcal{H} + \mathrm{i}\gamma]\, G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}') = \delta(\boldsymbol{r} - \boldsymbol{r}'). \tag{28.1}$$

As in Sect. 24.3, the infinitesimal term $\mathrm{i}\gamma$ is necessary to obtain convergence; its sign distinguishes between retarded and advanced GF.

A very general expression for $G^{\mathrm{r}}_{\epsilon}$ can be obtained in terms of eigenvalues and eigenfunctions of the Hamiltonian $\mathcal{H}$. In fact, let us expand G^{r}, as a function of $\boldsymbol{r}$, in eigenfunctions of $\mathcal{H}$:

$$G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}') = \sum_n A_n(\boldsymbol{r}')\phi_n(\boldsymbol{r}), \tag{28.2}$$

[1] The terms *retarded* and *advanced* are used also in this case, where no time dependence is considered. The reader will have no difficulties in realizing the analogies with the GFs of the time-dependent equation when it will be shown that the solutions of (28.1) are the energy eigenstates that, once the time dependence is added, leave the point of a δ excitation in $\boldsymbol{r}'$.

where $\mathcal{H}\phi_n(\boldsymbol{r}) = \epsilon_n \phi_n(\boldsymbol{r})$. Equation (28.1) becomes

$$[\epsilon - \mathcal{H} + \mathrm{i}\gamma] \sum_n A_n(\boldsymbol{r}')\phi_n(\boldsymbol{r}) = \sum_n A_n(\boldsymbol{r}') \left[\epsilon - \epsilon_n + \mathrm{i}\gamma\right] \phi_n(\boldsymbol{r}) = \delta(\boldsymbol{r} - \boldsymbol{r}').$$

If we now multiply by $\phi_m^*(\boldsymbol{r})$ and integrate in $\boldsymbol{r}$, using the orthonormality of the eigenfunctions, we obtain

$$A_m(\boldsymbol{r}') \left[\epsilon - \epsilon_m + \mathrm{i}\gamma\right] = \phi_m^*(\boldsymbol{r}').$$

Substituting the coefficients A_m obtained from this equation into the expression (28.2), we obtain the wanted result:

$$\boxed{G_\epsilon^{\mathrm{r}}(\boldsymbol{r}, \boldsymbol{r}') = \textstyle\sum_n \frac{\phi_n(\boldsymbol{r})\phi_n^*(\boldsymbol{r}')}{\epsilon - \epsilon_n + \mathrm{i}\gamma}} \tag{28.3}$$

In the limit of $\gamma \to 0$ this GF has a series of poles at energies equal to the eigenvalues. We already know from the previous chapters that if perturbations are present in the system, the convergence term $\mathrm{i}\gamma$ is substituted by the self-energy that brings information on the shifts of the eigenvalues and on the lifetimes of the states, and therefore on the broadening of their energies, due to the interactions.

With identical procedure we obtain the advanced G_ϵ^{a} which results to be the hermitian conjugate of G_ϵ^{r}.

To obtain a better insight on the physical meaning of $G_\epsilon^{\mathrm{r}}(\boldsymbol{r}, \boldsymbol{r}')$ let us multiply (28.3) by an arbitrary wavefunction $\psi(\boldsymbol{r}')$ and integrate, obtaining

$$\int G_\epsilon^{\mathrm{r}}(\boldsymbol{r}, \boldsymbol{r}')\psi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}' = \sum_n \frac{\phi_n(\boldsymbol{r})}{\epsilon - \epsilon_n + \mathrm{i}\gamma} \int \phi_n^*(\boldsymbol{r}')\psi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}' = \sum_n \frac{\phi_n(\boldsymbol{r})C_n}{\epsilon - \epsilon_n + \mathrm{i}\gamma},$$

where C_n is the coefficient which multiplies ϕ_n in the expansion of ψ. If we assume, for simplicity, a continuous, nondegenerate, spectrum of energy eigenvalues E, the above is written as

$$\int G_\epsilon^{\mathrm{r}}(\boldsymbol{r}, \boldsymbol{r}')\psi(\boldsymbol{r}')\mathrm{d}\boldsymbol{r}' = \int C(E) \frac{\phi_E(\boldsymbol{r})}{\epsilon - E + \mathrm{i}\gamma} g(E)\, \mathrm{d}E,$$

where $g(E)$ is the density of states. The integral can be evaluated in the complex E plane, yielding [2]

$$\int G_\epsilon(\boldsymbol{r}, \boldsymbol{r}')\psi(\boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}' = -2\pi\mathrm{i} C_\epsilon \phi_\epsilon(\boldsymbol{r}) g(\epsilon).$$

[2] In this simplified discussion devoted to understand the physical meaning of what is being done, no attention is paid to the positions of the poles. A more detailed calculation will be seen shortly for the case of interest to us.

Thus, we can say that $G^{\mathrm{r}}_{\epsilon}$, applied to any wavefunction ψ, yields the component of that wavefunction belonging to the energy ϵ, multiplied by the density of states at the same energy. If a $\delta(\boldsymbol{r}' - \boldsymbol{r}_1)$ is taken as $\psi(\boldsymbol{r}')$, we obtain

$$\int G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}')\delta(\boldsymbol{r}' - \boldsymbol{r}_1)\,\mathrm{d}\boldsymbol{r}' = G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}_1) = -2\pi\mathrm{i}C_{\epsilon}\phi_{\epsilon}(\boldsymbol{r})g(\epsilon).$$

Thus $G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}')$ is the component of energy ϵ of the wavefunction which is a δ in $\boldsymbol{r}'$, multiplied by the density of states in ϵ.

28.2 GF for a Perfect, Infinite, Two-Dimensional Wire

Often mesoscopic systems are fabricated starting from a two-dimensional electron gas. The third dimension, orthogonal to the quantum well, is extremely small, and it may be assumed that only the lowest energy level of the well is occupied by electrons. This dimension does not participate in the electron dynamics and may be neglected in this theory.

Let us start from the study of the GF for an isolated two-dimensional conducting wire with a confining potential $U(y)$ along the transverse dimension y, and without any applied potential along the longitudinal dimension x. Its length is assumed to be infinite, for the time being.

The time-independent Schrödinger equation is separable, and its y component is

$$\left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}y^2} + U(y)\right)\chi_n(y) = \epsilon_n\chi_n(y).$$

We know that the eigenfunctions $\chi_n(y)$ can be chosen to be real[3] and orthonormal. The eigenfunctions along x are plane waves, so that the total eigenfunctions are

$$\psi_{nk}(x, y) = \chi_n(y)\frac{1}{\sqrt{L}}\mathrm{e}^{\mathrm{i}kx}, \tag{28.4}$$

where a fictitious length L of the wire is assumed which will disappear in the final results, and the total energy eigenvalues are

$$\epsilon_{nk} = \epsilon_n + \frac{\hbar^2k^2}{2m}. \tag{28.5}$$

Now, using the general expression (28.3), we obtain

$$G^{\mathrm{r}}_{\epsilon}(x, y, x', y') = \frac{1}{L}\sum_{kn}\frac{\chi_n(y)\chi_n(y')\mathrm{e}^{\mathrm{i}k(x-x')}}{\epsilon - \epsilon_n - \hbar^2k^2/2m + \mathrm{i}\gamma}.$$

[3] The coefficients of the equation are real, and if a complex function is a solution, its real and imaginary parts are separately solutions.

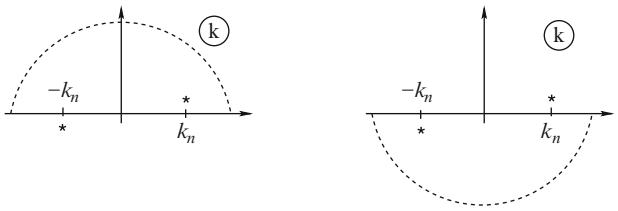


Fig. 28.2. Integration contours in the complex plane for the evaluation of $G^{\mathrm{r}}_{\epsilon}$

We may replace the sum over k with an integral with the density of states $L/2\pi$ (which cancels the fictitious wire length L):

$$G^{\mathrm{r}}_{\epsilon} = -\frac{1}{2\pi}\frac{1}{\hbar^2/2m}\sum_n \chi_n(y)\chi_n(y')\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}k(x-x')}}{k^2-(2m/\hbar^2)(\epsilon-\epsilon_n+\mathrm{i}\gamma)}\,\mathrm{d}k.$$

This integral can be evaluated using integration in the complex plane. The integrand has two poles in the values of k where the denominator vanishes:

$$k = \pm k_n\,(1+\mathrm{i}\delta/2)\,, \quad k_n = \frac{\sqrt{2m(\epsilon-\epsilon_n)}}{\hbar}, \quad \delta = \frac{\gamma}{(\epsilon-\epsilon_n)}, \tag{28.6}$$

where it has been taken into account that γ is infinitesimal. For $x > x'$, we may close the integration path on the upper complex half-plane as shown in Fig. 28.2, while for $x < x'$ the contour must be closed on the lower half-plane, and the residue theorem yields, after simple calculations,

$$G^{\mathrm{r}}_{\epsilon}(x,y,x',y') = -\frac{\mathrm{i}}{\hbar}\sum_n \frac{1}{v_n}\chi_n(y)\chi_n(y')\mathrm{e}^{\mathrm{i}k_n|x-x'|}, \tag{28.7}$$

where $v_n = \hbar k_n/m$. From the discussion at the end of the previous section and (28.4), the above (28.7) tells us that for $k = k_n$ the amplitude of the eigenfunction $\psi_{nk}(x,y)$ generated by the excitation in (x',y'), including the density of states, is, for $x > x'$,

$$-\frac{\mathrm{i}\sqrt{L}}{\hbar}\frac{1}{v_n}\chi_n(y')\mathrm{e}^{-\mathrm{i}k_n x'}. \tag{28.8}$$

28.3 From Green Function to S Matrix

Now we consider the conductor $\boldsymbol{S}$ connected to the leads and assume that the directions of the x coordinates along the leads are oriented toward the outgoing directions, as shown in Fig. 28.3. The general energy eigenfunctions

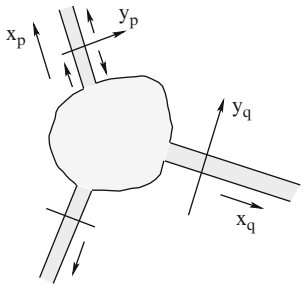


Fig. 28.3. Scattering states in a conductor with several leads

of the total system, conductor plus leads, contain incoming and outgoing terms in each lead. It is convenient, for our purposes, to consider the basis of the *scattering states*, seen in Sect. 21.1, formed by waves incoming and reflected in one lead, and outgoing in all other leads. A transverse state in a given lead is called a *channel*.

Let us consider the scattering state $\psi^{(p)}_{nk}$ entering from the channel n in lead p with longitudinal wavevector k. If (x_p, y_p) is a point in the lead p and (x_q, y_q) a point in the lead q, which may also be the lead p itself, this wavefunction is $(k_n > 0)$,

$$\psi^{(p)}_{nk_n}(x_q, y_q) = \chi_n(y_q)\frac{1}{\sqrt{L}}\mathrm{e}^{-\mathrm{i}k_n x_q}\delta_{pq} + \sum_m s'_{mn}\chi_m(y_q)\frac{1}{\sqrt{L}}\mathrm{e}^{\mathrm{i}k_m x_q},$$

where the first term is the incoming wave, present only in the lead p, and the second term represents the reflected wave in the lead p and the transmitted wave in the other leads. The above equation may be considered as the definition of the matrix s' (see Sect. 21.1), which connects the amplitudes of the wave in the incoming channel n in lead p to that of the channel m in lead q. The value of k_m, is determined by the energy of the eigenstate, given by (28.5).

Our present purpose is to find an expression of the GF connecting two points in two different leads, say lead p and lead q, in terms of the matrix s'. It was just seen at the end of Sect. 28.1 that $G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}')$ is the component of energy ϵ of the wavefunction which is a δ in $\boldsymbol{r}'$, multiplied by the density of states in ϵ. Then, in the previous section, we found that in a wire the amplitude of the eigenfunction $\psi_{nk}(x, y)$, including the density of states, generated by the excitation in (x', y') is given by (28.8). Thus, we expect that a δ excitation in (x'_p, y'_p) in lead p corresponds to an outgoing wave with the amplitude (28.8) plus a scattering state with the same amplitude. The retarded GF connecting the point (x'_p, y'_p) in lead p with the point (x_q, y_q) in any lead q (including p) is given by

$$G^{\mathrm{r}}_{\epsilon}(x_q, y_q, x'_p, y'_p) = -\frac{i\sqrt{L}}{\hbar}\sum_n \frac{1}{v_n}\chi_n(y'_p)\mathrm{e}^{-\mathrm{i}k_n x'_p} \times \left[\chi_n(y_q)\frac{1}{\sqrt{L}}\mathrm{e}^{\mathrm{i}k_n x_q}\delta_{pq} + \sum_m s'_{mn}\chi_m(y_q)\frac{1}{\sqrt{L}}\mathrm{e}^{\mathrm{i}k_m x_q}\right]. \tag{28.9}$$

Here we have eliminated the entering part of the scattering state since for $q = p$ we assume $x_q > x'_p$. The first term is the outgoing wave generated directly from the excitation; the second term contains the wave reflected by the conductor for $q = p$ and the transmitted waves for $q \neq p$.

In discussing Landauer theory of conductance (see Sect. 21.1), we saw that it is convenient to consider the matrix s, whose unitarity is consequence of current conservation, given by $s_{mn} = s'_{mn}\sqrt{v_m/v_n}$ With this definition, our GF (28.9) becomes

$$-\frac{\mathrm{i}}{\hbar}\sum_{nm}\left[\mathrm{e}^{\mathrm{i}k_n(x_q - x'_p)}\delta_{mn} + s_{mn}\mathrm{e}^{\mathrm{i}k_m x_q}\mathrm{e}^{\mathrm{i}k_n x'_p}\right]\frac{1}{\sqrt{v_n v_m}}\chi_m(y_q)\chi_n(y'_p).$$

If, without lost of generality, we take the origin of the x coordinates at the positions where the GF is evaluated, our expression becomes

$$G^{\mathrm{r}}_{\epsilon}(x_q = 0, y_q, x'_p = 0, y'_p) = -\frac{\mathrm{i}}{\hbar}\sum_{nm}\left[\delta_{mn} + s_{mn}\right]\frac{1}{\sqrt{v_n v_m}}\chi_m(y_q)\chi_n(y'_p).$$

Using the orthonormality relations of the χ_m, we obtain

$$s_{mn} = -\delta_{mn} + \mathrm{i}\hbar\sqrt{v_m v_n}\iint dy_p dy_q \chi_n(y_p)\chi_m(y_q)G^{\mathrm{r}}_{\epsilon}(y_q, y_p), \tag{28.10}$$

where the notation has been simplified by putting $G^{\mathrm{r}}_{\epsilon}(x_q = 0, y_q, x_p = 0, y_p) = G^{\mathrm{r}}_{\epsilon}(y_q, y_p)$. This is the wanted relation that yields the scattering matrix in terms of the GF.

Now that we have found the connection between the scattering matrix and the GF of the conductor, we have to find a way to calculate the latter.

28.4 Finite-Difference Scheme for the Conductor GF

Following again Datta [116], to proceed, it is convenient to consider the numerical operations we must perform to solve our problem. Let $\mathcal{H}$ be the Hamiltonian of an electron in our two-dimensional conductor with a confining potential $U(\boldsymbol{r})$, neglecting, for the time being, the contacts:

$$\mathcal{H} = -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right] + U(x, y).$$

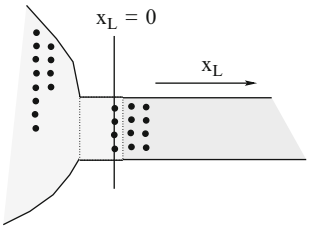


Fig. 28.4. Finite-difference grid inside the conductor and at the border between conductor and lead

For the actual computation, we may use a finite-difference scheme. Define a set of grid points[4] in our conductor, as shown in Fig. 28.4. For simplicity, we assume a square grid with spacing d. If all points in the grid are ordered and labeled with i, a function $f(x, y)$ is thus expressed as a list of complex numbers f_i, and the derivatives have the forms

$$\frac{\partial f}{\partial x} \approx \frac{1}{d}[f_i - f_j], \quad \frac{\partial^2 f}{\partial x^2} \approx \frac{1}{d^2}[f_i - 2f_j + f_k], \tag{28.11}$$

where i, j, and k are three consecutive grid point in the x direction. Similar expressions hold for the y direction.

Since inside the conductor we may have any potential, it is useful to extend the conductor to include the first part of the lead, so that outside the conductor the wavefunctions are those of perfect two-dimensional wires.

The Hamiltonian is now written as a matrix H_{ij}. The application of this matrix to the element j of the function has a two diagonal terms, one due to the potential, given by $U(\boldsymbol{r}_j) = U_j$, and a second due to the central term of the second derivative in (28.11), given by $2t$, where $t = \hbar^2/2md^2$. Nondiagonal terms come only from the lateral terms of the second derivative; they are present only for indices corresponding to adjacent grid points and have value $-t$. Symbolically, we may write the matrix of the Hamiltonian as

$$H_{ij} = \begin{bmatrix} U_1 + 2t & -t & 0 & 0 & \dots \\ -t & U_2 + 2t & 0 & 0 & \dots \\ 0 & -t & U_3 + 2t & 0 & \dots \\ 0 & 0 & -t & U_4 + 2t & \dots \\ \dots & & & & \end{bmatrix}. \tag{28.12}$$

The location of the off-diagonal terms in positions adjacent to the main diagonal is symbolic to remember that they are present only for adjacent points in the grid. Furthermore, to consider the Hamiltonian matrix only for points inside the conductor is equivalent to assume that outside the wavefunction is

[4] Sometimes the points of the grid are actual physical components of the conductor, such as atoms. In such a case, the method is considered a tight-binding model.

zero. This is not a restrictive condition, since the boundary can be taken far enough, where the wavefunction is vanishing small.

At this point also $G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}, \boldsymbol{r}')$ becomes a matrix, whose indices indicate the points in the grid for $\boldsymbol{r}$ and $\boldsymbol{r}'$, and from (28.1), we may write:[5]

$$G_{ij} = [(\epsilon + \mathrm{i}\gamma)I - \mathcal{H}]^{-1}\mathcal{I}, \tag{28.13}$$

where the matrix element G^{r}_{ij} stands for $G^{\mathrm{r}}_{\epsilon}(\boldsymbol{r}_i, \boldsymbol{r}_j)$.

28.5 The Effect of the Leads

Self-Energy

Now is time to take care of the effects of the leads. Let us start with only one contact. In the finite-difference scheme, we must add points inside the lead. Some terms of the Hamiltonian matrix H_{ij} will connect pairs of points inside the conductor (already considered); other terms $H_{\alpha\beta}$ will connect pairs of points inside the lead; finally, other terms $H_{i\alpha}(= H^*_{\alpha i})$ will connect points inside the conductor with points in the lead.

If the indices are ordered in such a way that all points in the lead have smaller indices than those in the conductor, the matrix to be inverted to obtain the Green matrix in (28.13), takes the form

$$\begin{bmatrix} (\epsilon + \mathrm{i}\gamma)\delta_{\alpha\beta} - H_{\alpha\beta} & H_{\alpha i} \\ H_{i\alpha} & (\epsilon + \mathrm{i}\gamma)\delta_{ij} - H_{ij} \end{bmatrix}. \tag{28.14}$$

Similarly, the resulting Green matrix connects points inside the lead, points inside the conductor, and points in the lead with points in the conductor. We may then write

$$\begin{bmatrix} G^{\mathrm{r}}_{\alpha\beta} & G^{\mathrm{r}}_{\alpha i} \\ G^{\mathrm{r}}_{i\alpha} & G^{\mathrm{r}}_{ij} \end{bmatrix} = \begin{bmatrix} (\epsilon + \mathrm{i}\gamma)\delta_{\alpha\beta} - H_{\alpha\beta} & H_{\alpha i} \\ H_{i\alpha} & (\epsilon + \mathrm{i}\gamma)\delta_{ij} - H_{ij} \end{bmatrix}^{-1} \mathcal{I}. \tag{28.15}$$

The product of the two matrices must give the unit matrix:

$$\begin{bmatrix} (\epsilon + \mathrm{i}\gamma)\delta_{\alpha\beta} - H_{\alpha\beta} & H_{\alpha j} \\ H_{i\beta} & (\epsilon + \mathrm{i}\gamma)\delta_{ij} - H_{ij} \end{bmatrix} \begin{bmatrix} G^{\mathrm{r}}_{\beta\gamma} & G^{\mathrm{r}}_{\beta k} \\ G^{\mathrm{r}}_{j\gamma} & G^{\mathrm{r}}_{jk} \end{bmatrix} = \mathcal{I}. \tag{28.16}$$

With the convention of the sum over repeated indices, we may write this equation in a more explicit form as the set of equations:

[5] In the Green equation (28.1), the δ function on the r.h.s. has dimension of the inverse of the volume under consideration (a surface for our quantum wire). Thus, the GF has dimensions of $(energy \times surface)^{-1}$. If, in the discretized problem, we substitute the Dirac δ with the Kroneker δ, the dimensions are lost. On the contrary, we cannot substitute for $\boldsymbol{r} = \boldsymbol{r}'$ the value of the diverging Dirac δ, but rather its mean value in the grid surface, given by d^{-2}. For this reason, we use here the symbol $\mathcal{I}$ equal to the unit matrix divided by d^2.

$$[(\epsilon + \mathrm{i}\gamma)\delta_{\alpha\beta} - H_{\alpha\beta}]G^{\mathrm{r}}_{\beta\gamma} + H_{\alpha j}G^{\mathrm{r}}_{\mathrm{i}\gamma} = \delta_{\alpha\gamma}/d^2, \tag{28.17}$$

$$[(\epsilon + \mathrm{i}\gamma)\delta_{\alpha\beta} - H_{\alpha\beta}]G^{\mathrm{r}}_{\beta k} + H_{\alpha j}G^{\mathrm{r}}_{jk} = 0, \tag{28.18}$$

$$H_{i\beta}G^{\mathrm{r}}_{\beta\gamma} + [(\epsilon + \mathrm{i}\gamma)\delta_{ij} - H_{ij}]G^{\mathrm{r}}_{j\gamma} = 0, \tag{28.19}$$

$$H_{i\beta}G^{\mathrm{r}}_{\beta k} + [(\epsilon + \mathrm{i}\gamma)\delta_{ij} - H_{ij}]G^{\mathrm{r}}_{jk} = \delta_{ik}/d^2. \tag{28.20}$$

At this point, it is useful to give short names to the above matrices:

$$G^{\mathrm{r}}_{ij} \to G^{\mathrm{r}}_{\mathrm{C}}, \quad H_{ij} \to H_{\mathrm{C}}, \quad G^{\mathrm{r}}_{\alpha\beta} \to G^{\mathrm{r}}_{\mathrm{L}}, \quad H_{\alpha\beta} \to H_{\mathrm{L}},$$

$$G^{\mathrm{r}}_{i\alpha} \to G^{\mathrm{r}}_{\mathrm{CL}}, \quad H_{i\alpha} \to H_{\mathrm{CL}}, \quad G^{\mathrm{r}}_{\alpha i} \to G^{\mathrm{r}}_{\mathrm{LC}}, \quad H_{\alpha i} \to H_{\mathrm{LC}},$$

and we know that

$$G^{\mathrm{r}}_{\mathrm{L}} = [(\epsilon + \mathrm{i}\gamma)\mathcal{I} - H_{\mathrm{L}}]^{-1}\mathcal{I} \tag{28.21}$$

is the retarded GF of the semi-infinite isolated lead. Applying this matrix to (28.18), we then obtain

$$\mathcal{I}G^{\mathrm{r}}_{\mathrm{LC}} = -G^{\mathrm{r}}_{\mathrm{L}}H_{\mathrm{LC}}G^{\mathrm{r}}_{\mathrm{C}}. \tag{28.22}$$

Substituting (28.22) into (28.20), we obtain

$$[(\epsilon + \mathrm{i}\gamma) - H_{\mathrm{C}} - H_{\mathrm{CL}}G^{\mathrm{r}}_{\mathrm{L}}H_{\mathrm{LC}}d^2]G^{\mathrm{r}}_{\mathrm{C}} = \mathcal{I}. \tag{28.23}$$

This equation can then be written as

$$G^{\mathrm{r}}_{\mathrm{C}} = [(\epsilon + \mathrm{i}\gamma) - H_C - H_{\mathrm{CL}}G^{\mathrm{r}}_{\mathrm{L}}H_{\mathrm{LC}}d^2]^{-1}\mathcal{I}.$$

This is the retarded GF of the conductor in presence of the lead. The last term in the square brackets represent the effect of the lead on the conductor, and, similarly to what is being done for the effect of the scattering, it is defined as retarded self-energy Σ^{r} due to the lead. If several leads are present and we may assume that their effects are independent, the corresponding self-energies may be summed:

$$\Sigma^{\mathrm{r}} = \sum_L H_{\mathrm{CL}}G^{\mathrm{r}}_{\mathrm{L}}H_{\mathrm{LC}}d^2. \tag{28.24}$$

In this way, the result just obtained can be written in the familiar form

$$\boxed{G^{\mathrm{r}}_{\mathrm{C}} = \mathcal{I}[\epsilon - H_{\mathrm{C}} - \Sigma^{\mathrm{r}}]^{-1}} \tag{28.25}$$

where the presence of the self-energy has made unnecessary the convergence factor γ.

To proceed we must find, now, explicit expressions for the factors that appear in the self-energy (28.24), i.e., the retarded GF of the semi-infinite lead $G^{\mathrm{r}}_{\mathrm{L}}$, and the coupling Hamiltonian G_{CL} (G_{LC} is its hermitian conjugate). Let us start from the latter point.

The Coupling Hamiltonian

Let us assume that the coupling between the leads and the conductor is present only in adjacent points (see Fig. 28.4). Assume also, for simplicity, that the coupling is only due to kinetic term.[6] Thus, with the same considerations made to obtain the Hamiltonian inside the conductors, we have that for each pair of adjacent points at the border between the conductor and the lead $H_{i\alpha_i} = t$, while the other elements of H_{CL} are zero. The self-energy in (28.24) becomes

$$\Sigma^{\mathrm{r}}_{ij} = \sum_L t^2 (G^{\mathrm{r}}_{\mathrm{L}})_{\alpha_i \beta_j} d^2, \tag{28.26}$$

where α_i and β_j are the indices of the grid point of the lead adjacent to the point i and j, respectively, of the conductor.

The Lead GF

The evaluation of the retarded GF of the semi-infinite lead requires more work. Owing to the expression (28.26) just found, we need only the matrix elements connecting points in the lead adjacent to the conductor. These values, however, must be found in analytical terms, since the wire must be considered of infinite length. The Hamiltonian H_{L} that appears in the expression (28.21) of the lead self-energy is that of an isolated semi-infinite two-dimensional wire, as indicated in Fig. 28.4. We may solve the Schrödinger equation of the wire with a separable Hamiltonian in the two directions, subject to the condition that the wavefunction is zero at $x = 0$. The eigenfunctions are then given by

$$\phi_{kn}(x, y) = \sqrt{\frac{2}{L}} \sin(kx) \chi_n(y),$$

again normalized to the fictitious length L of the wire. Here, as before, $\chi_n(y)$ is the n-th eigenfunction of the transverse confining potential, and ϵ_n the corresponding eigenvalue. The total energy eigenvalues are given in (28.5). Now, using the general expression (28.3) for the retarded GF, we obtain for our wire

$$G^{\mathrm{r}}_{\mathrm{L}}(\boldsymbol{r}, \boldsymbol{r}') = \frac{2}{L} \sum_{kn} \frac{\sin(kx)\chi_n(y) \sin(kx')\chi_n(y')}{\epsilon - \epsilon_n - (\hbar^2 k^2/2m) + \mathrm{i}\gamma}.$$

The values of interest are those with $x = x'$ (points adjacent to the conductor). Furthermore, we may convert the sum over k into an integral:

$$G^{\mathrm{r}}_{\mathrm{L}}(x, y, x, y') = \frac{L}{\pi} \frac{2}{L} \int_0^\infty \mathrm{d}k \sum_n \frac{\sin^2(kx)\chi_n(y)\chi_n(y')}{\epsilon - \epsilon_n - (\hbar^2 k^2/2m) + \mathrm{i}\gamma}. \tag{28.27}$$

[6] This assumption, together with the general treatment of the subject, makes the present approach substantially equivalent to the "quantum transmitting boundary method" [278]. Remember that the boundary between conductor and lead has been put inside the wire.

Now we remember that $\sin(kx) = (\mathrm{e}^{\mathrm{i}kx} - \mathrm{e}^{-\mathrm{i}kx})/2\mathrm{i}$. Thus,

$$\sin^2(kx) = |\sin(kx)|^2 = \frac{1 + 1 - \mathrm{e}^{2\mathrm{i}kx} - \mathrm{e}^{-2\mathrm{i}kx}}{4},$$

and we may transform the integral in (28.27) as follows

$$\int_0^\infty \frac{\sin^2(kx)}{\epsilon - \epsilon_n - (\hbar^2 k^2/2m) + \mathrm{i}\gamma}\,\mathrm{d}k = \frac{1}{4}\int_0^\infty \frac{(1 - \mathrm{e}^{2\mathrm{i}kx}) + (1 - \mathrm{e}^{-2\mathrm{i}kx})}{\epsilon - \epsilon_n - (\hbar^2 k^2/2m) + \mathrm{i}\gamma}\,\mathrm{d}k.$$

The second integrand function becomes equal to the first exchanging k into $-k$ so that the integral can be written as

$$\frac{1}{4\hbar^2/2m}\int_{-\infty}^{\infty} \frac{1 - \mathrm{e}^{2\mathrm{i}kx}}{(2m/\hbar^2)(\epsilon - \epsilon_n + \mathrm{i}\gamma) - k^2}\,\mathrm{d}k. \tag{28.28}$$

This integral can be evaluated using integration in the complex plane. The integrand has the two poles given in (28.6). This time we have only $x > 0$ and we close the integration path on the upper complex half-plane. The result of the application of the residue theorem is, after simple calculations,

$$-\frac{m\pi\mathrm{i}}{\hbar^2 2k_n}\mathrm{e}^{\mathrm{i}k_n x}[\mathrm{e}^{-\mathrm{i}k_n x} - \mathrm{e}^{\mathrm{i}k_n x}] = -\frac{\pi}{\hbar v_n}\mathrm{e}^{\mathrm{i}k_m x}\sin(k_n x).$$

Inserting this result in (28.27) for the GF, we obtain

$$G_{\mathrm{L}}^{\mathrm{r}}(x, y, x, y') = -\frac{2}{\hbar}\sum_n \frac{1}{v_n}\mathrm{e}^{\mathrm{i}k_n x}\sin(k_n x)\chi_n(y)\chi_n(y').$$

We need this expression for the grid points adjacent to the conductor, i.e., for points at a distance d from the edge of the lead. In fact, the line of the grid points where the lead wavefunction has been assumed zero ($x = 0$) is the line of the last points of the conductor, at distance d from the first line where the same wavefunction is nonzero. Thus, the last equation must be evaluated at $x = d$. For a sufficiently small d, $\sin(k_n d) \approx k_n d$. Remembering also that $t = \hbar^2/2md^2$, we may write

$$G_{\mathrm{L}}^{\mathrm{r}}(d, y, d, y') = -\frac{1}{td}\sum_n \mathrm{e}^{\mathrm{i}k_n d}\chi_n(y)\chi_n(y'). \tag{28.29}$$

At this point, we are in the condition to complete the calculation of the self-energy for the conductor. Inserting (28.29) into (28.26), we obtain, for a given lead, to be summed over the various leads,

$$(\Sigma_L^r)_{ij} = -td\sum_n \mathrm{e}^{\mathrm{i}k_n d}\chi_n(\alpha_i)\chi_n(\alpha_j). \tag{28.30}$$

This, summed over the various leads and inserted into (28.25), allows us to evaluate numerically the GF $G_{\mathrm{C}}^{\mathrm{r}}$ by means of matrix inversion.

28.6 Conductance

We have now all the necessary elements to evaluate the transmission coefficients T_{mn} and therefore the conductance of our mesoscopic conductor. From Landauer–Büttiker theory, we know that $T_{mn} = |s_{mn}|^2$. If we use the result in (28.10), expressing the integrals in the finite-difference scheme, for $n \neq m$:

$$T_{mn} = |s_{nm}|^2 = \hbar^2 v_n v_m \sum_{i,j} \chi_n(y_i)\chi_m(y_j) G^{\mathrm{r}}_{\epsilon,ji} d^2 \sum_{i',j'} \chi_n(y_{i'})\chi_m(y_{j'}) G^{r*}_{\epsilon,j'i'} d^2.$$

From this, using $G^a_{i,j} = G^{r*}_{j,i}$, defining

$$\Gamma_p(i,i') = \sum_{n\in p} \chi_n(y_i)\hbar v_n \chi_n(y_{i'}), \tag{28.31}$$

and remembering (21.14), we obtain

$$\overline{T}_{qp} = \sum_{n\in p, m\in q} T_{mn} = \sum_{i,j,j',i'} \Gamma_q(j',j) G^{\mathrm{r}}_{\epsilon,ji} \Gamma_p(i,i') G^a_{\epsilon,i'j'} d^4,$$

or

$$\overline{T}_{qp} = Tr\left[\Gamma_q G^{\mathrm{r}}_\epsilon \Gamma_p G^a_\epsilon\right]. \tag{28.32}$$

With this result for the transmission function $\overline{T}_{qp}$, the current is obtained by means of (21.15) of Landauer–Büttiker theory.

As it regards the coefficients $\Gamma_p(i,i')$ defined in (28.31), they have a simple relation with the self-energies which clarifies their physical meaning. In fact, from (28.30),

$$(\Sigma_p{}^r)_{ij} = -td \sum_{n\in p} \mathrm{e}^{\mathrm{i}k_n d} \chi_n(y_i)\chi_n(y_j),$$

and therefore

$$(\Sigma_p{}^a)_{ij} = (\Sigma_p{}^r)^*_{ji} = -td \sum_{n\in p} \mathrm{e}^{-\mathrm{i}k_n d} \chi_n(y_j)\chi_n(y_i).$$

We thus obtain

$$(\Sigma_p{}^r)_{ij} - (\Sigma_p{}^a)_{ij} = -td \sum_{n\in p} 2\mathrm{i}\sin(k_n d)\chi_n(y_i)\chi_n(y_j).$$

Remembering that $t = \hbar^2/2md^2$, for small d this becomes

$$(\Sigma_p{}^r)_{ij} - (\Sigma_p{}^a)_{ij} = -\mathrm{i} \sum_{n\in p} \chi_n(y_i)\hbar v_n \chi_n(y_j).$$

Comparing this with the definition of Γ_p in (28.31) above, we obtain

$$\Gamma_p = \mathrm{i}[\Sigma_p{}^r - \Sigma_p{}^a]. \tag{28.33}$$

The coefficients Γ are therefore the imaginary part of the self-energies, i.e., the electron lifetimes, as discussed in Chap. 26 (cf. (26.43)).

Scattering Mechanisms

In the previous pages, we have studied the effect of the leads on the electron dynamics inside a conductor. This effect has been described by a self-energy in the equation for the electron Green functions. A part from the effects of the leads, the electron dynamics has been considered as described by the conductor Schrödinger equation. In other words, we have assumed a coherent propagation of the electrons from one lead to the other, without scattering agents inside the conductor. Such scatterings, however, are unavoidable. Even if, in principle, impurity scattering may be included in the potential profile inside the conductor, phonon scattering and electron–electron interactions must be accounted for in a correct treatment of conduction in mesoscopic structures. In most cases, the different sources of scattering are considered independent from each other.

Self-energies due to these interactions can be included in the calculation of the GFs, and therefore on the conductance, as described in the previous chapters. Details can be found in Datta's book [116] and in the specialized literature.

In these last chapters, the basic principles of the Green-function approach to electron transport have been presented. We tried to concentrate on the most fundamental equations and on the physical ideas behind them, reducing, whenever possible, too cumbersome calculations. We hope that, having gone through these pages, the reader may be able to approach more detailed literature without the embarrassment that often arises in front of the variety and abstractness of the mathematical symbols.

Part VI

Appendices

A

Vector Spaces and Fourier Analysis

Hilbert Spaces

The basic elements of the theory of abstract vector spaces will be summarized here. For a more complete treatment, we refer the reader to textbooks of linear algebra, or to the classical text of Von Neumann [464] for a very rigorous treatise of the mathematical foundations of quantum mechanics. Messiah [306] gives a good summary of the theory of vector spaces as applied to quantum mechanics.

A vector space is a set of elements, *vectors*, indicated here with the Dirac symbol *ket* [120]:

$$|v\rangle$$

on which the sum and the scalar multiplication are defined:

$$|v\rangle + |w\rangle = |s\rangle, \qquad |v'\rangle = \lambda|v\rangle,$$

where λ is a complex number. An expression like

$$|w\rangle = \lambda_1|v_1\rangle + \lambda_2|v_2\rangle + \cdots$$

is called a *linear combination* of the vectors $|v_1\rangle$, $|v_2\rangle, \ldots$. A number of vectors are said to be *linearly independent* if none of them can be written as a linear combination of the other ones. A vector space is said to have *dimensionality* n is there exist at most n linearly independent vectors. If such finite n does not exist, the space is said to have infinite dimensions.

A *scalar product* is defined for each ordered pair of vectors, $|u\rangle$ and $|v\rangle$, indicated by

$$\langle u|v\rangle.$$

The scalar product is a complex number and has the following properties:

$$\langle u|v\rangle = \langle v|u\rangle^*, \qquad \langle u|u\rangle \geq 0 \quad (= 0 \text{ only if } \; |u\rangle = 0), \tag{A.1}$$

$$\langle u| \{\lambda_1|v_1\rangle + \lambda_2|v_2\rangle\} = \lambda_1\langle u|v_1\rangle + \lambda_2\langle u|v_2\rangle.$$

A vector $\langle u|$ appearing on the left of a scalar product is called a *bra*. Two vectors $\langle u|$ and $\langle v|$ are said to be orthogonal if their scalar product is zero. The square root of the scalar product

$$|u| = \sqrt{\langle u|u\rangle}$$

is called the *norm* of the vector $|u\rangle$. The norm of a vector is zero only for the null vector. The *Schwartz inequality* states that

$$|u|\,|v| \geq |\langle u|v\rangle|.$$

The equal sign holds only if the two vectors are proportional, or *parallel*, i.e., if one of them is given by the other times a complex number.

A vector space is said to be *complete* if any limit of a sequence of vectors is still a vector of the space; it is said to be *separable* if there is a sequence of vectors everywhere dense in the space, i.e., a sequence that gets as close as we want to any vector of the space. A vector space which is complete and separable is called a *Hilbert space*.

Linear Operators

An operator $\mathcal{A}$ in a vector space is a law that associates a precise new vector to each vector of the space:

$$\mathcal{A}|u\rangle = |v\rangle.$$

An operator is said to be linear if it preserves the linear combinations:

$$\mathcal{A}\{\lambda_1|v_1\rangle + \lambda_2|v_2\rangle\} = \lambda_1\mathcal{A}|v_1\rangle + \lambda_2\mathcal{A}|v_2\rangle.$$

Sum and product of linear operators are defined as follows:

$$(\mathcal{A}+\mathcal{B})|u\rangle = \mathcal{A}|u\rangle + \mathcal{B}|u\rangle, \qquad (\mathcal{A}\mathcal{B})\,|u\rangle = \mathcal{A}(\mathcal{B}|u\rangle).$$

In general, the product is not commutative. The *commutator* of two operators $\mathcal{A}$ and $\mathcal{B}$ is defined as

$$[\mathcal{A},\mathcal{B}] = [\mathcal{A},\mathcal{B}]_- = \mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}.$$

Commutators play a fundamental role in the formulation and theoretical elaboration of quantum mechanics. The minus sign as suffix is not usually written. It is explicitly indicated when it is necessary to distinguish the commutator

from the *anticommutator*:

$$[\mathcal{A}, \mathcal{B}]_+ = \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A}.$$

Two linear operators $\mathcal{A}$ and $\mathcal{A}^\dagger$ are said to be *Hermitian conjugate* if, for any pairs of vectors $|u\rangle$ and $|u\rangle$,

$$\langle u|\mathcal{A}|v\rangle = \langle v|\mathcal{A}^\dagger|u\rangle^*. \tag{A.2}$$

The Hermitian conjugate of a given linear operator is unique. The Hermitian conjugate of the product of two operators is the product of the Hermitian conjugates in reverse order:

$$(\mathcal{A}\mathcal{B})^\dagger = \mathcal{B}^\dagger\mathcal{A}^\dagger.$$

A linear operator $\mathcal{H}$ is said to be *Hermitian* if it coincides with its Hermitian conjugate. It is said to be anti-Hermitian if its Hermitian conjugate is equal to its opposite. Any linear operator $\mathcal{A}$ can always be decomposed into a Hermitian part and an anti-Hermitian part,

$$\mathcal{A} = \frac{1}{2}\left(\mathcal{A} + \mathcal{A}^\dagger\right) + \frac{1}{2}\left(\mathcal{A} - \mathcal{A}^\dagger\right),$$

just as any complex number can be decomposed into a real and an imaginary part.

The inverse $\mathcal{A}^{-1}$ of an operator $\mathcal{A}$ is defined by

$$\mathcal{A}^{-1}\mathcal{A} = \mathcal{A}\mathcal{A}^{-1} = \mathcal{I},$$

where $\mathcal{I}$ is the identity operator. When two operators $\mathcal{A}$ and $\mathcal{B}$ posses inverse, then

$$(\mathcal{A}\mathcal{B})^{-1} = \mathcal{B}^{-1}\mathcal{A}^{-1}.$$

An operator $\mathcal{U}$ is said to be *unitary* if its inverse coincides with its Hermitian conjugate:

$$\mathcal{U}^\dagger = \mathcal{U}^{-1}.$$

It is immediate to verify that a unitary operator preserves the scalar product of any pair of vectors:

$$\text{if } |u'\rangle = \mathcal{U}|u\rangle \text{ and } |v'\rangle = \mathcal{U}|v\rangle, \quad \text{then } \langle u'|v'\rangle = \langle u|v\rangle,$$

and that the product of two unitary operators is unitary:

$$(\mathcal{U}\mathcal{V})^\dagger = \mathcal{V}^\dagger\mathcal{U}^\dagger = \mathcal{V}^{-1}\mathcal{U}^{-1} = (\mathcal{U}\mathcal{V})^{-1}.$$

The *mean value* of an operator $\mathcal{A}$ in a state $|u\rangle$ is defined as

$$\langle A\rangle_u = \frac{\langle u|\mathcal{A}|u\rangle}{\langle u|u\rangle}.$$

From the definition (A.2) of Hermitian conjugate operators, it follows that the mean value of a Hermitian operator is real in any state.

Orthonormal Basis

An orthonormal basis of a vector space is a complete set of orthogonal vectors $|\varphi_i\rangle$ such that

$$\langle\varphi_i|\varphi_j\rangle = \delta_{ij}, \tag{A.3}$$

where δ_{ij} is the Kronecker delta. The fact that the set is complete means that any vector of the space can be given as a linear combination of vectors of the set:

$$|u\rangle = \sum_i u_i|\varphi_i\rangle, \qquad u_i = \langle\varphi_i|u\rangle. \tag{A.4}$$

The second expression above is obtained by a scalar multiplication of the first one by $\langle\varphi_j|$ and the use of (A.3). The numbers u_i form the *representation* of the vector $|u\rangle$ in the given basis. The representation of a linear operator is given by its *matrix elements*:

$$A_{ij} = \langle\varphi_i|\mathcal{A}|\varphi_j\rangle.$$

It is immediate to verify that

$$\text{if} \quad |v\rangle = \mathcal{A}|u\rangle, \quad \text{then} \quad v_i = \sum_j A_{ij}u_j.$$

The two expressions in (A.4) together yield

$$|u\rangle = \sum_i |\varphi_i\rangle\langle\varphi_i|u\rangle. \tag{A.5}$$

This form suggests a very powerful interpretation of the symbols. Given two vectors $|u\rangle$ and $|v\rangle$, an expression like

$$|u\rangle\langle v| \tag{A.6}$$

defines a linear operator $\mathcal{P}_{uv}$ that, applied to $|w\rangle$ yields, as graphically suggested, the scalar product of $|v\rangle$ and $|w\rangle$ times the vector $|u\rangle$:

$$\mathcal{P}_{uv}|w\rangle = |u\rangle\langle v|w\rangle.$$

If the vectors in (A.6) are two equal vectors $|\varphi\rangle$ of unit length, the corresponding operator applied to the vector $|w\rangle$ yields the projection of $|w\rangle$ on the axis defined by $|\varphi\rangle$.

$$\mathcal{P}_\varphi = |\varphi\rangle\langle\varphi| \tag{A.7}$$

is defined as the projection operator on the direction of $|\varphi\rangle$. Equation (A.5) assumes then the geometrical meaning of separating the vector $|u\rangle$ into the sum of its components along the different axes of the basis. If the basis is complete, the vector must be completely reconstructed. This means that the sum of the projection operators must be the identity:

$$\boxed{\textstyle\sum_i |\varphi_i\rangle\langle\varphi_i| = 1} \tag{A.8}$$

The above equation is the *completeness relation* of the basis and it is often called a *spectral decomposition of the identity*.

Unitary Transformations

A *unitary transformation* is a mathematical law that associates with each vector $|u\rangle$ and each linear operator $\mathcal{A}$ of a vector space a new vector $|u'\rangle$ and a new linear operator $\mathcal{A}'$ by means of a unitary operator $\mathcal{U}$ as follows:

$$|u'\rangle = \mathcal{U}|u\rangle, \qquad \mathcal{A}' = \mathcal{U}\mathcal{A}\mathcal{U}^{\dagger}.$$

A unitary operator preserves all scalar products (and therefore all vector norms) and preserves all linear operations, i.e.,

$$\text{if} \quad |v\rangle = \mathcal{A}|u\rangle, \quad \text{then} \quad |v'\rangle = \mathcal{A}'|u'\rangle.$$

Thus, the structure of the vector space is not modified, and a unitary transformation can be seen as a rotation of the whole vector space.

Eigenvalues and Eigenvectors

Given the linear operator $\mathcal{A}$, its *eigenvalue equation* is

$$\mathcal{A}|a\rangle = a|a\rangle. \tag{A.9}$$

Values of a that satisfy (A.9) are called *eigenvalues* of $\mathcal{A}$ and the vectors $|a\rangle$ are called *eigenvectors* belonging to the eigenvalue a. The application of a linear operator to one of its eigenvectors simply multiplies the vector by a constant, i.e., leaves the eigenvector parallel to itself. The mean value of an operator in one of its eigenvectors is given by the corresponding eigenvalue, thus *the eigenvalues of Hermitian operators are real*.

An eigenvalue is said to be *n-fold degenerate* if there are n independent eigenvectors belonging to it. A linear combination of eigenvectors belonging to a given eigenvalue is still an eigenvector belonging to the same eigenvalue. Thus, they can be substituted by n orthonormal vectors, for example by means of the Gram–Schmidt procedure [229].

A very important theorem, and very simple to prove, states that *eigenvectors of a Hermitian operator belonging to different eigenvalues are orthogonal*.

Observables

We have just seen that eigenvectors of a Hermitian operator belonging to the same eigenvalue may be chosen to be orthonormal, while when they belong to different eigenvalues are automatically orthogonal and can be easily normalized to unity, by multiplication of an appropriate constant. Thus, all eigenvectors of a linear Hermitian operator $\mathcal{H}$ are or can be chosen to be orthonormal. When such set of vectors form a complete basis $\mathcal{H}$ is said to be an *observable*; the set of vectors is called the basis of $\mathcal{H}$, and if vectors and operators are given in terms of their components and matrix elements in that basis we say that the representation of $\mathcal{H}$ is used.

A fundamental theorem of the theory of vector space, of extreme importance for its application in quantum physics states [306] that *if two observables commute, they posses a common complete set of eigenvectors.*

Now, if an eigenvalue a of an observable $\mathcal{A}$ is not degenerate, its knowledge is sufficient to identify the corresponding eigenvector $|a\rangle$. If, on the contrary, the eigenvalue is degenerate this is not true, and we may need the eigenvalue b of another observable, say $\mathcal{B}$, commuting with $\mathcal{A}$, to identify univocally the common eigenvector. If the two eigenvalues are still degenerate, as a pair, i.e., if there are more than one eigenvector belonging to the same pair of eigenvalues, a third commuting observable may be considered, and so on. A *complete set of commuting observables* is a set of commuting observables such that their eigenvalues univocally identify the corresponding eigenvector.

Square-Integrable Functions

A complex function ψ of the real variables $q_1, q_2, \ldots$ is said to be square-integrable if the integral

$$\int |\psi(q_1, q_2, \ldots)|^2 \mathrm{d}q_1 \mathrm{d}q_2 \ \ldots$$

exists finite. The set of square-integrable functions of a given domain form a functional realization of a vector Hilbert space. The linear combination of two such functions

$$\lambda_1 \psi_1(q_1, q_2, \ldots) + \lambda_2 \psi_2(q_1, q_2, \ldots)$$

is still a function of the same set. The scalar product of ψ_1 and ψ_2 is defined as the integral

$$\int \psi_1^*(q_1, q_2, \ldots) \psi_2(q_1, q_2, \ldots) \mathrm{d}q_1 \mathrm{d}q_2 \ \ldots,$$

which has the properties required for a scalar product. A linear operator may be, for example, the multiplication by a complex number or the differentiation. The mean value of the derivative, for example, in a function $\psi(x)$ of unit norm, is given by

$$\int f^*(x) \frac{\mathrm{d}}{\mathrm{d}x} f(x) \, \mathrm{d}x.$$

Of particular importance in quantum mechanics is, for example, the commutator of the multiplication by x and the differentiation with respect to x, given by

$$\left[x, \frac{\mathrm{d}}{\mathrm{d}x}\right] = -1.$$

In fact, applied to any function $f(x)$, yields

$$\left[x, \frac{\mathrm{d}}{\mathrm{d}x}\right] f(x) = x \frac{\mathrm{d}}{\mathrm{d}x} f(x) - \frac{\mathrm{d}}{\mathrm{d}x}[x f(x)] = x \frac{\mathrm{d}f}{\mathrm{d}x} - f - x \frac{\mathrm{d}f}{\mathrm{d}x} = -f \, .$$

Equation of Harmonic Motion

Let us now consider the linear operator "second derivative" defined in the space of square-integrable complex functions $f(x)$ of a real variable x defined in the interval $[-a/2, a/2]$. We leave as an exercise for the reader to show that this operator is Hermitian if we restrict the space to the functions that either are zero at the extremes of the interval or have equal values and equal derivatives in the same points. Let us adopt the second type of boundary conditions:

$$f\left(-\frac{a}{2}\right) = f\left(\frac{a}{2}\right), \qquad f'\left(-\frac{a}{2}\right) = f'\left(\frac{a}{2}\right). \tag{A.10}$$

The eigenvalue equation for such an operator is

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} f(x) = C f(x). \tag{A.11}$$

The solutions of this equation are real exponentials if C is positive, or imaginary exponentials

$$\mathrm{e}^{\mathrm{i}kx}, \qquad k = \pm\sqrt{-C},$$

if C is negative. It is easy to conclude that the boundary conditions in (A.10) cannot be satisfied if the exponentials are real. Thus, only negative eigenvalues exist for (A.11) with the boundary conditions (A.10). Equations (A.10) are satisfied for a discrete infinite set of *wavevectors* k:

$$k = k_n = \frac{2\pi}{a} n, \qquad n = 0,\ \pm 1,\ \pm 2,\ \ldots.$$

Except $k_\circ = 0$, that has only one eigenfunction, all other eigenvalues are doubly degenerate, since the eigenfunctions with $k = \pm k_n$ belong to the same eigenvalue $C_n = -k_n^2$. These eigenfunctions are

$$\varphi_n(x) = A_n \mathrm{e}^{\mathrm{i}k_n x}.$$

The normalization condition determines the constant A_n, to within a phase factor:

$$\int_{-a/2}^{a/2} |\varphi(x)|^2 \mathrm{d}x = 1 = |A_n|^2 a, \qquad A_n = \frac{1}{\sqrt{a}}, \qquad \varphi_n(x) = \frac{1}{\sqrt{a}} \mathrm{e}^{\mathrm{i}k_n x}. \tag{A.12}$$

Fourier Series

Let us consider a linear combination of the eigenfunctions just found:

$$F(x) = \sum_n c_n \varphi_n(x) = \sum_n c_n \frac{1}{\sqrt{a}} \mathrm{e}^{\mathrm{i}k_n x}, \tag{A.13}$$

or, in vector terms,

$$|F\rangle = \sum_n c_n |\varphi_n\rangle.$$

The coefficients c_n are given by

$$c_n = \langle \varphi_n | F \rangle = \frac{1}{\sqrt{a}} \int_{-a/2}^{a/2} \mathrm{e}^{-\mathrm{i}k_n x} F(x)\, \mathrm{d}x.$$

A fundamental theorem of Fourier analysis (see, for example, [14]) states that for any square-integrable function $F(x)$ the *Fourier series* given by the sum in (A.13) converges in the mean[1] to $F(x)$. Thus, according to the definition given above, the second derivative is an observable in the vector space of the square-integrable functions in the given interval with the given boundary conditions.

The Continuum Spectrum and the Dirac Delta Function

If the length a of the interval in which the functions φ_n in (A.12) are defined increases, the values of k_n become closer, and for $a \to \infty$ the Fourier series becomes an integral. However, each time the functions of a basis set depend on a continuous index a difficulty arises. An indication of this difficult is already found in the fact that the normalization constant A_n vanishes when a becomes infinite. As a confirmation, we note that the plane waves $\mathrm{e}^{\mathrm{i}kx}$ are not square-integrable in the entire real axis:

$$\int_{-\infty}^{+\infty} \left|\mathrm{e}^{\mathrm{i}kx}\right|^2 \mathrm{d}x = \int_{-\infty}^{+\infty} 1\, \mathrm{d}x \to \infty.$$

The difficulty is best evidenced if we extend the spectral decomposition of the identity in (A.8) to the continuous case:

$$\int |\varphi_k\rangle \mathrm{d}k \langle \varphi_k| = 1 \tag{A.14}$$

and we apply it to one of the basis functions:

$$|\varphi_{k'}\rangle = \int |\varphi_k\rangle \mathrm{d}k \langle \varphi_k | \varphi_{k'}\rangle\ .$$

[1] It means that

$$\lim_{N\to\infty} \int \left| F(x) - \sum_{n=0}^{N} \varphi_n(x) \right|^2 \mathrm{d}x = 0,$$

where the sum runs over positive and negative n. From the point of view of physical applications, the function $F(x)$ and the limit of its Fourier series are totally equivalent.

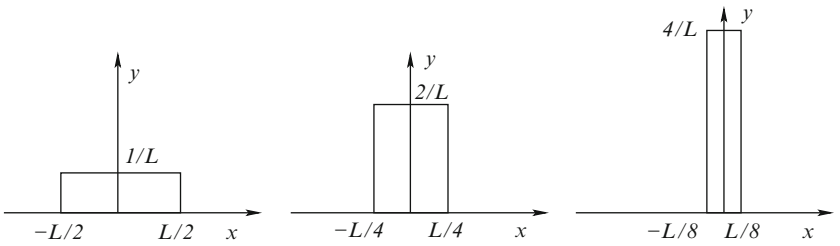


Fig. A.1. The succession of functions $\delta_n(x)$ in (A.16)

Thus, the scalar product

$$\langle \varphi_k | \varphi_{k'} \rangle, \tag{A.15}$$

which in the discrete case is the Kronecker delta in (A.3), should be a function of k zero everywhere except in $k = k'$, with the property that after multiplication by another function of k and integrated with respect to k, it yields a finite value. Such a function does not exist among the functions of the traditional mathematical analysis, that in this context are referred to as *proper functions*. The Dirac *delta function* is an extension of this concept of functions, called an *improper function*.

To introduce this concept, let us consider a succession of functions $\delta_L(x)$ defined by

$$\delta_n(x) = \begin{cases} 0 & if\ |x| > L/2^n \\ 2^{n-1}/L & if\ |x| \le L/2^n \end{cases} \qquad n = 1, 2, \ldots . \tag{A.16}$$

The first three functions of this succession are shown in Fig. A.1. All of them have the shape of a rectangle of unit area. The first one has basis L and height $1/L$; the second has basis $L/2$ and height $2/L$, and so on. Thus, the functions of the succession in (A.16) have the following properties:

$$\int_{-\infty}^{\infty} \delta_n(x)\, \mathrm{d}x = 1 \quad \forall n, \quad \text{thus} \quad \lim_{n\to\infty} \int_{-\infty}^{\infty} \delta_n(x)\, \mathrm{d}x = 1, \tag{A.17}$$

and

$$\lim_{n\to\infty} \delta_n(x) = \begin{cases} 0 & if\ x \neq 0 \\ \infty & if\ x = 0 \end{cases} . \tag{A.18}$$

Finally, and most important, if $f(x)$ is a function sufficiently regular around $x = 0$,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \delta_n(x) f(x)\, \mathrm{d}x = f(0). \tag{A.19}$$

In fact,

$$\lim_{n\to\infty}\int_{-\infty}^{\infty}\delta_n(x)f(x)\,\mathrm{d}x=\lim_{n\to\infty}\int_{-L/2^n}^{L/2^n}\left(\frac{2^{n-1}}{L}\right)f(x)\,\mathrm{d}x=\lim_{n\to\infty}f(\overline{x}_n)=f(0),$$

where the mean-value theorem has been used, and $\overline{x}_n$ is a point inside the interval where $\delta_n(x)$ is different from zero. According to (A.17)–(A.19), the function we are looking for should be the limit of the functions $\delta_n(x)$:

$$\delta(x)=\lim_{n\to\infty}\delta_n(x)\qquad(?)$$

and its characteristic properties would be

$$\boxed{\int_{-a}^{a}\delta(x)\,\mathrm{d}x=1,\qquad\int_{-a}^{a}\delta(x)f(x)\,\mathrm{d}x=f(0)}\tag{A.20}$$

where a is any positive number. This function, however, does not exist as proper function, and we must remember that its use must always be considered within an integral, and that a limit operation *outside* the integral is understood. The Dirac delta function in (A.20) is sometimes called an *improper function*. The concept of improper functions has been put in rigorous grounds by the *theory of distributions*.

It is now clear that the scalar product in (A.15) must be equal to the Dirac delta:

$$\langle\varphi_k|\varphi_{k'}\rangle=\delta(k-k'),\tag{A.21}$$

and we shall see in next section that this is exactly the case.

Some important properties of the δ-function are used in this text, and precisely

$$\delta(ax)=\frac{1}{|a|}\delta(x),\qquad\delta(f(x))=\sum_n\frac{1}{|f'(x_n)|}\delta(x-x_n),\tag{A.22}$$

where a is a real constant, f' is the derivative of f, and x_n are the solutions of the $f(x)=0$. The first of the above equations is clearly a particular case of the second, and they can be easily proved with a substitution of variables $y=f(x)$, taking carefully into account the signs of $\mathrm{d}y$ and $f'(x_n)$.

Integral Representation of the δ and the Fourier Integral

Let us now return to the problem of the eigenfunctions of the harmonic-motion equation for the infinite interval,

$$\varphi_k(x)=A_k\mathrm{e}^{\mathrm{i}kx},\tag{A.23}$$

and to the problem of the scalar product of two such functions, that we may write as

$$\begin{aligned}\lim_{L\to\infty}\int_{-L}^{L} A^*_{k'}\mathrm{e}^{-\mathrm{i}k'x}A_k\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}x &= \lim_{L\to\infty} A_kA^*_{k'}\int_{-L}^{L}\mathrm{e}^{\mathrm{i}(k-k')x}\,\mathrm{d}x\\ &= \lim_{L\to\infty} A_kA^*_{k'}\frac{1}{\Delta k}\left(\mathrm{e}^{\mathrm{i}\Delta kL}-\mathrm{e}^{-\mathrm{i}\Delta kL}\right)\\ &= \lim_{L\to\infty} A_kA^*_{k'}2L\frac{\sin(\Delta kL)}{\Delta kL}. \end{aligned} \tag{A.24}$$

Let us then consider the integral

$$\int_{-\infty}^{\infty} 2L\frac{\sin(\Delta kL)}{\Delta kL}\,\mathrm{d}\Delta k = 2\int_{-\infty}^{\infty}\frac{\sin y}{y}\,\mathrm{d}y = 2\pi. \tag{A.25}$$

Thus, if the constant A_k is taken equal to $1/\sqrt{2\pi}$, i.e., if we take

$$\boxed{\varphi_k(x) = \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathrm{i}kx}} \tag{A.26}$$

the integral of the scalar product in (A.24) is equal to 1 for any L. Furthermore, if we consider the function

$$\delta_L(k) = \frac{L}{\pi}\,\frac{\sin(kL)}{kL},$$

it has the following properties:

$$\lim_{L\to\infty}\int_{-\infty}^{\infty}\delta_L(k)\,\mathrm{d}k = 1,$$

since from (A.25), this integral has value one for any L. Furthermore,

$$\begin{aligned}\lim_{L\to\infty}\int_{-\infty}^{\infty} f(k)\delta_L(k)\,\mathrm{d}k &= \lim_{L\to\infty}\int_{-\infty}^{\infty} f(k)\frac{L}{\pi}\,\frac{\sin(kL)}{kL}\,\mathrm{d}k\\ &= \lim_{L\to\infty}\frac{1}{\pi}\int_{-\infty}^{\infty} f(\xi/L)L\,\frac{\sin\xi}{\xi}\,\mathrm{d}\xi = f(0),\end{aligned}$$

since, as L increases, the function f is evaluated at arguments closer and closer to zero, unless ξ is increasingly large, where the rest of the integrand becomes negligibly small. These results indicate that the limit for $L\to\infty$ of δ_L becomes the Dirac delta, or

$$\boxed{\tfrac{1}{2\pi}\textstyle\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}x = \delta(k)} \tag{A.27}$$

The integral representation of the δ function in (A.27) is significantly different from the simpler, pedagogical, example given above of the rectangular

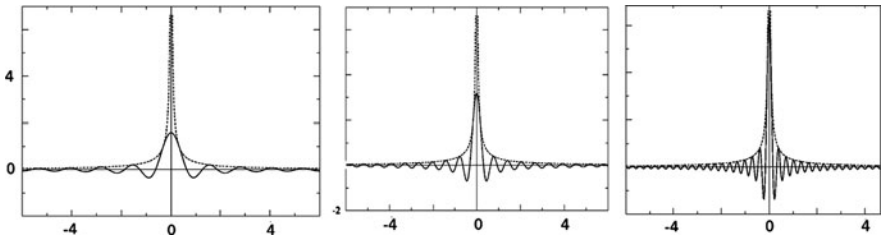


Fig. A.2. For the integral representation of the δ function, see text

functions. Figure A.2 shows the function $\delta_L(k)$ for three increasing values of L. These functions are oscillating more and more rapidly as L increases. If these functions are multiplied by a regular function $f(k)$ and integrated, the contributions outside the origin vanishes because the cancellation of the values at the adjacent positive and negative peaks is more and more complete as L increases, and only the central value remains, where δ_L increases indefinitely.

Equations (A.26) and (A.27), legitimate, in terms of improper functions, the scalar product in (A.15) as in (A.21).

With the integral representation of the δ in (A.27), it is immediate to obtain the Fourier integral of any square-integrable function:

$$f(x) = \int_{-\infty}^{\infty} f(x')\delta(x-x')\,\mathrm{d}x' = \int_{-\infty}^{\infty} f(x')\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}k(x-x')}\,\mathrm{d}k\,\mathrm{d}x'$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x')\mathrm{e}^{-\mathrm{i}kx'}\,\mathrm{d}x'\right]\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k,$$

or

$$\boxed{f(x) = \tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-\infty}^{\infty} A(k)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k, \quad A(k) = \tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-\infty}^{\infty} f(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x} \tag{A.28}$$

The above equation is the *Fourier integral* representation of the function f, and $A(k)$ is called the *Fourier transform* of f.

In vector terms, the application of the integral spectral decomposition in (A.14) yields

$$|f\rangle = \int |\varphi_k\rangle\,\mathrm{d}k\langle\varphi_k|f\rangle.$$

Parseval relation immediately follows:

$$\int_{-\infty}^{\infty} |f(x)|^2\,\mathrm{d}x = \langle f|f\rangle = \int \langle f|\varphi_k\rangle\,\mathrm{d}k\langle\varphi_k|f\rangle = \int_{-\infty}^{\infty} |A(k)|^2\,\mathrm{d}k.$$

B

One-Dimensional Potential Step, Barrier, and Well

In this appendix, a number of simple quantum systems will be considered. They are one-dimensional particles subject to a piecewise constant potential. For such problems, analytical solutions are easily found. In spite of their simplicity, they illustrate most of the major features of quantum dynamics with respect to classical dynamics: quantum reflection, tunnel effect, energy quantization, resonances. The results presented in this appendix are used in many quantum applications. In particular, they are good models for many semiconductor structures.

In Sect. 2.4, it was shown that for the solution of a quantum-mechanical problem it is necessary to find eigenvalues and eigenvectors of the Hamiltonian of the system. We are working here in the wave-mechanics representation, and the equation to be solved is the one-dimensional eigenvalue Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) + V(x)\psi = \epsilon\psi, \tag{B.1}$$

where $V(x)$ is the one-dimensional potential energy of the particle in x.

One-Dimensional Free Particle

The simplest case we may consider is that of a free particle for which $V(x) = cost = 0$, and the Schrödinger equation reduces to

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) = -k^2\psi, \quad \text{where} \quad k = \sqrt{\frac{2m\epsilon}{\hbar^2}} \quad \text{or} \quad \epsilon = \frac{\hbar^2 k^2}{2m}. \tag{B.2}$$

The general solution of the equation in (B.2) is

$$\psi(x) = A\mathrm{e}^{\mathrm{i}kx} + B\mathrm{e}^{-\mathrm{i}kx}. \tag{B.3}$$

If ϵ is negative, k is pure imaginary, the exponentials in (B.3) are real and diverge for $x \to \infty$ or $x \to -\infty$. Therefore, *there are no solutions for negative*

energy for a free particle, as in classical mechanics. If ϵ is positive k is real and the two solutions above can be combined in one expression (for $k = 0$ the two solutions coincide):

$$\psi_\epsilon(x) = A_k \mathrm{e}^{\mathrm{i}kx}.$$

Here, k is the wavevector; $\hbar k$ is the eigenvalue of the momentum, and the relation between energy and wavevector is given in (B.2). The spectrum of the energy eigenvalues is therefore formed by all positive values. They are doubly degenerate ($\pm k$), with the exception of $\epsilon = 0$. The spectrum is continuous, and therefore the orthonormalization must be performed with the Dirac delta. If the time dependence is included, the energy and momentum eigenstates are

$$\Psi_{\epsilon,p} = \frac{1}{\sqrt{2\pi\hbar}} \mathrm{e}^{\mathrm{i}(kx-\omega t)}, \quad p = \hbar k, \quad \epsilon = \frac{\hbar^2 k^2}{2m}, \quad \omega = \frac{\epsilon}{\hbar}. \tag{B.4}$$

Wavepackets, Phase and Group Velocities

The plane wave in (B.4) travels along the x axis with velocity

$$v_\phi = \frac{\omega}{k} = \frac{\hbar k}{2m}, \tag{B.5}$$

called *phase velocity*. With a linear combination of such plane waves it is possible to realized a *wavepacket*, i.e., a wavefunction with values significantly different from zero in a region of length Δx, and with k components significantly different from zero in a region of length Δk:

$$\Psi(x,t) = \int A(k) \mathrm{e}^{\mathrm{i}(kx-\omega t)} \mathrm{d}k = \int |A(k)| \mathrm{e}^{\mathrm{i}(kx-\omega t+\alpha(k))} \mathrm{d}k,$$

where the Fourier coefficient $A(k)$ has been explicitly written in terms of its modulus and its phase $\alpha(k)$. The mathematics of Fourier analysis requires that the following inequality is satisfied,

$$\Delta x \Delta k \geq \frac{1}{2},$$

on which uncertainty relations are based.

In the superposition which forms the wavepacket, each plane wave travels with its own phase velocity (B.5) so that the wavepacket is deformed and spreads with time. We may still define a velocity of the wavepacket as the velocity of the point in which all components are in phase, i.e., where

$$\frac{\partial(\text{phase})}{\partial k} = \frac{\partial}{\partial k}[kx - \omega t + \alpha(k)] = 0, \quad \text{or} \quad x = \frac{\partial \alpha}{\partial k} + \frac{\partial \omega}{\partial k} t.$$

This defines the *group velocity* of the wavepacket:

$$v_g = \frac{\partial \omega}{\partial k} = \frac{\partial \epsilon}{\partial p} = \frac{\hbar k}{m}. \tag{B.6}$$

This velocity depends upon k so that the velocity of the wavepacket is well defined only if the component $A(k)$ is significantly different from zero in a small enough region of k.

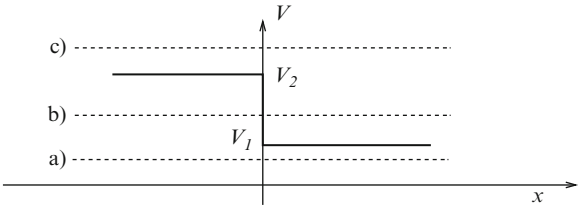


Fig. B.1. Potential step

Potential Step and Infinite Barrier

The next example is that of a potential step. The Schrödinger equation (B.1) is now written as

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) = -\frac{2m}{\hbar^2}[\epsilon - V(x)]\psi,$$

where

$$V(x) = \begin{cases} V_2 & \text{for } x < 0 \\ V_1 & \text{for } x > 0 \end{cases}.$$

as shown in Fig. B.1.

The procedure to solve this equation, as any other one-dimensional Schrödinger equation with piecewise constant potential, is the following: first solve the equation separately in the different regions where $V(x)$ is constant, and then require the continuity of the wavefunction and its derivative at the points of discontinuities of the potential energy. These regularity conditions are required both by physical reasons (the probability density and the current must have a unique value at these points) and by mathematical reasons (the Schrödinger equation shows that the second derivative must exist finite at all values of x). At this point, it is convenient to consider different regions of possible eigenvalues of the energy.

(a) $\epsilon < V_1$

It is easy to show [306] that in this case it is not possible to satisfy the continuity conditions for both the wavefunction and its derivative at the same time.

(b) $V_1 < \epsilon < V_2$

The solution is

$$\psi(x) = \begin{cases} A_1 \mathrm{e}^{-\mathrm{i}k_1 x} + B_1 \mathrm{e}^{\mathrm{i}k_1 x}, & x \geq 0 \\ A_2 \mathrm{e}^{-\kappa_2 x} + B_2 \mathrm{e}^{\kappa_2 x}, & x \leq 0 \end{cases},$$

with

$$k_1 = \sqrt{\frac{2m}{\hbar^2}(\epsilon - V_1)}, \quad \kappa_2 = \sqrt{\frac{2m}{\hbar^2}(V_2 - \epsilon)}.$$

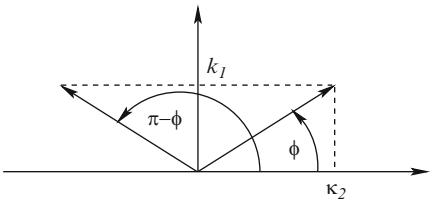


Fig. B.2. For the solution of Schrödinger equation with a potential step, see text

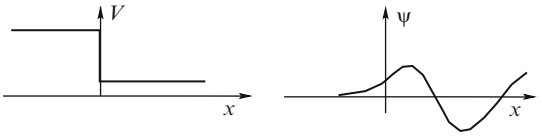


Fig. B.3. Eigenfunction of a potential step

Regularity at $x \to -\infty$ requires $A_2 = 0$. Continuities of the function and its derivative at $x = 0$ require

$$\begin{cases} 1 + \mathcal{B}_1 = \mathcal{B}_2 \\ -ik_1 + ik_1\mathcal{B}_1 = \kappa_2\mathcal{B}_2 \end{cases}, \quad \text{where} \quad \mathcal{B}_1 = B_1/A_1, \qquad \mathcal{B}_2 = B_2/A_1,$$

with solution

$$\mathcal{B}_1 = \frac{\kappa_2 + \mathrm{i}k_1}{-\kappa_2 + \mathrm{i}k_1}.$$

From Fig. B.2, we see that

$$|\mathcal{B}_1| = 1, \qquad \tan\phi = \frac{k_1}{\kappa_2}.$$

Thus,

$$\mathcal{B}_1 = \mathrm{e}^{\mathrm{i}[\phi-(\pi-\phi)]} = -\mathrm{e}^{2\mathrm{i}\phi}, \qquad \mathcal{B}_2 = 1 + \mathcal{B}_1 = 1 - \mathrm{e}^{2\mathrm{i}\phi}.$$

The wavefunction can then be written as

$$\psi(x) = A_1 \begin{cases} \mathrm{e}^{-\mathrm{i}k_1 x} - \mathrm{e}^{2\mathrm{i}\phi}\mathrm{e}^{\mathrm{i}k_1 x} \\ \left(1 - \mathrm{e}^{2\mathrm{i}\phi}\right)\mathrm{e}^{\kappa_2 x} \end{cases} = \begin{cases} A\sin(k_1 x + \phi) & x \geq 0 \\ B\mathrm{e}^{\kappa_2 x} & x \leq 0 \end{cases}, \tag{B.7}$$

shown in Fig. B.3, with

$$\frac{B}{A} = \sin\phi = \frac{k_1}{\sqrt{k_1^2 + \kappa_2^2}}.$$

The energy spectrum is continuous and nondegenerate.

Several comments are relevant at this point:

Comment I: For $x > 0$, we have a combination of an incoming wave and an outgoing wave with equal amplitudes, equal intensities, equal current densities. Thus, the *reflection coefficient* R *and the transmission coefficient* T are

$$R = 1, \qquad T = 0.$$

Comment II: $|\Psi|^2 \neq 0$ in $x < 0$. The particle can be found in a region where $\epsilon < V$, in contrast with classical mechanics.

Comment III: *Infinite barrier*. The exponential part in (B.7) decreases more rapidly, for decreasing $x < 0$, as κ_2 increases, i.e., as the potential V_2 increases. If $V_2 \to \infty$, the wavefunction vanishes at the position of the potential step ($\sin\phi = 0, \phi = 0$), while its derivative remains different from zero.

Comment IV: *Quantum delay*. If a wavepacket hits the potential step from the left, it is totally reflected. It can be easily shown, however, that the center of the reflected wavepacket leaves the position of the step with a certain delay with respect to the time at which the center of the incoming wavepacket hits the step [306]. This delay vanishes if the potential step is infinite.

(c) $\epsilon > V_2$

The solution is

$$\psi(x) = \begin{cases} A_1 \mathrm{e}^{-\mathrm{i}k_1 x} + B_1 \mathrm{e}^{\mathrm{i}k_1 x}, & x \geq 0 \\ A_2 \mathrm{e}^{-\mathrm{i}k_2 x} + B_2 \mathrm{e}^{\mathrm{i}k_2 x}, & x \leq 0 \end{cases},$$

with

$$k_1 = \sqrt{\frac{2m}{\hbar^2}(\epsilon - V_1)}, \qquad k_2 = \sqrt{\frac{2m}{\hbar^2}(\epsilon - V_2)}.$$

It is formed by incoming and outgoing plane waves in both sides. The continuities of the wavefunction and of its derivative require

$$\begin{cases} A_1 + B_1 = A_2 + B_2 \\ -\mathrm{i}k_1 A_1 + \mathrm{i}k_1 B_1 = -\mathrm{i}k_2 A_2 + \mathrm{i}k_2 B_2 \end{cases}.$$

These are two equations with four unknowns. The system is homogeneous and one unknown is taken care of by the normalization condition. Two linearly independent solutions remain. The energy spectrum is therefore continuous and doubly degenerate. The solutions can be chosen, for example, taking $A_1 = 0$ or $B_2 = 0$. The resulting solutions are the *scattering states*, given by plane waves which come from one side and at the step are split in one reflected part and one transmitted.

Let us consider the solution with $B_2 = 0$ and assume, for normalization, $A_1 = 1$ that corresponds to take an incoming plane wave of unit amplitude. The solution may be written as

$$\psi(x) = \begin{cases} \mathrm{e}^{-\mathrm{i}k_1 x} + \mathcal{R}\mathrm{e}^{\mathrm{i}k_1 x} & x \geq 0 \\ \mathcal{S}\mathrm{e}^{-\mathrm{i}k_2 x} & x \leq 0 \end{cases}.$$

The regularity conditions yield, after simple calculations,

$$\mathcal{R} = \frac{k_1 - k_2}{k_1 + k_2}, \qquad \mathcal{S} = 1 + \mathcal{R} = \frac{2k_1}{k_1 + k_2}.$$

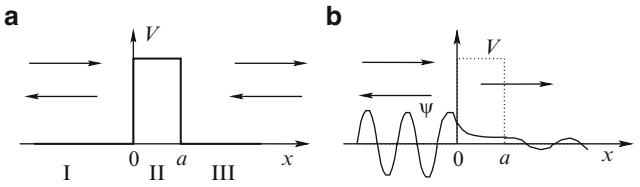


Fig. B.4. Tunnel through a barrier

The current density carried by the plane wave $A\mathrm{e}^{\mathrm{i}kx}$ is proportional to $|A|^2k$. Thus, in our case the incident, reflected, and transmitted current densities are

$$j_i \propto 1\ k_1, \quad j_r \propto |\mathcal{R}|^2 k_1, \quad j_t \propto |\mathcal{S}|^2 k_2.$$

The reflection and transmission coefficients are then

$$R = \frac{j_r}{j_i} = |\mathcal{R}|^2 = \frac{(k_1 - k_2)^2}{(k_1 + k_2)^2}, \quad T = \frac{j_t}{j_i} = \frac{|\mathcal{S}|^2 k_2}{k_1} = \frac{4k_1 k_2}{(k_1 + k_2)^2},$$

whose sum is one, as it must be.

Note that a *quantum reflection* exists also when the energy is above the potential step, contrary to classical mechanics.

Potential Barrier: Tunnel Effect

Let us now consider the case of a potential barrier,

$$V(x) = \begin{cases} 0, & x < 0 \\ V_\circ, & 0 \le x \le a \\ 0, & x > a \end{cases}$$

shown in Fig. B.4.

The solution is again a combination of plane waves with proper wavevectors (real or imaginary) in the different regions of the x axis:

$$\psi(x) = \begin{cases} A_1\mathrm{e}^{-\mathrm{i}k_1 x} + B_1\mathrm{e}^{\mathrm{i}k_1 x} & x \le 0 \\ A_2\mathrm{e}^{-\mathrm{i}k_2 x} + B_2\mathrm{e}^{\mathrm{i}k_2 x} & 0 \le x \le a \\ A_3\mathrm{e}^{-\mathrm{i}k_1 x} + B_3\mathrm{e}^{\mathrm{i}k_1 x} & x \ge 0 \end{cases}.$$

The regularity conditions at the points of discontinuity of the potential yield the coefficients of the combination in one part of the axis as linear combinations of the coefficients of the adjacent region:

$$\begin{pmatrix} A_2 \\ B_2 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}.$$

A second matrix connects the coefficients of the next region with those of the previous one. By matrix multiplication, the linear relation between the

coefficients of the last region with those of the first one is obtained. Again we have two independent solutions that can be chosen, for example, as the scattering states. The energy spectrum is continuous and doubly degenerate for $\epsilon > 0$. In the case shown in Fig. B.4a with $0 < \epsilon < V_\circ$, we have

$$\psi(x) = \begin{cases} A_1 \mathrm{e}^{\mathrm{i}kx} + B_1 \mathrm{e}^{-\mathrm{i}kx} & \text{in } I \\ A_2 \mathrm{e}^{\kappa x} + B_2 \mathrm{e}^{-\kappa x} & \text{in } II \\ A_3 \mathrm{e}^{\mathrm{i}kx} + B_3 \mathrm{e}^{-\mathrm{i}kx} & \text{in } III \end{cases} .$$

For the scattering state shown in Fig. B.4b, we put $B_3 = 0$. Particles coming from the left with energy smaller that the height of the barrier have a nonzero probability to be found on the right of the barrier. This is the famous *tunnel effect* which shows in a very clear way that in quantum mechanics it is not possible to assign to the particles a given trajectory between two positions where it can be found.

Infinite Potential Well

Next example considers a particle in a one-dimensional infinite potential well, as shown in Fig. B.5. In the interval between $-l/2$ and $l/2$ the potential is zero, and the solution has the form given in (B.3) for the free particle, with $k = \sqrt{2m\epsilon/\hbar^2}$. As we have just seen the wavefunction must vanish at the edges of the well. This condition selects the possible wavevectors:

$$k_n = \frac{\pi}{l} n, \quad \epsilon_n = \frac{\hbar^2 k_n^2}{2m} = \frac{\hbar^2 \pi^2}{2ml^2} n^2, \quad \lambda_n = \frac{2\pi}{k_n} = \frac{2}{n} l, \quad l = n\frac{\lambda_n}{2}, \quad n = 1, 2, \ldots .$$

They are such that the width of the well contains an integer number of half wavelengths, as expected. The energy spectrum in discrete and nondegenerate. The corresponding eigenfunctions are

$$\psi(x) = \begin{cases} A_n \cos(k_n x) & n \text{ odd} \\ A_n \sin(k_n x) & n \text{ even} \end{cases} .$$

The first three eigenfunctions are shown in Fig. B.5.

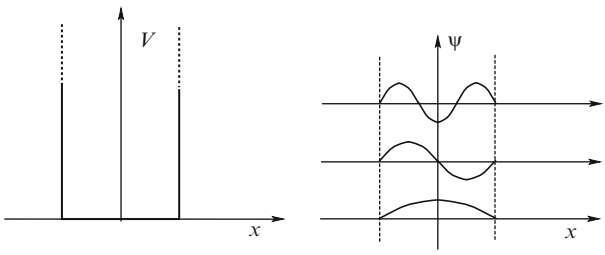


Fig. B.5. Infinite potential well and its first eigenfunctions

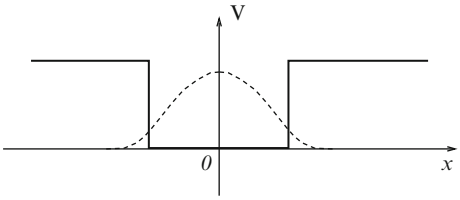


Fig. B.6. Finite potential well (*continuous line*) and its ground-state eigenfunction

The constant A_n is fixed by the normalization and, with an arbitrary choice of the phase, results to be independent of n:

$$A_n = \sqrt{2/l}.$$

Finite Potential Well: Resonances

If the potential energy corresponds to a well with finite confining steps, as shown in Fig. B.6, eigenvalues and eigenfunctions are found in the usual way: the solutions of the Schrödinger equation are plane waves or exponentials in the separate regions. Eigenvalues are obtained by the regularity conditions. A numerical or graphical procedure is required in this case [306]. Here, the symmetrical case is considered for simplicity. The energy spectrum is discrete and nondegenerate below the confining potential and continuous, doubly degenerate, above. In Fig. B.6, the ground state wavefunction is depicted. Note that the wavefunction penetrates somewhat in the region classically forbidden, as in the case of the potential step.

For values of the energy above the well equal to the eigenvalues of the infinite well, resonances appear: the wavefunction inside the well is much bigger than for eigenfunctions out of resonance. In physical terms, some insight into this important phenomenon may be obtained considering that quantum reflections at the steps generate positive interference when the width of the well contains an integer number of half wavelengths.

C

Quantum Theory of Harmonic Oscillator

The quantum theory of the harmonic oscillator is of particular importance for many reasons. In fact, harmonic forces are often present in nature and even more often used as model systems. Small oscillations about a point of stable equilibrium are described as independent harmonic oscillators, as we have seen in sect. 1.5. Most important, the theory of the harmonic oscillator introduces a formalism, through creation and annihilation operators, that is very frequently used in all fields of modern theoretical physics.

The Hamiltonian

If we define q and p as the position and momentum operators of the particle subject to a harmonic force, the Hamiltonian and the fundamental commutation relation are given by

$$\mathcal{H} = \frac{\mathsf{p}^2}{2m} + \frac{1}{2} m\omega^2 \mathsf{q}^2, \qquad [\mathsf{q}, \mathsf{p}] = \mathrm{i}\hbar, \tag{C.1}$$

where m is the mass of the particle and ω the classical frequency of the oscillator, $\omega = \sqrt{k/m}$, if k is the constant such that the classical force is given by $F = -kq$.

Creation and Annihilation Operators

Let us first define the dimensionaless, reduced Hamiltonian

$$\mathcal{H}' = \frac{\mathcal{H}}{\hbar\omega} = \frac{\mathsf{p}^2}{2m\hbar\omega} + \frac{1}{2}\frac{m\omega}{\hbar}\mathsf{q}^2.$$

Then two new dimensionless operators $\mathcal{Q}$ and $\mathcal{P}$ are defined as

$$\mathcal{Q} = \sqrt{\frac{m\omega}{\hbar}}\ \mathsf{q}, \qquad \mathcal{P} = \frac{1}{\sqrt{m\hbar\omega}}\ \mathsf{p},$$

with commutator

$$[\mathcal{Q}, \mathcal{P}] = \left[\sqrt{\frac{m\omega}{\hbar}}\ \mathsf{q}, \frac{1}{\sqrt{m\hbar\omega}}\ \mathsf{p} \right] = \mathrm{i}. \tag{C.2}$$

In terms of the new operators, the reduced Hamiltonian is

$$\mathcal{H}' = \frac{1}{2}\left(\mathcal{Q}^2 + \mathcal{P}^2\right).$$

Let us now define the *annihilation operator* a and its Hermitian conjugate, *creation operator* $\mathsf{a}^\dagger$, as

$$\begin{cases} \mathsf{a} = \frac{1}{\sqrt{2}}(\mathcal{Q} + \mathrm{i}\mathcal{P}) \\ \mathsf{a}^\dagger = \frac{1}{\sqrt{2}}(\mathcal{Q} - \mathrm{i}\mathcal{P}) \end{cases} \quad \text{or} \quad \begin{cases} \mathcal{Q} = \frac{1}{\sqrt{2}}(\mathsf{a} + \mathsf{a}^\dagger) \\ \mathcal{P} = -\frac{\mathrm{i}}{\sqrt{2}}(\mathsf{a} - \mathsf{a}^\dagger) \end{cases}. \tag{C.3}$$

The commutation relation of a and $\mathsf{a}^\dagger$ can be found immediately from the commutator in (C.2):

$$[\mathsf{a}, \mathsf{a}^\dagger] = \left[\frac{1}{\sqrt{2}}(\mathcal{Q} + \mathrm{i}\mathcal{P}), \frac{1}{\sqrt{2}}(\mathcal{Q} - \mathrm{i}\mathcal{P}) \right] = 1. \tag{C.4}$$

Note that the operators a and $\mathsf{a}^\dagger$ are not Hermitian and therefore are not observables. Furthermore, since $\mathcal{Q}$ and $\mathcal{P}$, and therefore q and p, can be expressed in terms of a and $\mathsf{a}^\dagger$ by means of the equations (C.3), any other physical quantity can be expressed in terms of these two operators. In particular,

$$\mathcal{H}' = \mathcal{N} + \frac{1}{2}, \quad \mathcal{N} \equiv \mathsf{a}^\dagger \mathsf{a}, \quad \mathcal{H} = \left(\mathsf{a}^\dagger \mathsf{a} + \frac{1}{2} \right) \hbar\omega = \left(\mathcal{N} + \frac{1}{2} \right) \hbar\omega. \tag{C.5}$$

$\mathcal{N}$ is called *number operator*, for reasons that will be made clear shortly.

Eigenvalues and Eigenvectors

It is possible to find eigenvalues and eigenvectors of the Hamiltonian of the harmonic oscillator using only the fundamental commutation relation. Let us call $|\nu\rangle$ one normalized eigenvector of $\mathcal{N}$ (and therefore of $\mathcal{H}$), belonging to the eigenvalue ν:

$$\mathcal{N}|\nu\rangle = \nu|\nu\rangle.$$

Then

(1) The eigenvalue ν is ≥ 0.

In fact, if we consider the squared modulus of the vector $\mathsf{a}|\nu\rangle$, we have

$$0 \leq \langle\nu|\mathsf{a}^\dagger \mathsf{a}|\nu\rangle = \langle\nu|\mathcal{N}|\nu\rangle = \nu.$$

(2) The application of a to $|\nu\rangle$ produces either the null vector, and this is so only when $\nu = 0$, or a new eigenvector of $\mathcal{N}$ belonging to the eigenvalue $\nu - 1$.
In fact, from the derivation of (1) it results that $a|\nu\rangle$ is the null vector only if $\nu = 0$. Furthermore,

$$\mathcal{N}\mathsf{a}|\nu\rangle = \mathsf{a}^\dagger\mathsf{a}\mathsf{a}|\nu\rangle = (\mathsf{a}\mathsf{a}^\dagger - 1)\mathsf{a}|\nu\rangle = (\mathsf{a}\mathsf{a}^\dagger\mathsf{a} - \mathsf{a})|\nu\rangle = \mathsf{a}(\mathcal{N} - 1)|\nu\rangle = (\nu - 1)\mathsf{a}|\nu\rangle,$$

as we wanted.

(3) The application of $\mathsf{a}^\dagger$ to $|\nu\rangle$ always produces a new eigenvector of $\mathcal{N}$ belonging to the eigenvalue $\nu + 1$.
In fact, the squared modulus of the vector $a^\dagger|\nu\rangle$ is

$$\langle\nu|\ \mathsf{a}\ \mathsf{a}^\dagger\ |\nu\rangle = \langle\nu|\mathcal{N} + 1|\nu\rangle = \nu + 1,$$

which is never zero owing to the result in (1). Furthermore,

$$\mathcal{N}\mathsf{a}^\dagger|\nu\rangle = \mathsf{a}^\dagger\mathsf{a}\ \mathsf{a}^\dagger|\nu\rangle = \mathsf{a}^\dagger(\mathcal{N} + 1)|\nu\rangle = (\nu + 1)\mathsf{a}^\dagger|\nu\rangle,$$

that completes the proof.

At this point, we can draw some important conclusions. Since we showed that any eigenvalue ν is nonnegative, the process of successive applications of a, that results in the reduction of the eigenvalue by a unity, must come to an end, i.e. there must be a value of ν' such that

$$\mathsf{a}|\nu'\rangle = 0.$$

For the point (2) above, this can be true only if $\nu' = 0$. Thus, the values of ν must be integers and the eigenstates of $\mathcal{N}$ are

$$|0\rangle \quad |1\rangle \quad |2\rangle \quad |3\rangle \quad \ldots,$$

with eigenvalues $0, 1, 2, \ldots$. These are also the eigenstates of the Hamiltonian, because of (C.5), and the corresponding eigenvalues are

$$\boxed{\varepsilon_n = \left(n + \tfrac{1}{2}\right)\hbar\omega} \tag{C.6}$$

From the values found above for the moduli of the vectors $\mathsf{a}|n\rangle$ and $\mathsf{a}^\dagger|n\rangle$, we have that the effects of a and $\mathsf{a}^\dagger$ on the eigenvectors of $\mathcal{N}$ are, apart from arbitrary phase factors,

$$\mathsf{a}\ |n\rangle = \sqrt{n}\ |n-1\rangle, \qquad \mathsf{a}^\dagger\ |n\rangle = \sqrt{n+1}\ |n+1\rangle. \tag{C.7}$$

The ground state $|0\rangle$ is called *vacuum*. Starting from it, all the other eigenstates of the Hamiltonian can be obtained by means of successive applications of $\mathsf{a}^\dagger$, as follows:

$$|0\rangle, \qquad \mathsf{a}^\dagger|0\rangle = |1\rangle, \qquad \mathsf{a}^\dagger|1\rangle = \sqrt{2}|2\rangle, \qquad \mathsf{a}^\dagger|2\rangle = \sqrt{3}|3\rangle, \ldots, \tag{C.8}$$

or

$$|0\rangle, \ \ |1\rangle = \mathsf{a}^\dagger|0\rangle, \ \ |2\rangle = \frac{1}{\sqrt{2}}\mathsf{a}^\dagger|1\rangle = \frac{1}{\sqrt{2}}(\mathsf{a}^\dagger)^2|0\rangle, \ldots \ |n\rangle = \frac{1}{\sqrt{n!}}(\mathsf{a}^\dagger)^n|0\rangle. \tag{C.9}$$

Zero-Point Energy

It is important to note that the eigenvalues of the Hamiltonian of the harmonic oscillator are equally spaced by the quantity $\hbar\omega$, i.e., the energy of a photon with a frequency equal to the frequency of the classical oscillations. The ground state corresponds to an energy $\epsilon_\circ = \frac{1}{2}\hbar\omega$, called *zero-point energy*. Such an energy means that even at zero absolute temperature the harmonic oscillator maintains a fine amount of vibration, which has important effects, experimentally observable. On the contrary, a particle at rest ($p = 0$) in the origin ($q = 0$) of the oscillator would violate the uncertainty relation between position and momentum.

The {q} Representation, Eigenfunctions

According to (2.19) the $\{\mathsf{q}\}$ representation of a state vector in the Schrödinger picture is the wavefunction of the standard wave mechanics. To find the eigenfunctions of the Hamiltonian of the harmonic oscillator, it is useful to remember that all the eigenstates can be obtained from the ground state by successive applications of the creation operator, as indicated in (C.9). To find the ground wavefunction, we use the equation

$$\mathsf{a}|0\rangle = 0 \quad \text{or} \quad \frac{1}{\sqrt{2}}(\mathcal{Q} + \mathrm{i}\mathcal{P})|0\rangle = \frac{1}{\sqrt{2}}\left[\sqrt{\frac{m\omega}{\hbar}}\ \mathsf{q} + \frac{\mathrm{i}}{\sqrt{m\hbar\omega}}\mathsf{p}\right]|0\rangle = 0.$$

In the $\{\mathsf{q}\}$ representation, this equation becomes the simple differential equation

$$\frac{\mathrm{d}}{\mathrm{d}q}\psi_\circ(q) = -\frac{m\omega}{\hbar}\ q\psi_\circ(q),$$

with solution

$$\psi_\circ(q) = Ce^{-\frac{1}{2}\frac{m\omega}{\hbar}q^2}, \quad C = \sqrt[4]{\frac{m\omega}{\pi\hbar}}, \tag{C.10}$$

where C has been determined by the normalization.

Excited states are obtained by successive applications of the creation operator, as indicated in (C.8). We recall that in the $\{\mathsf{q}\}$ representation $\mathsf{a}^\dagger$ is given by

$$\frac{1}{\sqrt{2}}\left[\sqrt{\frac{m\omega}{\hbar}}\ \mathsf{q} - \frac{\mathrm{i}}{\sqrt{m\hbar\omega}}\mathsf{p}\right] = \frac{1}{\sqrt{2}}\left[\sqrt{\frac{m\omega}{\hbar}}\ \mathsf{q} - \frac{\mathrm{i}}{\sqrt{m\hbar\omega}}(-\mathrm{i}\hbar)\frac{\mathrm{d}}{\mathrm{d}q}\right].$$

The resulting eigenfunctions of the harmonic oscillator can be written as [398]

$$\psi_n(q) = N_n H_n\left(\sqrt{\frac{m\omega}{\hbar}}q\right)e^{-\frac{1}{2}\frac{m\omega}{\hbar}q^2}, \qquad N_n = \frac{C}{\sqrt{2^n n!}}, \tag{C.11}$$

where $H_n(\xi)$ are the Hermite polynomials defined by

$$H_n(\xi) = (-1)^n \mathrm{e}^{\xi^2}\frac{d^n}{\mathrm{d}\xi^n}\mathrm{e}^{-\xi^2}.$$

D

Landau Levels

In Chap. 11, the effect of a magnetic field on the dynamics of a charged particle was studied in semiclassical terms. When the field is strong enough, however, such that the particles can close a cyclotron orbit without being scattered, quantum effects become essential since close orbits always involve quantization. In our case, this occurs when $\omega_c\tau$ is larger or comparable with one, where ω_c is the cyclotron frequency, and τ is the scattering time. In the following, we shall therefore present the quantum theory of a particle in a homogeneous material in presence of a uniform and constant magnetic field $\boldsymbol{B}$, in the simple approximation of a spherical effective mass m.

The Hamiltonian for our particle is given by (see Sect. 1.8)

$$\mathcal{H} = \frac{1}{2m}\left[\mathsf{p} - q\boldsymbol{A}\right]^2 \ ,$$

where q is the particle charge and $\boldsymbol{A}$ is the vector potential. If the z direction is taken parallel to $\boldsymbol{B}$, the latter may be taken as $\boldsymbol{A} = (-By, 0, 0)$. With this choice, called Landau gauge, the Hamiltonian becomes

$$\mathcal{H} = \frac{1}{2m}\left(\mathsf{p} + qBy\hat{x}\right)^2 = \frac{\mathsf{p}^2}{2m} + \frac{1}{2m}q^2B^2y^2 + \frac{1}{m}qBy\mathsf{p}_x \ ,$$

where we have taken into account that p_x and y commute.

The Schrödinger equation to be solved is then

$$\mathcal{H}\psi(\boldsymbol{r}) = \left(\frac{\mathsf{p}_x^2 + \mathsf{p}_y^2 + \mathsf{p}_z^2}{2m} + \frac{1}{2m}q^2B^2y^2 + \frac{1}{m}qBy\mathsf{p}_x\right)\psi(\boldsymbol{r}) = \epsilon\psi(\boldsymbol{r}) \ .$$

The z component appears only in the term $\mathsf{p}_z^2/2m$. This means that, as in classical physics, the dynamics of the particle along the direction of a constant and uniform magnetic field is the same as that of a free particle, and the Hamiltonian eigenfunctions contain a factor $\exp(\mathrm{i}k_z z)$. The other two components are instead mixed by the magnetic field. Nevertheless, we shall now

prove that the eigenfunctions can be written in the form

$$\psi(x,y,z) = \mathrm{e}^{\mathrm{i}k_z z}\mathrm{e}^{\mathrm{i}\alpha x}\phi(y) \ .$$

Substituting this function into the Schrödinger equation yields

$$\frac{\hbar^2 k_z^2}{2m}\psi + \frac{\hbar^2\alpha^2}{2m}\psi - \frac{\hbar^2}{2m}\frac{\partial^2\phi}{\partial y^2}\mathrm{e}^{\mathrm{i}k_z z}\mathrm{e}^{\mathrm{i}\alpha x} + \frac{q^2B^2y^2}{2m}\psi + \frac{qBy}{m}\hbar\alpha\psi = \epsilon\psi \ .$$

Now we divide by the exponentials and put

$$\omega_c = \frac{qB}{m} \ , \qquad \epsilon' = \epsilon - \frac{\hbar^2 k_z^2}{2m} \ .$$

The result is

$$-\frac{\hbar^2}{2m}\frac{d^2\phi}{dy^2} + \frac{1}{2}m\omega_c^2\left(y^2 + \frac{\hbar^2\alpha^2}{\omega_c^2 m^2} + 2y\frac{\hbar\alpha}{m\omega_c}\right)\phi = \epsilon'\phi \ .$$

Finally we make the transformation

$$y \ \rightarrow \ y' = y + \frac{\hbar\alpha}{m\omega_c} \ , \qquad \frac{d}{dy} = \frac{d}{dy'}$$

and obtain the equation

$$-\frac{\hbar^2}{2m}\frac{d}{dy'^2}\phi + \frac{1}{2}m\omega_c^2 y'^2\phi = \epsilon'\phi \ . \tag{D.1}$$

This equation is immediately recognized as the Schrödinger equation of a harmonic oscillator. Thus, the energy eigenvalues are

$$\epsilon_{nk_z} = \left(n + \frac{1}{2}\right)\hbar\omega_c + \frac{\hbar^2 k_z^2}{2m} \ ,$$

shown in Fig. D.1, and the eigenfunctions are (apart from a normalization constant),

$$\psi_{nk_z\alpha} = \mathrm{e}^{\mathrm{i}k_z z}\mathrm{e}^{\mathrm{i}\alpha x}\psi_n(y - y_\circ) \ , \tag{D.2}$$

where ψ_n are the eigenfunctions of the harmonic oscillator given in (C.11).

The center of the oscillation is

$$y_\circ = -\frac{\hbar\alpha}{m\omega_c} = -\frac{\hbar\alpha}{qB} = -\alpha\ell_M^2 \qquad \ell_M^2 = \frac{\hbar}{m\omega_c} = \frac{\hbar}{eB} \ , \tag{D.3}$$

which can therefore be considered the y coordinate of the center of the classical cyclotron orbits.[1] The quantity ℓ_M is named *magnetic length*.

[1] As it regards the x coordinate $x_\circ$ of the center of the classical orbit, it is represented by the operator $x + \mathsf{p}_y/B$ [261]. It must be noted, however, the $x_\circ$ and $y_\circ$ cannot take definite values simultaneously, since the corresponding operators do not commute.

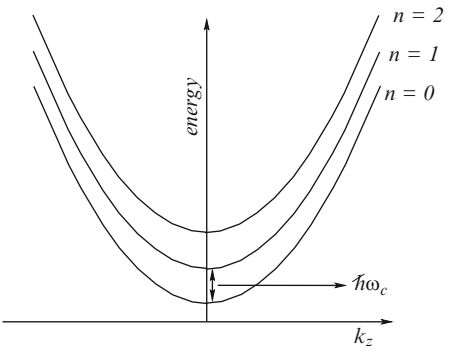


Fig. D.1. Energy bands with Landau levels

For simplicity, in what follows we shall neglect the longitudinal motion along z, of no interest for the present discussion.

Since, in an infinite sample, any value of α is allowed, and the energy eigenvalue does not depend upon α, all real values of the $y_\circ$ are available and all the corresponding wavefunctions belong to the same degenerate eigenvalue ϵ_n.

In a finite sample of y dimension L_y the maximum value of $y_\circ$ is L_y, so that the maximum value of α is $\alpha_M = L_y qB/\hbar$. If the dimension of the sample along x is L_x, periodic boundary conditions for the wavefunction require, as for the Bloch states, discrete values of α at a distance $\Delta\alpha = 2\pi/L_x$. Thus, the total number of Landau states[2] for any n is (apart from the longitudinal motion)

$$N = \frac{\alpha_M}{\Delta\alpha} = L_y \frac{qB}{\hbar} \Big/ \frac{2\pi}{L_x} = \frac{SqB}{2\pi\hbar} = \Phi(B)/\Phi_\circ \,, \tag{D.4}$$

where S is the area of the sample orthogonal to the magnetic field, $\Phi(B)$ is the total magnetic flux across this area, and $\Phi_\circ$ is, for electrons, a natural flux unit given by

$$\Phi_\circ \equiv \frac{h}{e} = 4.136 \times 10^{-15}\,\mathrm{Wb} \,.$$

From (D.4), we may also obtain an indication of the area occupied by each Landau state, assuming that they do not overlap:

$$A = \frac{S}{N} = \frac{2\pi\hbar}{eB} = 2\pi\ell_M^2 \,.$$

This shows that the magnetic length can be taken as an indication of the linear dimensions of the Landau states.

Let us finally consider an energy interval $\Delta\epsilon$, containing several landau levels. In the absence of magnetic field, the density of states for a two-dimensional electron gas is independent of energy and, ignoring the spin, is given by

[2] Spin degeneracy is removed by the magnetic field so that in this density of states only single spin states are considered.

$Sm/2\pi\hbar^2$ (see (19.5) of Chap. 19).The number of states in the interval of energy $\Delta\epsilon$ is then

$$N_{\text{free}} = \frac{Sm}{2\pi\hbar^2}\Delta\epsilon \ .$$

In the presence of the magnetic field, the number of Landau levels in the same interval of energy is $N_L = \Delta\epsilon/\hbar\omega_c$. Thus, each Landau level incorporates a number of states given by

$$N = N_{\text{free}}/N_L = \frac{Sm}{2\pi\hbar^2}\Delta\epsilon \Big/ \Delta\epsilon/\hbar\omega_c = \frac{\text{SqB}}{2\pi\hbar} \ ,$$

coherently with the result in (D.4). As the magnetic field increases, the Landau levels move away from each other, and each of them incorporates the levels which, in absence of $\boldsymbol{B}$ would be located in the interval between them.

E

Perturbation Theory

Very few quantum problems allow an exact solution, and in most cases approximate solutions must be searched for. Even though "exact" numerical solutions can be obtained for an increasing number of problems, simple analytical approximate solutions are often useful for a physical understanding of the dependence of the phenomenon upon the parameters defining the system under investigation. This is particularly true for the perturbation theory, the approximation method most frequently used in the present book, which will be presented in this appendix.

The General Idea

Perturbation theory is conveniently applied when the Hamiltonian of the system of interest can be separated into two parts. The "unperturbed" Hamiltonian $\mathcal{H}_\circ$ is supposed to be time independent and exactly solved, meaning that we know its eigenvalues $\epsilon_{\circ m}$ and its eigenvectors $|\phi_{\circ m}\rangle$. A second term $\mathcal{H}'$ in the Hamiltonian describes the "small perturbation". A *coupling constant* λ is "extracted" from $\mathcal{H}'$ to keep track of the order in the power expansion which will be performed:

$$\mathcal{H} = \mathcal{H}_\circ + \mathcal{H}' = \mathcal{H}_\circ + \lambda \mathcal{V}'. \tag{E.1}$$

The property of the perturbation of being "small" will be clarified in the development of the theory.

The basic idea of the perturbation theory is to develop the unknown quantities in powers of λ, i.e., of the perturbation, and to solve the resulting equations separately for the different orders. The convergence of the perturbation expansion is a serious problem. It will not be considered here, assuming that the perturbation is weak enough to ensure the convergence.

The perturbation Hamiltonian may be constant or time dependent. Correspondingly, two different developments of the theory are performed.

E.1 Time-Independent Perturbations

As indicated above, eigenvalues and eigenvectors of the unperturbed Hamiltonian are supposed to be known:

$$\mathcal{H}_\circ|\phi_{\circ m}\rangle = \epsilon_{\circ m}|\phi_{\circ m}\rangle, \quad \langle\phi_{\circ n}|\phi_{\circ m}\rangle = \delta_{nm}. \tag{E.2}$$

Our problem is to find eigenvalues ϵ_m and eigestates $|\phi_m\rangle$ of the total Hamiltonian:

$$\mathcal{H}|\phi_m\rangle = \epsilon_m|\phi_m\rangle. \tag{E.3}$$

Perturbation Expansion

Let us expand the unknowns in powers of λ:

$$|\phi_m\rangle = |\phi_m^{(0)}\rangle + \lambda|\phi_m^{(1)}\rangle + \lambda^2|\phi_m^{(2)}\rangle + \ldots, \quad \epsilon_m = \epsilon_m^{(0)} + \lambda\epsilon_m^{(1)} + \lambda^2\epsilon_m^{(2)} + \ldots. \tag{E.4}$$

Now substitute (E.4) into (E.3) and separate the various orders, obtaining

$$\begin{cases} \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}|\phi_m^{(0)}\rangle = 0 \\ \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}|\phi_m^{(1)}\rangle = [\epsilon_m^{(1)} - \mathcal{V}']|\phi_m^{(0)}\rangle \\ \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}|\phi_m^{(2)}\rangle = [\epsilon_m^{(1)} - \mathcal{V}']|\phi_m^{(1)}\rangle + \epsilon_m^{(2)}|\phi_m^{(0)}\rangle \\ \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}|\phi_m^{(3)}\rangle = [\epsilon_m^{(1)} - \mathcal{V}']|\phi_m^{(2)}\rangle + \epsilon_m^{(2)}|\phi_m^{(1)}\rangle + \epsilon_m^{(3)}|\phi_m^{(0)}\rangle \\ \ldots\ldots \end{cases}. \tag{E.5}$$

Representation of the Unperturbed Hamiltonian

The vectors in the above equations are now represented on the basis given by the eigenvectors of the unperturbed Hamiltonian in (E.2):

$$|\phi_m^{(i)}\rangle = \sum_n C_n^{(i)}(m)|\phi_{\circ n}\rangle \ .$$

The first three equations in (E.5) become

$$\begin{cases} \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}\sum_n C_n^{(0)}(m)|\phi_{\circ n}\rangle = 0 \\ \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}\sum_n C_n^{(1)}(m)|\phi_{\circ n}\rangle = [\epsilon_m^{(1)} - \mathcal{V}']\sum_n C_n^{(0)}(m)|\phi_{\circ n}\rangle \\ \{\mathcal{H}_\circ - \epsilon_m^{(0)}\}\sum_n C_n^{(2)}(m)|\phi_{\circ n}\rangle \\ \qquad = [\epsilon_m^{(1)} - \mathcal{V}']\sum_n C_n^{(1)}(m)|\phi_{\circ n}\rangle + \epsilon_m^{(2)}\sum_n C_n^{(0)}(m)|\phi_{\circ n}\rangle \end{cases}. \tag{E.6}$$

Zero Order: No Degeneracy

The first equation (E.6) may be written as

$$\sum_n \{\epsilon_{\circ n} - \epsilon_{\circ m}\}C_n^{(0)}(m)|\phi_{\circ n}\rangle = 0,$$

where it has been taken into account that the eigenvalues at zero order are the unperturbed eigenvalues. Because of the orthogonality of the basis vectors, the linear combination above requires that all coefficients be zero. If there is no degeneracy, $\epsilon_{\circ n} \neq \epsilon_{\circ m}$ for $n \neq m$. This means that in absence of degeneracy all coefficients $C_n^{(0)}(m)$ must vanish for $n \neq m$, while $C_m^{(0)}(m)$ has an undetermined value, which is taken equal to one for normalization. The obvious result is obtained that, in absence of degeneracy, the zero order eigenvalues and eigenvectors are the unperturbed ones:

$$\epsilon_m^{(0)} = \epsilon_{\circ m}, \quad C_n^{(0)}(m) = \delta_{nm}, \quad |\phi_m^{(0)}\rangle = |\phi_{\circ m}\rangle.$$

First-Order Eigenvalues: No Degeneracy

In absence of degeneracy, the perturbation theory now proceeds as follows: each equation in (E.6), starting from the second one, is projected first on the m-th axis, i.e., is multiplied by $\langle\phi_{\circ m}|$ on the left; the resulting equation yields the correction of the energy eigenvalue at the corresponding order; then, with this result, the same equation is projected on the other axes, i.e., it is multiplied by $\langle\phi_{\circ n}|$ with $n \neq m$; the resulting equations provide the coefficients $C_n^{(i)}(m)$ for the correction of the m-th eigenstate. The coefficient $C_m^{(i)}(m)$ remains undetermined, and may be fixed by the normalization. Actually, only its real part is fixed by this condition; it may be shown that the freedom of its imaginary part is related to the freedom in the phase of the eigenvector.

Following the procedure just indicated, the first-order correction of the energy eigenvalues is given by

$$\Delta\epsilon_m^{(1)} = \lambda\epsilon_m^{(1)} = \lambda\mathcal{V}'_{mm} = \langle\phi_{\circ m}|\mathcal{H}'|\phi_{\circ m}\rangle.$$

This result has a simple and intuitive physical interpretation: *The first-order correction of an energy level is given by the mean value of the perturbation in the corresponding unperturbed state*. The amount of this correction gives also an idea of whether the perturbation may be considered "small".

First-Order Eigenvectors: No Degeneracy

Following again the indicated prescription, the eigenvectors corrected to first-order in the perturbation are found to be

$$|\phi_m\rangle \approx |\phi_{\circ m}\rangle + \sum_{k\neq m} \frac{H'_{km}}{\epsilon_{\circ m} - \epsilon_{\circ k}} |\phi_{\circ k}\rangle.$$

The main term is given by the unperturbed state and is corrected by contributions coming from all other states. The corrections are large if the difference between the unperturbed energies of the state under examination and of the correcting state is small and if the corresponding matrix element of the perturbation is large. The value of their ratio gives also an indication about the applicability of the perturbation expansion.

In a similar way, we proceed with higher-order terms.

Removal of the Degeneracy

If some of the unperturbed eigenvalues are degenerate, it is necessary, as first step of the perturbation theory, to diagonalize the perturbation Hamiltonian between the states belonging to the degenerate eigenvalues. The combinations of unperturbed states, which diagonalize $\mathcal{V}'$ in the subspace of the degenerate eigenvalue, are still eigenstates of $\mathcal{H}_\circ$ and must be chosen as elements of the working basis. Having done so, everything proceeds as described above. Details can be found in all good textbooks of quantum mechanics as, e.g., [398].

E.2 Time-Dependent Perturbations

Let us now assume that the perturbation Hamiltonian $\mathcal{H}' = \lambda\mathcal{V}'$ in (E.1) is a function of time. The unperturbed Hamiltonian $\mathcal{H}_\circ$ is again supposed to be time independent, and its eigenvalues and eigenvectors are known, as indicated in (E.2).

Since the total Hamiltonian is now time dependent, we are not looking for its eigenstates but for the solution of the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = \mathcal{H}|\Psi(t)\rangle.$$

It is convenient, for this purpose, to use as basis the eigenstates of the unperturbed Hamiltonian, including their time dependence:

$$|\Phi_{\circ n}(t)\rangle = |\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)}\ , \quad \omega_{\circ n} = \epsilon_{\circ n}/\hbar. \tag{E.7}$$

The use of such a basis is equivalent to working in the interaction picture (see Sect. 2.2.4). On this basis, the state vector is written as

$$|\Psi(t)\rangle = \sum_n a_n(t)|\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)}, \tag{E.8}$$

and the Schrödinger equation becomes

$$i\hbar\sum_n \dot{a}_n(t)|\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)} + i\hbar\sum_n a_n(t)|\phi_{\circ n}\rangle(-i\omega_{\circ n})\mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)}$$

$$= \sum_n a_n(t)\epsilon_{\circ n}|\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)} + \sum_n \mathcal{H}'(t)a_n(t)|\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)}.$$

The second term of the l.h.s. cancels the first term of the r.h.s.: the unperturbed dynamics is totally described by the time-dependent basis, and disappears from the equation. It remains

$$i\hbar\sum_n \dot{a}_n(t)|\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)} = \sum_n \mathcal{H}'(t)a_n(t)|\phi_{\circ n}\rangle \mathrm{e}^{-\mathrm{i}\omega_{\circ n}(t-t_\circ)}.$$

Multiplying this equation on the left by $\langle\phi_{\circ m}|\mathrm{e}^{\mathrm{i}\omega_{\circ m}(t-t_\circ)}$, i.e., projecting it on the m-th axis, the result is

$$\dot{a}_m(t) = \frac{1}{i\hbar}\sum_n H'_{mn}(t)a_n(t)\mathrm{e}^{\mathrm{i}\omega_{mn}(t-t_\circ)}, \tag{E.9}$$

where

$$\omega_{mn} = \omega_{\circ m} - \omega_{\circ n} \qquad \text{and} \qquad H'_{mn}(t) = \langle\phi_{\circ m}|\mathcal{H}'(t)|\phi_{\circ n}\rangle.$$

Equation (E.9) is exact and shows how the Hamiltonian generates the evolution of the system mixing the different states according to its matrix elements.

At this point, we consider the perturbation expansion of the coefficients that define the unknown state vector in (E.8),

$$a_m(t) = a_m^{(0)}(t) + \lambda a_m^{(1)}(t) + \lambda^2 a_m^{(2)}(t) + \ldots,$$

and insert it into (E.9). After separating the various orders, we obtain

$$\dot{a}_m^{(0)}(t) = 0\ , \qquad \dot{a}_m^{(s+1)}(t) = \frac{1}{i\hbar}\sum_n V'_{mn}(t)a_m^{(s)}(t)\mathrm{e}^{\mathrm{i}\omega_{mn}(t-t_\circ)}.$$

The first of these equations shows that to zero order the state remains in its initial condition:

$$a_m^{(0)}(t) = \ cost\ = \ a_m^{(0)}(t_\circ),$$

while the other equations provide the coefficients at the successive orders.

Thus, the problem is formally solved. In particular, if we assume that at the initial time $t_\circ$ the system is in the initial state $|\phi_{\circ i}\rangle$ and we want to know the probability that at time t it is found in the final state $|\phi_{\circ f}\rangle$, the answer, to first order in the perturbation, is

$$P_{if}(t) = |a_f(t)|^2 = \left|\frac{1}{i\hbar}\int_{t_\circ}^{t}\langle\phi_{\circ f}|\mathcal{H}'(t')|\phi_{\circ i}\rangle\,\mathrm{e}^{\mathrm{i}\omega_{fi}t'}\,dt'\right|^2.$$

Harmonic Perturbations: Fermi Golden Rule

To proceed, it is necessary to assume a given form for $\mathcal{H}'(t)$. Taking into account that any function of t can be developed in Fourier series, and considering the practical importance of harmonic perturbations, we assume

$$\mathcal{H}' = \mathcal{F}\mathrm{e}^{-\mathrm{i}\omega t} + h.c. = \mathcal{F}\mathrm{e}^{-\mathrm{i}\omega t} + \mathcal{F}^\dagger\mathrm{e}^{\mathrm{i}\omega t},$$

where the presence of the Hermitian conjugate ensures the Hermiticity of the Hamiltonian. The matrix elements are

$$H'_{fi} = \langle\phi_{\circ f}|\mathcal{H}'(t)|\phi_{\circ i}\rangle = \langle\phi_{\circ f}|\mathcal{F}\mathrm{e}^{-\mathrm{i}\omega t} + \mathcal{F}^\dagger\mathrm{e}^{\mathrm{i}\omega t}|\phi_{\circ i}\rangle = F_{fi}\mathrm{e}^{-\mathrm{i}\omega t} + F_{if}^*\mathrm{e}^{\mathrm{i}\omega t}.$$

Thus, to first order,

$$a_f(t) = \frac{1}{i\hbar}\int_{t_\circ}^{t}\left\{F_{fi}\mathrm{e}^{-\mathrm{i}\omega t'} + F_{if}^{*}\mathrm{e}^{\mathrm{i}\omega t'}\right\}\mathrm{e}^{\mathrm{i}\omega_{fi}t'}\,dt'$$

$$= \frac{1}{i\hbar}\left\{F_{fi}\frac{\mathrm{e}^{\mathrm{i}(\omega_{fi}-\omega)t} - \mathrm{e}^{\mathrm{i}(\omega_{fi}-\omega)t_\circ}}{\mathrm{i}(\omega_{fi}-\omega)} + F_{if}^{*}\frac{\mathrm{e}^{\mathrm{i}(\omega_{fi}+\omega)t} - \mathrm{e}^{\mathrm{i}(\omega_{fi}+\omega)t_\circ}}{\mathrm{i}(\omega_{fi}+\omega)}\right\}. \quad \text{(E.10)}$$

As a function of the final state, one of the two terms becomes dominant when the denominator vanishes, i.e., when

$$\omega_{fi} \mp \omega \approx 0\ , \qquad \epsilon_f \approx \epsilon_i \pm \hbar\omega.$$

It is evident from energy considerations that the first term corresponds to the absorption of a quantum of the perturbing field, and the second to the emission.

Let us then assume that the final state has energy close to the one corresponding to an absorption or an emission, and keep only one term. We shall see that this approximation is in general reasonable, since the two terms become sharply peaked, and where one is essentially different from zero the other is practically zero.

The probability of finding the system at time t in the final state $|\phi_{\circ f}\rangle$ after an absorption is then

$$|a_f^{(ass)}(t)|^2 = \frac{1}{\hbar^2}|F_{fi}|^2\frac{|\mathrm{e}^{\mathrm{i}(\omega_{fi}-\omega)t} - \mathrm{e}^{\mathrm{i}(\omega_{fi}-\omega)t_\circ}|^2}{(\omega_{fi}-\omega)^2}$$

$$= \frac{2}{\hbar^2}|F_{fi}|^2\frac{1-\cos((\omega_{fi}-\omega)\Delta t)}{(\omega_{fi}-\omega)^2},$$

where $\Delta t = t - t_\circ$, or

$$|a_f^{(ass)}(t)|^2 = \frac{1}{\hbar^2}|F_{fi}|^2\frac{\sin^2\left[\frac{(\omega_{fi}-\omega)}{2}\Delta t\right]}{\left(\frac{\omega_{fi}-\omega}{2}\right)^2}. \quad \text{(E.11)}$$

The above formula gives the probability, to first order, of finding the system in the state $|\phi_{\circ f}\rangle$ Δt seconds after it was prepared in the state $|\phi_{\circ i}\rangle$, as a consequence of the absorption of a quantum of the perturbing field. Note that this probability is different from zero not only at exact energy conservation, but also "close" to it. A similar result is obtained for emission.

The second-order probability can be obtained with similar calculations. It corresponds to two absorptions, or two emissions, or one absorption and one emission in the two possible time orders.

It is convenient, at this point, to analyze the function that is present in the result (E.11). More precisely, let us consider the function

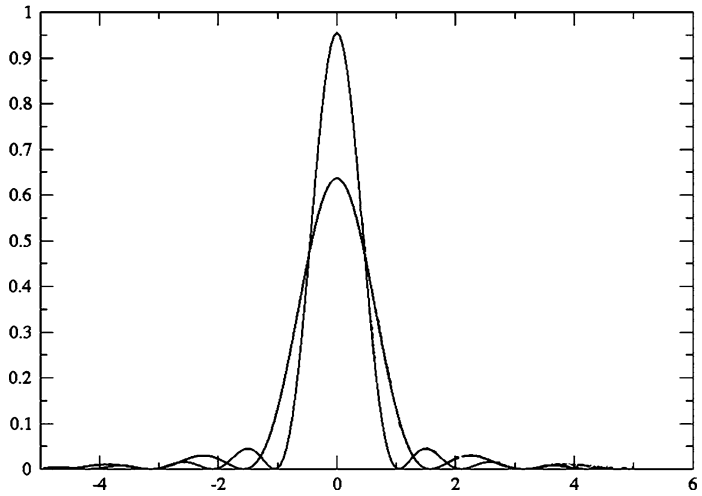


Fig. E.1. Function (E.12) appearing in the transition probability. At increasing time the function is more peaked. See text

$$f_t(\alpha) = \frac{\sin^2(\alpha t)}{\pi \alpha^2 t} \tag{E.12}$$

as a function of α, with t as a parameter. For $\alpha = 0$, the function is continuous and its value is t/π. Thus, the value in the origin diverges if $t \to \infty$. For $\alpha \neq 0$, on the contrary, the function oscillates between 0 and 1, as shown in Fig. E.1, enveloped by the curve $1/(\pi t \alpha^2)$. Thus,

$$\lim_{t\to\infty} f_t(\alpha) = \begin{cases} \infty & \text{if} \quad \alpha = 0 \\ 0 & \text{if} \quad \alpha \neq 0 \end{cases} .$$

Furthermore,

$$\int_{-\infty}^{\infty} f_t(\alpha)\, d\alpha = \int_{-\infty}^{\infty} \frac{\sin^2(\alpha t)}{\pi \alpha^2 t}\, d\alpha = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin^2 \xi}{\xi^2}\, \mathrm{d}\xi = 1.$$

For sufficiently long times Δt, we may thus approximate

$$f_t(\alpha) \approx \delta(\alpha).$$

Then (E.11) becomes

$$|a_f^{(ass)}(t)|^2 = \frac{\pi}{\hbar^2} |F_{fi}|^2 \Delta t \delta \left(\frac{\omega_{fi} - \omega}{2} \right) = \frac{2\pi}{\hbar} |F_{fi}|^2 \Delta t \delta(\epsilon_f - \epsilon_i - \hbar\omega).$$

and exact energy conservation is recovered. (The fact that this function is strongly peaked justifies considering the two terms in (E.10) separately.) The probability in (E.11) is proportional to Δt and we can define a probability per unit time, or *transition rate*, given by

$$P(i \to f) = \frac{2\pi}{\hbar} |F_{fi}|^2 \delta(\epsilon_f - \epsilon_i - \hbar\omega).$$

This is the famous Fermi golden rule. If the final state belongs to a continuum of states (which, for normalization purposes, will be considered as a very dense discrete spectrum), we may consider the probability of transition to any of the states with energy in the interval $d\epsilon$. The number dN of such states is given by $g(\epsilon)d\epsilon$, $g(\epsilon)$ being the density of states in energy. The Fermi golden rule takes then the form

$$\boxed{P(i \to f) = \tfrac{2\pi}{\hbar} |F_{fi}|^2 \delta(\epsilon_f - \epsilon_i - \hbar\omega) g(\epsilon) d\epsilon} \tag{E.13}$$

Discussion on the Validity of the Fermi Golden Rule and Collisional Broadening

From the above derivation of the Fermi golden rule, it is clear that it holds true, with the δ of energy conservation, for sufficiently long times Δt. The width of the central peak of the function considered in (E.12) may be estimated as the distance between the two central zeros. These occur when $\alpha \Delta t = \pm\pi$. The transition may thus end to a state with energy in an interval of the order of $\Delta\epsilon$ such that

$$\Delta\epsilon \Delta t \approx \pi\hbar,$$

coherently with the uncertainty relation between the uncertainty of the energy of the state, also called *collisional broadening*, and its lifetime, both due to the perturbation Hamiltonian.

A final point to be considered is the following. The transition probability has been found proportional to the time Δt. For the validity of the perturbation theory to first order, it is necessary that the probability remains much smaller than unity. Thus, Δt must be short enough. On the contrary, for the approximation of the δ of energy conservation, the same Δt has been assumed infinitely long. There seems to be a severe contradiction. It can be noted, however, that the uncertainty of the energy of the state, i.e., the width of the central peek of the function $f_t(\alpha)$, is independent of the strength of the perturbation, while the transition probability depends upon it quadratically.

We may thus conclude that the Fermi golden rule with the δ of energy conservation is valid for a time sufficiently long to neglect the uncertainty on the energy of the final state, if the perturbation is sufficiently weak to reach this time before the transition probability becomes comparable with one. In the opposite case, it is necessary to apply the theory at higher orders.

References

1. A. Abramo, Int. J. High Speed Electron. Syst. **13**, 701, (2003)
2. A. Abramo, R. Brunetti, C. Jacoboni, F. Venturi, E. Sangiorgi, J. Appl. Phys. **76**, 5786 (1994)
3. A.A. Abrikosov, L.P. Gorkov, I.E. Dzyaloshinski, *Methods of Quantum Field Theory in Statistical Physics* (Dover Publications, New York, 1975)
4. I. Adawi, Phys. Rev. **115**, 1152 (1959)
5. Y. Aharonov, D. Bohm, Phys. Rev. **115**, 485 (1959)
6. H. Akera, T. Ando, Phys. Rev. B **41**, 11967 (1990)
7. H. Akera, T. Ando, Phys. Rev. B **43**, 11676 (1991)
8. M.P. Anantram, Appl. Phys. Lett. **78**, 2055 (2001)
9. M.P. Anantram, S. Datta, Y. Xue, Phys. Rev. B **61**, 14219 (2000)
10. P.W. Anderson, Phys. Rev. **109**, 1492 (1958)
11. T. Ando, J. Phys. Soc. Jpn. **43**, 1616 (1977)
12. T. Ando, A.B. Fowler, F. Stern, Rev. Mod. Phys. **54**, 437 (1982)
13. H. Aoki, T. Ando, Solid State Commun. **38**, 1079 (1981)
14. G.Arfken, *Mathematical Methods for Physicists*, 3rd edn. (Academic, Boston, 1985)
15. P.N. Argyres, Phys. Rev. **132**, 1527 (1963)
16. V.K. Arora, Phys. Rev. B **23**, 5611 (1981)
17. A. Asenov, A.R. Brown, J.H. Davies, S. Kaya, G. Slavcheva, IEEE Trans. Electron. Dev. **50**, 1837 (2003)
18. N.W. Ashcroft, N.D. Mermin, *Solid State Physics*, (Saunders College, Philadelphia, 1976)
19. J.E. Avron, D. Osadchy, R. Seiler, Phys. Today **56**, 39 (2003)
20. G. Baccarani, C. Jacoboni, A.M. Mazzone, Solid State Electron. **20**, 5 (1977)
21. G. Baccarani, P. Ostoja, Solid State Electron. **18**, 579 (1975)
22. A. Baldereschi, S. Baroni, R. Resta, Phys. Rev. Lett. **61**, 734 (1988)
23. N. Balkan (ed.), *Hot Electrons in Semiconductors*, (Clarendon, Oxford, 1998)
24. J. Bardeen, W.H. Brattain, Phys. Rev. **74**, 230 (1948)
25. J.R. Barker, J. Phys. C Solid State Phys. **6**, 2663 (1973)
26. J.R. Barker, Solid State Electron. **21**, 267 (1978)
27. C.H.W. Barnes, J.M. Shilton, A.M. Robinson, Phys. Rev. B **62**, 8410 (2000)
28. K. Barnham, D. Vvedensky (eds.), *Low-Dimensional Semiconductor Structures: Fundamentals and Device Applications* (Cambridge University Press, Cambridge, 2001)

29. S. Baroni, S. de Gironcoli, A. Dal Corso, P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001)
30. G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures* (Les éditions de physique, Les Ulis Cedex, 1990)
31. R. Battisti, A. Corrado, Energy **30**, 952 (2005)
32. J.C. Bean, *Materials and Technologies for High-Speed Devices* in *High-Speed Semiconductor Devices* (Wiley, New York, 1990)
33. C.W.J. Beenakker, H. van Houten, Solid State Phys. **44**, 1 (1991)
34. G. Bergmann, Phys. Rep. **107**, 1 (1984)
35. A. Bertoni, P. Bordone, R. Brunetti, C. Jacoboni, S. Reggiani, Phys. Rev. Lett. **84**, 5912 (2000)
36. A. Bertoni, P. Bordone, R. Brunetti, C. Jacoboni, S. Reggiani, J. Mod. Opt. **49**, 1219 (2002)
37. A. Bertoni, P. Bordone, R. Brunetti, C. Jacoboni, J. Phys. Condens. Matter **11**, 5999 (1999)
38. A. Bertoni, P. Bordone, G. Ferrari, N. Giacobbi, C. Jacoboni, J. Comput. Electron. **2**, 137 (2003)
39. B.A. Biegel, J.D. Plummer, Phys. Rev. B **54**, 8070 (1996)
40. K. Binder, D.W. Heermann, *Monte Carlo Simulation in Statistical Physics*, 3rd edn. (Springer, Berlin, 1997)
41. G.L. Bir, G.E. Pikus, Sov. Phys. Solid State **2**, 2039 (1961)
42. G.L. Bir, G.E. Pikus, *Symmetry and Strain-Induced Effects in Semiconductors* (Wiley, New York, 1974)
43. J.L. Birman, Phys. Rev. **127**, 1093 (1962)
44. J.L. Birman, M. Lax, R. Loudon, Phys. Rev. **145**, 620 (1966)
45. D.J. Bishop, D.C. Tsui, R.C. Dynes, Phys. Rev. Lett. **44**, 1153 (1980)
46. A. Blacha, H. Presting, M. Cardona, Phys. Stat. Sol. (b) **126**, 11 (1984)
47. J.S. Blakemore, *Semiconductor Statistics* (Dover Publications, New York, 1987)
48. X. Blase, L.X. Benedict, E.L. Shirley, S.G. Louie, Phys. Rev. Lett. **72**, 1878 (1994)
49. A.D. Boardman, W. Fawcett, H.D. Rees, Solid State Commun. **6**, 305 (1968)
50. U. Bockelmann, G. Bastard, Phys. Rev. B **42**, 8947 (1990)
51. P. Bordone, A. Bertoni, R. Brunetti, C. Jacoboni, VLSI Des. **13**, 211 (2001)
52. P. Bordone, C. Jacoboni, J. Comp. Electron. **1**, 67 (2002)
53. P. Bordone, C. Jacoboni, P. Lugli, L. Reggiani, P. Kocevar, Physica **134B**, 69 (1985), J. Appl. Phys. **61**, 1460 (1987)
54. P. Bordone, P. Lugli, Phys. Rev. B **49**, 8178 (1994)
55. P. Bordone, M. Pascoli, R. Brunetti, A. Bertoni, C. Jacoboni, A. Abramo, Phys. Rev. B **59**, 3060 (1999)
56. P. Bordone, D. Vasileska, D.K. Ferry, Phys. Rev. B **53**, 3846 (1996)
57. M. Born, K. Huang, *Dynamical Theory of Crystal Lattice* (Oxford University Press, Oxford, 1988)
58. M. Born, J.R. Oppenheimer, Ann. Phys. (Leipzig) **84**, 457 (1927)
59. V. Borsari, C. Jacoboni, Phys. Stat. Sol. (b) **54**, 649 (1972)
60. S. Bosi, C. Jacoboni, J. Phys. C **9**, 315 1976
61. S. Bosi, C. Jacoboni, L. Reggiani, J. Phys. C **12**, 1525 (1979)
62. M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, K. Stokbro, Phys. Rev. B **65**, 165401 (2002)

63. R. Bray, W.E. Pinson, Phys. Rev. Lett. **11**, 268 (1963)
64. K.F. Brennan, P.P. Ruden (eds.), *Topics in High Field Transport in Semiconductors* (World Scientific, Singapore, 2001)
65. H. Brooks, C. Herring, Phys. Rev. **83**, 879 (1951)
66. M.R. Brozel, G.E. Stillman (eds.), *Properties of Gallium Arsenide*, 3rd edn. (EMIS Datareviews Series No. 16, INSPEC, IEE London, UK, 1996)
67. T. Brudevoll, B. Lund, T.A. Fjeldly, J. Appl. Phys. **71**, 4972 (1992)
68. S. Bruers, C. Maes, K. Netocny, J. Stat. Phys. **129**, 725 (2007)
69. R. Brunetti, A. Bertoni, P. Bordone, C. Jacoboni, VLSI Des. **13**, 375 (2001)
70. R. Brunetti, C. Jacoboni, Phys. Rev. Lett. **50**, 1164 (1983)
71. R. Brunetti, C. Jacoboni, in *Semiconductors Probed by Ultrafast Laser Spectroscopy, Vol I*, ed. by R.R. Alfano (Academic, New York, 1985), p. 367
72. R. Brunetti, C. Jacoboni, F. Nava, L.Reggiani, G. Bosman, R.J.J. Zijlstra, J. Appl. Phys. **52**, 6713 (1981)
73. R. Brunetti, C. Jacoboni, F. Rossi, Phys. Rev. B **39**, 10781 (1989)
74. R. Brunetti, C. Jacoboni, Phys. Rev. B **57**, 1723 (1998)
75. W. Brütting (ed.), *Physics of Organic Semiconductors* (Wiley-VHC, Weinheim, 2005)
76. H. Budd, in *Proceedings of the International Conference on the Physics of Semiconductors, Kyoto*, J. Phys. Soc. Jpn. Suppl. **21**, 420 (1966)
77. J. Bude, in *Monte Carlo Device Simulation: Full Band and Beyond*, ed. by K. Hess (Kluwer Academic, Boston, 1991), p.27
78. A.K. Buin, A. Verma, M.P. Anantram, J. Appl. Phys. **104**, 053716 (2008)
79. F.A. Buot, K.L. Jensen, Phys. Rev. B **42**, 9429 (1990)
80. M.G. Burt, Phys. Rev. B **50**, 7518 (1994)
81. M.G. Burt, Appl. Phys. Lett. **65**, 717 (1994)
82. P.N. Butcher, Rep. Prog. Phys. **30**, 97 (1967)
83. M. Büttiker, Phys. Rev. Lett. **57**, 1761 (1986)
84. M. Büttiker, Phys. Rev. B **38**, 9375 (1988)
85. M. Büttiker, Y. Imry, R. Landauer, S. Pinhas, Phys. Rev. B **31**, 6207 (1985)
86. H.B. Callen, Phys. Rev. **76**, 1394 (1949)
87. H.B. Callen, T. Welton, Phys. Rev. **83**, 34 (1951)
88. C. Canali, C. Jacoboni, F. Nava, G. Ottaviani, A. Alberigi Quaranta, Phys. Rev. B **12**, 2265 (1975)
89. E. Cancellieri, P. Bordone, A. Bertoni, G. Ferrari, C. Jacoboni, J. Comput. Electron. **3**, 411 (2004)
90. E. Cancellieri, P. Bordone, C. Jacoboni, Phys. Rev. B **76**, 214301 (2007)
91. F. Capasso, Science **235**, 172 (1987)
92. F. Capasso, A.Y. Cho, Surf. Sci. **299–300**, 878 (1994)
93. M. Cardona, F.H. Pollak, Phys. Rev. **142**, 530 (1966)
94. C. Caroli, R. Combescot, P. Nozières, D. Saint-James, J. Phys. C **4**, 916 (1971)
95. C. Caroli, R. Combescot, D. Lederer, P. Nozières, D. Saint-James, J. Phys. C **4**, 2598 (1971)
96. P. Carruthers, F. Zachariasen, Rev. Mod. Phys. **55**, 245 (1983)
97. A.H. Castro Neto, F. Guinea, N.M.R. Peres, K.S. Novoselov, A.K. Geim, Rev. Mod. Phys. **81**, 109 (2009)
98. D.M. Caughey, R.F. Thomas, Proc. IEEE **55**, 2192 (1967)
99. R.G. Chambers, Proc. Phys. Soc. Lond. A **65**, 458 (1952)
100. L.L. Chang, L. Esaki, R. Tsu, Appl. Phys. Lett. **24**, 593 (1974)

101. D.M. Chapin, C.S. Fuller, G.L. Pearson, J. Appl. Phys. **25**, 676 (1954)
102. J.-C. Charlier, X. Blase, S. Roche, Rev. Mod. Phys. **79**, 677 (2007)
103. D. Chattopadhyay, H.J. Queisser, Rev. Mod. Phys. **53**, 745 (1981)
104. J.R. Chelikowsky, M.L. Cohen, Phys. Rev. B **14**, 556 (1976)
105. E. Ciancio, R.C. Iotti, F. Rossi, Semicond. Sci. Technol. **19**, S212 (2004)
106. M.L. Cohen, T.K. Bergstresser, Phys. Rev. **141**, 789 (1966)
107. E.M. Conwell, *High Field Transport in Semiconductors* (Academic, New York, 1967)
108. E.M. Conwell, M.O. Vassell, Phys. Rev. **166**, 797 (1968)
109. E.M. Conwell, V. Weisskopf, Phys. Rev. **77**, 388 (1950)
110. M. Costato, L. Reggiani, Lettere al Nuovo Cimento **4**, 848 (1970)
111. M. Costato, C. Jacoboni, L. Reggiani, Phys. Stat. Sol (b) **52**, 461 (1972)
112. M. Costato, L. Reggiani, Phys. Stat. Sol (b) **58**, 47 and 471 (1973)
113. R.A. Craig, J. Math. Phys. **9**, 605 (1968)
114. V.L. Dalal, A.B. Dreeben, A. Triano, J. Appl. Phys. **42**, 2864 (1971)
115. C.G. Darwin, Proc. Roy. Soc. **154**, 61 (1936)
116. S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge, 1995)
117. S. Datta, M.R. Melloch, S. Bandyopadhyay, R. Noren, M. Vaziri, M. Miller, R. Reifenberger, Phys. Rev. Lett. **55**, 2344 (1985)
118. R. Dewar, J. Phys. A Math. Gen. **38**, L371 (2005)
119. R. Dingle, H.L. Störmer, A.C. Gossard, W. Wiegmann, Appl. Phys. Lett. **33**, 665 (1978)
120. P.A.M. Dirac, *The Principles of Quantum Mechanics*, 4th edn. (Oxford University Press, London, 1958)
121. R.M. Dreizler, E.K.W. Gross, *Density Functional Theory* (Springer, Berlin, 1990)
122. G. Dresselhaus, A.F. Kip, C. Kittel, Phys. Rev. **98**, 368 (1955)
123. O. Dubay, G. Kresse, Phys. Rev. B **67**, 035401 (2003)
124. H.B. Dwight, *Tables of Integrals and Other Mathematical Data*, 4th edn. (The Macmillan Company, New York, 1961)
125. D.J. Eaglesham, M. Cerullo, Phys. Rev. Lett. **64**, 1943 (1990)
126. L.R.A. Eaves, R.A. Hoult, R.A. Stradling, R.J. Tidey, J.C. Portal, S. Askenazy, J. Phys. C. **8**, 1034 (1975)
127. R. Egger, A. Bachtold, M.S. Fuhrer, M. Bockrath, D.H. Cobden, P.L. McEuen, in *Luttinger Liquid Behavior in Metallic Carbon Nanotubes*, eds. R. Haug, H. Schoeller. *Interacting Electrons in Nanostructures* (Springer, Berlin, 2001)
128. H. Ehrenreich, J. Phys. Chem. Solids **2**, 131 (1957)
129. H. Ehrenreich, J. Phys. Chem. Solids **9**, 129 (1959)
130. H. Ehrenreich, A.W. Overhauser, Phys. Rev. **104**, 331 (1956)
131. C. Erginsoy, Phys. Rev. **79**, 1013 (1950)
132. L. Esaki, R. Tsu, IBM J. Res. Dev. **14**, 61 (1970)
133. L. Esaki, L.L. Chang, Phys. Rev. Lett. **33**, 495 (1974)
134. J. Faist, F. Capasso, D.L. Sivco, C. Sirtori, A.L. Hutchinson, A.Y. Cho, Science **264**, 553 (1994)
135. J.-P. Farges (ed.), *Organic Conductors: Fundamentals and Applications* (Marcel Dekker, New York, 1994)
136. G. Fasol, H. Sakaki, Phys. Rev. Lett. **70**, 3643 (1993)
137. W. Fawcett, A.D. Boardman, S. Swain, J. Phys. Chem. Solids **31**, 1963 (1970)

138. W. Fawcett, H.D. Rees, Phys. Lett. A **29**, 578 (1969)
139. E. Fermi, Nuovo Cimento **11**, 157 (1934)
140. G. Ferrari, N. Giacobbi, P. Bordone, A. Bertoni, C. Jacoboni, Semicond. Sci. Technol. **19**, S254 (2004)
141. D.K. Ferry, Phys. Rev. B **14**, 1605 (1976)
142. D.K. Ferry, Surface Sci. **75**, 86 (1978)
143. D.K. Ferry, S.M. Goodnick, *Transport in Nanostructures* (Cambridge University Press, Cambridge, 1997)
144. D.K. Ferry, I. Knezevic, S.M. Ramey, L. Shifren, 2003 in *Progress in Nonequilibrium Green's functions II* (World Scientific, Singapore, 2003), p. 127
145. A.L. Fetter, J.D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971)
146. R.P. Feynman, F.L. Vernon, Ann. Phys. **24**, 118 (1963)
147. R.P. Feynman, A.R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965)
148. M.V. Fischetti, S.E. Laux, Phys. Rev. B **38**, 9721 (1988)
149. M.V. Fischetti, Z. Ren, P.M. Solomon, M. Yang, K. Rim, J. Appl. Phys. **94**, 1079 (2003)
150. G. Fishman, Phys. Rev. B **34**, 2394 (1986)
151. V.I. Fistul, *Heavily Doped Semiconductors* (Plenum, New York, 1969)
152. K. Fletcher, P.N. Butcher, J. Phys. C Solid State Phys. **5**, 212 (1972)
153. A. Forghieri, R. Guerrieri, P. Ciampolini, A. Gnudi, M. Rudan, G. Baccarani, IEEE Trans. Comput. Aided Des. **CAD-7**, 231 (1988)
154. E.Z. Francis, *Quantum Hall Effects: Field Theoretical Approach and Related Topics*, 3rd edn. (World Scientific, Singapore, 2008)
155. W.R. Frensley, Rev. Mod. Phys. **62**, 745 (1990)
156. H. Frölich, Proc. Roy. Soc. **A160**, 230 (1937), **A188**, 521 (1947)
157. H. Frölich, B.V. Paranjape, Proc. Phys. Soc. **B69**, 21 (1956)
158. H. Fu, L.-W. Wang, A. Zunger, Phys. Rev. B **57**, 9971 (1998)
159. T. Furuta, M. Tomizawa, Appl. Phys. Lett. **56**, 824 (1990)
160. J. García-García, X. Oriols, F. Martín, J. Suñé, Solid State Electron. **39** 1795 (1996)
161. L. Gherardi, A. Pellacani, C. Jacoboni, Lettere al Nuovo Cimento **14**, 225 (1975)
162. A.K. Geim, K.S. Novoselov, Nat. Mater. **6**, 183 (2007)
163. P. Giannozzi, S. de Girancoli, P. Pavone, S. Baroni, Phys. Rev. B **43**, 7231 (1991)
164. G. Gilat, L.J. Raubenheimer, Phys. Rev. **144**, 390 (1966)
165. N. Giordano, Phys. Rev. B **22**, 5635 (1980)
166. D. Giulini, E. Joos, C. Kiefer, J. Kupsch, I.-O. Stamatescu, H.D. Zeh, *Decoherence and the Appearance of a Classical World in Quantum Theory* (Springer, Berlin, 1996)
167. V.J. Goldman, D.C. Tsui, J.E. Cunningham, Phys. Rev. Lett. **58**, 1256 (1987)
168. H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, 1959)
169. T. González Sánchez, J.E. Velázquez Pérez, P.M. Gutierrez Conde, D. Pardo Collantes, Semicond. Sci. Technol. **6**, 862 (1991)
170. S.M. Goodnick, R.G. Gann, J.R. Sites, D.K. Ferry, C.W. Wilmsen, D. Fathy, O.L. Krivanek, J. Vac. Sci. Technol. B **1**, 803 (1983)
171. S.M. Goodnick, P. Lugli, Phys. Rev. B **37**, 2578 (1988)

172. W. Greiner, *Quantum Mechanics, an Introduction*, 3rd edn. (Springer, Berlin, 2007)
173. E.K.U. Gross, E. Runge, O. Heinonen, *Many-Particle Theory* (Adam Hilger, Bristol, 1991)
174. J.B. Gunn, Solid State Commun. **1**, 88 (1963)
175. M. Hadjazi, J.F. Palmier, A. Sibille, H. Wang, E. Paris, F. Mollot, Electron. Lett. **29**, 648 (1993)
176. C. Hamaguchi, *Basic Semiconductor Physics* (Springer, Berlin, 2001)
177. C. Hammar, Phys. Rev. B **4**, 417 (1971)
178. J.M. Hammersley, D.C. Handscomb, *Monte Carlo Methods* (Methuen & Co Ltd, London, 1967)
179. W. Hänsch, G.D. Mahan, Phys. Rev. B **28**, 1886 (1983)
180. W. Hänsch, G.D. Mahan, Phys. Rev. B **28**, 1902 (1983)
181. W.A. Harrison, Phys. Rev. **104**, 1281 (1956)
182. J.W. Harrison, J.R. Hauser, Phys. Rev. B **13**, 5347 (1976)
183. J.W. Harrison, J.R. Hauser, J. Appl. Phys. **47**, 292 (1976)
184. H. Haug, A.-P. Jauho, *Quantum Kinetics in Transport and Optics of Semiconductors*, 2nd edn. (Springer, Berlin, 2008)
185. C. Herring, E. Vogt, Phys. Rev. **101**, 944 (1956)
186. K. Hess, H. Morkoc, H. Shishijo, B.G. Streetman, Appl. Phys. Lett. **35**, 469 (1979)
187. K. Hess (ed.), *Monte Carlo Device Simulation: Full Band and Beyond* (Kluwer Academic, Boston, 1991)
188. D.E. Hill, J. Appl. Phys. **41**, 1815 (1970)
189. M. Hillery, R.F. O'Connell, M.O. Scully, E.P. Wigner, Phys. Rep. **106**, 121 (1984)
190. C. Hilsum, Electron. Lett. **10**, 259 (1974)
191. P. Hohenberg, W. Kohn, Phys. Rev. **136**, B864 (1964)
192. W.V. Houston, Phys. Rev. **57**, 184 (1940)
193. D.J. Howarth, E.H. Sondheimer, Proc. Roy. Soc. **A219**, 406 (1953)
194. K. Huang, *Statistical Mechanics* (Wiley, New York, 1988)
195. K. Husimi, Proc. Phys. Math. Soc. Jpn. **22**, 264 (1940)
196. H. Ibach, H. Lüth, *Solid-State Physics* (Springer, Berlin, 1991)
197. S. Iijima, Nature (London) **354**, 56 (1991)
198. J.C. Inkson, *Many-Body Theory of Solids* (Plenum, New York, 1984)
199. R. Ionicioiu, G. Amaratunga, F. Udrea, Int. J. Mod. Phys. B **15**, 125 (2001)
200. J.C. Irvin, Bell Syst. Tech. J. **41**, 387 (1962)
201. K. Ismail, M. Arafa, K.L. Saenger, J.O. Chu, B.S. Meyerson, Appl. Phys. Lett. **66**, 1077 (1995)
202. J.D. Jackson, *Classical Electrodynamics*, 2nd edn. (Wiley, New York, 1975)
203. C. Jacoboni, Phys. Stat. Sol. (b) **65**, 61 (1974)
204. C. Jacoboni, in *Proceedings of the 13th International Conference on the Physics of Semiconductors*, ed. by G. Fumi (Marves, Roma, 1976), p. 1195
205. C. Jacoboni, J. Lumin. **30**, 120 (1985)
206. C. Jacoboni, Proc. SPIE **942**, 32 (1988)
207. C. Jacoboni, R. Brunetti, S. Monastra, Phys. Rev. B **68** 125205 (2003)
208. C. Jacoboni, C. Canali, G. Ottaviani, A. Alberigi Quaranta, Solid State Electron. **20**, 77 (1977)
209. C. Jacoboni, P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation* (Springer, Wien, 1989)

210. C. Jacoboni, R. Minder, G. Majni, J. Phys. Chem. Solids **36**, 1129 (1975)
211. C. Jacoboni, P. Poli, L. Rota, Solid State Electron. **31**, 523 (1988)
212. C. Jacoboni, L. Reggiani, Phys. Lett. **33A**, 333 (1970)
213. C. Jacoboni, L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983)
214. C. Jacoboni, L. Reggiani, Adv. Phys. **4**, 493 (1979)
215. C. Jacoboni, L. Reggiani, R. Brunetti, in *Proceedings of the 3rd International Conference on Hot Carriers in Semiconductors*, J. Phys. (Paris) Colloq. **42**, C7-123 (1981)
216. C. Jacoboni, R. Brunetti, P. Bordone, A. Bertoni, in *Topics in High Field Transport in Semiconductors* (World Scientific, Singapore, 2001), p. 25
217. A.-P. Jauho, *Green's Function Methods: Nonequilibrium, High-Field Transport* in: *Quantum Transport in Semiconductors*, ed. by D.K. Ferry, C. Jacoboni (Plenum, New York, 1992), p. 101
218. A.-P. Jauho, J.W. Wilkins, Phys. Rev. B **29**, 1919 (1984)
219. A. Janner, L. Van Hove, E. Verboven, Physica **28**, 1341 (1962)
220. B. Jeckelmann, B. Jeanneret, Meas. Sci. Technol. **14**, 1229 (2003)
221. K.L. Jensen, F.A. Bout, J. Appl. Phys. **65**, 5248 (1989)
222. K.L. Jensen, F.A. Buot, Phys. Rev. Lett. **66**, 1078 (1991)
223. J.C. Johnson, H.-J. Choi, K.P. Knutsen, R.D. Schaller, P. Yang, R.J. Saykally, Nat. Mat. **1**, 106 (2002)
224. A. Jorio, M.S. Dresselhaus, G. Dresselhaus (eds.), *Carbon Nanotubes: Advanced Topics in the Synthesis, Structure, Properties and Applications* (Springer, Berlin, 2008)
225. R. Joshi, R.O. Grondin, Appl. Phys. Lett. **54**, 2438 (1989)
226. C. Jungemann, S. Keith, M. Bartels, B. Meinerzhagen, IEICE Trans. Electron. **E82-C**, 870 (1999)
227. C. Jungemann, B. Meinerzhagen, *Hierarchical Device Simulation – The Monte Carlo Perspective* (Springer, Wien, 2003)
228. L.P. Kadanoff, G. Baym, *Quantum Statistical Mechanics* (Benjamin/ Cummings, Reading, 1962)
229. P.B. Kahn, *Mathematical Methods for Scientists and Engineers* (Wiley, New York 1990, Dover 2004)
230. M.H. Kalos, P.A. Whitlock, *Monte Carlo Methods, Volume I Basics* (Wiley, New York, 1986)
231. E.O. Kane, J. Phys. Chem. Solids **1**, 82 (1956)
232. E. Kapon, D.M. Hwang, R. Bhat, Phys. Rev. Lett. **63**, 430 (1989)
233. A. Kastalsky, S. Luryi, IEEE Electron. Dev. Lett. **4**, 334 (1983)
234. M.A. Kastner, Rev. Mod. Phys. **64**, 849 (1992)
235. M. Keever, H. Shishijo, K. Hess, S. Banerjee, L. Witkowski, H. Morkoc, B.G. Streetman, Appl. Phys. Lett. **38**, 36 (1981)
236. M.J. Kelly, *Low-Dimensional Semiconductors* (Oxford University Press, New York, 1995)
237. F.S. Khan, J.H. Davies, J.W. Wilkins, Phys. Rev. B **36**, 2578 (1987)
238. R.A. Kiehl, T.C.L.G. Sollner (eds.), *High Speed Heterostructure Devices*, Semiconductors and Semimetals **41** (Academic, San Diego, 1994)
239. K.W. Kim, M.A. Stroscio, A. Bhatt, R. Mickevicius, V.V. Mitin, J. Appl. Phys. **70**, 319 (1991)
240. J. Kirkwood, J. Chem. Phys. **7**, 39 (1939)
241. C. Kittel, *Introduction to Solid State Physics*, 8th edn. (Wiley, New York, 2004)

242. M.E. Klausmeier-Brown, M.R. Meloch, M.S. Lundstrom, Appl. Phys. Lett. **56**, 160 (1990)
243. G. Klimeck, S.S. Ahmed, N. Kharche, M. Korkusinski, M. Usman, M. Prada, T.B. Boykin, IEEE Trans. Electron. Dev. **54**, 2090 (2007)
244. K.V. Klitzing, G. Dorda, M. Pepper, Phys. Rev. Lett. **45**, 494 (1980)
245. N.C. Kluksdahl, A.M. Kriman, D.K. Ferry, C. Ringhofer, Phys. Rev. B **39**, 7720 (1989)
246. M. Kociak, A.Yu. Kasumov, S. Guéron, B. Reulet, I.I. Khodos, Yu.B. Gorbatov, V.T. Volkov, L. Vaccarini, H. Bouchiat, Phys. Rev. Lett. **86**, 2416 (2001)
247. M. Kohler, Z. Phys. **124**, 772 (1948)
248. M. Kohler, Z. Phys. **125**, 679 (1949)
249. W. Kohn, L.J. Sham, Phys. Rev. A **140**, 1133 (1965)
250. K. Kometer, G. Zandler, P. Vogl, Phys. Rev. B **46**, 1382 (1992)
251. D. Kranzer, J. Phys. C Solid State Phys. **6**, 2967 (1973)
252. R. de L. Kronig, W.G. Penney, Proc. Roy. Soc. **A130**, 499 (1930)
253. R. Kubo, J. Phys. Soc. Jpn. **12**, 570 (1957)
254. R. Kubo, Rep. Prog. Phys. **29**, 255 (1966)
255. R. Kubo, M. Toda, N. Hashitsume, *Statistical Physics II, Nonequilibrium Statistical Mechanics*, 2nd edn. (Springer, Berlin, 1992)
256. T. Kuhn, F. Rossi, Phys. Rev. B **46**, 7496 (1992)
257. T. Kunikiyo, M. Takenaka, Y. Kamakura, M. Yamaji, H. Mizuno, M. Morifuji, K. Taniguchi, C. Hamaguchi, J. Appl. Phys. **75**, 297 (1994)
258. T. Kurosawa, in *Proceedings of the International Conference on the Physics of Semiconductors, Kyoto*, J. Phys. Soc. Jpn. Suppl. **21**, 424 (1966)
259. T. Kurosawa, H. Maeda, J. Phys. Soc. Jpn. **31**, 668 (1971)
260. D.P. Landau, K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cabridge, 2000)
261. L.D. Landau, E.M. Lifshitz, *Quantum Mechanics, Non-Relativistic Theory* (Pergamon, Oxford, 1958)
262. R. Landauer, IBM J. Res. Dev. **1**, 223 (1957)
263. R. Landauer, Philos. Mag. **21**, 863 (1970)
264. R.B. Laughlin, Phys. Rev. B **23**, 5632 (1981)
265. R.B. Laughlin, Phys. Rev. Lett. **50**, 1395 (1983)
266. P. Lawaetz, Phys. Rev. **174**, 867 (1968)
267. P. Lawaetz, Phys. Rev. **183**, 730 (1969)
268. M. Lax, J.J. Hopfield, Phys. Rev. **124**, 115 (1961)
269. M. Lax, J.L. Birman, Phys. Stat. Sol. (b) **49**, K153 (1972)
270. J.P. Leburton, Phys. Rev. B **45**, 11022 (1992)
271. P.A. Lebwhol, J. Appl. Phys. **44**, 1744 (1973)
272. P.A. Lebwhol, P.J. Price, Solid State Commun. **9**, 1221 (1971)
273. P.A. Lebwhol, P.J. Price, Appl. Phys. Lett. **19**, 530 (1971)
274. P.A. Lee, Physica **140A**, 169 (1986)
275. J. Lee, H.N. Spector, J. Appl. Phys. **54**, 3921 (1983)
276. J. Lee, H.N. Spector, J. Appl. Phys. **57**, 366 (1985)
277. J. Lee, M.O. Vassel, J. Phys. C Solid State Phys. **17**, 2525 (1984)
278. C.S. Lent, D.J. Kirkner, J. Appl. Phys. **67**, 6353 (1990)
279. K. Leo, P.H. Bolivar, F. Brüggermann, R. Schwedler, K. Köhler, Solid State Commun. **84**, 943 (1992)

280. Z.M. Li, Z.K. Tang, H.J. Liu, N. Wang, C.T. Chan, R. Saito, S. Okada, G.D. Li, J.S. Chen, N. Nagasawa, S. Tsuda, Phys. Rev. Lett. **87**, 127401 (2001)
281. D. Liebig, *Proceedings of 6th International Conference on SISDEP*, Erlangen, Germany (Springer, Wien, 1995)
282. J. Lin, L.C. Chiu, J. Appl. Phys. **57**, 1373 (1985)
283. P. Lipavský, F.S. Kahn, A. Kalvovà, J.W. Wilkins, Phys. Rev. B **43**, 6650 (1991)
284. P. Lipavský, V. Špička, B. Velický, Phys. Rev. B **34**, 6933 (1986)
285. R.A. Logan, A.J. Peters, J. Appl. Phys. **31**, 122 (1960)
286. J.R. Lowney, H.S. Bennet, J. Appl. Phys. **69**, 7102 (1991)
287. P. Lugli, C. Erlen, A. Pecchia, F. Brunetti, L. Latessa, A. Bolognesi, G. Csaba, G. Scarpa, A. Di Carlo, Appl. Phys. A **87**, 593 (2007)
288. P. Lugli, P. Bordone, L. Reggiani, M. Rieger, P. Kocevar, S. Goodnick, Phys. Rev. B **39**, 7852 (1989)
289. P. Lugli, P. Bordone, L. Reggiani, M. Rieger, P. Kocevar, S. Goodnick, Phys. Rev. B **39**, 7866 (1989)
290. P. Lugli, D.K. Ferry, IEEE Trans. Electron. Dev. **32**, 2431 (1985)
291. M. Lundstrom, *Fundamentals of carrier transport*, 3rd edn. (Cambridge University Press, Cambridge, 2000)
292. S. Luryi, A. Kastalsky, Superlattice Microst. **1**, 389 (1985)
293. S.K. Lyo, D. Huang, Phys. Rev. B **73**, 205336 (2006)
294. G.D. Mahan, Phys. Rep. **145**, 251 (1987)
295. G.D. Mahan, *Many-Particle Physics*, 2nd edn. (Plenum, New York, 1990)
296. G.D. Mahan, *Green's Function Methods: Quantum Boltzmann Equation for Linear Transport* in *Quantum Transport in Semiconductors*, ed. by D.K. Ferry, C. Jacoboni (Plenum, New York, 1992), p. 101
297. G.D. Mahan, Phys. Rev. B **68**, 125409 (2003)
298. G.D. Mahan, W. Hansch, J. Phys. F **13**, L47 (1983)
299. H.M. Manasevit, I.S. Gergis, A.B. Jones, Appl. Phys. Lett. **41**, 464 (1982)
300. M.J. Manfra, L.N. Pfeiffer, K.W. West, R. de Picciotto, K.W. Baldwin, Appl. Phys. Lett. **86**, 162106 (2005)
301. R.D. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem*, 2nd edn. (The McGraw-Hill Book Company, New York, 1976)
302. A. Matulionis, J. Požela, A. Reklaitis, Solid State Commun. **16**, 1133 (1975)
303. A.L. Mears, R.A. Stradling, J. Phys. C **4**, L22 (1972)
304. T.C. McGill, R. Baron, Phys. Rev. B **11**, 5208 (1975)
305. Y. Meir, N.S. Wingreen, Phys. Rev. Lett. **68**, 2512 (1992)
306. A. Messiah, *Quantum Mechanics* (North Holland, Amsterdam, 1975)
307. N. Metropolis, S. Ulam, J. Am. Stat. Assoc. **247**, 335 (1949)
308. J.R. Meyer, F.J. Bartoli, Phys. Rev. B **24**, 2089 (1981)
309. R. Mickevičius, V. Mitin, Phys. Rev. B **48**, 17194 (1993)
310. V.V. Mitin, V.A. Kochelap, M.A. Stroscio, *Quantum Heterostructures* (Cambridge University Press, Cambridge, 1999)
311. A.M. Morales, C.M. Lieber, Science **279**, 208 (1998)
312. F.J. Morin, J.P. Maita, Phys. Rev. **96**, 28 (1954)
313. P.M. Morse, H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953)
314. J. Motohisa, H. Sakaki, Appl. Phys. Lett. **60**, 1315 (1992)
315. F. Mousty, P. Ostoja, L. Passari, J. Appl. Phys. **45**, 576 (1974)

316. J.E. Moyal, Proc. Camb. Phil. Soc. **45**, 99 (1949)
317. J.G. Muga, R. Sala, S. Brouard, Solid State Commun. **94**, 877 (1995)
318. R.S. Muller, T.I. Kamins, *Device Electronics for Integrated Circuits* (Wiley, New York, 1977)
319. M. Nawaz, J.P. Leburton, J. Jin, Appl. Phys. Lett. **90**, 183505 (2007)
320. M. Nedjalkov, I. Dimov, F. Rossi, C. Jacoboni, Math. Comput. Model. **23**, 159 (1996)
321. M. Nedjalkov, H. Kosina, S. Selberherr, C. Ringhofer, D.K. Ferry, Phys. Rev. B **70**, 115319 (2004)
322. M. Nedjalkov, P. Vitanov, Soild State Electron. **31**, 1065 (1988)
323. M. Nedjalkov, P. Vitanov, Soild State Electron. **32**, 893 (1989)
324. J. Nelson, *The Physics of Solar Cells* (Imperial College, London, 2003)
325. G.W. Neudeck, *The PN Junction Diode* (Addison-Wesley, Reading, 1989)
326. N. Nintunze, M.A. Osman, Semic. Sci. Technol. **10**, 11 (1995)
327. K. Nishiohara, N. Shiguo, T. Wada, IEEE Trans. Electron. Dev. **39**, 634 (1992)
328. L. Nordheim, Ann. Physik **9**, 607 (1931)
329. P. Norton, T. Braggins, H. Levinstein, Phys. Rev. B **8**, 5632 (1973)
330. K.S. Novoselov, A.K. Geim, S.V. Morozov, D. Jiang, Y. Zhang, S.V. Dubonos, I.V. Grigorieva, A.A. Firsov, Science **306**, 666 (2004)
331. K.S. Novoselov, A.K. Geim, S.V. Morozov, D. Jiang, M.I. Katsnelson, I.V. Grigorieva, S.V. Dubonos, A.A. Firsov, Nature **438**, 197 (2005)
332. T. Ohmi, S. Hasuo, in *Proceedings of the 10th International Conference on the Physics of Semiconductors* (Cambridge, Mass. 1970), p. 60
333. G. Ottaviani, L. Reggiani, C. Canali, F. Nava, A. Alberigi-Quaranta, Phys. Rev. B **12**, 3318 (1975)
334. G. Paulavičius, R. Mickevičius, V. Mitin, M.A. Stroscio, J. Appl. Phys. **82**, 3392 (1997)
335. G. Pennington, N. Goldsman, Phys. Rev. B **68**, 045426 (2003)
336. V. Perebeinos, J. Tersoff, P. Avouris, Phys. Rev. Lett. **94**, 086802 (2005)
337. L. Pfeiffer, K.W. West, H.L. Stormer, K.W. Baldwin, Appl. Phys. Lett. **55**, 1888 (1989)
338. A. Philips Jr., P.J. Price, Appl. Phys. Lett. **30**, 528 (1977)
339. J.C. Phillips, Phys. Rev. **112**, 685 (1959)
340. J.C. Phillips, L. Kleinman, Phys. Rev. **116**, 287 (1959)
341. J.C. Phillips, K.C. Pandey, Phys. Rev. Lett. **17**, 287 (1973)
342. R.F. Pierret, *Advanced Semiconductor Fundamentals* (Addison-Wesley, Reading, 1987)
343. R.F. Pierret, *Field Effect Devices*, 2nd edn. (Addison-Wesley, Reading, 1990)
344. P. Poli, L. Rota, C. Jacoboni, Appl. Phys. Lett. **55**, 1026 (1989)
345. J. Pozhela, A. Reklaitis, Solid State Commun. **27**, 1073 (1978)
346. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipies in Fortran*, 2nd edn. (Cambridge University Press, Cambridge, 1992)
347. P.J. Price, J. Appl. Phys. **31**, 949 (1960)
348. P.J. Price, in *Fluctuation Phenomena in Solids*, ed. by R.E. Burgess, Ch. 8 (Academic, New York, 1965)
349. P.J. Price, in *Proceedings of the 9th International Conference on the Physics of Semiconductors*, ed. by S.M. Ryvkin (Nauka, Leningrad, 1968), p. 753
350. P.J. Price, IBM J. Res. Dev. **14**, 12 (1970)
351. P.J. Price, Semiconduct. Semimet. **14**, 249 (1979)

352. P.J. Price, Ann. Phys. **133**, 217 (1981)
353. P.J. Price, J. Vac. Sci. Technol. **19**, 599 (1981)
354. P.J. Price, J. Appl. Phys. **53**, 6863 (1982)
355. P.J. Price, Surf. Sci. **143**, 145 (1984)
356. S. Ragazzi, A. Di Carlo, P. Lugli, F. Rossi, Phys. Stat. Sol. (b) **204**, 339 (1997)
357. A. Ramamoorthy, R. Akis, J.P. Bird, IEEE Trans. Nanotech. **5**, 712 (2006)
358. S. Ramo, Proc. IRE **27**, 584 (1939)
359. A.J. Read, R.J. Needs, K.J. Nash, L.T. Canham, P.D.J. Calcott, A. Qteish, Phys. Rev. Lett. **69**, 1232 (1992)
360. H.D. Rees, Phys. Lett. A **26**, 416 (1968)
361. H.D. Rees, J. Phys. Chem. Solids **30**, 643 (1969)
362. L. Reggiani, Phys. Rev. B **17**, 2800 (1978)
363. L. Reggiani, *General Theory* in *Hot-Electron Transport in Semiconductors*, ed. by L. Reggiani (Springer, Berlin, 1985), p. 7
364. L. Reggiani (ed.), *Hot-Electron Transport in Semiconductors* (Springer, Berlin, 1985)
365. L. Reggiani, C. Canali, F. Nava, G. Ottaviani, Phys. Rev. B **16**, 2781 (1977)
366. S. Reggiani, L. Silvestri, A. Cacciatori, E. Gnani, A. Gnudi, G. Baccarani, IEDM2007, Tech. Dig., p. 557
367. S. Reich, C. Thomsen, J. Maultzsch, *Carbon Nanotubes: Basic Concepts and Physical Properties* (Wiley-VCH, Berlin, 2004)
368. S. Reich, J. Maultzsch, C. Thomsen, P. Ordejón, Phys. Rev. B **66**, 035412 (2002)
369. L.E. Reichl, *A Modern Course in Statistical Physics* (University of Texas Press, Austin, 1980)
370. F. Reif, *Fundamentals of statistical and thermal physics* (McGraw-Hill, New York, 1965)
371. G. Rickayzen, *Green's Functions and Condensed Matter* (Academic, London, 1980)
372. B.K. Ridley, *Quantum Processes in Semiconductors*, 2nd edn. (Clarendon, Oxford, 1988)
373. B.K. Ridley, J. Appl. Phys. **48**, 754 (1977)
374. B.K. Ridley, Rep. Prog. Phys. **54**, 169 (1991)
375. D.L. Rode, Phys. Rev. B **2**, 1012 (1970)
376. D.L. Rode, S. Knight, Phys. Rev. B **3**, 2534 (1971)
377. F. Rossi (ed.), *Semiconductor Macroatoms* (Imperial College, London, 2005)
378. F. Rossi, C. Jacoboni, Semicond. Sci. Technol. **7**, B383 (1992)
379. F. Rossi, C. Jacoboni, Europhys. Lett. **18**, 169 (1992)
380. F. Rossi, P. Poli, C. Jacoboni, Semicond. Sci. Technol. **7**, 1017 (1992)
381. F. Rossi, T. Kuhn, Rev. Mod. Phys. **74**, 895 (2002)
382. L. Rota, C. Jacoboni, P. Poli, Solid State Electron. **32**, 1417 (1989)
383. G. Ruch, G.S. Kino, Phys. Rev. **174**, 921 (1968)
384. H. Rücker, E. Molinari, P. Lugli, Phys. Rev. B **45**, 6747 (1992)
385. M. Rudan, M.C. Vecchi, D. Ventura, *The Hydrodynamic Model in Semiconductors – Coefficient Calculation for the Conduction Band of Silicon* in *The Pitman Research Notes in Mathematical Series 340* (Longman, New York, 1995), p. 186
386. M. Rudan, M. Lorenzini, R. Brunetti, *Hydrodynamic Simulation of Semiconductor Devices* in *Theory of Transport Properties of Semiconductor Nanostructures*, ed. by E. Schöll (Chapman & Hall, London, 1998), p. 2

387. E.J. Ryder, Phys. Rev. **90**, 766 (1953)
388. E.J. Ryder, W. Shockley, Phys. Rev. **81**, 139 (1951)
389. R. Saito, G. Dresselhaus, M.S. Dresselhaus, *Physical Properties of Carbon Nanotubes* (Imperial College, London, 1998)
390. H. Sakaki, Jpn. J. Appl. Phys. **19**, L735 (1980)
391. H. Sakaki, T. Noda, K. Hirakawa, M. Tanaka, T. Matsusue, Appl. Phys. Lett. **51**, 1934 (1987)
392. T. Saku, Y. Horikoshi, Y. Tokura, Jpn. J. Appl. Phys. **35**, 34 (1996)
393. N. Sano, T. Aoki, M. Tomizawa, A. Yoshii, Phys. Rev. B **41**, 12122 (1990)
394. N. Sano, A. Yoshii, Phys. Rev. B **45**, 4171 (1992)
395. M. Saraniti, S.M. Goodnick, IEEE Trans. Electron. Dev. **47**, 1909 (2000)
396. W. Sasaki, M. Shibuya, K. Mizuguchi, G.M. Hatoyama, J. Phys. Chem. Solids **8**, 250 (1959)
397. D.L. Scharfetter, H.K. Gummel, IEEE Trans. Electron. Dev. **ED-16**, 64 (1969)
398. L.I. Schiff, *Quantum Mechanics*, 3rd edn. (McGraw-Hill, New York, 1968)
399. N. Sclar, Phys. Rev. **104**, 1559 (1956)
400. T.C. Schmidt, K. Möhring, Phys. Rev. A **48**, R3418 (1993)
401. R. Scholz, J. Appl. Phys. **77**, 3219 (1995)
402. K. Seeger, *Semiconductor Physics*, 3rd edn. (Springer, Berlin, 1964)
403. F. Seitz, Phys. Rev. **73**, 549 (1948)
404. F. Seitz, N.G. Einspruch, *Electronic genie: The Tangled History of Silicon* (Univ of Illinois Pr, Urbana, 1998)
405. S. Selberherr, *Analysis and Simulation of Semiconductor Devices* (Springer, Wien, 1984)
406. J. Shah (ed.), *Hot Carriers in Semiconductor Microstructures* (Academic, New York, 1992)
407. J. Shah, *Ultrafast Spectroscopy of Semicondutors and Semiconductor Nanostructures* (Springer, Berlin, 1999)
408. J. Shah, R.C.C. Leite, Phys. Rev. Lett. **22**, 1304 (1969)
409. C.V. Shank, R.L. Fork, R.F. Leheny, J. Shah, Phys. Rev. Lett. **42**, 112 (1979)
410. M.P. Shaw, V.V. Mitin, E. Schöll, H.L. Grubin, *The Physics of Instabilities in Solid State Electron Devices* (Plenum, New York, 1992)
411. M. Shibuya, Phys. Rev. **99**, 1189 (1955)
412. L. Shifren, D.K. Ferry, Phys. Lett. A **285**, 217 (2001)
413. H. Shichijio, K. Hess, Phys. Rev. B **23**, 4197 (1981)
414. H. Shirakawa, E.J. Louis, A.G. Macdiarmid, K. Chwan, A.J. Heeger, Chem. Commun. **16**, 578 (1977)
415. W. Shockley, J. Appl. Phys. **9**, 635 (1938)
416. W. Shockley, Bell Syst. Tech. J. **28**, 435 (1949)
417. W. Shockley, *Electrons and Holes in Semiconductors* (D. Van Nostrand, Princeton, 1950)
418. W. Shockley, Bell Syst. Tech. J. **30**, 990 (1951)
419. W. Shockley, J.A. Copeland, R.P. James, *Quantum Theory of Atoms, Molecules, and the Solid State* (Academic, New York, 1966)
420. A. Sibille, J.F. Palmier, H. Wang, F. Mollot, Phys. Rev. Lett. **64**, 52 (1990)
421. J. Sing, *Electronic and Optoelectronic Properties of Semiconductor Structures* (Cambridge University Press, Cambridge, 2003)
422. J.M. Soler, E. Artacho, J.D. Gale, A. García, J. Junquera, P. Ordejón, D. Sánchez-Portal, J. Phys. Condens. Matter **14**, 2745 (2002)

423. E.H. Sondheimer, Proc. R. Soc. A **203**, 75 (1950)
424. G.A. Steele, G. Gotz, L.P. Kouwenhoven, Nat. Nanotech. **4**, 363 (2009)
425. F. Stern, W.E. Howard, Phys. Rev. **163**, 816 (1967)
426. M. Stone, *Quantum Hall Effect* (World Scientific, Singapore, 1992)
427. R. Stratton, Proc. Roy. Soc. **A246**, 406 (1958)
428. G.B. Stringfellow, *Organometallic Vapor-Phase Epitaxy: Theory and Practice*, 2nd ed. (Academic, San Diego, 1999)
429. M.A. Stroscio, Phys. Rev. B **40**, 6428 (1989)
430. T. Sugaya, J.P. Bird, D.K. Ferry, A. Segeev, V. Mitin, K.-Y. Jang, M. Ogura, Y. Sugiyama, Appl. Phys. Lett. **81**, 727 (2002)
431. S.M. Sze, *Physics of Semiconductor Devices*, 2nd edn. (Wiley, New York, 1981)
432. K. Takeda, N. Matsumoto, A. Taguchi, H. Taki, E. Ohta, M. Sakata, Phys. Rev. B **32**, 1101 (1985)
433. I. Takesue, J. Haruyama, N. Kobayashi, S. Chiashi, S. Maruyama, T. Sugai, H. Shinohara, Phys. Rev. Lett. **96**, 057001 (2006)
434. J.Y. Tang, K. Hess, J. Appl. Phys. **54**, 5139 (1983)
435. Z.K. Tang, L. Zhang, N. Wang, X.X. Zhang, G.H. Wen, G.D. Li, J.N. Wang, C.T. Chan, P. Sheng, Science **292**, 2462 (2001)
436. S.J. Tans, M.H. Devoret, H. Dai, A. Thess, R.E. Smalley, L.J. Geerlings, C. Dekker, Nature (London) **386**, 474 (1997)
437. S. Tarucha, D.G. Austing, T. Honda, R.J. van der Hage, L.P. Kouwenhoven, Phys. Rev. Lett. **77**, 3613 (1996)
438. S. Taschini, M. Rudan, R. Brunetti, Phys. Rev. B **60**, 13582 (1999)
439. V.I. Tatarskii, Sov. Phys. Usp. **26**, 311 (1983)
440. D. Ter Haar, Rev. Mod. Phys. **27**, 289 (1955)
441. D. Ter Haar, Rep. Progr. Phys. **24**, 304 (1961)
442. R. Thoma, H.J. Peifer, W.L. Engl, W. Quade, R. Brunetti, C. Jacoboni, J. Appl. Phys. **69**, 2300 (1991)
443. S.E. Thompson, G. Sun, Y.S. Choi, T. Nishida, IEEE Trans. Electron. Dev. **53**, 1010 (2006)
444. M. Tiersten, IBM J. Res. Dev. **5**, 122 (1961)
445. M. Tiersten, J. Phys. Chem. Solids **25**, 1151 (1964)
446. G. Timp, A.M. Chang, P. Mankiewich, R. Behringer, J.E. Cunningham, T.Y. Chang, R.E. Howard, Phys. Rev. Lett. **59**, 732 (1987)
447. M. Tinkham, *Group Theory and Quantum Mechanics* (McGraw-Hill, New York, 1964)
448. R.C. Tolman, *The Principles of Statistical Mechanics* (Oxford University Press, London 1938, Dover 1980)
449. J.Y. Tsao, *Materials Fundamentals of Molecular Beam Epitaxy* (Academic, San Diego, 1993)
450. M. Tsetseri, G.P. Triberis, Phys. Rev. B **69**, 075313 (2004)
451. M. Tsetseri, G.P. Triberis, M. Tsaousidou, Superl. Microstr. **43**, 340 (2008)
452. D.C. Tsui, H.L. Stormer, A.C. Gossard, Phys. Rev. Lett. **48**, 1559 (1982)
453. R. Ulbrich, Solid State Electron. **21**, 51 (1977)
454. C.G. Van de Walle, Phys. Rev. B **39**, 1871 (1989)
455. W.G. van der Wiel, S. De Franceschi, J.M. Elzerman, T. Fujisawa, S. Tarucha, L.P. Kouwenhoven, Rev. Mod. Phys. **75**, 1 (2003)
456. H. van Houten, C.W.J. Beenakker, Phys. Rev. Lett. **63**, 1893 (1989)
457. L. Van Hove, Physica **21**, 517 (1955)

458. L. Van Hove, Physica **23**, 441 (1957)
459. L. Van Hove, Physica **25**, 268 (1959)
460. L. Van Hove, E. Verboven, Physica **27**, 418 (1961)
461. B.J. van Wees, L.P. Kouwenhoven, E.M.M. Willems, C.J.P.M. Harmans, J.E. Moij, H. van Houten, J.G. Williamson, C.T. Foxon, Phys. Rev. B **43**, 12431 (1991)
462. F. Venturi, E. Sangiorgi, S. Luryi, P. Poli, L. Rota, C. Jacoboni, IEEE Trans. Electron. Dev. **38**, 611 (1991)
463. P. Vogl, Phys. Rev. B **13**, 694 (1976)
464. J. Von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, 1955)
465. L.-W. Wang, Energy Environ. Sci. **2**, 944 (2009)
466. Z.M. Wang, K. Holmes, Yu.I. Mazur, G.J. Salamo, Appl. Phys. Lett. **84**, 1931 (2004)
467. C. Waschke, H.G. Roskos, R. Schwedler, K. Leo, H. Kurz, K. Köhler, Phys. Rev. Lett. **70**, 3319 (1993)
468. S. Washburn, R.A. Webb, Adv. Phys. **35**, 375 (1986)
469. R.A. Webb, S. Washburn, C.P. Umbach, R.B. Laibowitz, Phys. Rev. Lett. **54**, 2696 (1985)
470. W. Wegscheider, L.N. Pfeiffer, M.M. Dignam, A. Pinczuk, K.W. West, S.L. McCall, R. Hull, Phys. Rev. Lett. **71**, 4071 (1993)
471. H. Weyl, Z. Phys. **46**, 1 (1927)
472. E. Wigner, Z. Phys. Chem. **B19**, 203 (1932)
473. E. Wigner, Phys. Rev. **40**, 749 (1932)
474. J.D. Wiley, Phys. Rev. B **4**, 2485 (1971)
475. J.D. Wiley, in *Semiconductors and Semimetals*, R.K. Willardson, A.C. Beer (eds.), vol. 10 (Academic, New York, 1974), p. 91
476. A.H. Wilson, *The Theory of Metals*, 2nd edn. (Cambridge University Press, Cambridge, 1953)
477. T.H. Wonnacott, R.J. Wonnacott, *Introductory Statistics*, 5th edn. (Wiley, New York, 1990)
478. Q. Xie, A. Madhukar, P. Chen, N. Kobayashi, Phys. Rev. Lett. **75**, 2542 (1995)
479. S. Yamakawa, H. Ueno, K. Taniguchi, C. Hamaguchi, K. Miyatsuji, K. Masaki, U. Ravaioli, J. Appl. Phys. **79**, 911 (1996)
480. Z. Yao, C.L. Kane, C. Dekker, Phys. Rev. Lett. **84**, 2941 (2000)
481. Y. Yin, P. Alivisatos, Nature **437**, 664 (2005)
482. D. Yoshioka, *The Quantum Hall Effect* (Springer, Berlin, 2002)
483. P.Y. Yu, M. Cardona, *Fundamentals of Semiconductors*, 3rd edn. (Springer, Berlin, 2001)
484. C. Zener, Proc. Roy. Soc. **145**, 523 (1934)
485. X. Zhao, C.M. Wei, L. Yang, M.Y. Chou, Phys. Rev. Lett. **92**, 236805 (2004)
486. Y. Zhang, Y-W. Tan, H.L. Stormer, P. Kim, Nature **438**, 201 (2005)
487. Y. Zheng, T. Ando, Phys. Rev. B **65**, 245420 (2002)
488. J.-R. Zhou, D.K. Ferry, in *Proceedings of 3rd International Workshop on Computational Electronics*, New York, 1994
489. J.M. Ziman, *Electrons and Phonons* (Oxford University Press, London, 1960)
490. J. Zimmermann, Y. Leroy, E. Constant, J. Appl. Phys. **49**, 3378 (1979)
491. P.S. Zory Jr. (ed.), *Quantum Well Lasers* (Academic, San Diego, 1991)
492. K.H. Zschauser, *Proceedings of 4th International Symposium of GaASs and Related Compounds*, ed. by C. Hilsum (Wright, London, 1972)

493. W.H. Zurek, Rev. Mod. Phys. **75**, 715 (2003)
494. R. Zwanzig, J. Chem. Phys. **33**, 1338 (1960)
495. R. Zwanzig, Physica **30**, 1109 (1964)
496. See, for example, [96]; T.S. Monteiro, J. Phys. A Mat. Gen. **27**, 787 (1994); G. Manfredi and M.R. Feix: Phys. Rev. E **62**, 4665 (2000); H-y. Fan and J. Chen: Eur. Phys. J. D **23**, 437 (2003). The Journal of Optics B has published a Wigner Centennial issue (J. Optics B **5**(3), (2003)), where many references can be found.

Index